Short-term lagged interactions between freight and passenger volumes in urban traffic: inter- and intra-modal effects with explainable machine

Transportation Research Part A 206 (2026) 104927

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Short-term lagged interactions between freight and passenger volumes in urban traffic: inter- and intra-modal effects with explainable machine learning

E. Amirnazmiafshar [a], D.P. Song [a,*], B. Kenny [b], J.M. Wu [c], B. Kulcsár [c], Y.Z. Liu [d], C. Olaverri-Monreal [d]

[a] University of Liverpool, UK
[b] ESG Consultants Ltd, UK
[c] Chalmers University of Technology, Sweden
[d] Johannes Kepler University, Austria

ARTICLE INFO

Keywords:
Freight-passenger interaction
Urban traffic management
Multimodal transport planning
Short-term demand forecasting
Lagged traffic volumes
Explainable machine learning

ABSTRACT

Urban transport systems face increasing complexity as freight and passenger flows compete for limited road capacity. While multimodal forecasting methods have progressed, short-term interactions between vehicle classes remain underexplored, particularly in real-world operational settings. This study addresses that gap by examining whether recent freight or passenger volumes are significantly associated with current traffic conditions across modes. Using 6,003 hourly records from Liverpool, UK, we develop an interpretable machine learning framework combining K-means clustering, XGBoost classification, and the DALEX explainability toolkit. Results show that one-hour lagged freight volume significantly improves the classification of current passenger traffic states, while the reverse effect is limited. Global feature importance and local interpretability analyses consistently identify freight volume as the most influential predictor. Partial dependence plots (PDPs) reveal a nonlinear inflexion point, where freight volumes exceeding roughly 500 vehicles per hour in this Liverpool case study are associated with reduced passenger flow. McNemar's test confirms a statistically significant improvement, and robustness checks, including alternative lag structures, interaction terms, and reciprocal models, reinforce the stability of this finding. These insights offer practical value for short-term forecasting, corridor-level coordination, and longer-term multimodal planning. The observed directional asymmetry, wherein freight volumes more reliably predict passenger conditions than the reverse, highlights the potential benefits of incorporating freight data into real-time traffic management systems. More broadly, the study demonstrates how interpretable machine learning can uncover cross-modal dependencies and support the development of more integrated, responsive, and equitable urban mobility systems.

1. Introduction

Urban transport networks today face increasingly complex challenges due to the simultaneous rise in passenger and freight

* Corresponding author.
E-mail address: Dongping.Song@liverpool.ac.uk (D.P. Song).

https://doi.org/10.1016/j.tra.2026.104927
Received 6 August 2025; Received in revised form 20 January 2026; Accepted 10 February 2026
Available online 14 February 2026

mobility demands. As e-commerce and just-in-time logistics continue to expand, freight movement (e.g., via light goods vehicles and heavy goods vehicles) has become more temporally dispersed and spatially embedded within the urban landscape (Safikhani et al., 2020). At the same time, cities must accommodate growing levels of passenger travel (e.g. via cars, taxis, buses, and coaches), often using the same road infrastructure. This co-existence intensifies congestion, reduces operational efficiency, and complicates both short-term traffic control and long-term infrastructure planning (Khiari and Olaverri-Monreal, 2020; Alessandretti et al., 2023; Tanwar and Agarwal, 2024).

Despite ongoing advancements in traffic modelling and demand forecasting, most approaches continue to treat freight and passenger traffic as separate domains. However, in dense urban environments, where infrastructure is shared, these flows inevitably interact with one another. For instance, delivery vehicles may obstruct traffic lanes, reducing the speeds of passenger vehicles or affecting mode choice. Conversely, peak passenger demand may hinder freight delivery operations during time-sensitive windows. A limited understanding of these short-term, reciprocal relationships constrains the effectiveness of operational interventions and predictive traffic systems (Kim and Cho, 2022; Huang et al., 2023).

This research addresses a critical yet understudied issue in urban transportation by examining whether a short-term increase in the volume of one vehicle class, such as freight or passenger, is significantly associated with changes in the volume of the other vehicle class. Although recent studies have advanced multimodal demand forecasting, most focus on within-mode patterns, such as autoregressive behaviour or cross-effects between public transport services, like subways and bike-share systems (Hua et al., 2024; Liang et al., 2022a, 2022b). In contrast, the short-term interdependence between freight vehicles and passenger flows on shared roadways remains unexamined, especially in real-world operational settings.

To bridge this gap, this paper develops a data-driven framework to investigate the temporal associations between lagged freight traffic and subsequent passenger vehicle volumes, and vice versa, using a case study of major roads in Liverpool, UK. The proposed approach integrates unsupervised clustering, supervised machine learning (XGBoost), and interpretable AI techniques (DALEX (Descriptive mAchine Learning EXplanations)) to assess whether lagged volumes in one category can improve the prediction of traffic levels in the other. This study not only contributes to improved short-term traffic forecasting accuracy but also provides actionable policy insights for congestion management, delivery scheduling, and multimodal infrastructure planning.

This study uniquely integrates the DALEX explainability framework with XGBoost classification to interpret cross-modal interactions between freight and passenger traffic. To our knowledge, this is among the first applications of DALEX in multimodal traffic modelling, and the first such application using real-world urban traffic data in the UK. This approach yields interpretable outputs that are critical for transparent, policy-relevant decision-making in multimodal transport planning.

To address this gap, the study is guided by two research questions:

- Does a short-term lag in freight volume (i.e., one hour earlier) significantly associate with current passenger or freight traffic states?
- Does lagged passenger activity (i.e., one hour earlier) significantly associate with current passenger or freight traffic states?

These questions aim to uncover directional dependencies between vehicle classes and support more responsive and transparent traffic forecasting in shared urban road environments.

The remainder of this article is structured as follows. Section 2 reviews the relevant literature on traffic classification, time-series forecasting with lagged inputs, and multimodal interaction modelling. Section 3 details the data sources and feature engineering processes used to construct the analytical dataset. Section 4 outlines the methodology, including clustering, supervised classification, and interpretability analysis. Section 5 presents the results, covering model performance, statistical validation, and robustness checks. Section 6 discusses the findings in terms of methodological, operational, and policy implications. Finally, Section 7 concludes the paper and suggests directions for future research.

2. Literature review

This section reviews three areas of relevant research that underpin the present study: traffic classification techniques, time-series forecasting with lagged inputs, and modelling of multimodal or cross-modal traffic interactions. These strands of literature collectively inform the framework developed to investigate short-term freight-passenger dependencies in shared urban road infrastructure.

2.1. Traffic classification

Infrastructure investment, congestion control, and operational responsiveness are central to effective transport system management. These functions can be enhanced through continuous traffic monitoring, such as the collection and analysis of video streaming data in traffic management centres (Allamehzadeh et al., 2017). Classification of traffic volume levels is foundational to transport system management, informing decisions related to infrastructure investment, congestion mitigation, and operational planning. Traditional discretisation methods, such as equal-interval and equal-frequency (quantile-based) binning, are straightforward to implement but often fail to capture real-world variability, particularly in the presence of skewed or multimodal data distributions (Dougherty et al., 1995). A variety of clustering algorithms have been applied to traffic flow classification, including hierarchical clustering (Li and Rose, 2011), Gaussian mixture models (Liu et al., 2016), and density-based methods such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Yang et al., 2022). Partition-based algorithms, particularly K-means, have also been widely used for empirical grouping of traffic observations and congestion-state detection (Esfahani et al., 2019; Rao et al., 2019; Montazeri-Gh and Fotouhi, 2011). Other studies have extended clustering frameworks through fuzzy or time-series-based methods,

such as Fuzzy C-Means (Silgu and Celikoglu, 2015) and Gaussian-Weighted Dynamic Time Warping combined with k-medoids (Li et al., 2022). These clustering approaches collectively enable data-driven identification of latent traffic states that better capture dynamic patterns than rule-based classification.

Following unsupervised segmentation, a range of supervised machine learning algorithms has been employed for traffic-class prediction, including Random Forests (Zafar and Ul Haq, 2020), Support Vector Machines (Cheng et al., 2020), deep-learning frameworks such as CNN, LSTM, and CoAtNet (Wang et al., 2022; Yuan et al., 2025), and gradient-boosted tree ensembles like XGBoost (Chen and Guestrin, 2016; Zhang and Haghani, 2015; Ke et al., 2021). These methods have become common tools for modelling complex, nonlinear relationships in structured traffic datasets, enabling both accurate prediction and interpretable insights into multimodal flow dynamics.

2.2. Time-series forecasting and lag effects

Short-term traffic demand is characterised by strong temporal dependencies, where recent patterns often shape current conditions. Incorporating lagged features into forecasting models is a common strategy for improving predictive accuracy. Lag variables, such as traffic volumes from the previous hour or the same hour on a prior day, help capture autocorrelation and reflect temporal inertia in driver behaviour. Multiple studies have demonstrated the value of such lagged inputs in enhancing model performance across various methods (Koesdwiady et al., 2016; Jiber et al., 2020; Kumar and Vanajakshi, 2015).

Huang et al. (2023) developed a hybrid Wavelet-LSTM model that incorporated vehicle functional attributes, such as distinguishing between freight and passenger vehicles, and lagged flow data to enhance short-term traffic forecasts. While their model targeted total traffic flow, their approach underscores the potential predictive value of class-specific vehicle dynamics. Similarly, Liang et al. (2022a) and Kim and Cho (2022) confirmed that temporal lags capture fluctuations in travel demand more effectively than static features alone.

These studies underscore the value of lagged inputs in short-term urban traffic forecasting and motivate the current study's use of freight and passenger lags to model cross-modal interactions.

2.3. Multimodal interaction modelling

Beyond within-mode autoregressive models, there is growing interest in capturing cross-modal dependencies, particularly relevant in multimodal transport systems and shared-use corridors. Hua et al. (2024) found that prior-hour metro and taxi usage substantially improved bike-share demand prediction, indicating substitutive or complementary modal relationships. Kim and Cho (2022) similarly identified causal connections among subway, bus, and bike-share usage patterns.

Recent advances in deep learning and network science have further extended this line of inquiry. Graph neural networks (GNNs) have been utilised to model spatiotemporal dependencies across various transport modes, yielding notable improvements in predictive accuracy (Liang et al., 2022b; Wang et al., 2022). Other approaches include multi-task learning and memory networks to handle demand imbalances between high- and low-frequency services such as buses and ferries (Li et al., 2021). These developments contribute to a growing body of work focused on temporal, spatial, and intermodal modelling within complex transport systems (Ma et al., 2022; Wang et al., 2024). Alessandretti et al. (2023) further frame these systems as multilayer mobility networks, reinforcing the need for methods that can account for real-time interactions between vehicle classes.

Despite ongoing efforts to model multimodal interactions, most existing studies prioritise predictive accuracy over interpretability, making it difficult for planners to extract actionable insights from model outputs. While techniques like XGBoost and deep learning have improved performance, they often function as "black boxes." Very few studies have explored explainable AI methods, such as DALEX, to interpret the directional associations of lagged freight or passenger volumes on multimodal traffic conditions. This gap limits the integration of advanced forecasting into policy-relevant decision-making frameworks. Our study addresses this by combining supervised learning and explainable AI to offer both accurate and interpretable forecasts of freight-passenger interactions on urban roads.

2.4. Model interpretability and behavioural insights

Model interpretability has become increasingly important for understanding the internal logic of machine learning models used in transport applications. SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016) are two widely used algorithms that generate local explanations of predictions, quantifying each feature's contribution to a specific output instance. In contrast, DALEX (Biecek, 2018) is a *meta*-framework that integrates multiple interpretability techniques, including SHAP, LIME, and permutation-based importance, within a unified and model-agnostic environment. DALEX allows global and local analyses to be applied consistently across different models, facilitating transparent interpretation of feature effects in complex predictive frameworks.

Additional behavioural research complements these computational tools by offering insight into user movement across modal boundaries. For instance, Lanza et al. (2022) highlighted that micromobility users and pedestrians often deviate from expected patterns, which can complicate forecasting based purely on mode-level aggregation. Vision-based studies, such as those by Yang et al. (2021), have demonstrated that congestion at urban intersections can be accurately detected using roadside surveillance video. Their method combines vehicle detection from YOLOv3 (a real-time object detection algorithm) with optical flow analysis based on bounding box corners to estimate vehicle speeds and assess traffic states within the region of interest.

These findings underscore the importance of integrating predictive modelling with transparent explanations and behavioural

understanding, particularly in systems where multiple vehicle classes interact over shared space and time.

3. Data source and feature engineering

This study integrates official traffic statistics with spatial and environmental variables to construct an empirical dataset for examining the temporal interactions between freight and passenger vehicle flows on key urban corridors.

Liverpool was selected as the case study. This city hosts the Port of Liverpool, one of the country's largest gateways for both international and domestic freight. Simultaneously, it is a densely populated urban area with high passenger mobility and chronic congestion. This duality makes Liverpool an ideal context for analysing the interdependence between freight and passenger transport. Liverpool, as a major port city in the UK, presents a unique urban setting with a high mix of freight and passenger traffic. This characteristic makes it an ideal case study for examining freight-passenger interactions on shared road infrastructure, particularly in corridors affected by port-related logistics activity. According to the UK Department for Transport, the combined vehicle miles travelled on Liverpool's roads reached 1.35 billion in 2023 (UK Department for Transport, 2025).

The primary dataset was sourced from the Road Traffic Statistics portal maintained by the Department for Transport, specifically the "Raw Counts" dataset for the Liverpool local authority. This dataset includes 35,076 hourly records spanning from 2000 to 2023, collected across 394 counting sites. Each record comprises hourly vehicle counts, along with road segment metadata such as road name, type, segment length, traffic direction, and geographic coordinates (provided for one end of each segment). Vehicle types are reported in a disaggregated format, including pedal cycles, two-wheeled motor vehicles, cars and taxis, buses and coaches, light goods vehicles (LGVs), and multiple categories of heavy goods vehicles (HGVs), differentiated by axle configuration.

To ensure temporal and spatial consistency, preprocessing steps filtered the data to include only weekday (Monday-Friday, excluding holidays) observations during daytime working hours (07:00–18:00). Minor roads were excluded due to a lack of reliable land-use metadata, resulting in a refined dataset of 17,340 hourly records covering only major roads.

To enhance model interpretability and analytical tractability, vehicle types were grouped into two broad categories: freight vehicles (including LGVs and all HGVs) and passenger vehicles (comprising cars, taxis, buses, coaches, two-wheeled motor vehicles, and pedal cycles). This aggregation strikes a balance between practical and theoretical considerations. Practically, it mitigates data sparsity and reduces model dimensionality. Theoretically, it aligns with the conceptual distinction between goods movement and people mobility, the central focus of this research.

Land-use characteristics were extracted from OpenStreetMap (OSM) polygon data using a bounding box covering the Liverpool metropolitan area. Spatial processing was performed in R, where the origin and destination coordinates of each traffic record were matched to the nearest OSM land-use polygon within a 500-meter radius using geospatial proximity queries. When no polygon was found within this range, the most frequently occurring land-use category in the dataset was used as a proxy for the missing polygon. The original OSM land-use labels were then reclassified into nine general categories: residential, commercial, industrial, green spaces, transport, public services, agriculture, miscellaneous, and other. This geospatial enrichment process was conducted using reproducible workflows in R, primarily based on the sf and osmdata packages.

Traffic speed data were obtained using the TomTom Traffic Flow API, based on the origin coordinates of each road segment. Due to the lack of historical speed data, a proxy approach was employed: real-time speed data were collected across two consecutive work weeks (ten weekdays total), with hourly readings from 07:00 to 18:00. Given the observed stability in traffic conditions during the collection period, the average speed for each hour across the ten days was computed and used as a representative value for that hour and location throughout the dataset.

Weather data were sourced from the Open-Meteo Archive API. For each trip date, daily maximum temperature, minimum temperature, and total precipitation were collected based on Liverpool city centre coordinates (latitude 53.4084, longitude −2.9916). Although these data reflect daily rather than hourly conditions, they provide a reasonable approximation of the environmental influences on travel behaviour.

As only the origin coordinates were available for each segment, destination points were estimated using spherical geometry, based on segment length and direction. These estimates enabled the generation of additional spatial features, including trip direction (inward vs. outward) and proximity to the city centre. Trip direction was classified as "Towards City" if the destination was closer to the geographic centre of Liverpool (latitude 53.40478, longitude −2.98104) than the origin; otherwise, it was marked as "Outward." A binary variable was also created to flag peak travel periods, defined as 07:00–09:00 and 16:00–19:00.

Several engineered features were introduced to capture temporal dependencies in traffic behaviour. Total passenger volume was computed as the sum of all relevant vehicle types and used as the response variable in clustering and classification tasks. In the main specification, the passenger-traffic variable aggregates cars and taxis, buses and coaches, two-wheeled motor vehicles, and pedal cycles, representing total person-movement on shared road space. Pedal cycles were included to maintain consistency with the UK Department for Transport (DfT) traffic-count classification, in which vehicle flows are measured as the number of vehicles per hour per location, and to reflect the mixed-traffic conditions typical of Liverpool, where cyclists commonly share the same carriageway with motorised vehicles and can influence overall traffic speed and flow capacity. To model autocorrelation, lagged features were created, most notably, previous_hour_goods_total and previous_hour_passenger_total, which represent the total number of freight and passenger vehicles, respectively, during the previous hour. These variables were constructed by sorting the data by road segment, traffic direction, and time, and then shifting counts by one time step.

As lagging introduces missing values at the beginning of each road segment group, an imputation strategy was applied. Missing lagged values were filled using the median freight or passenger volume for that road segment to preserve temporal structure without discarding data. In addition to one-hour lags, rolling two-hour averages (previous_2hr_avg_goods and previous_2hr_avg_passenger)

were also created. These were retained only if the corresponding one-hour lag showed a statistically significant association with the target variable, supporting a parsimonious model (i.e., one that maintains simplicity without sacrificing predictive power).

Categorical variables, such as land-use at origin and destination, peak-period status, and trip direction, were one-hot encoded to ensure compatibility with supervised learning algorithms. These categorical variables were combined with continuous variables, including vehicle counts, average traffic speeds, weather indicators, and spatial metrics, to form the final feature set. The resulting dataset comprises 6,003 hourly observations, each enriched with a diverse set of structured and engineered features. These include temporal indicators, vehicle flows, spatial attributes, weather conditions, and contextual road characteristics. This dataset forms the empirical basis for investigating both contemporaneous and lagged interactions between freight and passenger traffic on Liverpool's principal roads.

4. Methodology

This section outlines the modelling framework developed to investigate short-term interactions between freight and passenger

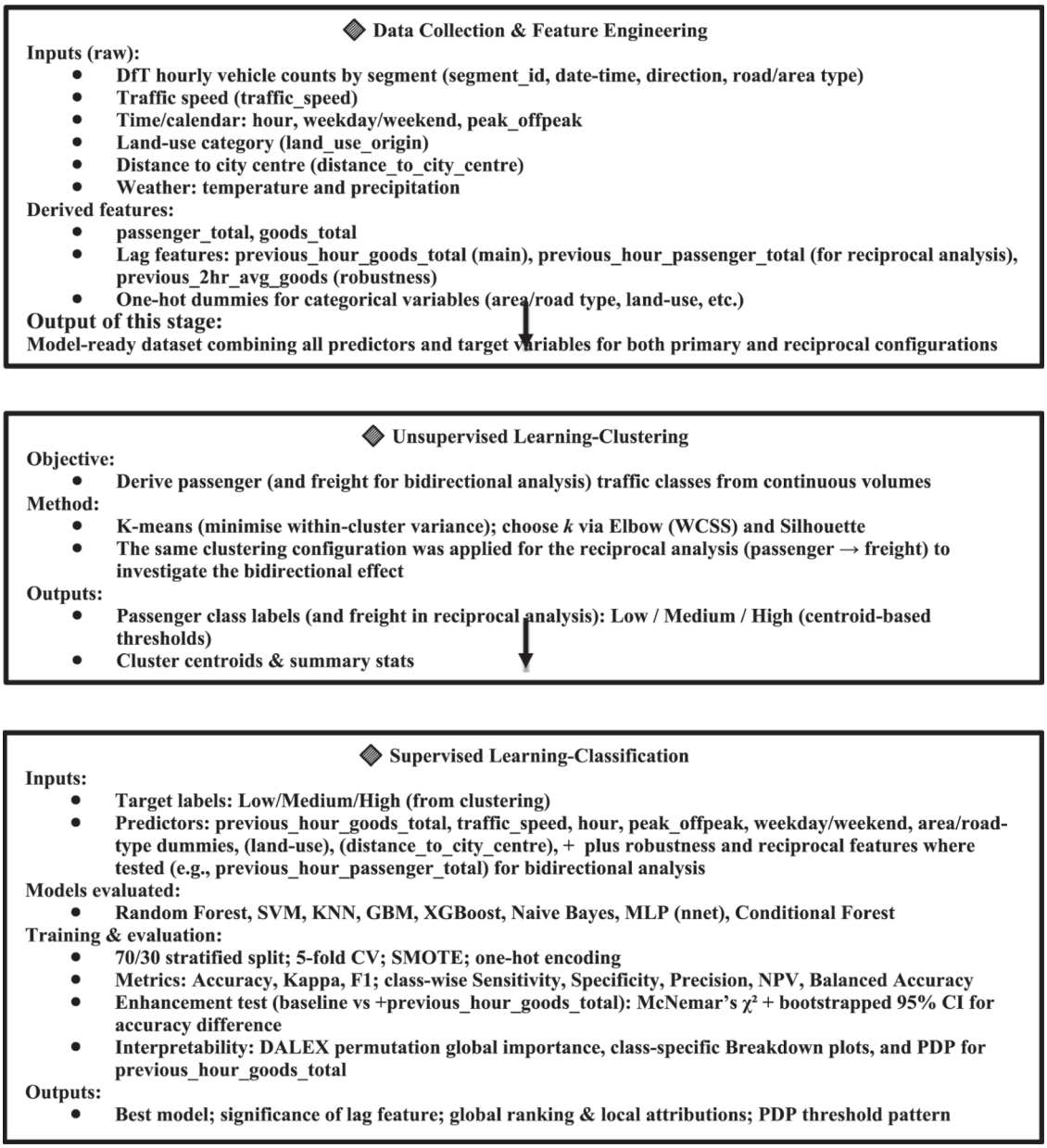

Fig. 1. Overview of the data processing and modelling framework.

traffic using machine learning techniques. Short-term interaction is defined here as the temporal association between past freight activity (freight traffic volume in the previous hour) and current passenger traffic conditions. As illustrated in Fig. 1, the methodology proceeds through several sequential stages. It begins with an unsupervised learning step, where K-means clustering is used to transform continuous passenger volume data into interpretable traffic categories. These categories serve not as ground-truth labels, but as meaningful typologies that reflect latent structure in traffic patterns.

Next, supervised classification models are trained to predict these categories based on explanatory features such as freight traffic and traffic speed. This enables operational forecasting and interpretability analysis by linking real-time variables to the previously defined traffic states. The methodology then incorporates lagged variables to examine whether freight traffic from a previous hour is associated with current passenger conditions. This is followed by a statistical evaluation to determine whether the inclusion of such lagged features yields a measurable improvement in classification performance. Finally, model interpretability is assessed using the DALEX framework, and robustness is tested through additional model extensions. To assess the robustness of the modelling results, a sensitivity analysis was also conducted by redefining the passenger target to include only motorised modes (excluding pedal cycles) while keeping the same feature set, XGBoost configuration, cross-validation, and evaluation process unchanged.

Fig. 1 provides an overview of the data processing and modelling framework.

Fig. 2 illustrates the conceptual framework guiding the investigation of bidirectional short-term dependencies between passenger and freight traffic, consistent with the study's two research questions. The first question examines whether freight activity one hour earlier (t-1) is significantly associated with current passenger or freight traffic states, capturing both cross-modal and intra-modal effects originating from freight flows. The second question mirrors this analysis from the opposite perspective, assessing whether lagged passenger activity influences current passenger or freight conditions. As shown in Fig. 2, the framework models four pathways:

- Freight (t-1) → Passenger (t);
- Passenger (t-1) → Freight (t);
- Passenger (t-1) → Passenger (t);
- Freight (t-1) → Freight (t)

By jointly analysing these intra- and cross-modal dependencies, the model captures both the reciprocal coupling and self-reinforcing temporal continuity that shape short-term multimodal traffic dynamics in urban networks.

4.1. *Discretising traffic volume using k-means clustering*

Among alternative clustering algorithms (e.g., Gaussian Mixture Models, DBSCAN, fuzzy C-means), K-means was selected for its computational efficiency, interpretability, and established performance in traffic segmentation. This choice aligns with the study's objective of producing compact, operationally meaningful traffic states rather than probabilistic cluster memberships.

To enable the application of supervised classification models and enhance interpretability, the continuous variable representing passenger traffic volume was discretised into categorical levels at the road segment level and on an hourly basis. The exact process was applied separately to freight traffic volumes. Discretisation facilitates the use of multi-class classification algorithms and captures stronger nonlinear relationships between explanatory features and traffic conditions. Moreover, categorical traffic levels provide more actionable and policy-relevant insights than raw volume values, as discrete categories are easier to communicate and respond to in operational contexts.

Since the optimal number of traffic categories was not known a priori, an unsupervised learning approach was used to identify natural groupings within the data. Specifically, the K-means clustering algorithm was selected due to its wide use and proven effectiveness in transportation segmentation tasks. K-means seeks to partition the data into k clusters by minimising the total within-cluster variance. Unlike equal-interval or quantile-based binning methods, K-means is a data-driven, scalable approach that is well-suited for identifying latent structure in skewed or multimodal distributions, which are common characteristics in urban traffic data.

The K-means algorithm seeks to minimise within-cluster variance by solving the following objective function, formally defined in

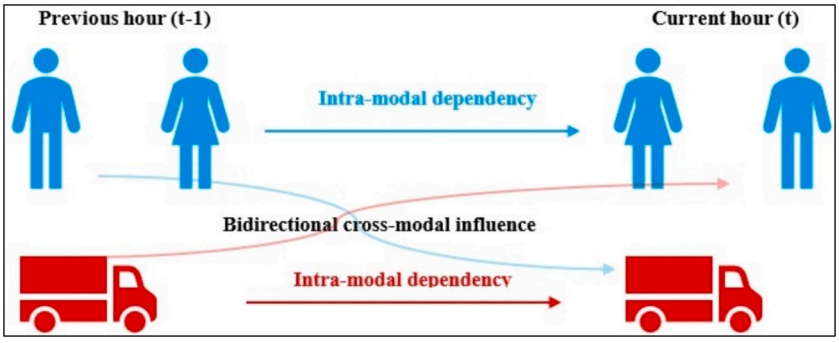

Fig. 2. Conceptual framework illustrating the bidirectional short-term (one-hour lag) interactions between passenger and freight traffic across intra- and cross-modal pathways.

Equation (1) (MacQueen, 1967) as:

$$\underset{S}{argmin} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{1}$$

where S_i represents the set of data points assigned to the cluster i and μ_i is the centroid of that cluster. x denotes an individual data point, k is the number of clusters, $S = \{S_1, \cdots, S_k\}$ is the full partition of the dataset, $\|\cdot\|$ indicates the Euclidean norm, and the argmin operator selects the clustering configuration that minimises the total within-cluster variance. This objective minimises intra-cluster distances, resulting in compact and well-separated groupings.

To determine the ideal number of clusters, two unsupervised validation methods were employed: the Elbow Method and Silhouette Analysis. The Elbow Method evaluates the within-cluster sum of squares (WCSS) across different values of k, while the Silhouette Analysis assesses average cohesion and separation between clusters. The empirical results and final clustering configuration are presented in Section 5.1. Based on these results, the identified clusters were used to define traffic classes for subsequent supervised learning tasks.

To ensure a comprehensive assessment of bidirectional and intra-modal dependencies, the modelling framework was applied across four short-term configurations: (1) passenger(t-1) → passenger(t), (2) passenger(t-1) → freight(t), (3) freight(t-1) → passenger(t), and (4) freight(t-1) → freight(t). Each configuration was implemented using the same two-step (clustering-classification) structure, with consistent preprocessing, evaluation, and statistical comparison procedures.

4.2. Predicting traffic categories via supervised classification

Although the clustering step assigns each observation to a traffic category, these labels are generated solely based on passenger volume, without reference to explanatory variables. The supervised models do not aim to replicate these labels but rather to learn how other features, such as freight traffic volume and traffic speed, relate to the cluster-defined categories. This enables predictive modelling using real-time inputs and allows for further exploration of relationships, such as the associative contribution of lagged freight traffic on passenger volume predictions. In this way, the unsupervised step provides meaningful class definitions, while the supervised models translate them into an operational framework for short-term forecasting and interpretability analysis.

To train and evaluate the supervised models, the dataset was randomly split into a training set (70%) and a test set (30%) using stratified sampling to preserve the distribution of traffic categories. The training process included 5-fold cross-validation, with SMOTE (Synthetic Minority Over-sampling Technique) applied to address class imbalance. Multiple models were evaluated using both overall and per-class metrics, and final performance was assessed on the held-out test set (Fernández et al., 2018). Although the learned models were trained with SMOTE on the training folds to mitigate minority under-representation, the natural class imbalance among the clustered traffic categories (see Section 5.1) remains in the data and in all held-out evaluations; therefore, the task is imbalanced by definition. We accordingly report macro- and per-class metrics, including balanced accuracy, and confirmed similar conclusions with class-weighted XGBoost and under-sampling sensitivity checks.

Building on these clusters, a range of supervised classification models was tested to predict the membership of traffic levels. The models included a diverse set of algorithms commonly used in traffic demand forecasting: tree-based models (Random Forest, GBM, XGBoost, cforest), kernel-based models (SVM), ensemble methods, probabilistic models (Naive Bayes), and neural networks (multi-layer perceptron with a single hidden layer, implemented via the nnet package in R).

The regularised objective function minimised by XGBoost, which governs classification model training, is defined in Equation (2) (Chen and Guestrin, 2016):

$$\mathscr{L}(\phi) = \sum_{i=1}^{n} \ell\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{t=1}^{T} \Omega(f_t)$$

$$where \Omega(f_t) = \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} w_j^2 \tag{2}$$

where,

ℓ is the loss function (e.g., softmax for multi-class classification), T_t is the number of leaves in the tree f_t, Ω is the regularisation term, w_j are the leaf weights of the tree f_t, γ and λ are regularisation coefficients. Here, n denotes the number of training samples, y_i is the true label for sample i, $\hat{y}_i^{(t)}$ is the model's prediction at iteration t, T is the total number of boosted trees, f_t represents the t-th tree in the ensemble, and ϕ denotes the full set of model parameters. This regularised framework enables XGBoost to control overfitting and maintain computational efficiency, making it particularly well-suited for high-dimensional, structured datasets such as urban traffic observations.

To benchmark predictive performance, the candidate models were evaluated using both overall and per-class metrics, including Accuracy, F1 Score, Cohen's Kappa, Sensitivity, Specificity, Precision, Negative Predictive Value (NPV), and Balanced Accuracy. Full results are presented in Section 5.2. Final model selection prioritised not only raw accuracy but also consistency across traffic classes, interpretability, and operational applicability. Based on this benchmarking framework, XGBoost was subsequently selected as the core predictive model for the lag-effect and interpretability analyses presented in later sections.

Because this is a three-class problem with a moderately imbalanced distribution (see Section 5.1), overall Accuracy alone can mask class-specific behaviour. We therefore emphasise Cohen's Kappa (chance-corrected agreement), macro-averaged F1 (balances Precision and Recall across classes), and Balanced Accuracy (averages Sensitivity across classes). Sensitivity reflects the model's ability to avoid missed congestion events, while Precision and Specificity capture its capacity to limit false alarms.

Negative Predictive Value (NPV) is formally defined as the proportion of true negatives among all predicted negatives. While originally developed for binary classification, it is extended here to the three-class problem by computing NPV separately for each traffic category (one-vs-all formulation). This provides a class-specific measure of how reliably the model identifies the absence of a given traffic condition. For example, a high NPV for the Medium class indicates that when the model predicts non-Medium conditions (i.e., free-flow or heavy-congestion states), that prediction can be trusted, helping to avoid unnecessary or premature operational responses.

Balanced Accuracy, together with Cohen's Kappa, ensures consistent reliability and chance-corrected agreement under imbalanced conditions. Accuracy provides an overall measure of correct classification across all traffic states, whereas the F1 Score integrates Precision and Recall to represent the harmonic balance between false negatives and false positives. These evaluation metrics were selected not only for statistical completeness but also for their relevance to both methodological rigour and operational decision-making. All held-out evaluations preserve the natural class distribution; SMOTE is used only within training folds.

4.3. Feature engineering for lagged variables

To assess the temporal association between freight traffic in a previous hour and subsequent passenger demand, a new feature, *previous_hour_goods_total*, was created to represent the one-hour lag of total goods vehicle volume. Out of the full dataset of 6,003 hourly observations, 5,795 were eligible for lag computation. As expected, the first entry in each group lacked a prior observation, resulting in 208 missing values, or 3.59% of the lag-eligible sample.

To address this, missing values were imputed using the median goods vehicle volume for the corresponding road segment. The completed feature was then incorporated into the supervised learning framework for model training and evaluation.

4.4. Evaluation strategy for model enhancement

To evaluate whether the inclusion of the *previous_hour_goods_total* feature significantly improved model performance, two statistical methods were applied.

First, McNemar's test was used to compare the classification outcomes of the baseline model (excluding the lagged variable) with those of the enhanced model (including the lagged variable). This test identifies the number of observations correctly predicted by one model but not the other. The test statistic is defined in Equation (3) (McNemar, 1947; Dietterich, 1998):

Table 1
Model comparison framework for McNemar-based hypothesis testing and robustness analysis.

Test category	Model comparison	Target variable	Reference specification	Compared specification	Purpose
First-order tests (research questions)	Freight → Passenger (main effect)	Passenger class (Low/Med/High)	All explanatory features excluding previous_hour_goods_total	+ previous_hour_goods_total	Assess short-term freight → passenger effect (RQ1)
	Passenger autoregressive model	Passenger class (Low/Med/High)	All explanatory features excluding previous_hour_passenger_total	+ previous_hour_passenger_total	Assess passenger → passenger temporal dependency (RQ2)
	Passenger → Freight reciprocal model	Freight class (Low/Med/High)	Freight-target model excluding previous_hour_passenger_total	+ previous_hour_passenger_total	Assess passenger → freight cross-modal effect (RQ2)
	Freight self-predictive model	Freight class (Low/Med/High)	Freight-target model excluding previous_hour_goods_total	+ previous_hour_goods_total	Assess freight → freight temporal dependency (intra-modal benchmark)
Second-order tests (extensions and robustness)	Interaction with speed	Passenger class (Low/Med/High)	Model including previous_hour_goods_total	+ interaction term previous_hour_goods_total × traffic_speed	Test whether speed moderates the freight lag effect
	Interaction with peak/off-peak	Passenger class (Low/Med/High)	Model including previous_hour_goods_total	+ interaction term previous_hour_goods_total × peak_offpeak	Test whether time of day moderates the freight lag effect
	Longer lag (two-hour rolling average)	Passenger class (Low/Med/High)	Model including previous_hour_goods_total	Replace with previous_2hr_avg_goods	Test the persistence of the freight effect over a longer lag
	Motorised-only passenger definition (excluding pedal cycles)	Passenger class (Low/Med/High)	Passenger model including pedal cycles	Passenger model excluding pedal cycles (motorised-only)	Robustness check on passenger traffic definition

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{3}$$

Where,

b represents observations correctly classified only by the enhanced model, and c represents those correctly classified only by the baseline model. The statistic χ^2 measures whether the number of disagreements between the two models is larger than would be expected by chance. A statistically significant result provides evidence that the enhanced model performs better than chance alone.

To avoid ambiguity across different experiments, it is important to clarify how the baseline and enhanced models are defined. For all McNemar's tests, the baseline model refers to the XGBoost classifier trained using the same feature set described in Section 4.2, excluding the specific lagged or interaction feature being tested. The enhanced (or test) model adds that single new feature, or, in the case of alternative specifications, replaces it (e.g., the two-hour lag replacing the one-hour lag). For reciprocal or freight-focused analyses, the baseline and enhanced models are analogously defined within the corresponding target variable domain (i.e., passenger or freight). Table 1 summarises all model comparison specifications evaluated using McNemar's test, distinguishing between first-order tests that directly address the research questions and second-order extensions and robustness checks. All models used identical preprocessing, stratified sampling, and XGBoost classification settings.

Second, a bootstrapped 95% confidence interval was estimated for the difference in classification accuracy between the two models. This approach quantifies the likelihood that observed improvements are not due to sampling variability, further supporting the robustness of the enhancement.

To operationalise this comparison, two separate XGBoost models were trained using the caret package: a baseline model and an enhanced model incorporating the lagged feature, *previous_hour_goods_total*. All categorical variables were one-hot encoded using the dummyVars function to ensure compatibility with XGBoost. A bootstrapping procedure with 1,000 resamples was used to compute the confidence interval for the difference in classification accuracy.

In preparation for interpretability analysis, global feature importance scores were aggregated across dummy variables to reflect their original feature groups. The DALEX framework was then employed to generate permutation-based global importance plots, class-specific Breakdown plots, and partial dependence plots (PDPs). Class labels were relabeled in the visualisations to enhance clarity for decision-makers.

4.5. Model interpretability via DALEX

Before applying the DALEX interpretability framework, it is important to distinguish between the two types of feature-importance measures employed in this study. The first is algorithm-specific importance, derived internally from XGBoost as Total Gain, which quantifies how much each feature contributes to reducing loss across the ensemble's decision trees. This measure reflects how the model allocates predictive weight during training. The second is model-agnostic permutation-based importance, computed through the DALEX package, which assesses the increase in prediction loss (cross-entropy) when a feature's values are randomly permuted. Unlike the internal metric, the permutation-based approach provides an independent and reproducible estimate of each variable's true predictive contribution, regardless of model structure. These two measures serve distinct interpretive purposes: the gain-based metric explains how the model internally distributes predictive influence during training, whereas the permutation-based metric evaluates how sensitive overall model predictions are to perturbations in each variable, providing an external validation of feature relevance. Presenting both perspectives ensures that feature relevance is validated both internally, within the learning algorithm, and externally through a transparent, algorithm-independent framework.

To support transparent and policy-relevant interpretation, the DALEX package was used for model-agnostic explainability. DALEX is not a single algorithm but a unified toolkit that implements several global and local explanation methods, including SHAP, LIME, and permutation-based feature importance. This study employs the permutation-based importance method. It quantifies the contribution of each feature by measuring the increase in cross-entropy loss when that feature is randomly permuted. The method originates from Breiman's (2001) work on random forests and was later generalised to any predictive model by Fisher et al. (2019). It allows variable relevance to be compared across different algorithms without accessing internal model parameters. Permutation importance was selected for three reasons. It provides model-agnostic transparency, avoids algorithm-specific heuristics such as gain or split frequency, and remains robust in the presence of correlated or nonlinear features. These characteristics align with the objective of identifying the directional and relative influence of lagged freight volumes on passenger traffic states. The method offers a clear and reproducible measure of importance that is particularly suitable for complex, mixed urban traffic datasets.

DALEX also supports class-specific and instance-level analyses, which help reveal both general feature behaviour and local prediction logic. Its compatibility with tree-based models such as XGBoost makes it an appropriate and interpretable tool for traffic analytics workflows. In addition, it is important to recognise that permutation-based feature importance primarily reflects the main effects of individual predictors and may underestimate higher-order interaction effects. This characteristic is well established in the interpretability literature. Nevertheless, it is appropriate for the aims of this study, which focus on assessing the overall and directional contribution of lagged freight volume relative to other variables rather than decomposing complex interaction structures. The permutation-based approach therefore provides a clear, model-agnostic estimate of how strongly each feature influences predictive performance, complementing the algorithm-specific Total Gain measure from XGBoost. To ensure transparency, we now also clarify that tree-based gain importance, while useful, is less interpretable and can introduce biases linked to variable frequency, scale, and split structure, making permutation-based importance a fairer and more context-robust choice for this analysis. Taken together, these two perspectives offer a balanced interpretability framework that combines internal model behaviour with an external, algorithm-

independent assessment of variable relevance, strengthening the transparency and robustness of the analysis. The results of this interpretability analysis are presented in Sections 5.4 and 5.5.

4.6. Partial dependence plots

To investigate the marginal effect of lagged freight volume on class-specific predictions, a PDP was generated for the feature *previous_hour_goods_total*. PDPs illustrate the average predicted probability of each output class as a function of a single input feature, while averaging out the effects of all other variables (Friedman, 2001). This technique is widely used to reveal nonlinear relationships, saturation points, and threshold effects in complex machine learning models.

The resulting plot provides a global view of how different levels of freight traffic relate to the probability of observing Low, Medium, or High passenger volumes. The findings are presented in Section 5.6.

4.7. Model extensions and robustness testing

As summarised in Table 1, each extension or reciprocal model was evaluated against its corresponding baseline specification using McNemar's test and bootstrapped confidence intervals to ensure consistent statistical comparison. To test the robustness and boundaries of the observed lagged freight effect, several model extensions and diagnostic procedures were designed. The first set of tests introduced interaction terms to assess whether contextual conditions modulate the predictive contribution of *previous_hour_goods_total*. Specifically, two interaction variables were constructed: one between *previous_hour_goods_total* and traffic_speed, and another between *previous_hour_goods_total* and peak/offpeak status. These tests aimed to determine whether traffic speed or time-of-day dynamics amplify or attenuate the predictive contribution of lagged freight volume.

In a second extension, a longer lag structure was tested by replacing the one-hour lag with a two-hour rolling average of freight traffic, represented by the variable previous_2hr_avg_goods. This formulation was intended to capture potential cumulative exposure or temporal decay effects.

To evaluate possible bidirectional influences, a reciprocal cross-modal model was constructed in which previous_hour_passenger_total was used to predict current freight traffic levels. This test examined whether lagged passenger volume provides explanatory power for freight traffic dynamics. In addition, a passenger autoregressive specification was evaluated, in which previous_hour_passenger_total was used to predict current passenger traffic categories (Passenger(t − 1) → Passenger(t)), to test for within-mode temporal dependence.

Finally, a self-predictive model was tested to examine whether *previous_hour_goods_total* could improve the prediction of current freight traffic conditions. This analysis provided a comparative baseline for understanding whether the lagged variable is more informative within its class or in predicting cross-modal behaviour.

Each of these model variants was evaluated using McNemar's test to quantify prediction disagreements and a bootstrapped 95% confidence interval to assess the statistical significance of any observed improvements. Only enhancements confirmed by these tests were considered robust and meaningful. The outcomes of these experiments are presented in Section 5.7.

5. Results

This section presents the key findings of the study, structured across multiple analytical stages. Section 5.1 reports the results of unsupervised clustering, including the rationale for selecting the final number of clusters and the resulting thresholds used to categorise traffic levels. Section 5.2 evaluates the performance of multiple supervised learning models, comparing overall and class-specific

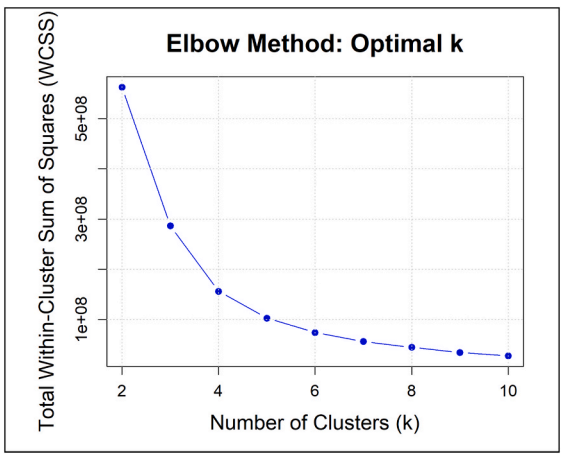

Fig. 3. Elbow method plot showing the WCSS across different values of k.

accuracy. Section 5.3 assesses model improvements achieved by introducing the lagged freight-volume feature and examines their statistical significance. Section 5.4 presents the global interpretability analysis using permutation-based feature importance (DALEX), while 5.5 explores local feature contributions through class-specific Breakdown plots. Section 5.6 investigates nonlinear threshold effects via PDPs. Finally, Section 5.7 summarises robustness checks and model extensions that test the generalisability of the predictive relationships. Together, these analyses provide a comprehensive view of both model performance and behavioural insights into urban traffic dynamics.

5.1. *Data-driven categorisation of traffic volume using K-means clustering*

This section presents the results of the unsupervised clustering process used to define categorical traffic levels, including validation metrics, final cluster selection, and class thresholds for downstream modelling.

Fig. 3 displays the Elbow Method plot, which evaluates the WCSS as k increases. Fig. 4 shows the average silhouette scores for varying values of k.

Based on the Elbow and Silhouette Methods and the Table 5 results (see section 5.7), k = 3 was selected, as it provided a balanced trade-off between cluster compactness, interpretability, and predictive stability. Because the Elbow and Silhouette methods indicated that k values between 2 and 4 were plausible, we also conducted a robustness check (Section 5.7) to verify that our main findings were stable across k = 2, 3, and 4. The lagged-freight feature improved classifier performance in all cases, with the strongest and most stable improvement observed at k = 3. A detailed explanation of this choice is provided in Section 6.1. The Total Sum of Squares (TSS) was 1,580,018,640; the WCSS was 286,500,774; and the Between-Cluster Sum of Squares (BetweenSS) was 1,293,517,866. The BetweenSS/TSS ratio was 0.8187, indicating that approximately 82% of the variance in the passenger volume data was explained by the cluster structure. Based on the resulting cluster centroids, the traffic classes were defined as follows:

- Low: 6–662 vehicles (average = 381; n = 2,763)
- Medium: 663–1,364 vehicles (average = 945; n = 2,460)
- High: 1,365–3,136 vehicles (average = 1,785; n = 780)

The three clusters accounted for a total of 6,003 hourly observations and were used as target categories for the supervised classification models described in the following sections. This corresponds to approximately 46% Low, 41% Medium, and 13% High traffic observations, indicating a moderately imbalanced class distribution that reflects real-world traffic variability.

5.2. *Evaluation of supervised learning models for traffic state classification*

The results follow this multiclass-and-imbalance-aware evaluation framework, balancing overall and per-class measures for fair model comparison. The models discussed in Section 4.2 were selected for their established use in transportation research and their capacity to handle multi-class classification problems. Table 2 reports the overall performance of each model using three standard evaluation metrics: Accuracy, Kappa, and F1 Score. Furthermore, Table 3 provides per-class metrics, including Sensitivity, Specificity, Precision, NPV, and Balanced Accuracy, offering a granular assessment of each model's classification ability. In both tables, the highest scores for each metric are highlighted in bold to indicate the top-performing models.

Among all models, XGBoost (xgbTree) achieved high performance across multiple metrics. It yielded an overall Kappa of 0.6309 and an F1 Score of 0.7581, closely trailing GBM and cforest. For the "High" traffic class, XGBoost achieved:

- Precision: 0.8496

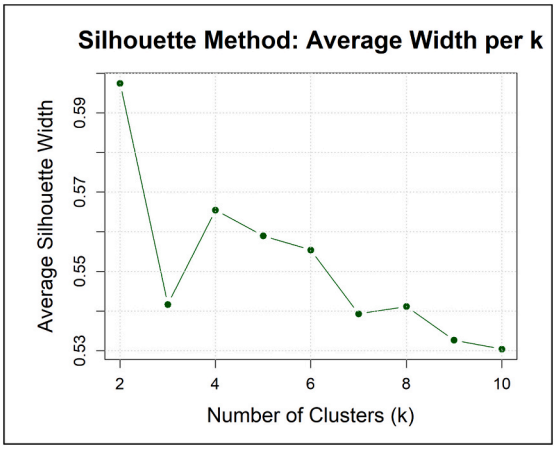

Fig. 4. Silhouette plot showing the average silhouette width for different numbers of clusters (k).

Table 2
Overall accuracy, Kappa, and F1 scores of classification models for predicting traffic categories of passenger-carrying vehicles.

Model	Accuracy	Kappa	F1 Score
rf	0.762778	0.610362	0.745889
svmLinear	0.671111	0.470293	0.656695
knn	0.737222	0.574215	0.718742
gbm	**0.788889**	**0.655747**	**0.770935**
xgbTree	0.774444	0.630859	0.758104
naive_bayes	0.632222	0.410208	0.593528
nnet	0.611667	0.382181	0.582841
cforest	0.778333	0.637345	0.760379

Table 3
Per-class performance metrics of classification models for predicting traffic categories of passenger-carrying vehicles.

Model	Class	Sensitivity	Specificity	Precision	NPV	Balanced Accuracy
rf	Low	**0.714092**	0.797552	0.710243	0.800567	0.755822
rf	Medium	0.739316	**0.942529**	**0.657795**	0.960312	0.840923
rf	High	0.812802	0.874486	0.846541	0.845771	0.843644
svmLinear	Low	0.483740	0.831450	0.666045	0.698576	0.657595
svmLinear	Medium	0.824786	0.893359	0.536111	0.971528	0.859073
svmLinear	High	0.794686	0.746914	0.727876	0.810268	0.770800
knn	Low	0.643631	0.807910	0.699558	0.765388	0.725770
knn	Medium	0.786325	0.914432	0.578616	0.966262	0.850378
knn	High	0.806763	0.861111	0.831880	0.839519	0.833937
gbm	Low	0.700542	0.855932	**0.771642**	**0.804425**	**0.778237**
gbm	Medium	0.820513	0.932312	0.644295	0.972037	**0.876412**
gbm	High	0.858696	**0.875514**	**0.854567**	**0.879132**	**0.867105**
xgbTree	Low	0.707317	0.823917	0.736248	0.802017	0.765617
xgbTree	Medium	0.782051	0.938059	0.653571	0.966447	0.860055
xgbTree	High	0.832126	0.874486	0.849568	0.859454	0.853306
naive_bayes	Low	0.300813	**0.930320**	0.750000	0.656915	0.615567
naive_bayes	Medium	0.824786	0.860792	0.469586	0.970482	0.842789
naive_bayes	High	**0.873188**	0.619342	0.661482	0.851485	0.746265
nnet	Low	0.299458	0.847458	0.577024	0.635145	0.573458
nnet	Medium	**0.854701**	0.856322	0.470588	**0.975273**	0.855511
nnet	High	0.821256	0.679012	0.685484	0.816832	0.750134
cforest	Low	0.701897	0.838041	0.750725	0.801802	0.769969
cforest	Medium	0.786325	0.936143	0.647887	0.967019	0.861234
cforest	High	0.844203	0.869342	0.846247	0.867556	0.856772

- Sensitivity: 0.8321

These values show a high degree of accuracy in identifying peak congestion scenarios, making XGBoost a strong candidate for further analysis. Beyond the aggregate accuracy reported in Table 2, the per-class metrics presented in Table 3 provide further insight into the model's practical reliability across traffic states. The relatively high Sensitivity for the High class (0.83) demonstrates that peak-congestion conditions are seldom missed, supporting timely activation of stronger control measures such as rerouting advisories or adaptive signal coordination. The consistently strong NPV for the Medium class (0.97) indicates that when the system predicts a non-Medium state, it is generally correct, thereby reducing unnecessary deployment of interventions during mid-level traffic conditions, such as moderate signal retiming or other adaptive control actions. High Specificity (0.82–0.94) and Precision (0.65–0.85) across classes help limit false positives, conserving operator attention and avoiding intervention fatigue. Finally, the stability of Balanced Accuracy (0.76–0.86), macro-F1 (0.76), and Kappa (0.63) confirms that performance is not driven by any single traffic category, an important criterion for equitable and dependable decision support in naturally imbalanced settings. Collectively, these results illustrate how statistical performance translates into actionable, class-aware traffic management insights.

5.3. Performance comparison and statistical significance

This section evaluates the predictive contribution of including *previous_hour_goods_total* by comparing the enhanced model to the baseline using McNemar's test, as described in Section 4.4. The test showed that 101 observations were correctly predicted only by the enhanced model, versus 68 by the baseline. The resulting p-value was < 0.05.

The bootstrapped 95% confidence interval for the accuracy gain was [0.0039, 0.0328], excluding the value of zero.

Table 4 compares the baseline and enhanced models across key performance metrics. The enhanced model showed improvements in overall Accuracy, F1 Score, and Kappa. Class-specific gains were observed in Sensitivity (Low and High), Specificity (all classes),

Precision (all classes), NPV (Low and High), and Balanced Accuracy (all classes). Minor decreases were noted in Sensitivity and NPV for the Medium traffic class. Overall, the enhancements indicate consistent performance improvements across most evaluation dimensions. To evaluate whether the inclusion of non-motorised traffic affected performance, a secondary model used a passenger target excluding pedal cycles. With pedal cycles included, accuracy improved from 0.7872 to 0.8056 (McNemar $p = 0.0138$; 95% CI [0.0039, 0.0328]). Using the motorised-only definition, accuracy increased from 0.7883 to 0.8050 (McNemar $p = 0.0219$; 95% CI [0.0044, 0.0311]), indicating that the improvement remains statistically significant even without pedal cycles. This demonstrates that the observed enhancement is not driven by the presence of cyclists but reflects a genuine relationship between lagged freight activity and passenger-vehicle flow dynamics. Pedal cycles were retained in the final model to maintain consistency with the UK Department for Transport (DfT) traffic-count classification and to represent the full spectrum of person-mobility using shared road space in Liverpool, where cyclists commonly travel within mixed traffic and can influence aggregate flow conditions.

As shown in Table 4, although the improvement in overall accuracy (around 1.8 percentage points) is modest in magnitude, it is statistically significant and operationally meaningful. In ensemble-based traffic models with already high baseline performance, such gains typically indicate the presence of a stable and interpretable relationship rather than a purely predictive uplift. Accordingly, the added value of the lagged freight feature lies less in increasing raw accuracy and more in revealing a consistent, policy-relevant linkage between short-term freight intensity and passenger traffic dynamics.

Fig. 5 illustrates feature importance rankings for both models. previous_hour_goods_total emerged as one of the top-ranked predictors in the enhanced model. Hence, these results are summarised in Fig. 5, which presents the algorithm-specific (Total Gain) feature importance from XGBoost for both the base and enhanced models. The subsequent section extends this analysis using a model-agnostic interpretability approach (DALEX) to confirm and contextualise these findings.

5.4. Global feature importance and model interpretation

Building on the gain-based results reported in Fig. 5, Fig. 6 presents the complementary model-agnostic permutation-based feature importance derived from DALEX, offering an independent validation of the enhanced model's explanatory structure. To assess the behaviour of the enhanced model, global feature importance was calculated using DALEX's permutation-based method, as described in Section 4.5. Fig. 6 shows that previous_hour_goods_total ranked as the most informative feature. Its importance exceeded that of other key predictors such as distance_origin_city, peak_offpeak, and traffic_speed, and it outperformed the model's internal gain-based rankings.

The observed classification improvements were modest: accuracy increased from 0.7872 to 0.8056, and the macro F1 score rose from 0.7608 to 0.7898. McNemar's test for model comparison produced a $\chi 2$ value of 6.06 ($p = 0.0138$), and the 95% bootstrapped confidence interval for the accuracy difference was [0.0039, 0.0328], excluding zero.

6. Local interpretability and class-specific feature contributions

To further interpret the model's internal reasoning, class-specific Breakdown plots were generated using the DALEX framework to analyse local feature contributions for individual predictions. Each Breakdown plot decomposes the model's raw logit output for a given observation into additive effects attributed to all input variables, showing how each feature increases or decreases the model's confidence in a particular passenger-traffic category (High, Medium, or Low) before the softmax transformation. The x-axis represents the change in the class-specific logit (Δ log-odds), while the y-axis lists features in the order of their contribution.

Figs. 7-9 present representative examples predicted respectively as High, Medium, and Low passenger traffic flow, illustrating how

Table 4
Comparison of performance metrics between the base model and the enhanced model incorporating *previous_hour_goods_total*.

Metric	Base Model	With previous_hour_goods
Accuracy	0.7872	0.8056
F1 score	0.7608	0.7898
Kappa	0.6476	0.6778
Sensitivity (Low)	0.6880	0.7393
Sensitivity (Medium)	0.8587	0.8478
Sensitivity (High)	0.7385	0.7791
Specificity (Low)	0.9515	0.9630
Specificity (Medium)	0.8786	0.8920
Specificity (High)	0.8220	0.8239
Precision (Low)	0.6793	0.7489
Precision (Medium)	0.8577	0.8699
Precision (High)	0.7425	0.7546
NPV (Low)	0.9533	0.9611
NPV (Medium)	0.8795	0.8731
NPV (High)	0.8189	0.8430
Balanced Acc. (Low)	0.8198	0.8511
Balanced Acc. (Medium)	0.8686	0.8699
Balanced Acc. (High)	0.7803	0.8015

Table 5
Robustness of model performance across different cluster granularities (k = 2, 3, 4).

k	Accuracy (Base)	Accuracy (Enhanced)	95% Bootstrap CI for Δ Accuracy	McNemar p-value	Interpretation
2	0.89	0.90	[−0.002, +0.017]	0.141	Very high accuracy but overly coarse binary split
3	0.78	0.80	[+0.0039, +0.0328]	0.0138	Balanced accuracy–interpretability trade-off; significant gain
4	0.76	0.77	[−0.011, +0.018]	0.715	Slightly lower accuracy; added model complexity

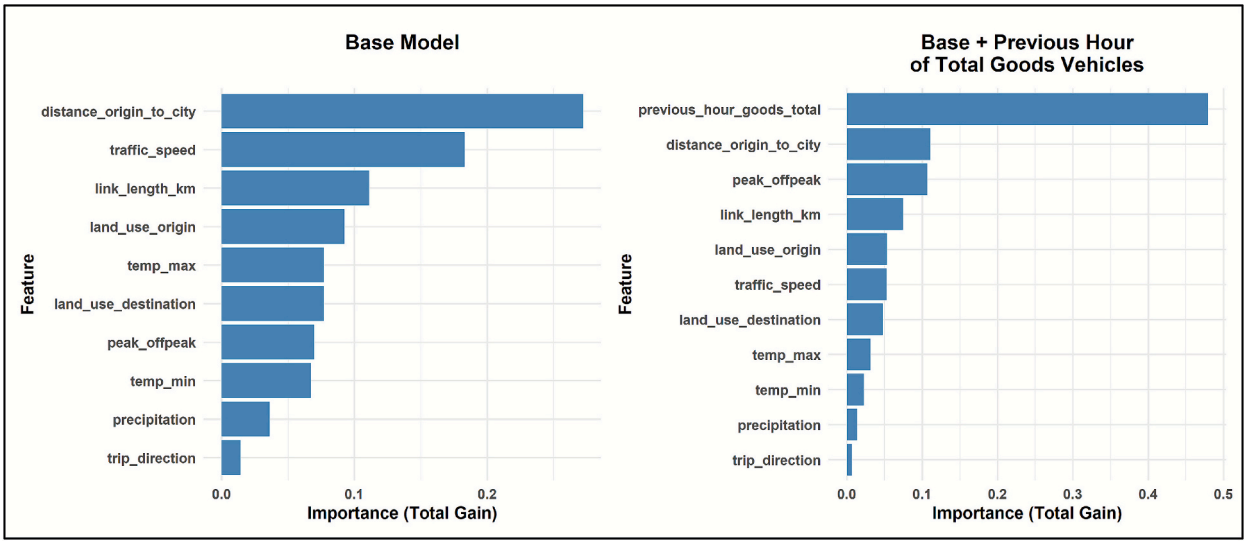

Fig. 5. XGBoost feature importance (Total Gain) for the base and enhanced models, showing previous_hour_goods_total as a key predictor.

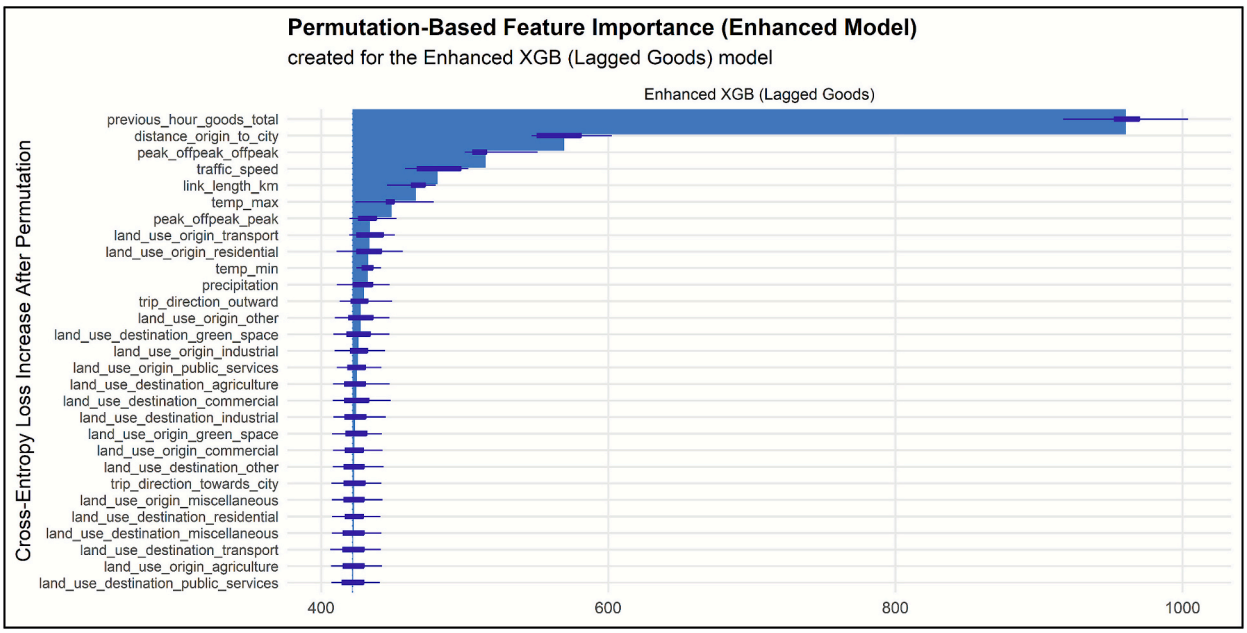

Fig. 6. Model-agnostic permutation-based feature importance (DALEX) for the enhanced model, confirming previous_hour_goods_total as the most influential variable.

multiple contextual factors, such as temperature, distance, and previous-hour (lagged) freight traffic flow, collectively shape the model's classification across different traffic states. Each figure corresponds to a distinct real-world observation from the test data, reflecting feature values that typify high, medium, and low passenger-traffic conditions.

In these plots, the model's baseline logit (intercept) represents its initial confidence before any feature effects are applied (for instance, 0.416 for the High class in Fig. 7). Individual variables then sequentially increase or decrease this baseline; for example, previous_hour_goods_total = 317 increases the logit by + 0.163 in the same case. Positive effects are displayed as green bars and negative effects as red bars, with their magnitudes shown numerically to the right of each bar (not above them), indicating the signed contribution of each feature to the cumulative logit. Hence, for instance, in the High class of Fig. 7, the first green bar shows + 0.163, while the value 0.416 written above it corresponds to the intercept level from which the bar begins to change.

The corresponding interpretability insights derived from these Breakdown plots are further discussed in Section 6.7.

6.1. Partial dependence and nonlinear threshold effects

To explore how the model's predictions respond to variations in a specific feature, a PDP was generated for previous_hour_goods_total. Fig. 10 displays how the predicted probabilities for each passenger traffic class vary across different levels of lagged freight volume. The results indicate a nonlinear relationship and class-specific shifts across different ranges of the feature.

At lower values of *previous_hour_goods_total*, the model assigned high probabilities to the Medium passenger traffic class. As the freight volume increases into an intermediate range (approximately 250–450 vehicles), the probability of the High passenger class becomes dominant. Beyond a context-specific inflexion point, approximately 500 goods vehicles per hour in this Liverpool dataset, the predicted probability of the Low passenger traffic class rises sharply. This value should be interpreted as a local example of a nonlinear response rather than a universal threshold applicable to all urban networks. These transitions illustrate distinct zones where the model adjusts its predictions according to the level of freight activity.

6.2. Robustness, generalisability, and model comparison

This subsection examines the robustness and generalisability of the model across alternative configurations. We tested different numbers of passenger-traffic clusters (k = 2–4), evaluated four directional lag pathways between freight and passenger flows, and introduced interaction terms with contextual variables such as traffic speed and peak/off-peak status. Additional analyses replaced the one-hour lag with a two-hour lag (representing freight traffic flow two hours earlier) and assessed reciprocal models using lagged passenger flow.

Because the Elbow method suggested k = 3 or 4 and the Silhouette analysis favoured k = 2, a robustness assessment was conducted to ensure that the key findings were not sensitive to the chosen number of passenger-traffic clusters. Table 5 compares model performance across these configurations. While the magnitude of improvement varied, the inclusion of the lagged-freight variable

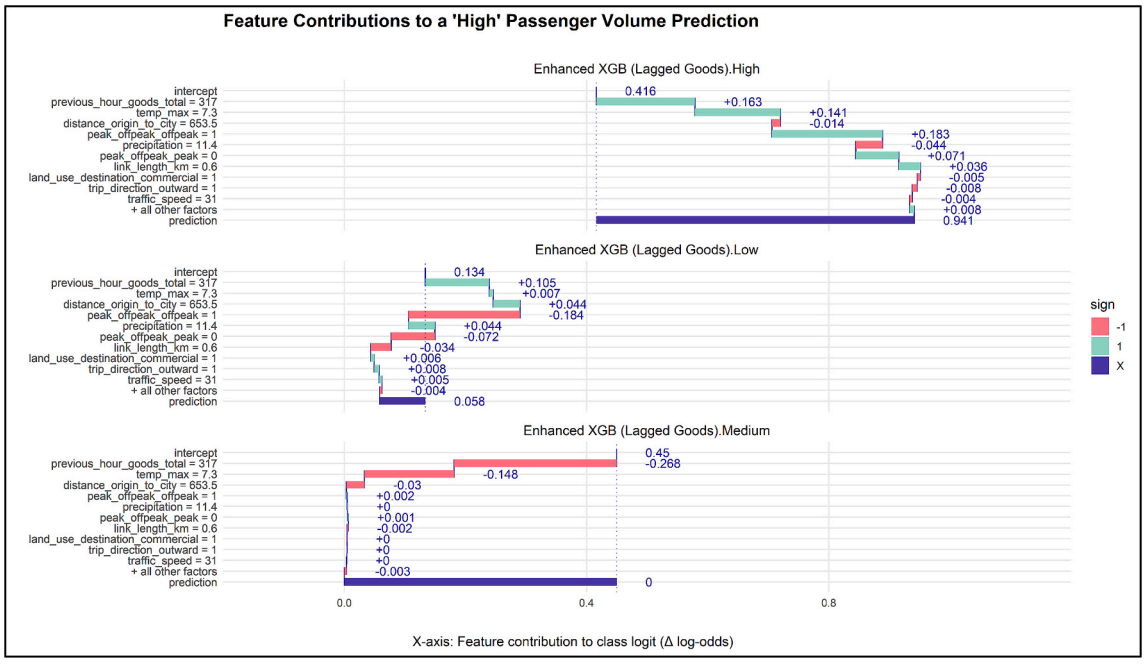

Fig. 7. Class-specific Breakdown plots for an observation predicted as high passenger volume, showing feature-level contributions to the prediction (Δ log-odds scale).

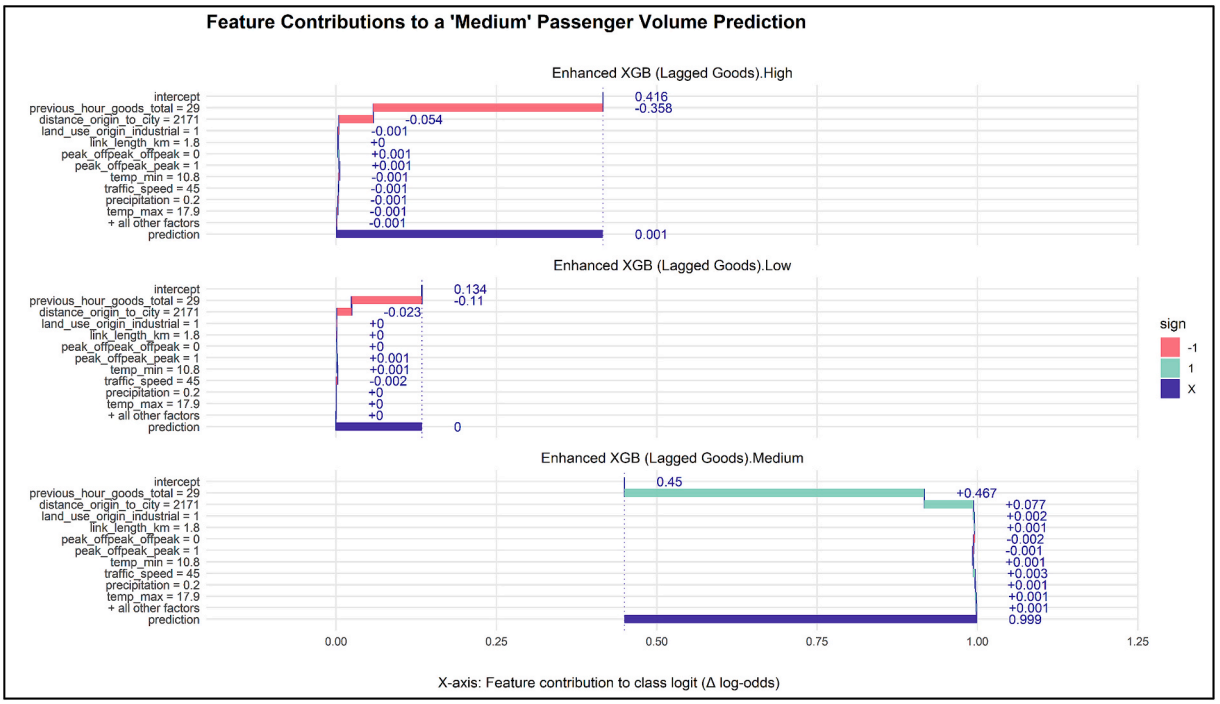

Fig. 8. Class-specific Breakdown plots for an observation predicted as medium passenger volume, illustrating context-dependent feature impacts (Δ log-odds scale).

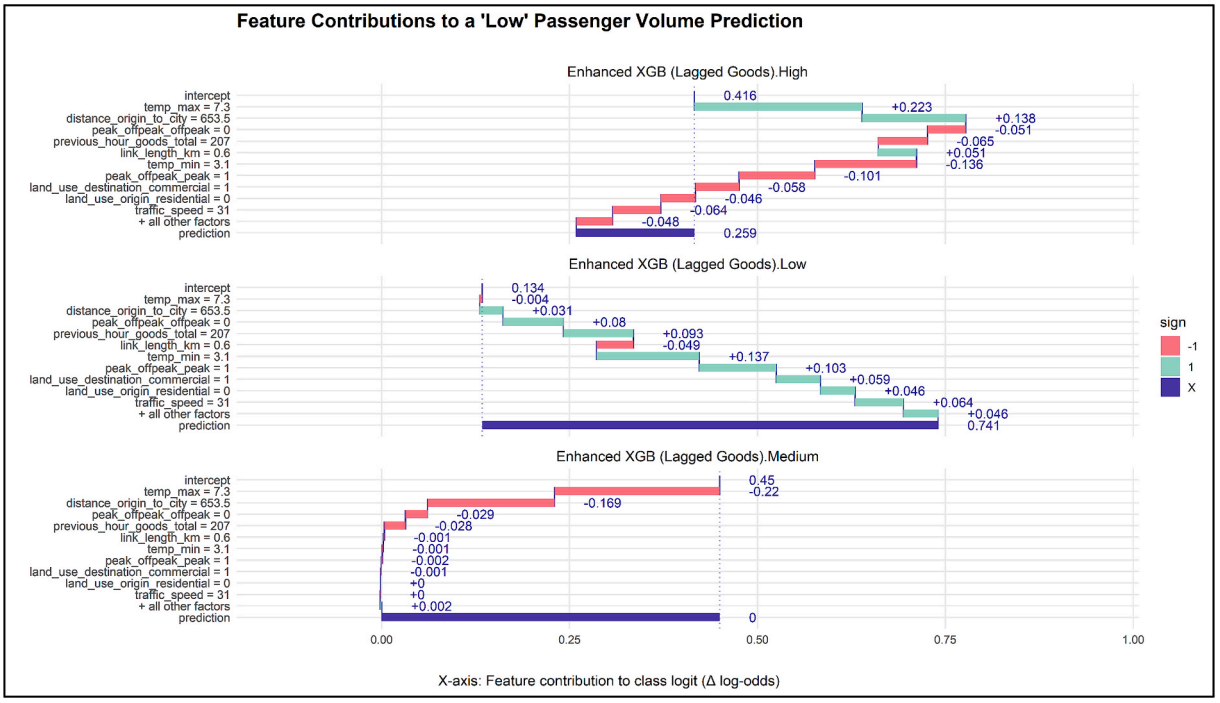

Fig. 9. Class-specific Breakdown plots for an observation predicted as low passenger volume, revealing the nuanced local contribution of key variables (Δ log-odds scale).

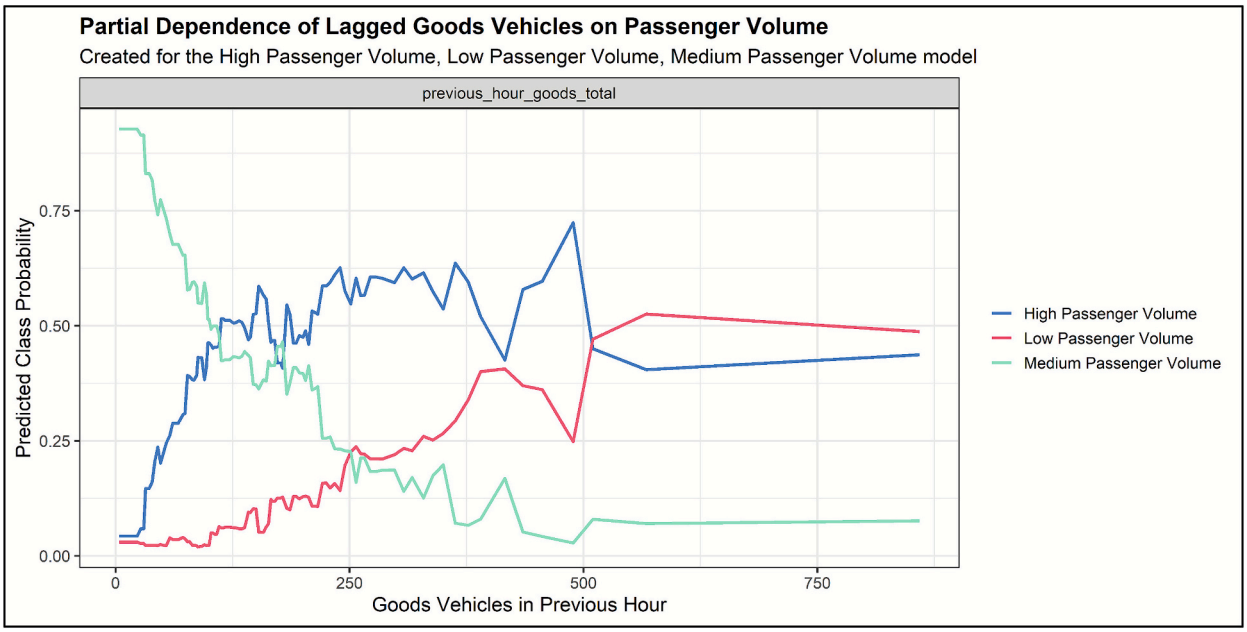

Fig. 10. PDP illustrating the relationship between previous-hour goods vehicle volume and the predicted probability of each passenger traffic category.

consistently enhanced classification accuracy across all tested k values, with the improvement being noticeably stronger for k = 3 compared with the smaller gains at k = 2 and k = 4. At k = 2, the coarse binary split produced very high baseline accuracy but also masked meaningful behavioural distinctions among traffic states, which increased variance in predictor effects and reduced interpretability. At k = 4, the finer granularity created unbalanced clusters and noisier boundaries, resulting in less stable improvements. These patterns highlight that both under- and over-clustering introduce weaknesses not present in the k = 3 configuration. The interpretation of this analysis is also discussed in Section 6.1.

As illustrated in Fig. 1, this stage builds upon the data processing and modelling workflow, where the bidirectional lag hypotheses shown conceptually in Fig. 2 are empirically evaluated. The four pathways, freight (t-1) → passenger (t), passenger (t-1) → freight (t), passenger (t-1) → passenger (t), and freight (t-1) → freight (t), were each implemented and statistically compared through the baseline-enhanced model framework described in Section 4.4. McNemar's test was then applied to assess whether introducing lagged variables or interaction terms produced statistically significant improvements in predictive performance. As detailed in Section 4.4, all statistical comparisons between baseline and enhanced models were evaluated using McNemar's test, with $\chi 2$ and p-values reported for each case. This non-parametric test identifies the number of discordant predictions (b, c) between paired models and is recommended for assessing classification improvements on matched samples (Dietterich, 1998). In addition, 95% bootstrapped confidence intervals for accuracy differences were computed to quantify the uncertainty of observed improvements.

To examine the robustness of the lagged freight relationship, a series of model extensions was applied as outlined in Section 4.7. The one-hour lag of *previous_hour_goods_total* consistently emerged as the only statistically significant and stable predictor across variations. Introducing interaction terms between *previous_hour_goods_total* and traffic_speed yielded no meaningful improvement. McNemar's test yielded a chi-squared statistic of 0.4551 (p = 0.4999), along with a bootstrapped confidence interval for the accuracy difference of [-0.0078, 0.0200]. Similarly, the interaction with peak/off-peak status was non-significant ($\chi 2 = 1.805$, p = 0.1791), with a confidence interval of [-0.0050, 0.0267].

Replacing the one-hour lag with a two-hour rolling average (previous_2hr_avg_goods) yielded a minor increase in accuracy (0.7906 vs. 0.7872) and F1 score (0.7675 vs. 0.7608), though the improvement was not statistically significant (p = 0.4118; CI = [-0.0050, 0.0267]). A reciprocal cross-modal model, in which lagged passenger volume (previous_hour_passenger_total) was used to predict current freight traffic categories (Passenger(t − 1) → Freight(t)), did not improve performance: accuracy decreased slightly from 0.8317 to 0.8267 and macro-F1 from 0.8221 to 0.8118, with no statistically significant difference (McNemar $\chi^2 = 0.5203$, p = 0.4707; 95% CI for ΔAccuracy [−0.0161, 0.0067]). Also, a passenger autoregressive specification, in which lagged passenger volume (previous_hour_passenger_total) was used to predict current passenger traffic categories (Passenger(t − 1) → Passenger(t)), increased accuracy from 0.7872 to 0.7961 and macro-F1 from 0.7608 to 0.7766; however, the improvement was not statistically significant (McNemar $\chi 2 = 1.0135$, p = 0.3141; 95% CI for ΔAccuracy [−0.0083, 0.0261]). To test the role of previous_hour_goods_total in predicting current freight traffic categories, a freight-focused model was evaluated. While accuracy improved from 0.8126 to 0.8344 and macro F1 from 0.8039 to 0.8226, McNemar's test showed no significant difference ($\chi 2 = 0$, p = 1; CI = [-0.0145, 0.0150]). Overall, the McNemar-based comparisons indicate that the one-hour lag of freight traffic (previous_hour_goods_total) is the only specification that yields a statistically significant and robust improvement in predictive performance ($\chi 2 = 6.06$, p = 0.0138). All alternative

specifications, including interaction terms, longer lag structures, passenger autoregressive models, reciprocal cross-modal models, and freight self-predictive models, did not produce statistically significant gains ($p > 0.05$), confirming the specificity and stability of the identified short-term freight-to-passenger effect. These findings validate that the observed gains are statistically meaningful and not due to random variation.

Table 6 summarises the statistical outcomes of all model comparisons described in Table 1, providing a consolidated view of McNemar-based significance tests and robustness checks across the core analyses and model extensions. In Table 6, χ^2 statistics are reported only for paired comparisons on identical target labels. Robustness checks involving alternative lag structures or target definitions are evaluated using bootstrap confidence intervals and associated p-values.

7. Discussion

This section interprets the study's main findings through thematic discussions that connect the technical results to methodological, operational, and policy implications. Section 6.1 explains the rationale behind the clustering design, evaluates alternative values of k, and clarifies the balance between statistical fit and practical applicability. Section 6.2 compares alternative classifiers and justifies the selection of XGBoost for deployment considerations. Section 6.3 evaluates the predictive value of lagged freight volume and its contribution across traffic classes. Section 6.4 discusses class-specific and contextual insights derived from local interpretability analysis. Section 6.5 examines nonlinear and threshold-like patterns that describe freight-passenger mobility dynamics. Section 6.6 assesses robustness, parsimony, and the asymmetric behaviour observed in reciprocal modelling. Section 6.7 summarises global and local interpretability findings using DALEX. Before deriving practical and policy implications, Section 6.8 outlines alternative explanations and clarifies the causal limitations of the study. Building on these findings, Section 6.9 discusses policy and operational implications for urban traffic management and multimodal planning. Finally, Section 6.10 identifies limitations and proposes several directions for future research to strengthen generalisability, methodological depth, and practical relevance.

7.1. Clustering design and practical applicability

The Elbow and Silhouette methods provide complementary perspectives on determining the optimal number of passenger-traffic clusters (k), as outlined in Section 5.1. The Elbow method indicated $k = 3$ or $k = 4$ as suitable options, as the WCSS dropped sharply up to these points and then levelled off, suggesting diminishing returns from additional clusters. In contrast, the Silhouette method exhibited the highest average silhouette width at $k = 2$, indicating the greatest inter-cluster separation. This divergence is expected, as the Elbow method emphasises internal compactness while the Silhouette method prioritises separation between clusters. Although $k = 4$ produced a marginally better Elbow fit, $k = 3$ was ultimately selected as a balanced and operationally meaningful configuration.

To further validate this choice, a robustness analysis was conducted in Section 5.7 to test the sensitivity of the classification results to different cluster granularities ($k = 2, 3,$ and 4). Table 5 in section 5.7 presents the robustness assessment of model performance across different cluster granularities (k = 2, 3, 4). The results show that incorporating the lagged-freight variable consistently enhanced accuracy across all k values, with the clearest and most robust improvement occurring at k = 3. At k = 2, the base model already achieved very high accuracy, leaving limited room for detectable improvement, while k = 4 introduced additional granularity and class noise, resulting in smaller gains. The k = 2 structure over-aggregated the data and obscured key behavioural differences, while the k = 4 configuration fragmented the data into smaller, unbalanced clusters that reduced predictive stability. These weaknesses were absent in the k = 3 solution, which provided the clearest, most interpretable, and operationally meaningful structure. These patterns confirm that k = 3 provides the most balanced and interpretable clustering for subsequent analyses. Therefore, $k = 3$ was retained as the most balanced and practically interpretable configuration, combining strong predictive performance, statistical robustness, and operational relevance. This choice aligns with established traffic management conventions, low, medium, and high demand,

Table 6
Statistical summary of model comparison results.

Test category	Model comparison	χ^2 statistic	p-value	95% Bootstrap CI for Δ Accuracy	Interpretation
First-order tests (research questions)					
	Freight → Passenger (one-hour lag)	6.06	0.0138	[0.0039, 0.0328]	**Statistically significant improvement**
	Passenger autoregressive model	1.0135	0.3141	[−0.0083, 0.0261]	No significant passenger temporal dependency
	Passenger → Freight reciprocal model	0.5203	0.4707	[−0.0161, 0.0067]	No significant cross-modal effect
	Freight self-predictive model	0	1.0000	[−0.0145, 0.0150]	No significant intra-modal effect
Second-order tests (extensions and robustness)					
	Interaction with traffic speed	0.4551	0.4999	[−0.0078, 0.0200]	No moderation effect
	Interaction with peak / off-peak	1.805	0.1791	[−0.0050, 0.0267]	No moderation effect
	Two-hour rolling freight lag	–	0.4118	[−0.0050, 0.0267]	No temporal persistence
	Motorised-only passenger definition	–	0.0219	[0.0044, 0.0311]	Main effect robust to passenger definition

representing a deliberate trade-off between quantitative fit and real-world applicability.

7.2. Model selection and deployment considerations

Although GBM and cforest marginally outperformed XGBoost in terms of raw predictive metrics, the final selection favoured XGBoost due to several practical advantages. XGBoost delivered consistent performance across traffic classes, exhibited strong computational efficiency, and offered built-in regularisation features to reduce overfitting. Furthermore, its scalability and compatibility with parallel processing make it highly suitable for real-world deployment in transportation systems.

In contrast, cforest was more computationally intensive and less interpretable, while GBM lacked the diagnostic flexibility provided by XGBoost. Thus, the selection of XGBoost was based not only on accuracy but also on interpretability, ease of implementation, and robustness, making it the most balanced and actionable model for this application.

The class-specific metrics further clarify how model performance aligns with operational decision-making. High Sensitivity for the High traffic class reduces the risk of overlooking peak-congestion periods, supporting the timely deployment of mitigation measures such as dynamic signal offsets, ramp metering, or targeted rerouting. Conversely, the high NPV for the Medium class (0.97) helps prevent unnecessary interventions during mid-level traffic conditions by providing confidence that a non-Medium prediction truly represents stable conditions. Strong Precision and Specificity ensure that interventions are not triggered without sufficient evidence, thereby avoiding over-response and resource inefficiency. In addition, Balanced Accuracy and Cohen's Kappa indicate that predictive reliability is consistent across all traffic categories, which is essential for maintaining fairness and transparency in model-supported operations. Taken together, these properties demonstrate that XGBoost not only achieves strong predictive accuracy but also delivers dependable, interpretable performance aligned with real-world control-room priorities.

7.3. Value of lagged freight volume as a predictor

The inclusion of the *previous_hour_goods_total* variable significantly enhanced the model's performance across nearly all evaluation metrics, especially for the Low and High traffic classes. McNemar's test and the bootstrapped confidence interval for the accuracy gain confirmed the statistical significance of this improvement. These findings suggest that short-term freight activity serves as a meaningful predictor of passenger traffic levels, consistent with hypotheses involving shared infrastructure or temporal coordination in travel behaviour.

Despite ranking as the most informative global predictor, *previous_hour_goods_total* resulted in only a modest increase in overall accuracy. This is typical in ensemble models, such as XGBoost, where features interact in nonlinear ways. Moreover, permutation-based feature importance measures highlight a variable's overall contribution to decision structure rather than isolated predictive strength. Partial collinearity with other spatio-temporal features (e.g., traffic_speed, peak_offpeak, land_use_origin) may also have reduced its apparent standalone contribution. Nevertheless, the statistically significant gains underscore its relevance and robustness in predicting urban traffic.

While the absolute improvement in accuracy was relatively modest (around 1.8 percentage points), this outcome is typical for mature ensemble models operating on complex, high-baseline systems such as urban traffic. In such contexts, even small yet statistically significant performance increases indicate that the new feature contributes stable, non-redundant information. The result, therefore, demonstrates explanatory robustness rather than incremental overfitting, highlighting that the lagged freight variable captures a genuine, policy-relevant behavioural linkage between freight intensity and short-term passenger flow dynamics.

These findings underscore the robustness of lagged freight volume as a predictive feature. Prior work by Huang et al. (2023) demonstrated that incorporating vehicle-level functional data, such as distinguishing trucks from cars, can improve urban traffic flow forecasting, supporting the value of freight-related signals. Our approach builds on this idea by showing that even aggregated lagged freight volumes offer significant predictive value, particularly in a multimodal road context. Moreover, our use of interpretable machine learning responds to the call by Liang et al. (2022b) for future cross-modal forecasting models that are both accurate and transparent, addressing a key limitation of existing deep learning-based approaches.

7.4. Class-specific and contextual insights

Break Down plots confirmed that the contribution of *previous_hour_goods_total* varies across traffic classes and local feature contexts. In High-volume cases, it contributed significantly to correct classification, while its role was more nuanced in Medium and Low cases. This directional and class-discriminative behaviour highlights the model's ability to learn context-sensitive relationships, critical for operational traffic modelling where response strategies may vary by congestion level.

7.5. Threshold patterns and mobility dynamics

Partial dependence analysis revealed nonlinear and threshold-like relationships between freight volume and predicted passenger traffic. At low levels of freight traffic, Medium passenger volumes dominated, consistent with minimal interaction. In the mid-range (250–450 vehicles), the likelihood of High passenger volumes increased, which is consistent with overlapping delivery and commuting peaks. However, beyond a context-specific inflexion point, approximately 500 freight vehicles per hour in the Liverpool case study, the model increasingly predicted Low passenger volumes, a pattern consistent with system-saturation hypotheses in which higher freight intensity coincides with reduced passenger volumes. These insights emphasise the need to incorporate freight-passenger interactions

into multimodal planning and traffic forecasting systems. While no directly comparable freight-passenger threshold studies were located, related work by Kim and Cho (2022) observed time-varying intermodal substitution patterns, indicating that shared infrastructure dynamics can shift sharply depending on contextual conditions. Our findings contribute to this literature by quantifying case-specific nonlinear threshold effects using interpretable model outputs, offering a replicable framework rather than a universal threshold for integrated transport planning. This inflexion should therefore be interpreted as a case-dependent signal shaped by Liverpool's network capacity and land-use mix, which may differ across urban contexts but can be re-identified through the same analytical approach.

7.6. Robustness and model parsimony

Additional robustness checks confirmed that the predictive association of *previous_hour_goods_total* is both time-sensitive and directionally asymmetric. The one-hour lag provided the most informative signal; extending this to longer lags or including interaction terms did not yield statistically significant improvements. Moreover, reciprocal modelling using lagged passenger data to predict freight traffic produced weaker results. To address the second research question, we also tested whether lagged passenger volume improved the prediction of current passenger or freight traffic states. Although a modest accuracy gain was observed when using previous_hour_passenger_total as a predictor, the improvement was not statistically significant. This indicates that, within the observed context, passenger activity has limited short-term predictive power compared to freight volumes.

To ensure that this asymmetric relationship was not an artefact of spatial or temporal bias, two complementary checks were undertaken. First, spatial composition was statistically controlled through the inclusion of both origin- and destination-land-use variables (land_use_origin, land_use_destination). The freight → passenger association remained robust in the presence of these variables, and DALEX-based feature-importance analyses consistently ranked lagged-freight volume above land-use predictors, indicating that land use provides contextual but not explanatory dominance. Second, potential temporal aggregation bias was evaluated by replacing the one-hour lag with a two-hour rolling average (previous_2hr_avg_goods). This specification produced a minor but statistically insignificant improvement in performance (Accuracy = 0.7906 vs. 0.7872; F1 = 0.7675 vs. 0.7608; p = 0.4118; CI = [-0.0050, 0.0267]), confirming that the directional relationship is not a by-product of hourly aggregation.

This asymmetry was somewhat unexpected, given the bidirectional nature of road use in shared corridors. One might assume that high passenger volumes, especially during peak periods, could constrain freight operations. However, our findings indicate a stronger directional predictive association (freight(t-1) → passenger(t)) than the reverse. Taken together, these results suggest that the asymmetry reflects a behavioural regularity in freight activity, characterised by relatively fixed delivery windows and scheduling routines, rather than data or spatial artefacts. This interpretation remains associative and does not imply causality, but highlights the relative temporal stability of freight operations compared with passenger flows. These findings reinforce the principle of parsimony; model extensions that increase complexity without clear performance gains may add computational cost without providing meaningful benefits. The strong performance of the core model underscores its practicality, robustness, and operational relevance for short-term passenger-traffic classification.

7.7. Global and local interpretability insights

The interpretability analysis using the DALEX framework confirmed that previous_hour_goods_total remained the dominant predictor across both global and local analyses (Figs. 5-10). Figs. 7-9 demonstrate that the influence of lagged freight traffic flow is strong yet context-dependent across different passenger-flow regimes:

- High case (Fig. 7; previous-hour freight = 317): The model assigns the highest confidence to the High category (cumulative logit = 0.941). The lagged-freight feature contributes + 0.163 to the High logit, while reducing Medium by − 0.268 and slightly increasing Low by + 0.105 before other features offset it. This corresponds to the intermediate-freight zone (around 250–450 vehicles) in Fig. 10, where the High passenger class becomes dominant.
- Medium case (Fig. 8; previous-hour freight = 29): The model's highest confidence is Medium (cumulative logit = 0.999). Lagged freight contributes + 0.467 to Medium, while − 0.358 to High and − 0.110 to Low, consistent with the low-freight range of Fig. 10, where Medium probabilities are highest.
- Low case (Fig. 9; previous-hour freight = 207): The model favours Low (cumulative logit = 0.741). Lagged freight adds + 0.093 to Low and reduces High and Medium by − 0.065 and − 0.028, respectively; temp_min (+0.137) and distance effects further reinforce this outcome. Although the PDP (Fig. 10) shows Low becoming dominant beyond around 500 vehicles, this local instance indicates that contextual factors, such as temperature, can shift the transition earlier for specific conditions.

Incorporating previous-hour (lagged) freight traffic flow materially improves class separation: its contribution is positive under higher observed freight levels, reinforces Medium under lower levels, and, when combined with contextual variables, helps differentiate Low-volume situations under certain mid-range conditions.

Beyond this key variable, contextual features such as traffic_speed, peak_offpeak, distance_origin_city, and land-use categories consistently ranked among the next most informative predictors. This pattern indicates that the model captures not only short-term freight-passenger dynamics but also their broader spatial and temporal environments. These findings broadly confirm the empirical evidence reported by Bao et al. (2022) and Wang et al. (2023), who showed that mixed or high-intensity land-use configurations intensify congestion in urban settings. Similarly, the influence of traffic_speed and peak_offpeak corroborates Zhang et al. (2011),

demonstrating that congestion effects amplify under slower, peak-hour conditions. The integration of spatial and temporal predictors complements the methodological insights of Cui et al. (2020), extending their argument from prediction accuracy to model interpretability in multimodal contexts. Finally, the alignment between our interpretable outputs and the governance-oriented discussion of Masik et al. (2021) extends the debate from modelling outcomes to the policy domain, showing how interpretable features such as land use and proximity can inform Smart City mobility strategies.

Collectively, these connections confirm that the results of this study are theoretically consistent with established research while broadening their scope to encompass multimodal passenger-freight interactions, thereby reinforcing both the validity and the applied relevance of the model's explanatory structure.

8. Alternative explanations and causal limitations

The relationships identified in this study should be interpreted as predictive associations rather than causal effects. The modelling framework establishes that lagged freight volumes are statistically associated with subsequent passenger-traffic states, but it does not demonstrate that freight activity causes these changes. Several mechanisms, such as land-use co-location, shared infrastructure constraints, or scheduling rigidity, could jointly explain the observed patterns. Accordingly, all interpretations are presented in associative terms, and the results are intended to support short-term forecasting and hypothesis generation, not causal inference. Future research could employ causal inference techniques such as Granger causality analysis, structural equation modelling, or quasi-experimental designs to better isolate underlying mechanisms and establish directionality. Combining these approaches with higher-frequency or intervention-based data (e.g., freight scheduling changes or adaptive signal control pilots) would enable more robust causal testing and strengthen the policy relevance of the findings.

8.1. Policy and operational implications

This section outlines the operational implications, which concern short-term interventions such as real-time traffic management, freight scheduling, and adaptive signal control, as well as the policy implications, which involve longer-term strategies including collaborative delivery frameworks, freight-prioritisation policies, infrastructure planning, and pricing mechanisms. It is important to note that the following operational and policy implications are not prescriptive recommendations but hypothesis-generating insights derived from associative modelling results. The proposed strategies, such as congestion pricing reform, adaptive signal control, coordinated freight scheduling, and shared-use corridor pilots, should be viewed as illustrative applications and potential directions for future pilot studies and empirical validation. These examples aim to demonstrate how the identified short-term associations could inform future research and experimental policy development rather than prescribe immediate interventions.

8.1.1. Operational implications

The identification of one-hour lagged freight volume as a strong and statistically significant predictor of subsequent passenger traffic class has clear operational relevance for urban traffic management. This feature consistently ranked among the most influential variables in the XGBoost model, exhibiting nonlinear patterns across various traffic levels, which supports its use in real-time monitoring and forecasting systems.

The class-wise performance metrics can be directly embedded into operational protocols. The high NPV for the Medium class (0.97) can function as a negative decision gate, signalling that when the system predicts a non-Medium state, interventions during mid-level traffic conditions can safely be withheld, reducing unnecessary signal retiming or freight restrictions. In contrast, the strong Sensitivity for the High class (0.83) supports a positive gate for rapid activation of peak-period response bundles, including temporary offset plans or traveller-information updates. Grounding such decisions in class-aware metrics minimises both under-reaction (missed peaks) and over-reaction (false alarms), optimising the use of limited operational resources. In practice, these metric-based gates can be formalised as standard operating procedures that reflect the model's validated behavioural patterns.

Partial dependence analysis revealed a threshold-like pattern beyond which increases in freight volume were associated with a sharp decline in predicted passenger traffic categories. This is consistent with congestion spillover hypotheses in high-freight conditions, which could be mitigated through dynamic control strategies. Transport authorities may consider integrating near-real-time freight volume data into traffic management centres to enable more responsive decision-making during periods of increased freight demand.

The lack of statistically significant interaction effects between lagged freight volume and contextual features such as average speed or peak-hour indicators implies that the predictive association from freight activity to passenger flow is relatively stable across different traffic conditions. This finding supports the design of simplified, scalable forecasting tools that rely solely on freight data, eliminating the need for complex contextual adjustments.

The evaluation of longer lag periods showed that extending the window to two-hour averages offered only marginal gains in predictive performance. This underscores the importance of short-horizon responsiveness and supports the use of hourly freight monitoring to inform immediate operational actions.

Based on these findings, traffic control centres could develop targeted response strategies for locations where freight volumes consistently exceed critical thresholds. Although not tested in this study, the modelling framework may also facilitate scenario-based sensitivity analyses, such as assessing how unusual freight spikes during specific time windows are associated with changes in downstream passenger conditions.

In summary, the results support the integration of freight activity monitoring into short-term traffic forecasting and urban control

systems. These operational applications can enhance situational awareness, improve traffic flow coordination, and reduce the impact of freight-associated congestion on passenger travel.

8.1.2. Policy implications

This study offers actionable insights for urban transport policy by documenting a short-term, asymmetric interaction between freight and passenger traffic. Specifically, the consistent predictive power of one-hour lagged freight volumes for subsequent passenger traffic categories is consistent with the structural rigidity in freight movement (e.g., fixed delivery schedules or contractual obligations). These dynamics suggest that freight activity should be systematically incorporated into long-term multimodal planning and demand management frameworks. Class-aware measures such as Balanced Accuracy (0.76–0.86), macro-F1 (0.76), and Kappa (0.63) provide a transparent performance audit that policymakers can use to ensure decision-support tools perform equitably across traffic states and locations. Reporting these alongside Sensitivity, Specificity, and predictive values (NPV and PPV) helps verify that policy interventions, such as congestion pricing or delivery-window regulation, are informed by balanced, not class-biased, model performance. Because the model reliably identifies when mid-level traffic conditions are absent (as reflected by the high NPV for the Medium class), policy measures such as temporary freight restrictions or curb-use controls can be activated only when genuinely required, thereby avoiding unnecessary or overly stringent regulation.

The observed nonlinear inflexion point, where our Liverpool-based model predicted a sharp increase in the probability of low passenger traffic beyond approximately 500 freight vehicles per hour in the previous hour, offers a context-specific indication of potential network stress points. While this value is not intended as a transferable threshold, it illustrates how interpretable machine learning methods, such as PDPs, can reveal critical inflexion zones in urban traffic dynamics. Such tools can support infrastructure planning by identifying corridors where capacity upgrades, protected lanes, or freight-passenger separation may be warranted. Moreover, the consistent predictive value of lagged freight volumes highlights the potential for proactive anticipatory delivery coordination based on real-time or forecasted freight activity. Although the interaction between freight intensity and peak/off-peak status was not statistically significant in our model, off-peak freight operations are widely implemented in practice. They may still be worth exploring where local evidence supports their effectiveness. Zheng et al. (2025), for example, show through game-theoretic modelling that time-sensitive delivery strategies can reduce intermodal conflicts in shared-use networks.

While pricing mechanisms were not directly modelled in this study, the directional predictive association between freight activity and subsequent passenger traffic carries important implications for congestion pricing. Uniform pricing schemes may overlook the indirect role freight flows play in shaping peak-period passenger demand. Jing et al. (2024) similarly caution that neglecting intermodal dynamics in agent-based simulations can lead to equity concerns, particularly for essential logistics services. Although our model found no significant time-of-day effects, differentiated congestion pricing, mode-sensitive fees, or dynamic scheduling remain relevant strategies for achieving more balanced, efficient, and equitable multimodal transport systems.

These findings also support coordinated operations and shared-use corridor planning. The directional lag relationships between freight and passenger volumes indicate opportunities for dynamic signal timing and route optimisation based on real-time freight data. Simulation-based studies on modular vehicle routing (Hatzenbühler et al., 2023) have highlighted the operational benefits of integrating passenger and freight schedules. While our study does not model shared vehicle operations, the predictive passenger-freight dynamics we identify offer a foundation for corridor-based pilot studies, particularly in low-density or suburban settings.

Finally, policy design must also consider equity. Uniform pricing strategies may disproportionately burden small carriers or low-margin freight operators if their contribution to congestion is misunderstood. Cross-subsidies, targeted incentives, or freight-sensitive exemptions could help ensure fair treatment while maintaining system efficiency. In summary, embedding freight sensitivity into scheduling, pricing, and infrastructure policies may help cities better balance capacity, reduce inefficiencies, and move toward more integrated, equitable transport governance.

The observed Liverpool-specific threshold effect, where passenger flow drops sharply once freight volumes exceed roughly 500 vehicles per hour, illustrates a case-specific nonlinear pattern rather than a transferable benchmark. This local inflexion point reinforces the potential of interpretable models to uncover corridor-specific capacity sensitivities that can inform targeted infrastructure and policy measures.

Moreover, the consistently high NPV for the Medium class (0.97) has direct policy relevance. It indicates that the model rarely recommends mid-intensity actions when they are unwarranted, reducing unintended burdens on logistics operators and road users. This reliability supports proportional policy responses, such as adaptive curb allocation or targeted freight scheduling, and helps mitigate risks of over-regulation stemming from false alarms or misclassified demand levels.

Building on the study's associative findings, the following hypotheses and directions are proposed for future research and pilot studies:

- Hypothesis 1 Incorporate freight signals into multimodal planning: One-hour lagged freight volumes can serve as a leading indicator for short-term passenger traffic conditions. Planners should integrate these variables into travel demand forecasting models, congestion mitigation strategies, and real-time operational decision tools.
- Hypothesis 2: Design infrastructure interventions based on threshold effects: In our Liverpool-based model, a sharp nonlinear shift was observed at approximately 500 freight vehicles per hour, where the probability of low passenger traffic increased substantially. While this threshold is illustrative of local network behaviour, analogous inflexion points could vary widely across cities depending on infrastructure capacity, modal mix, and scheduling practices. While this threshold is not universally applicable, it demonstrates how interpretable machine learning can identify corridor-specific tipping points that justify dedicated freight lanes, temporal separation policies, or targeted design interventions in high-volume areas.

- Hypothesis 3 Encourage off-peak freight activity: Regulatory incentives, zoning for delivery windows, and congestion-sensitive pricing could help shift freight flows away from peak passenger periods, alleviating system pressure without the need for major infrastructure expansion. This aligns with Zheng et al. (2025), who used game-theoretic modelling to show that delivery rescheduling can effectively reduce intermodal conflict.
- Hypothesis 4: Reform congestion pricing to reflect intermodal dynamics: The more substantial directional association from freight to passenger traffic challenges the fairness of uniform pricing schemes. If passenger congestion is partly associated with patterns in earlier freight activity, pricing models should account for this asymmetry. As shown by Jing et al. (2024), agent-based simulations demonstrate the equity risks of neglecting passenger-freight interactions. Mode-sensitive fees or differentiated incentives could improve fairness and system performance.
- Hypothesis 5 Develop collaborative data-sharing frameworks: Real-time freight data, often privately held, is essential for dynamic signal control, demand forecasting, and coordinated operations. Cities should prioritise API-based data agreements or logistics partnerships to enable freight-aware transport management.
- Hypothesis 6: Promote inter-agency coordination and explore shared-use corridor pilots: Integrating freight considerations into operational planning requires collaboration across transport, logistics, and land-use authorities. Joint mobility governance bodies or regional coordination platforms could facilitate this integration. While this study does not directly simulate shared vehicle operations, the predictive passenger-freight dynamics identified here support the case for modular vehicle pilots in suburban corridors, echoing findings from Hatzenbühler et al. (2023).
- Hypothesis 7 Ensure equity in freight-related policy design: Policies that overlook freight's contribution to congestion may inadvertently penalise small carriers or essential logistics services. Cross-subsidies, targeted incentives, or freight-sensitive exemptions should be considered to promote both equity and efficiency.

Although the interaction term between previous_hour_goods_total and peak/off-peak hours was not statistically significant (Section 5.7), the policy implication of encouraging off-peak freight activity remains policy-relevant. The statistical insignificance simply indicates that, within the analysed dataset, the predictive strength of lagged freight volume did not vary systematically by time of day. However, empirical studies such as Holguín-Veras et al. (2018) show that shifting deliveries to off-hour periods can effectively reduce congestion and emissions. Hence, this proposition reflects established best practice and complementary evidence rather than a direct effect estimated by the model.

In sum, embedding freight sensitivity into scheduling, pricing, infrastructure planning, and governance design may help cities build more integrated, responsive, and equitable urban mobility systems. These hypotheses and directions for future research are grounded in statistically validated, interpretable machine learning insights and provide a practical foundation for continued policy experimentation in multimodal transport systems.

8.2. Limitations and future directions

This study makes a significant contribution to urban traffic modelling by identifying short-term predictive associations between freight and passenger flows on shared infrastructure and by applying the DALEX framework to enhance the transparency and interpretability of machine learning predictions. While the results are robust and offer both methodological and operational insights, several directions remain open for further enhancement and generalisation.

First, the analysis focused on a single urban setting in Liverpool, which has specific infrastructure characteristics, scheduling practices, and land-use patterns. Although this provided a valuable environment for uncovering multimodal dynamics, future research could apply the same approach in other cities with different transport systems and spatial layouts. This would help validate the findings and support broader generalisation.

Second, the use of hourly aggregated traffic volumes was well-aligned with the modelling objectives and enabled practical analysis. However, higher-frequency data, such as real-time sensor feeds or probe vehicle information, could help capture short-term fluctuations more accurately. These data sources may improve model responsiveness in applications involving incident detection or adaptive traffic control. A further limitation concerns the use of a 10-day averaged traffic-speed variable, which was necessitated by data availability. While this aggregation provides a stable representation of typical link conditions, it inevitably smooths short-term fluctuations caused by incidents, weather, or day-specific demand peaks. Consequently, the model may under-represent transient congestion effects and slightly bias the estimated freight-passenger association toward more persistent structural patterns. Nonetheless, the inclusion of temporal (hour, peak/off-peak) and spatial (land-use, distance-to-city) features helped to partially mitigate this effect. Future research should therefore employ higher-frequency sensor data to capture true intra-day variability and validate the stability of these relationships under real-time conditions.

Third, the model included a concise and interpretable set of features, such as lagged freight volumes, traffic speed, land-use categories, and daily weather indicators, including temperature and precipitation. This limited feature space was selected to promote transparency and minimise model complexity. Future studies may consider incorporating additional variables, such as road incidents, signal timings, socioeconomic activity measures, or sub-hourly weather data, to further enhance predictive accuracy and contextual understanding.

The use of the XGBoost algorithm provided strong classification performance and compatibility with interpretability tools such as DALEX. Although the modelling process involved several technical steps, including variable imputation, one-hot encoding, and clustering, these are common in transport-focused machine learning workflows. Future work may explore streamlined alternatives to improve computational efficiency and facilitate real-time implementation.

It is also worth noting that the relationships identified in this study are associative. Future research could employ causal inference techniques, such as Granger causality tests or structural modelling, to investigate the underlying mechanisms that drive these observed patterns.

Lastly, while the policy implications presented here are based on robust model outputs, they have not yet been tested in operational environments. Collaborations with transportation agencies and logistics operators will be crucial for translating these insights into practical, real-world strategies. Possible directions include pilot programs focused on adaptive freight scheduling, dynamic curbside allocation, or integrated multimodal traffic coordination.

In conclusion, future studies should aim to test the proposed framework in diverse urban contexts, incorporate higher-resolution and real-time data sources, expand the range of explanatory features, and integrate causal or experimental designs. These steps will enhance the generalizability, operational relevance, and scientific value of freight-passenger interaction modelling.

9. Conclusion

This study examined short-term, directional interactions between freight and passenger traffic volumes on urban roads, addressing a key gap in multimodal transport modelling. Focusing on Liverpool, a major UK port city with high freight intensity, we developed a machine learning framework combining K-means clustering for traffic state definition, XGBoost for classification, and DALEX for model explainability. Two research questions guided the analysis: (1) Does a short-term lag in freight volume (i.e., one hour) help predict current passenger or freight traffic states? If so, how does this association vary across different traffic classes and conditions? (2) Does lagged passenger activity (i.e., one hour earlier) help predict current passenger or freight traffic states? If so, what is the nature and significance of this relationship?

Results indicate that previous-hour freight volume significantly improves the classification of current passenger traffic states, particularly under low and high congestion scenarios. PDPs revealed nonlinear, threshold-like relationships, where moderate freight volumes align with peak passenger flow, while excessive freight levels are associated with lower passenger activity. DALEX analysis confirmed this variable's predictive importance and provided interpretable, policy-relevant insights. In contrast, lagged passenger volume and longer-term freight lags offered minimal explanatory value, indicating a directionally asymmetric predictive relationship from freight to passenger traffic.

While lagged freight volumes proved to be a significant and robust predictor of passenger traffic, the data did not support a reciprocal predictive relationship whereby lagged passenger volumes would meaningfully explain subsequent traffic conditions. These insights have operational relevance for congestion management, adaptive signal control, and dynamic curb allocation. Interpretable model outputs could be embedded into intelligent transport systems (ITS) for real-time interventions.

Limitations of this study include the use of broad vehicle classifications, static speed data, the reliance on 10-day averaged link-speed data, and the exclusion of minor roads due to metadata constraints. The analysis was also limited to a single city context, Liverpool, which may affect generalizability. This study should therefore be interpreted as a proof-of-concept analysis designed to demonstrate the feasibility and explanatory potential of the proposed modelling framework. The directional asymmetry observed between freight and passenger flows reflects the unique characteristics of Liverpool's port-centric economy and road network configuration. While these findings are valuable for understanding multimodal interactions in similar contexts, their generalisation to other urban environments requires validation through comparative applications across cities with different spatial structures, economic bases, and mobility patterns. In particular, the nonlinear relationship and the presence of a context-specific threshold observed in the Liverpool case may warrant investigation in other cities. Exploring whether similar inflexion patterns emerge under different infrastructure capacities, modal mixes, or scheduling practices would help determine the framework's broader applicability and refine understanding of freight-passenger interactions across diverse urban settings.

Future research should test the framework in diverse urban environments with varying infrastructure and modal profiles. Given that the present study used 10-day averaged link-speed data, future applications should employ higher-frequency or real-time speed measurements to capture day-to-day congestion variability more accurately. Incorporating higher-frequency or real-time data, such as sensor-based traffic speeds or incident reports, could improve responsiveness for adaptive control. Expanding the feature set to include signal timing, socioeconomic indicators, or more granular weather data may also enhance predictive accuracy. Ultimately, applying causal inference techniques and validating the framework through real-world pilot projects will be crucial for operational deployment and effective policy integration.

CRediT authorship contribution statement

E. Amirnazmiafshar: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **D.P. Song:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **B. Kenny:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition, Data curation, Conceptualization. **J.M. Wu:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation. **B. Kulcsár:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Data curation, Conceptualization. **Y.Z. Liu:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Data curation. **C. Olaverri-Monreal:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgements

This work is supported by the UK ESRC (Grant No. ES/Y010574/1) under the Driving Urban Transitions project ERGODIC (Project No. 10092870).

Data availability

Data will be made available on request.

References

Alessandretti, L., Natera Orozco, L.G., Saberi, M., Szell, M., Battiston, F., 2023. Multimodal urban mobility and multilayer transport networks. Environ. Plann. B: Urban Anal. City Sci. 50 (8), 2038–2070.
Allamehzadeh, A., Aminian, M. S., Mostaed, M., & Olaverri-Monreal, C. (2017, February). Automatic Vehicle Counting Approach Through Computer Vision for Traffic Management. In International Conference on Computer Aided Systems Theory (pp. 405-412). Cham: Springer International Publishing.
Bao, Z., Ou, Y., Chen, S., Wang, T., 2022. Land use impacts on traffic congestion patterns: a tale of a Northwestern Chinese City. Land 11 (12), 2295.
Biecek, P., 2018. DALEX: Explainers for complex predictive models in R. J. Mach. Learn. Res. 19 (84), 1–5.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://link.springer.com/article/10.1023/a:1010933404324.
Chen, T., Guestrin, C., 2016. August). Xgboost: a scalable tree boosting system. In: In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
Cheng, Z., Wang, W., Lu, J., Xing, X., 2020. Classifying the traffic state of urban expressways: a machine-learning approach. Transp. Res. A Policy Pract. 137, 411–428.
Cui, Z., Lin, L., Pu, Z., Wang, Y., 2020. Graph Markov network for traffic forecasting with missing data. Transp. Res. Part C Emerging Technol. 117, 102671.
Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10 (7), 1895–1923.
Dougherty, J., Kohavi, R., Sahami, M., 1995. In: Supervised and Unsupervised Discretisation of Continuous Features. Morgan Kaufmann, pp. 194–202.
Esfahani, R.K., Shahbazi, F., Akbarzadeh, M., 2019. Three-phase classification of an uninterrupted traffic flow: a k-means clustering study. Transportmetrica b: Transport Dynamics.
Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets (Vol. 10, No. 2018, p. 4). Cham: Springer.
Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. 20 (177), 1–81.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Hatzenbühler, J., Jenelius, E., Gidófalvi, G., Cats, O., 2023. Modular vehicle routing for combined passenger and freight transport. Transp. Res. A Policy Pract. 173, 103688.
Holguín-Veras, J., Encarnación, T., González-Calderón, C.A., Winebrake, J., Wang, C., Kyle, S., Garrido, R., 2018. Direct impacts of off-hour deliveries on urban freight emissions. Transp. Res. Part D: Transp. Environ. 61, 84–103.
Hua, M., Pereira, F.C., Jiang, Y., Chen, X., Chen, J., 2024. Transfer learning for cross-modal demand prediction of bike-share and public transit. J. Intell. Transp. Syst. 1–14.
Huang, S., Sun, D., Zhao, M., Chen, J., Chen, R., 2023. Short-term traffic flow prediction approach incorporating vehicle functions from RFID-ELP data for urban road sections. IET Intel. Transport Syst. 17 (1), 144–164.
Jiber, M., Mbarek, A., Yahyaouy, A., Sabri, M.A., Boumhidi, J., 2020. Road traffic prediction model using extreme learning machine: the case study of Tangier. Morocco. Information 11 (12), 542.
Jing, P., Seshadri, R., Sakai, T., Shamshiripour, A., Alho, A.R., Lentzakis, A., Ben-Akiva, M.E., 2024. Evaluating congestion pricing schemes using agent-based passenger and freight microsimulation. Transp. Res. A Policy Pract. 186, 104118.
Ke, J., Feng, S., Zhu, Z., Yang, H., Ye, J., 2021. Joint predictions of multimodal ride-hailing demands: a deep multi-task multi-graph learning-based approach. Transp. Res. Part C Emerging Technol. 127, 103063.
Khiari, J., & Olaverri-Monreal, C. (2020, November). Boosting algorithms for delivery time prediction in transportation logistics. In 2020 International Conference on Data Mining Workshops (ICDMW) (pp. 251-258). IEEE.
Kim, M., Cho, G.H., 2022. Examining the causal relationship between bike-share and public transit in response to the COVID-19 pandemic. Cities 131, 104024.
Koesdwiady, A., Soua, R., Karray, F., 2016. Improving traffic flow prediction with weather information in connected cars: a deep learning approach. IEEE Trans. Veh. Technol. 65 (12), 9508–9517.
Kumar, S.V., Vanajakshi, L., 2015. Short-term traffic flow prediction using seasonal ARIMA model with limited input data. Eur. Transp. Res. Rev. 7, 1–9.
Lanza, K., Burford, K., Ganzar, L.A., 2022. Who travels where: Behavior of pedestrians and micromobility users on transportation infrastructure. J. Transp. Geogr. 98, 103269.
Li, C., Bai, L., Liu, W., Yao, L., Waller, S.T., 2021. A multi-task memory network with knowledge adaptation for multimodal demand forecasting. Transp. Res. Part C Emerging Technol. 131, 103352.
Li, M., Zhu, Y., Zhao, T., Angelova, M., 2022. Weighted dynamic time warping for traffic flow clustering. Neurocomputing 472, 266–279.
Li, R., Rose, G., 2011. Incorporating uncertainty into short-term travel time predictions. Transp. Res. Part C Emerging Technol. 19 (6), 1006–1018.
Liang, Y., Huang, G., & Zhao, Z. (2022a). Bike sharing demand prediction based on knowledge sharing across modes: A graph-based deep learning approach. In 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC) (pp. 857-862). IEEE.
Liang, Y., Huang, G., Zhao, Z., 2022b. Joint demand prediction for multimodal systems: a multi-task multi-relational spatiotemporal graph neural network approach. Transp. Res. Part C Emerging Technol. 140, 103731.
Liu, X., Pan, L., & Sun, X. (2016, June). Real-time traffic status classification based on Gaussian mixture model. In 2016 IEEE First International Conference on Data Science in Cyberspace (DSC) (pp. 573-578). IEEE.
Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, p. 30.
Ma, H., Zhou, M., Ouyang, X., Yin, D., Jiang, R., Song, X., 2022. In: October). Forecasting Regional Multimodal Transportation Demand with Graph Neural Networks: an Open Dataset. IEEE, pp. 3263–3268.
MacQueen, J., 1967. January). Some Methods for Classification and Analysis of Multivariate Observations Vol. 5, 281–298.
Masik, G., Sagan, I., Scott, J.W., 2021. Smart City strategies and new urban development policies in the polish context. Cities 108, 102970.
McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12 (2), 153–157.
Montazeri-Gh, M., Fotouhi, A., 2011. Traffic condition recognition using the k-means clustering method. Sci. Iran. 18 (4), 930–937.

Rao, W., Xia, J., Lyu, W., Lu, Z., 2019. Interval data-based k-means clustering method for traffic state identification at urban intersections. IET Intel. Transport Syst. 13 (5), 807–813. https://doi.org/10.1049/iet-its.2018.5379.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Safikhani, A., Kamga, C., Mudigonda, S., Faghih, S.S., Moghimi, B., 2020. Spatio-temporal modeling of yellow taxi demands in New York City using generalised STAR models. Int. J. Forecast. 36 (3), 1138–1148.

Silgu, M.A., Celikoglu, H.B., 2015. In: February). Clustering Traffic Flow Patterns by Fuzzy C-Means Method: Some Preliminary Findings. Springer International Publishing, Cham, pp. 756–764.

Tanwar, R., Agarwal, P.K., 2024. Assessing travel time performance of multimodal transportation systems using fuzzy-analytic hierarchy process: a case study of Bhopal City. Heliyon 10 (17).

UK Department for Transport. (2025, May 26). Road traffic statistics - Liverpool (Local authority 161). Retrieved from https://roadtraffic.dft.gov.uk/local-authorities/161.

Wang, B., Leng, Y., Wang, G., & Wang, Y. (2024). Fusiontransnet for smart urban mobility: Spatiotemporal traffic forecasting through multimodal network integration. arXiv preprint arXiv:2405.05786.

Wang, D., Chen, H., Li, C., Liu, E., 2023. Exploring the relationship between land use and congestion source in xi'an: a multisource data analysis approach. Sustainability 15 (12), 9328.

Wang, Z., Jiang, R., Xue , H., Salim, F.D., Song, X., & Shibasaki, R. (2022, June). Event-aware multimodal mobility nowcasting . In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 4, pp. 4228 - 4236).

Wang, C., Zhang, W., Wu, C., Hu, H., Ding, H., Zhu, W., 2022. A traffic state recognition model based on feature map and deep learning. Physica A: Statistical Mechanics and its Applications 607, 128198.

Yang, J., Liu, Y., Ma, L., Ji, C., 2022. Maritime traffic flow clustering analysis by density based trajectory clustering with noise. Ocean Eng. 249, 111001.

Yang, X., Wang, F., Bai, Z., Xun, F., Zhang, Y., Zhao, X., 2021. Deep learning-based congestion detection at urban intersections. Sensors 21 (6), 2052.

Yuan, Y., Zhang, W., Yang, X., Liu, Y., Liu, Z., Wang, W., 2025. Traffic state classification and prediction based on trajectory data. J. Intell. Transp. Syst. 29 (4), 365–379.

Zafar, N., Ul Haq, I., 2020. Traffic congestion prediction based on estimated Time of Arrival. PLoS One 15 (12), e0238200.

Zhang, K., Batterman, S., Dion, F., 2011. Vehicle emissions in congestion: Comparison of work zone, rush hour and free-flow conditions. Atmos. Environ. 45 (11), 1929–1939.

Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transp. Res. Part C Emerging Technol. 58, 308–324.

Zheng, Y., Yang, J., Zhang, X., 2025. Multi-Period operations optimization for passenger-freight shared transport: a game-theoretic approach. Transp. Res. A Policy Pract. 199, 104601.