Supporting Information for

Estimating annual average daily traffic on local roads: Integrating spatial insights with machine learning

Liang Ma*, Rami Al-Shukairi, Marc E. J. Stettler and Daniel J. Graham

*Corresponding author: Liang Ma
Email: liang.ma13@imperial.ac.uk

This PDF file includes:

Supporting text: 21 pages

Number of Figures: 10

Number of Tables: 4

Supporting Information

S1. Data description

The annual average daily traffic (AADT) data for England and Wales is sourced from publicly available road traffic estimates in the UK[1]. These estimates reflect the number of vehicles passing designated "count point" locations. In 2021, the source dataset included 19,720 count points, categorised into major roads (A roads and motorways) and minor roads (B, C, and unclassified roads). To associate AADT data with road attributes and account for network distance, each AADT count point is linked to a corresponding road segment based on geographical location, road class, and road name. Road segment information is obtained from Ordnance Survey Open Roads[2], which provides geographical information and road attributes for all road segments within the UK. After data merging and excluding points located on islands without direct road connections to the mainland (such as the Isles of Scilly), the final dataset includes 19,560 count points.

S1.1 Feature design

To predict AADT at unmeasured locations, we extract 905 contextual features from various open-source government datasets, capturing both on-network and off-network characteristics for individual count points. The design of the contextual feature set mainly follows Sfyridis & Agnolucci[3]. To represent the spatial autocorrelation structure of AADT, 144 supplementary spatial statistical features were created. **Table S1-1** shows the data sources and descriptions of these features. A feature is included in the analysis only if at least 75% of the count points have valid data.

Table S1-1. Spatial resolution and source of variables used in the AADT prediction model.

Category [(a)]	Description	Original data resolution [(b)]	Source
Rural/Urban classification (1)	Categorical label indicating rural or urban classification (10 levels)	Lower Layer Super Output Area (LSOA) (34,753)	Department for Environment Food & Rural Affairs[4]
Built-up area (BUA) (1)	Boolean indicator for presence in a BUA	BUAs (7,723)	Office for National Statistics[5]
Access to major towns and cities (4)	- Accessibility metrics to major towns and cities using population weights and different impedance functions - Boolean indicator for presence within a major town or city	Major towns and cities (112)	Office for National Statistics[6,7]
Access to functional urban areas (FUA) (5)	- Boolean indicator for presence within an FUA - Distance to the nearest FUA boundary (0 if inside)	FUAs (41)	Office for National Statistics[8]
	- Accessibility metrics to FUAs using population weights and different impedance functions	Local administrative units level 1 (348); FUAs (41)	Office for National Statistics[7,8]

Road attributes (6)	- Road class - Road function - Primary road (boolean) - Trunk road (boolean) - Form of way - Road length	Road segments (3,503,488)	Ordnance Survey[2]
Access to motorway junctions (6)	Number of motorway junctions within service areas at multiple radii	Motorway junctions (579)	Ordnance Survey[2]
Business counts (390)	Calculated within service areas at multiple radii: - Business counts by employment size band and industry sector [(c,d)] - Total business counts by employment size band and by industry sector - Mutual proportions between employment size bands and industry sectors	Middle Layer Super Output Area (MSOA) (7,201)	Office for National Statistics[9]
Earnings (48)	- Median gross annual earnings by place of residence and workplace, calculated within service areas at multiple radii	Parliamentary constituency (573)	Office for National Statistics[10,11]
	- Earnings inequality ratios (P90/10, P80/20, P75/25) by place of residence and workplace, calculated within service areas at multiple radii	Local authority (331)	Office for National Statistics[10,11]
Employment (306)	Calculated within service areas at multiple radii: - Total employment - Employment by industry section and by industry sector [(d)] - Proportion of each industry sector and industry section relative to total employment	LSOA (34,753)	Office for National Statistics[12]
Population (60)	Calculated within service areas at multiple radii - Population total - Population density - Number of households	LSOA (35,672)	Office for National Statistics[13]
	- Population by age group [(e)] (count and proportion), calculated within service areas at multiple radii	LSOA (34,753)	Office for National Statistics[7]
Car ownership (42)	Car ownership by body type [(f)] (count and proportion), calculated within service areas at multiple radii	LSOA (34,753)	Department for Transport & Driver and Vehicle Licensing Agency[14]

Accessibility to public transport stations (24)	Number of public transport stations within service areas at multiple radii, by type of stop [g]	Public Transport Access Nodes (436,503)	Department for Transport[15]
Accessibility to major ports (6)	Accessibility metrics to major ports, calculated using different impedance functions and weightings (overall goods and unitised goods)	Major ports [h] (36)	Department for Transport[16]
Accessibility to major airports (6)	Accessibility metrics to major airports, calculated using different impedance functions and weightings (terminal passenger volume and freight volume)	Major airports [h] (27)	Civil Aviation Authority[17]; Borsetti[18]
Coordinates (28)	- Geographical coordinates (Northing and Easting) of count points - Rotated coordinate transformations of count points using multiple oblique angles	AADT count points (19,560)	Department for Transport[1]
Spatial smooth surface [j] (1)	Fitted spatial trend of AADT using a generalised additive model (GAM) with smooth functions of Northing and Easting	AADT count points (19,560)	Department for Transport[1]
Eigenvector spatial filtering (100)	Selected eigenvectors of spatial weights matrix capturing spatial structure among count points	AADT count points (19,560)	Department for Transport[1]
Spatial lag of target variable [j] (2)	- Weighted average of AADT from spatially nearby count points - The number of effective neighbours [k] contributing to the lag calculation	AADT count points (19,560)	Department for Transport[1]
Local Moran's I of key predictors (10)	Local Moran's I statistic and cluster label for - Population density - Total employment - Median income by place of residence and workplace - Total business counts	Same as base variables listed above	Same as base variables listed above
Traversable total road length (3)	Total length of road segments reachable within fixed distances (500 m, 800 m, 1 km) from each count point	Road segments (3,503,488)	Ordnance Survey[2]

(a) The number of features within each category is shown in the bracket.

(b) The number of spatial units within England and Wales is shown in the bracket.

(c) Employment size band: Micro (0 to 9), Small (10 to 49), Medium-sized (50 to 249), Large (250+).

(d) Industry section: 21 sections (highest level) in the UK Standard Industrial Classification (SIC) hierarchy. Industry sector: Agriculture, Production, Construction, and Service; aggregated based on SIC code.

(e) Age group: Aged 0 to 15, 16 to 64, 65+. Mid-year population estimates in 2020 was used due to data availability by the time of analysis.
(f) Vehicle body types: cars, motorcycles, others.
(g) Public transport station type: bus stop, bus/coach station, railway station, metro/tram station.
(h) Accessibility to ports is calculated only for ports with available and non-zero freight data.
(i) Accessibility to airports is calculated only for airports that report data to Civil Aviation Authority and have available and non-zero passenger or freight volume.
(j) This feature is fitted using only the training set in cross-validation and then predicted for both training and test sets.
(k) Effective neighbours: nearby count points within a maximum distance threshold whose weights (based on road class compatibility and Gaussian distance decay) exceed a minimal influence threshold.

S1.2 Assignment of contextual features

Contextual features are categorised into two types based on the method used to assign feature values from the spatial units in the source datasets to individual AADT count points (see **Figure S1-1**).

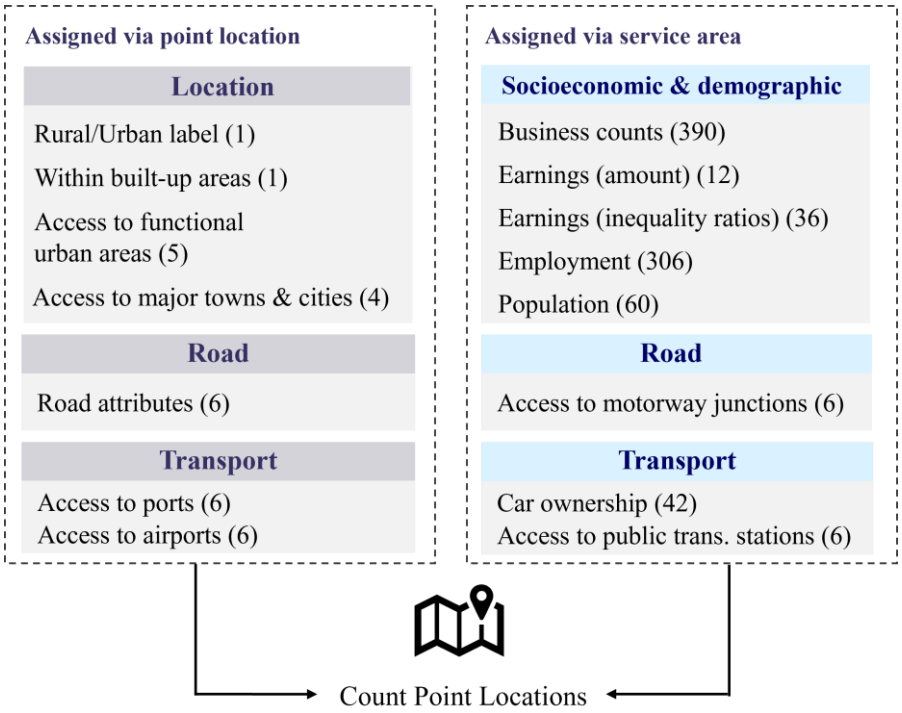

Figure S1-1. Assignement methods of contextual features used in AADT prediction. Features are assigned to count points based on either the precise coordinates of point location or service areas constructed using network distance buffers.

Location-based features

Several features are assigned based on the geographical location of the count points. These include settlement context features, accessibility to different tiers of urban areas, road attributes, and proximity to ports and airports.

Our study incorporates a range of accessibility features that go beyond simple distance calculations. The accessibility of a count point i to a set of infrastructure locations or urban areas $\{j: j = 1,2, \dots, n\}$ is derived by

$$\rho_i = \frac{1}{n}\sum_{j=1}^{n} w_j f(d_{ij})$$

where w_j is the weight of location j, d_{ij} is the Euclidean distance between i and j, and $f(\cdot)$ is a decreasing function of distance, commonly referred to as an impedance function[19]. In this study, three widely used impedance functions are applied: inverse distance $f(d_{ij}) = 1/d_{ij}$, inverse squared distance $f(d_{ij}) = 1/d_{ij}^2$, and negative exponential $f(d_{ij}) = \exp(-d_{ij})$. The weights w_j depend on the context of the accessibility calculation. For accessibility to urban areas, w_j corresponds to the population at location j; for airports, it reflects either terminal passenger volume or freight volume; and for ports, it is derived from reported throughput of overall goods or unitised goods.

Service area-based features

Assigning off-network characteristics, such as socioeconomic and demographic factors, to roads for AADT estimation commonly involves defining buffers around count points[20]. Following Sfyridis & Agnolucci[3], we compute six service areas with varying sizes (500 m, 800 m, 1,000 m, 1,600 m, 2,000 m, and 3,200 m) around each count point. Unlike conventional Euclidian distance-based buffers, these service areas are generated using network-based distance, providing a more realistic representation under real-world road conditions[21].

Particularly, the service area of count point i at radius r is derived by generating a surrounding polygon of the road segments traversable within a network distance r from that point. A polygon enclosing these traversable roads is then formed using Delaunay triangulation of road endpoints and subsequently trimmed to remain within 100 meters of the traversable road network. This trimming process holds particular value when dealing with sparse networks by preventing unrealistic and biased coverage. The process mainly follow the service area function in ArcGIS Pro. However, given the extensive scale of the road network under consideration, our implementation is fully conducted in python. Specifically, the identification of traversable roads is implemented with the ***NetworkX*** library[22]. Subsequent steps involving the generation of surrounding polygons are implemented using the ***Shapely*** library[23].

Once the service areas are created, we overlay them onto the spatial units of the source datasets to assign feature values to count points. For polygon-based datasets, such as those representing socioeconomic or demographic characteristics, feature values are computed as area-weighted averages based on the intersecting areas between polygons and the service area. For point-based datasets, such as public transport stations, features are assigned by counting the number of points that fall within each service area. An illustrative example of the service area construction is provided in **Figure S1-2**.

Figure S1-2. Illustrative example of service area construction. Panel (a) indicate the shape of the service area of a particular count point (red point) before trimming (blue polygon). The trimming process involves generating and merging a 100-meter buffer (green polygon) around all traversable roads (yellow lines). Panel (b) shows the resulting service area after trimming (purple polygon), which is the intersection of the blue and green polygons in Panel (a).

S1.3 Methods on spatial statistical features

The spatial statistical features supplement the contextual data by capturing spatial dependencies and latent structures that may influence traffic volumes. This subsection provides technical details on generating some of these features not fully described in the manuscript.

Oblique coordinate transformation

In our study, we include a set of oblique coordinate transformations to capture directional spatial trends and anisotropic patterns that may not align with standard cardinal directions. These transformations are computed by projecting coordinates onto rotated axes across 26 distinct angles[24]. The angles are determined by the 180° semicircle into 28 equal segments, excluding the standard directions of 0°, 90°, and 180°. The features are calculated by:

$$\mathrm{OGC}^{\theta} = x\cos\theta + y\sin\theta$$

where x and y are Easting and Northing coordinates, respectively, and θ is the oblique angle.

Spatial lag

To reflect local spatial dependencies, we incorporate spatial lag features, computed as weighted averages of AADT values from nearby points. For a particular count point i, the lag feature is derived as

$$\mathrm{lag}_i = \frac{\sum_{j\epsilon N_i} w_{ij} y_j}{\sum_{j\epsilon N_i} w_{ij}}$$

where N_i is the set of neighbouring count points within a maximum distance $d_{\max}$, y_j is the AADT value at count point j, and w_{ij} is the weight between count point i and j.

The weight combines a Gaussian distance-decay kernel with a compatibility adjustment based on road class similarity:

$$w_{ij} = \delta_{ij} \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right)$$

where d_{ij} is the Euclidean distance between count point i and j, and σ is the kernal bandwidth parameter controlling the spatial decay rate. The adjustment term δ_{ij} encodes the similarity between road classes between point i and j, and takes the value 1 when the road classes are the same, 0.5 when they differ by one level, 0.1 when they differ by two levels, and 0 otherwise. Count point j is regarded as a effective neighbour of point i if w_{ij} is greater than 1e-8.

The parameters σ and $d_{\max}$ are specific to the road class of each point. For major roads, we set σ =1500 m and $d_{\max}$ = 4500 m. For minor roads, we set σ =1000 m and $d_{\max}$ = 2500 m. These values are determined by empirical variogram analysis of AADT values, see section S2.

Spatial clustering of key predictors

To further account for spatial structure in the predictors themselves, we include spatial clustering indicators for key contextual variables, including population density, total employment, median gross annual pay by place of residence and and by workplace, and total business counts. For each selected predictor, we compute two clustering metrics at their original spatial resolutions: the local Moran's I statistic and its associated cluster type. The cluster type (High-High, Low-Low, High-Low, or Low-High) indicates the relationship between the value at a given location and the values at neighbouring locations. The local Moran's I and corresponding cluster type are then assigned to count points according to their geographical coordinates.

Local Moran's I is one of the most commonly used metric to quantify spatial autocorrelation[25]. The local Moran's I statistic for variable x at location i is derived by:

$$I_i = \frac{(x_i - \bar{x}) \sum_{j=1}^{n} w_{ij} (x_j - \bar{x})}{s^2}$$

where $\bar{x}$ and s^2 are the mean and variance of variable x, and w_{ij} represents the spatial weight between location i and j. As the spatial units of the selected predictors are polygons, we compute w_{ij} using Queen's contiguity spatial weight matrix. A high positive I value indicates a spatial cluster of similar values (High-High or Low-Low), while a negative value suggests spatial outliers (High-Low or Low-High).

Eigenvector spatial filtering

In spatial analysis, a spatial weight matrix W is commonly used to represent the dependency between pairs of locations, capturing both short- and long-distance spatial autocorrelation. However,

incorporating the full n × n spatial weight matrix directly into a model is often infeasible due to its high dimensionality. To address this, ESF is employed to decompose the spatial weight matrix into a set of orthogonal eigenvectors. These eigenvectors represent latent spatial patterns at various scales and can be included in the model to account for spatial autocorrelation while reducing complexity[26].

Formally, the matrix to be decomposed is:

$$C = MWM$$

where $M = I - \frac{1}{n}11^T$ is the centring matrix, 1 is an n × 1 vector of ones, W is the spatial weight matrix[27].

In our study, the spatial weight matrix is constructed separately for major and minor roads using a Gaussian kernel with a distance cutoff. The kernel bandwidth σ and cutoff distance $d_{\max}$ are selected based on empirical variogram analysis (see section S2) and are set larger than those used for spatial lag features. This is because ESF aims to capture spatial autocorrelation patterns across multiple scales, including broader regional trends, while spatial lags are inherently more localised. For major roads, we set $\sigma = 5000$ m and $d_{\max} = 9000$ m. For minor roads, we set $\sigma = 1000$ m and $d_{\max} = 500$ m. The eigen-decomposition is performed using the ***scipy***[28] library in python.

S2. Analysis of AADT values

In this section, we explore the characteristics of AADT values by analysing their distribution by road class and evaluating their level of spatial autocorrelation through an empirical variogram analysis. These analyses inform parameter choices for spatial feature construction and guide model selection by revealing the scale and structure of spatial dependencies and heterogeneity in traffic volumes.

Figure S2-1 shows that AADT distributions are highly skewed and heterogeneous across road classes. Motorways exhibit a relatively broad and near-symmetric distribution, while A roads display a clear right-skewed distribution with a long tail extending toward very high traffic volumes. Both road classes demonstrate relatively good coverage across medium to high traffic levels. In contrast, B roads, C roads, and especially unclassified roads tend to have lower AADT values, with steep drop-offs, heavier skew, and sparse high-volume observations. This increased skewness and heterogeneity present challenges for accurately learning and predicting outliers. These differences in distributional shape and scale motivate the use of stratified modelling approaches and justify applying a logarithm transformation to stabilise variance and improve model robustness. The distribution of AADT values after logarithm transformation is shown in **Figure S2-2.** The log-transformed AADT distributions show clear improvements in symmetry and shape across all road classes compared to the raw scale, especially for major roads.

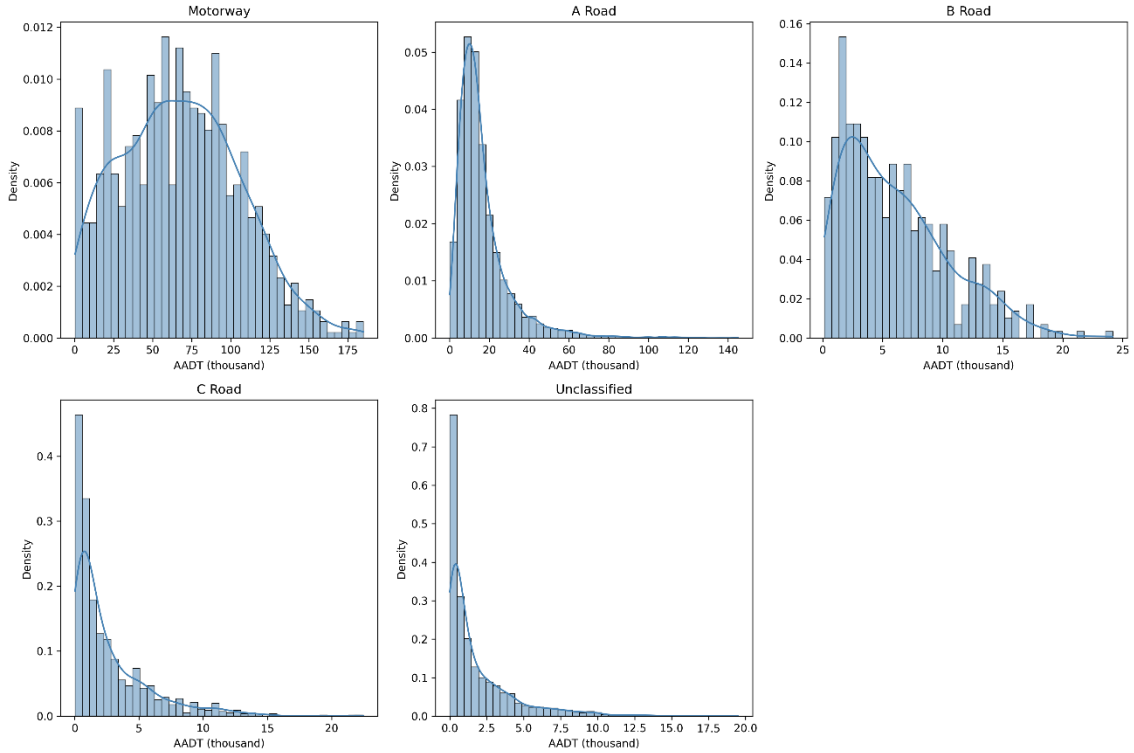

Figure S2-1. Distribution of AADT values by road class. Histograms and kernel density estimates of AADT (in thousands) are shown separately for motorways, A roads, B roads, C roads, and unclassified roads.

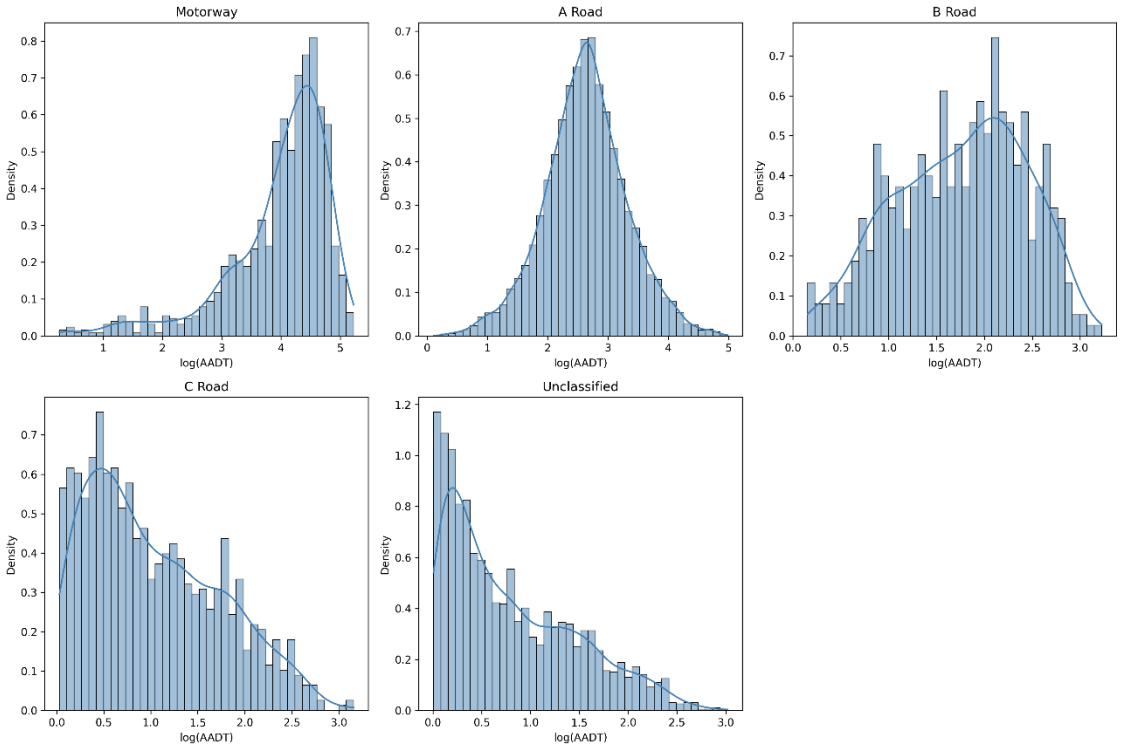

Figure S2-2. Distribution of log-transformed AADT values by road class. Histograms and kernel density estimates of log-transformed AADT values are shown separately for motorways, A roads, B roads, C roads, and unclassified roads.

To support the specification of spatial weight matrices for AADT observations, we analyse the spatial structure of AADT separately for motorways, A roads, and minor roads by fitting empirical variograms to grouped data using the ***scikit-gstat*** library[29] in python. We experiment with different combinations of theoretical models, binning strategies, and maximum lag distances, and apply Cressie's robust estimator to mitigate the influence of outliers. For each road type, the final variogram model is selected based on the highest pseudo-R^2 score, indicating the best fit to the empirical semivariances. After model selection, variograms for A roads and motorways are computed using the exponential model with a maximum lag of 9 km and uniform binning, while for minor roads, the Gaussian model is used with the same lag setting but ward binning strategy. The resulting fitted variograms are presented in **Figure S2-3**, **Figure S2-4**, and **Figure S2-5** for motorways, A roads, and minor roads, respectively.

These variograms generally demonstrate increasing spatial dissimilarity with distance, indicating positive spatial autocorrelation in AADT values. For motorways and A roads, the fitted theoretical models closely follow the empirical semivariances, showing strong spatial correlation at short distances and a smooth, gradual increase up to ~ 9 km. In contrast, the variogram for minor roads flattens much earlier, suggesting that spatial correlation diminishes rapidly beyond short distances.

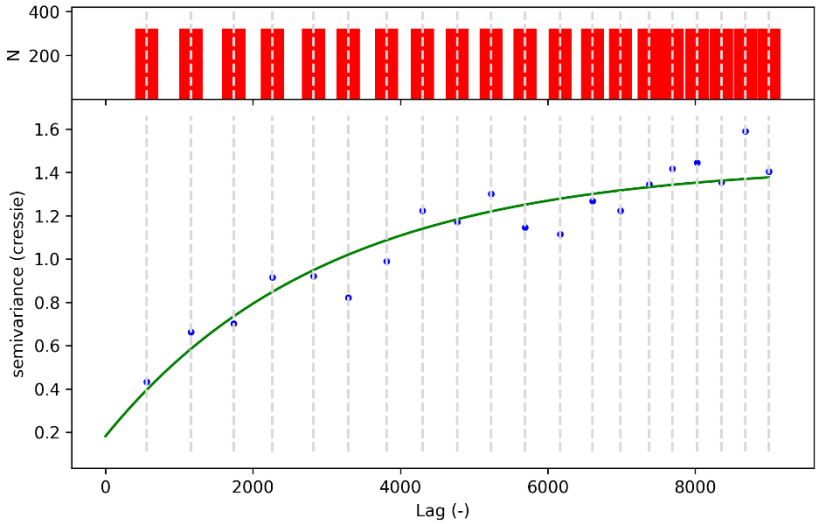

Figure S2-3. Empirical variogram of AADT for motorways. The variogram is fitted with an exponential model using Cressie's robust estimator. Blue points represent empirical semivariances calculated across 20 lag bins with approximately uniform sample sizes, up to a maximum lag distance of 9 km. The green curve indicates the fitted theoretical variogram. Red bars show the number of point pairs contributing to each bin.

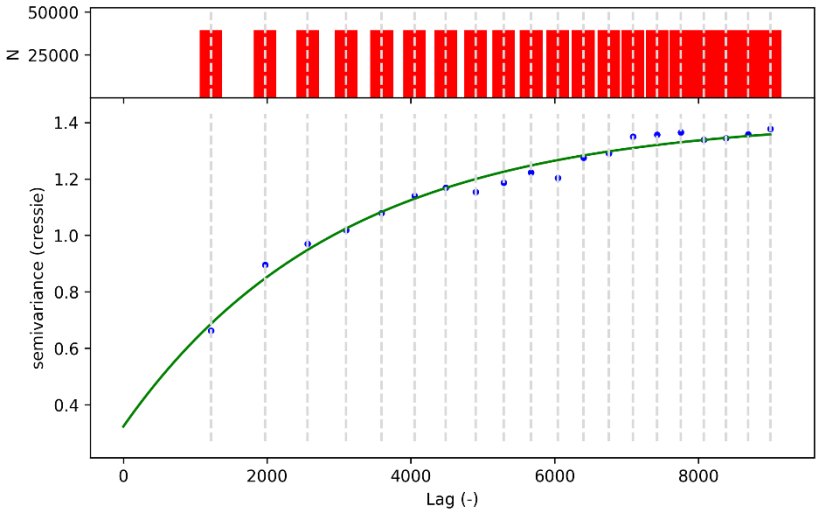

Figure S2-4. Empirical variogram of AADT for A roads. The variogram is fitted with an exponential model using Cressie's robust estimator. Blue points represent empirical semivariances calculated across 20 lag bins with approximately uniform sample sizes, up to a maximum lag distance of 9 km. The green curve indicates the fitted theoretical variogram. Red bars show the number of point pairs contributing to each bin.

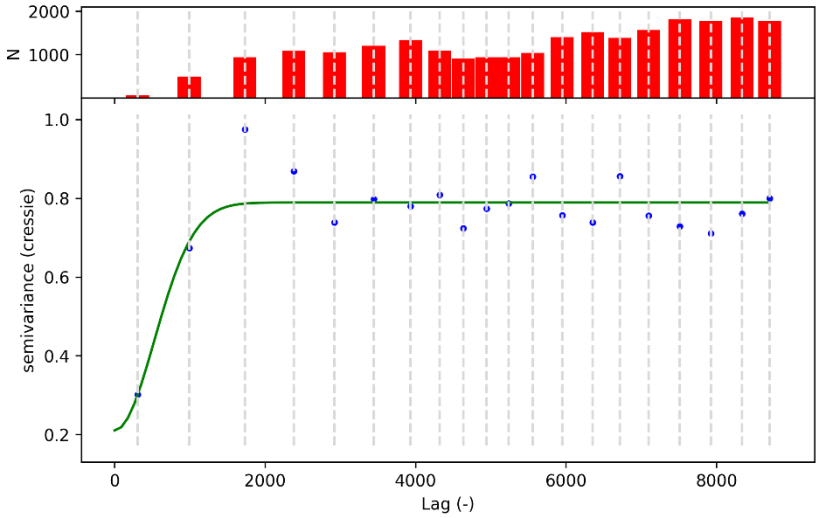

Figure S2-5. Empirical variogram of AADT for minor roads. The variogram is fitted with a Gaussian model using Cressie's robust estimator. Blue points represent empirical semivariances calculated across 20 lag bins derived using a ward-based binning strategy, with a maximum lag distance of 9 km. The green curve indicates the fitted theoretical variogram. Red bars show the number of point pairs contributing to each bin.

S3. Additional model results and diagnostics

This section presents additional results to complement the main manuscript. The first subsection reports summary statistics such as the number of selected features and the performance gap between training and test sets, providing insight into model complexity and generalisation. The second subsection includes supplementary figures based on spatial block cross-validation (CV), where training and test sets are spatially disjoint. These results are provided in addition to the main manuscript, which primarily focuses on results from sampling-intensity-weighted CV.

S3.1 Summary statistics on model complexity and generalisation

Table S3-1 summarises the performance gap across a range of model configurations. By comparing the difference in performance between training and test sets for each configuration, we assess the model's ability to generalise. These statistics complement model selection by highlighting the trade-offs between complexity, accuracy, and generalisability.

Table S3-2 reports the number of features selected for models of major and minor roads under two CV strategies. Results are shown for models using only spatial features, only contextual features, and a combined feature set prior to feature selection. Across both road types, models using combined contextual and spatial features retain the largest number of features, suggesting additive value in combining both types of information. Minor road models generally require more contextual features than major road models, reflecting the increased heterogeneity and complexity of local environments associated with less prominent road classes.

Table S3-1. Performance gap across model configurations.

Configuration					Performance gap [(a)]			
ID	Tuning	Encoding	Custom loss	Spatial features	CV type	nRMSE (%)	R^2 (%)	MAPE (%)
0	-	-	-	-	Spatial	-20.9	17.0	-30.2
					Weighted	-12.0	9.1	-24.3
1	-	-	-	✓	Spatial	-27.3	22.8	-28.4
					Weighted	-12.5	9.1	-24.2
2	-	✓	-	✓	Spatial	-26.9	22.5	-28.2
					Weighted	-12.4	9.0	-24.1
3	-	✓	✓	✓	Spatial	-27.2	21.7	-27.6
					Weighted	-13.4	9.4	-25.3
4	✓	✓	-	✓	Spatial	-36.9	27.4	-36.0
					Weighted	-32.7	17.6	-36.5
5	✓	✓	✓	✓	Spatial	-29.1	22.7	-27.9
					Weighted	-17.3	11.4	-26.5

(a) The performance gap is defined as the difference between the performance metric on the training set and that on the test set.

Table S3-2. Number of selected features by model group, feature set, and CV strategy.

Model group	CV type	Number of features		
		Spatial only	Contextual only	Contextual + Spatial
Major road	weighted	47	221	272
	spatial	49	223	264
Minor road	weighted	39	338	378
	spatial	38	338	378

S3.2 Additional results under spatial block CV

Figure S3-1 illustrates the predictive performance by road class under spatial block CV. Compared with Figure 2 in the manuscript, which presents corresponding results under sampling-intensity-weighted CV, model performance generally declines under spatial block CV, as expected. The reduction in predictive performance reflects the challenge of spatial extrapolation to regions that are entirely disjoint from the training data. The performance drop is particularly significant for major roads, with test R^2 decreased by ~ 25 percentage points. This highlights the spatial heterogeneity and complexity of traffic patterns on these networks. In contrast, minor road classes show smaller performance differences between the two CV strategies, suggesting more locally stable or spatially consistent traffic patterns. Overall, spatial block CV provides a conservative benchmark for evaluating model robustness under the most stringent generalisation conditions.

Figure S3-2 shows the contribution evaluation of spatial features associated with spatial block CV. Consistent with the results observed under sampling-intensity-weighted CV (c.f. Figure 3 and Figure 4 in manuscript), spatial features primarily act as complements to contextual features and the benefits of incorporating spatial features vary notably across road types. Interestingly, the contextual-only scenario achieves the best performance for major roads under spatial block CV, as opposed to results under weighted CV. Additionally, the relative importance of local dependency features diminishes under spatial block CV, implying reduced effectiveness when training and test regions are spatially disjoint. In contrast, ESF features remains valuable for major roads and coordinates-based features still contribute for minor roads.

Figure S3-3 presents the top 20 most important features for major and minor roads, respectively, based on mean absolute SHAP values derived from the model under spatial block CV. Notably, the top five features in each group are consistent with those identified under sampling-intensity-weighted CV, indicating the stability of key predictors across different validation strategies.

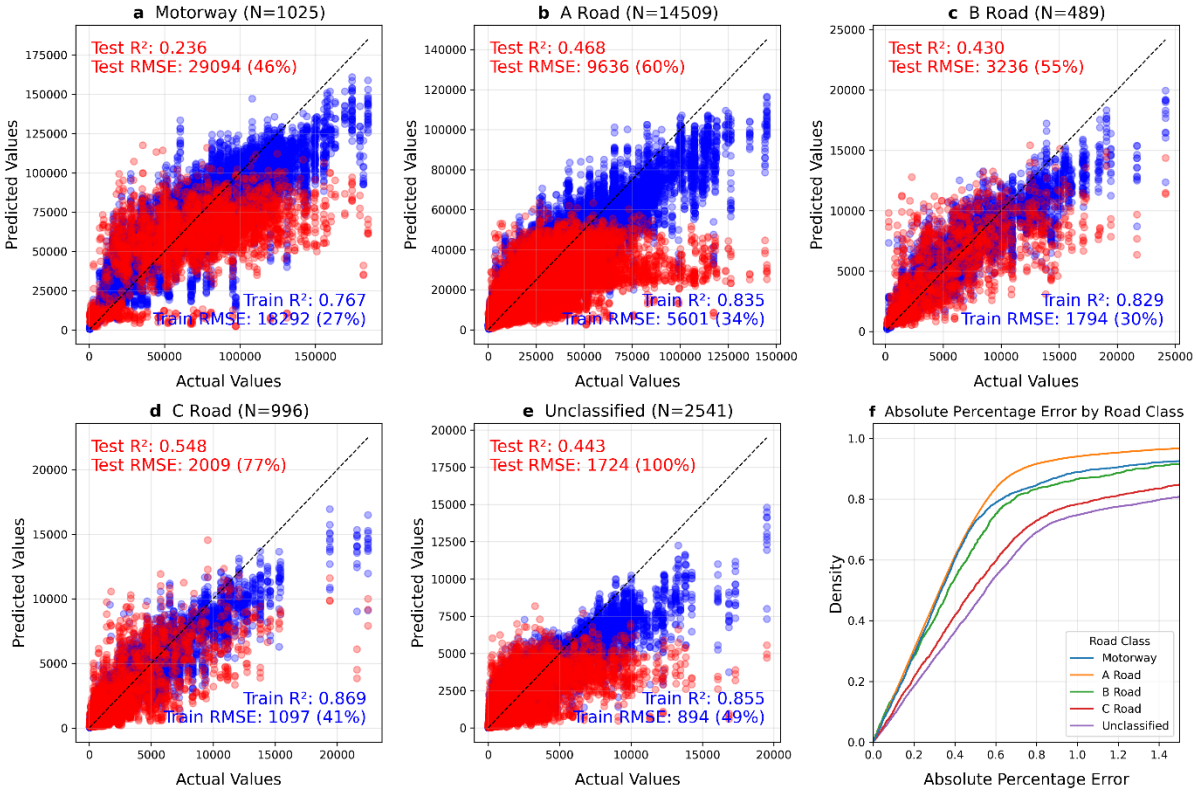

Figure S3-1. Predictive performance by road class under spatial block CV. Scatter plots (panels **a**-**e**) compare predicted and observed AADT on the training (red) and test (blue) sets for each road class, based on the final model. The dashed diagonal line indicates perfect prediction. Panel **f** shows the cumulative distribution of absolute percentage error by road class.

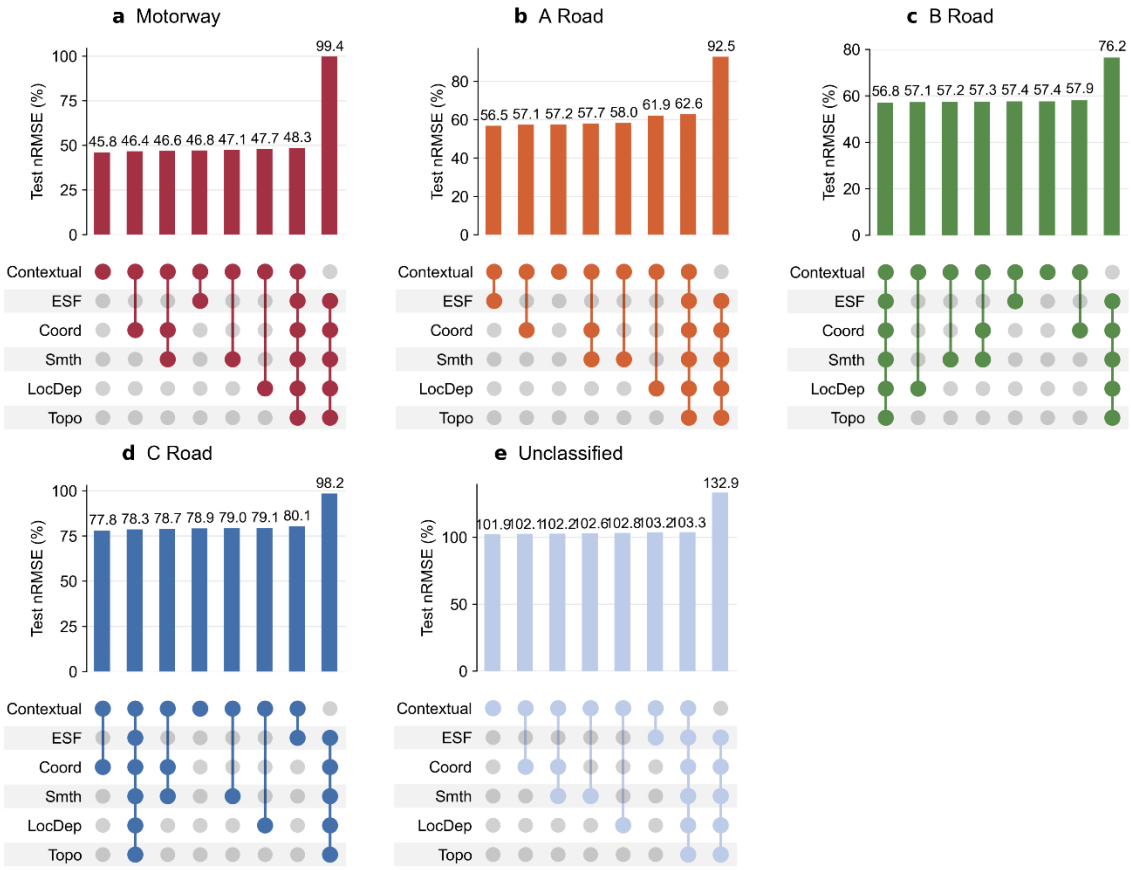

Figure S3-2. Test performance across spatial feature combinations by road class using spatial block CV. Each panel shows the test normalised RMSE (%) for models trained with different combinations of input features prior to feature selection, evaluated separately by road class under spatial block CV. All models are trained without hyperparameter tuning, target encoding, or custom loss to isolate the effect of spatial features. The matrix beneath each bar chart indicates the combinations of input features, including contextual, eigenvector spatial filtering (ESF), geographical coordinates and oblique coordinates (Coord), coordinate-based spatial smooth surface (Smth), local dependence including spatial lag of AADT and local Moran's I for key predictors (LocDep), and road network topology (Topo).

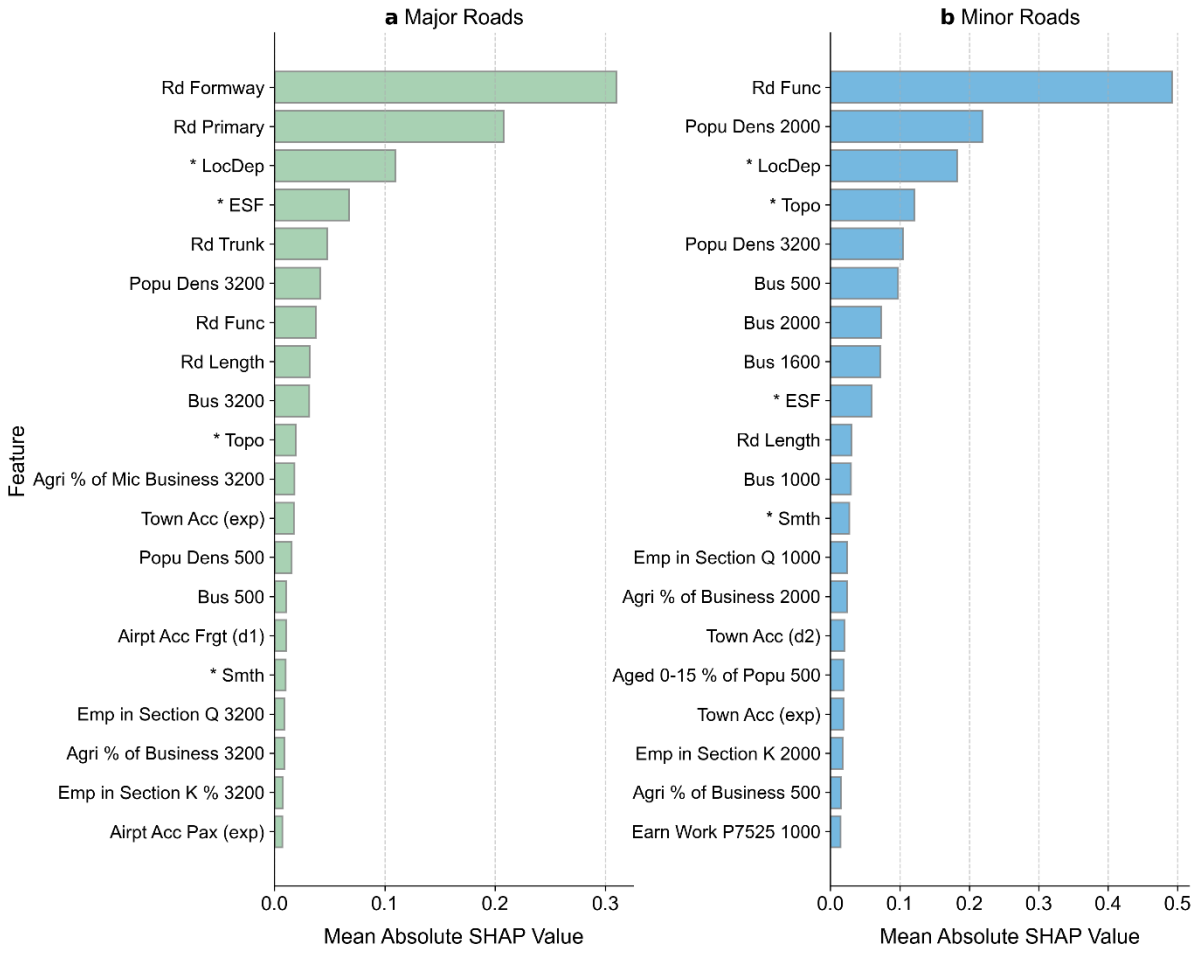

Figure S3-3. Global feature importance for predicting AADT on major roads (motorway and A roads) and minor roads (B, C, and unclassified roads) under spatial block CV. Bar plots show the top 20 features ranked by mean absolute SHAP value for (a) major roads and (b) minor roads. Spatial features are grouped by type, with SHAP values aggregated across all constituent features. Asterisks (*) denote spatial statistical features. *ESF* includes multiple eigenvector spatial filtering features, *LocDep* includes spatial lag of AADT and clustering metrics for key predictors, *Smth* refers to the coordinate-based spatial smooth surface, and *Topo* refers to topological indicators from the road network.

S4. Custom Loss Function and Hyperparameter Tuning

This section provides details on the custom loss function and the parameter tuning procedures used to optimise the performance of the LightGBM regression model.

To better reflect the practical objectives of traffic volume prediction, we implement a custom asymmetric loss function that penalises underestimation more heavily than overestimation. The function is implemented using the custom objective interface in the ***lightGBM*** library in python[30]. The loss function modifies the standard mean squared error by applying different weights to errors depending on whether the prediction is higher or lower than the true value:

$$\text{Loss}(y, \hat{y}) = \begin{cases} \lambda_{\text{under}} \cdot (y - \hat{y})^2 & if\ \hat{y} < y \\ \lambda_{\text{over}} \cdot (y - \hat{y})^2 & if\ \hat{y} \geq y \end{cases}$$

where y and $\hat{y}$ are the observed and predicted AADT, respectively; λ_{under} and λ_{over} are the penalty weights for underestimation and overestimation, respectively. In our study, we set $\lambda_{\text{under}} = 2$ and $\lambda_{\text{over}} = 1$, placing twice as much penalty on underestimation.

In addition to specifying a custom loss function, we apply Bayesian optimisation to tune a selected set of key hyperparameters in the LightGBM model. Specifically, we focus on 8 key hyperparameters, as shown in **Table S4-1**. The optimisation objective is to minimise the RMSE of the LightGBM model evaluated on an internal held-out set. Specifically, for each CV fold, Bayesian optimisation is performed on an 80/20 split of the training set, ensuring that hyperparameter tuning is conducted independently of the final test set to avoid information leakage. The optimisation procedure is implemented using the ***scikit-learn*** library in python[31].

In addition to the parameters shown in **Table S4-1**, we treat the number of boosted trees (*n_estimators*) as a dynamic hyperparameter. Instead of fixing or explicitly tuning its value, we rely on the early stopping function in the ***lightGBM*** library. Particularly, we set *early_stopping_rounds* = 70, meaning the model stops adding new trees if the RMSE does not improve for 70 consecutive iterations. This adaptive approach ensures that the ensemble size is optimally chosen based on validation performance, preventing overfitting while reducing unnecessary computational cost.

Table S4-1. Description and search domain of key hyperparameters for tuning.

Parameter [(a)]	Description	Search domain
learning_rate	Boosting step size	0.005 – 0.1, float
num_leaves	Max number of leaf nodes per tree	20 – 100, integer
max_depth	Maximum depth of each tree	3 – 10, integer
min_child_samples	Minimum samples per leaf	20 – 100, integer
subsample	Row sampling ratio per tree	0.6 – 1.0, float
colsample_bytree	Feature sampling ratio per tree	0.6 – 1.0, float
reg_alpha	L1 regularisation strength	1e-8 – 10, float
reg_lambda	L2 regularisation strength	1e-8 – 10, float

(a) The name of hyperparameters is aligned with those used in the lightGBM library in python.

References

1. Department for Transport. Road traffic statistics. https://roadtraffic.dft.gov.uk/downloads (2021).
2. Ordnance Survey. OS Open Roads. https://osdatahub.os.uk/downloads/open/OpenRoads (2021).
3. Sfyridis, A. & Agnolucci, P. Annual average daily traffic estimation in England and Wales: An application of clustering and regression modelling. *J. Transp. Geogr.* **83**, 102658 (2020).
4. Department for Environment Food & Rural Affairs. 2011 Rural Urban Classification. https://www.gov.uk/government/statistics/2011-rural-urban-classification-lookup-tables-for-all-geographies (2021).
5. Office for National Statistics. Built Up Areas (2022) GB BGG. https://geoportal.statistics.gov.uk/datasets/ons::built-up-areas-2022-gb-bgg/about (2023).
6. Office for National Statistics. Major Towns and Cities (Dec 2015) Boundaries V2. https://geoportal.statistics.gov.uk/datasets/ons::major-towns-and-cities-dec-2015-boundaries-v2/about (2023).
7. Office for National Statistics. Population estimates - small area based by single year of age - England and Wales. https://www.nomisweb.co.uk/datasets/pestsyoaoa (2021).
8. Office for National Statistics. Urban Audit FUA (Dec 2016) Full Extent Boundaries in the UK. https://geoportal.statistics.gov.uk/datasets/ons::urban-audit-fua-dec-2016-full-extent-boundaries-in-the-uk/about (2022).
9. Office for National Statistics. UK Business Counts - local units by industry and employment size band. https://www.nomisweb.co.uk/datasets/idbrlu (2021).
10. Office for National Statistics. Annual survey of hours and earnings - resident analysis. c (2021).
11. Office for National Statistics. Annual survey of hours and earnings - workplace analysis. https://www.nomisweb.co.uk/datasets/ashe (2021).
12. Office for National Statistics. Business Register and Employment Survey. https://www.nomisweb.co.uk/datasets/newbres6pub (2022).
13. Office for National Statistics. Population and household estimates, England and Wales: Census 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/populationandhouseholdestimatesenglandandwales/census2021unroundeddata (2022).
14. Department for Transport & Driver and Vehicle Licensing Agency. Vehicle licensing statistics. https://www.gov.uk/government/statistical-data-sets/vehicle-licensing-statistics-data-files (2022).
15. Department for Transport. National Public Transport Access Nodes (NaPTAN) and National Public Transport Gazetteer (NPTG). https://beta-naptan.dft.gov.uk/ (2023).

16. Department for Transport. Port and domestic waterborne freight statistics. https://www.gov.uk/government/statistical-data-sets/port-and-domestic-waterborne-freight-statistics-port (2022).

17. Civil Aviation Authority. Annual airport data 2021. https://www.caa.co.uk/data-and-analysis/uk-aviation-market/airports/uk-airport-data/uk-airport-data-2021/annual-2021/ (2023).

18. Borsetti, M. airportsdata. https://github.com/mborsetti/airportsdata (2023).

19. Graham, D. J. & Gibbons, S. Quantifying Wider Economic Impacts of agglomeration for transport appraisal: Existing evidence and future directions. *Econ. Transp.* **19**, 100121 (2019).

20. Pulugurtha, S. & Kusam, P. Modeling annual average daily traffic with integrated spatial data from multiple network buffer bandwidths. *Transp. Res. Rec.* **2291**, 53–60 (2012).

21. Gutiérrez, J. & García-Palomares, J. C. Distance-measure impacts on the calculation of transport service areas using GIS. *Environ. Plan. B Plan. Des.* **35**, 480–503 (2008).

22. Hagberg, A., Swart, P. & S Chult, D. Exploring network structure, dynamics, and function using networkx. in *Proceedings of the 7th Python in Science Conference (SciPy2008)* 11–15 (2008).

23. Gillies, S. *et al.* Shapely. (2023) doi:10.5281/ZENODO.7583915.

24. Bjørn Møller, A., Beucher, A. M., Pouladi, N. & Humlekrog Greve, M. Oblique geographic coordinates as covariates for digital soil mapping. *SOIL* **6**, 269–289 (2020).

25. Singh, U. *et al.* Hybrid multi-model ensemble learning for reconstructing gridded runoff of Europe for 500 years. *Inf. Fusion* **97**, 101807 (2023).

26. Jemeļjanova, M., Kmoch, A. & Uuemaa, E. Adapting machine learning for environmental spatial data - A reviews. *Ecological Informatics* vol. 81 102634 at https://doi.org/10.1016/j.ecoinf.2024.102634 (2024).

27. Griffith, D. & Chun, Y. Spatial autocorrelation and spatial filtering. in *Handbook of Regional Science* 1477–1507 (Springer, Berlin, Heidelberg, 2014). doi:10.1007/978-3-642-23430-9_72.

28. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

29. Mälicke, M., Möller, E., Schneider, H. D. & Müller, S. mmaelicke/scikit-gstat: A scipy flavoured geostatistical variogram analysis toolbox. (2021) doi:10.5281/ZENODO.4835779.

30. Ke, G. *et al.* LightGBM: A highly efficient gradient boosting decision tree. in *Advances in Neural Information Processing Systems* vols 2017-Decem 3147–3155 (2017).

31. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).