Using Machine Learning to Estimate Annual Average Daily Traffic by Vehicle Type on Local Roads with High-dimensional Geospatial Data

Liang Ma[a,*], Daniel J. Graham[a], Marc E.J. Stettler[a]

[a] Department of Civil and Environmental Engineering, Imperial College London, London, UK
liang.ma13@imperial.ac.uk, d.j.graham@imperial.ac.uk, m.stettler@imperial.ac.uk
* Corresponding author

Extended abstract submitted for presentation at the Conference in Emerging Technologies in Transportation Systems (TRC-30)
September 02-03, 2024, Crete, Greece

April 25, 2024

Keywords: Machine Learning; Spatial Prediction; Annual Average Daily Traffic; Feature Selection; High-dimensional Data.

1. INTRODUCTION

Understanding street-level annual average daily traffic (AADT) is important for various applications, such as developing air pollution and greenhouse gas emissions inventories (Ganji *et al.*, 2020). However, an exhaustive collection of AADT data across a country's road network is resource intensive. This often leads to a sampling approach in practice that prioritises major roads while underrepresenting minor roads (Department for Transport, 2022).

Various studies have explored methods to predict AADT in areas that have no direct data collection (Mathew & Pulugurtha, 2021; Selby & Kockelman, 2013). However, achieving reliable street-level AADT estimates remains challenging due to factors such as complex interactions among variables, insufficient input data, and model scalability to large networks. Moreover, existing methods often fail to simultaneously address the spatial heterogeneity in relationships among variables and spatial autocorrelation in road traffic data, further complicating accurate predictions and robust inferences.

Recent advances have been achieved in applying machine learning (ML) methods for AADT estimation (Ganji *et al.*, 2020; Sfyridis & Agnolucci, 2020). ML approaches generally have advantages in greater flexibility in model assumptions, enhanced capability to capture complex and nonlinear relationships, and better model scalability for large-scale applications (Fouedjio & Klump, 2019). However, ML algorithms do not recognise spatial context by default. Therefore, directly applying ML to geospatial data and using a traditional model evaluation metric may lead to biased results (Fouedjio & Klump, 2019).

This paper aims to apply a methodology that combines ML, spatial statistics, and extensive geospatial data to enhance AADT estimation and the assessment of spatial predictive ML models. Our approach uses a lightGBM model to estimate AADT in England and Wales (EW), incorporating over 900 spatial features with additional variables to account for spatial autocorrelation. The Boruta algorithm is applied to remove redundant features, proving effective in enhancing model performance. Unlike traditional evaluation methods, we use a cross-validation process tailored for spatial models. Our study demonstrates the potential of combining ML and spatial insights to provide effective and efficient AADT estimates at unmeasured locations and a reliable model evaluation. The AADT estimates are further split by vehicle and fuel type, thereby supporting pollution and carbon emissions estimation and offering insights for sustainable development.

2. METHODOLOGY

In this paper, we apply a ML framework to estimate street-level AADT in EW in 2021. The AADT data is obtained from open-source road traffic estimates, representing the number of vehicles passing through designated "count point" locations on an average day of the year (Department for Transport, 2021). The 2021 dataset includes 19,720 count points, divided into major roads (A roads and motorways) and minor roads (B, C, and unclassified roads). Among them, only 20% of count points pertained to minor roads, while minor roads constitute over 87% of the total road length in EW (Department for Transport, 2021). After mapping to the Ordnance Survey Road Network, the count points are further divided into training and test sets by a cross-validation (CV) approach to evaluate the performance of our model in locations without direct data collection.

2.1 Feature design and selection

To construct the predictive model, we extract over 900 spatial features from publicly available government data. The feature design (Figure 1) mainly follows Sfyridis & Agnolucci (2020) and is further enhanced by including spatial lags of AADT to consider spatial autocorrelation in road traffic. The spatial lag for count point i at order l is derived as a weighted sum of AADT values at its nearest l neighbours, with weights determined by a Gaussian kernel function (Liu, Kounadi & Zurita-Milla, 2022). Neighbour selection is determined by road class and Euclidean distance, considering computational efficiency and data availability. We calculate spatial lags for orders up to 4, noting that higher orders significantly expand the neighbourhood region.

In Figure 1, we outline two approaches to assign features to count points (i.e. roads). The first approach derives features based on the geographical location of count points, such as features for accessibility to urban areas and transport facilities with various impedance functions. The second approach assigns off-network characteristics, such as socioeconomic factors, to roads by creating buffers around count points. Following Sfyridis & Agnolucci (2020), we calculate six service areas of varying sizes around each count point using network distance rather than conventional Euclidean distance to better reflect real-world conditions. Specifically, the service area of each count point at radius r is derived by generating and then trimming a surrounding polygon of the roads that are reachable within a network distance r from the point. This process is closely adherent to the relevant function in ArcGIS Pro, yet is implemented using the ***NetworkX*** and ***Shapely*** library in python to handle the extensive road network in our study. Note that a feature is included in the model only if at least 75% of the count points have valid data.

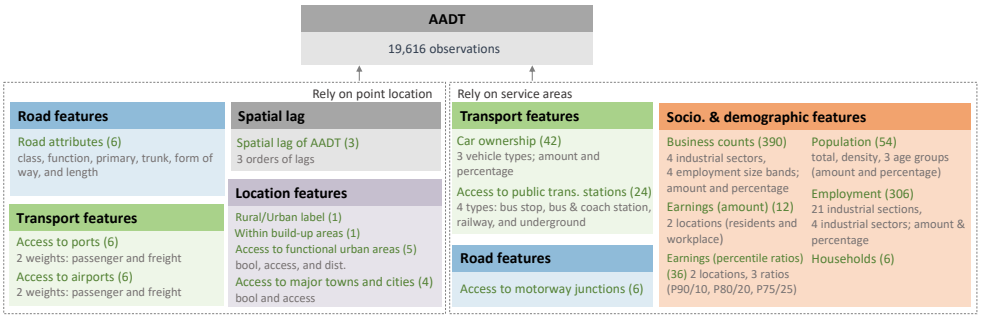

Figure 1 –*Feature design for the lightGBM model. Green text denotes feature categories, with the number of features shown in parentheses. Brief descriptions are provided in grey text.*

Dimension reduction is crucial in constructing high-dimensional ML models, as increased dimensionality of input features can compromise model accuracy, generalisation performance, and computational efficiency (Liu *et al.*, 2022). In this study, we utilise the Boruta algorithm for feature selection. The algorithm assesses feature importance by iteratively comparing original features with randomly shuffled features ("shadow features"). A feature is deemed important if it consistently outperforms all or the majority of shadow features (Kursa & Rudnicki, 2010). Our results highlight the efficacy of the feature selection process in improving model performance (see Section 3).

2.2 Model building and evaluation

Considering the nonlinearity and complexity of the relationships among model variables, we use lightGBM to predict AADT. As an advanced form of gradient boosting decision trees, lightGBM has demonstrated better performance than similar algorithms like XGBoost in computational speed and memory consumption (Ke *et al.*, 2017). We further enhance model performance by automatically tuning its key hyperparameters with Bayesian optimisation. To consider the heterogeneity in influencing factors for road traffic among different road classes, we develop separate models for major and minor roads. Our results show that tailored models for different groups of roads yield better performance compared to a single universal model (see Section 3).

Evaluating spatial models with a conventional random split between training and test sets has demonstrated overly optimistic results due to spatial autocorrelation (Hoffimann *et al.*, 2021). To assess model performance more accurately, we employ an *h*-block CV. We randomly divide the data into 10 folds and use each fold as a test set in turn. In contrast to the standard 10-fold CV, which uses the rest of the 9 folds as the training set, we further exclude data within a 4-step neighbouring radius of test data points from the training set for each test set. The model's performance averaging across all test sets is then considered as its performance in locations without direct data collection.

Furthermore, the output of the ML models provides AADT estimates for all motor vehicles, and the final step involves deriving conversion factors and applying them to the ML outputs. The conversion factors are determined by aggregating data from road traffic statistics and vehicle fleet composition projections (Department for Transport, 2021; National Atmospheric Emissions Inventory, 2023).

3. RESULTS AND DISCUSSION

Table 1 summarises the predictive performance across the four modelling scenarios in our study (with/without feature selection; universal/separate models). Optimal models for individual road classes show improved predictive accuracy when using feature selection and tailored models, especially for low-volume roads. We compare the performances of an identical model evaluated by distinct processes: the conventional 10-fold CV and the designated CV in our study. Our findings validate that conventional metrics tend to indicate more optimistic performance, especially for major roads, and emphasise the need for novel error estimation methods in spatial ML applications.

Table 1 – *Predictive Performance of Models (R-squared: %)*

Model	Feature set	Spatial CV (h-block) [a, b]					Random CV (k-fold) [a, b]				
		M Road	A Road	B Road	C Road	Unclass-ified	M Road	A Road	B Road	C Road	Unclass-ified
Universal	Full	43.9	**66.4**	48.5	43.9	29.3	44.8	69.5	48.9	50.3	27.3
	Selected[c]	33.2	63.8	20.9	46.7	-33.7	36.2	70.0	22.1	57.6	-30.8
Separate	Full	43.3	66.0	46.7	57.6	48.3	50.0	68.2	46.5	57.6	48.9
	Selected[c]	**47.6**	64.8	**49.5**	**58.7**	**49.4**	**58.3**	**71.1**	**49.4**	**59.7**	**50.3**

(a) R-squared is calculated at each test fold and then averaged across all folds, weighted by the total AADT of each fold.
(b) The optimal R-squared values among the modelling scenarios are highlighted in bold for each road class.
(c) Number of selected features: 140 in universal model; 160 for major roads and 107 for minor roads in separate models.

The sample size and complexity of network relationships in our study significantly exceed those typically found in existing literature. Our model performance is comparable to a similar-scale study (Mathew & Pulugurtha, 2021), which analysed over 12,000 roads in North Carolina with various geospatial and statistical models. As shown in Figure 2, the majority of count points in our study achieve better accuracy than the best mean absolute percentage error (82%) in Mathew & Pulugurtha (2021). Additionally, Sfyridis & Agnolucci (2020) predicted AADT in EW for 2016 using a clustering algorithm to categorise roads into groups before applying ML models to each group. Although their model showed better performance compared to ours, their use of the AADT value of the roads themselves in the clustering process restricts its application in truly unmeasured locations.

Our study introduces a ML framework to effectively and efficiently estimate street-level AADT by vehicle type across extensive road networks and emphasises the importance of spatial considerations in ML applications. We propose future efforts to enhance road clustering using data that is available at both locations with and without direct data collection and to address the stochastic components of AADT that remain unexplained by the ML model.

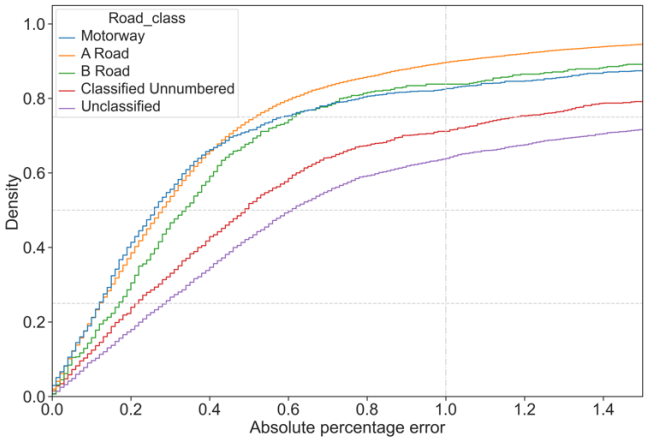

Figure 2 – *Cumulative histogram of absolute percentage error at individual count points (separate model; with feature selection). The absolute percentage errors are aggregated by road class.*

REFERENCES

Department for Transport (2022) *Minor road traffic estimates review: technical report*.

Department for Transport (2021) *Road traffic statistics*. [Online]. 2021. Available from: https://roadtraffic.dft.gov.uk/downloads [Accessed: 28 September 2022].

Fouedjio, F. & Klump, J. (2019) Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. *Environmental Earth Sciences*. [Online] 78 (1), 1–24. Available from: doi:10.1007/s12665-018-8032-z.

Ganji, A., Shekarrizfard, M., Harpalani, A., Coleman, J., et al. (2020) Methodology for spatio-temporal predictions of traffic counts across an urban road network and generation of an on-road greenhouse gas emission inventory. *Computer-Aided Civil and Infrastructure Engineering*. [Online] 35 (10), 1063–1084. Available from: doi:10.1111/mice.12508.

Hoffimann, J., Zortea, M., de Carvalho, B. & Zadrozny, B. (2021) Geostatistical Learning: Challenges and Opportunities. *Frontiers in Applied Mathematics and Statistics*. [Online] 7, 689393. Available from: doi:10.3389/fams.2021.689393.

Ke, G., Meng, Q., Finley, T., Wang, T., et al. (2017) LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems*. 2017 pp. 3147–3155.

Kursa, M.B. & Rudnicki, W.R. (2010) Feature selection with the boruta package. *Journal of Statistical Software*. [Online] 36 (11), 1–13. Available from: doi:10.18637/jss.v036.i11.

Liu, X., Kounadi, O. & Zurita-Milla, R. (2022) Incorporating Spatial Autocorrelation in Machine Learning Models Using Spatial Lag and Eigenvector Spatial Filtering Features. *ISPRS International Journal of Geo-Information*. [Online] 11 (4), 242. Available from: doi:10.3390/ijgi11040242.

Liu, X., Tang, H., Ding, Y. & Yan, D. (2022) Investigating the performance of machine learning models combined with different feature selection methods to estimate the energy consumption of buildings. *Energy and Buildings*. [Online] 273, 112408. Available from: doi:10.1016/j.enbuild.2022.112408.

Mathew, S. & Pulugurtha, S.S. (2021) Comparative Assessment of Geospatial and Statistical Methods to Estimate Local Road Annual Average Daily Traffic. *Journal of Transportation Engineering, Part A: Systems*. [Online] 147 (7), 04021035. Available from: doi:10.1061/jtepbs.0000542.

National Atmospheric Emissions Inventory (2023) *Emission factors for transport*. [Online]. 2023. Available from: https://naei.beis.gov.uk/data/ef-transport [Accessed: 14 August 2023].

Selby, B. & Kockelman, K.M. (2013) Spatial prediction of traffic levels in unmeasured locations: Applications of universal kriging and geographically weighted regression. *Journal of Transport Geography*. [Online] 29, 24–32. Available from: doi:10.1016/j.jtrangeo.2012.12.009.

Sfyridis, A. & Agnolucci, P. (2020) Annual average daily traffic estimation in England and Wales: An application of clustering and regression modelling. *Journal of Transport Geography*. [Online] 83, 102658. Available from: doi:10.1016/j.jtrangeo.2020.102658.