DECOMPOSING URBAN TRAFFIC FLOWS: A MULTI-STAGE APPROACH TO MODEL HEAVY VEHICLE MOVEMENTS IN GREATER SYDNEY

Patrick Fernandez
Ph.D. Candidate
School of Engineering
The Australian National University, Canberra ACT 2601, Australia
patrick.fernandez@anu.edu.au

Surendra Reddy Kancharla, Ph.D.
Postdoctoral Researcher
"Friedrich List" Faculty of Transport and Traffic Sciences
Technische Universität Dresden, Saxony, Germany, 01069
surendrareddy.kancharla@gmail.com

Dung-Ying Lin, Ph.D.
Associate Chairman and Professor
Department of Industrial Engineering and Engineering Management
National Tsing-Hua University (NTHU), Taiwan
dylin@ie.nthu.edu.tw

S. Travis Waller, Ph.D., Corresponding Author
Lighthouse Professor and Chair of Transport Modelling and Simulation
"Friedrich List" Faculty of Transport and Traffic Sciences
Technische Universität Dresden, Saxony, Germany, 01069
Professor
The Australian National University, Canberra ACT 2601, Australia
steven_travis.waller@tu-dresden.de

Word Count: 4908 words + 1 table(s) $\times$ 250 = 5158 words

Submission Date: October 29, 2024

ABSTRACT

This study presents an innovative multi-stage methodology for decomposing urban traffic flows into light vehicle (LV) and heavy vehicle (HV) categories, addressing a critical gap in transportation network analysis. Utilizing data from Greater Sydney's road network, we develop a comprehensive approach comprising three main stages: origin-destination (OD) matrix estimation using RapidEx [1], quadratic programming optimization for HV/LV proportion estimation, and XGBoost regression for generalization. Our analysis examines associations between HV trips and various urban characteristics, including points of interest (POI), nightlight intensity, zonal population, and zonal area. This approach yields novel insights into urban logistics, commercial activities, and public transportation patterns. The XGBoost machine learning model, a key component of our methodology, achieves a test R-squared of 0.637 and identifies critical factors influencing HV proportions. Key findings include the significant impact of destination characteristics, particularly complex non-linear relationships between nightlight intensity and HV proportions, as well as, service-related activities and area size, on HV traffic patterns.

Keywords: urban transportation, traffic decomposition, heavy vehicles, light vehicles, origin-destination flows, machine learning

INTRODUCTION AND BACKGROUND

Urban transportation networks are complex systems that significantly influence economic efficiency, environmental sustainability, and quality of life in cities worldwide. These networks are shaped by diverse vehicle types and their movement patterns, with a crucial distinction between Heavy Vehicles (HV) and Light Vehicles (LV). HVs, encompassing trucks, buses, and coaches, constitute a significant portion of total traffic volume. Trucks usually account for up to 15% of traffic, while buses may represent approximately 3% [2]. However, these figures exhibit substantial variability across different regions, influenced by factors such as local industrial activity, transportation infrastructure, and the extent of public transit systems.

The ability to distinguish between LV and HV movements is crucial for several reasons. It allows for a more detailed understanding of urban logistics and commercial activities, as HVs are often indicative of goods movement and industrial operations [3, 4]. It also provides insights into public transportation patterns as buses belong to the HV category. This decomposition by vehicle type enables more targeted approaches to traffic management, infrastructure planning, and policy-making. Despite constituting a small proportion of total traffic, HVs contribute disproportionately to various urban challenges. Commercial vehicles significantly impact traffic congestion and emissions [5, 6]. In terms of infrastructure wear, HVs are responsible for a disproportionate share of the road maintenance burden [7]. This underscores the need for accurate identification and disaggregation of HV movements to generate effective policies and manage their impact on urban areas.

Even though the importance of differentiating between vehicle types is recognized, achieving a fine level of disaggregation in traffic flows has remained a significant challenge in transportation research. The primary obstacles have been the lack of comprehensive flow data at the link level and the scarcity of disaggregated flow observations. While the increased use of sensors such as loop detectors and cameras has provided disaggregated flow data at select locations, estimating potential flow at every link of a network has been largely unattainable.

Recent advancements in transportation modeling tools, such as the RapidEx [1], have created new opportunities to address these research gaps. By leveraging travel time data from sources such as Google or TomTom, in conjunction with road network information from OpenStreetMap, RapidEx can now estimate potential flow at every link in a network, achieving travel times that closely align with observed data. This capability, combined with the tool's ability to generate OD patterns consistent with these flows, provides an opportunity for fine-grained traffic flow disaggregation.

However, the application of such advanced tools often faces challenges due to the lack of detailed, decomposed data needed to generate accurate OD patterns for all links in a network. Even in developed nations, the availability of such data frequently lacks the necessary granularity and network coverage [8]. While sensor deployment for obtaining decomposed flow data is increasing in some areas, it is not yet commonplace worldwide. This data scarcity highlights the need for models that can be developed in data-rich areas and then applied to other contexts where specific, detailed transportation data is limited.

To address this challenge, understanding the relationship between transportation and economic activity becomes crucial. This relationship is fundamental for urban planning and policy-making [3, 4], as it allows for the creation of transferable models that can estimate traffic patterns based on more commonly available economic indicators. While this applies to all types of vehicle movements, it is particularly relevant for HVs, which include both freight transport and public

transit. The movement of HVs, especially in freight transport, plays a vital role in bridging the gap between production and consumption locations. This necessitates the development of models that integrate economic activity, logistics decision-making, and traffic flows [9, 10]. By focusing on these relationships, we can create more comprehensive models that account for both LV and HV movements, providing a better understanding of urban transportation dynamics.

In recent years, there has been increased interest in freight transport research, driven by the need to better understand freight flows and volumes within transportation networks [10, 11]. However, obtaining reliable freight trip generation data remains challenging, especially where large-scale freight transport surveys are infrequent [12, 13]. To address these limitations, researchers have explored various proxies for economic activity in trip generation models [14–18]. Among these proxies, POI data and nighttime light intensity data have gained significant attention as indicators of economic activity [17, 18]. POI data, representing entities with geolocation information, has shown strong potential in transport research [19, 20]. These studies demonstrate the effectiveness of such proxies in estimating patterns of socio-economic change and transport demand at various geographical scales [21].

To fully capitalize on these advancements and explore relationships between vehicle movement patterns and urban socio-economic indicators, machine learning techniques have emerged as a promising approach. In particular, XGBoost has shown impressive results in traffic prediction tasks [22–24]. Its ability to handle complex non-linear relationships and robustness to overfitting makes it particularly suitable for transportation data analysis, enabling the development of models that link socio-demographic factors and other proxies to vehicle type splits.

Our research leverages these techniques to develop an innovative approach for decomposing urban traffic flows into LV and HV categories. Focusing on Greater Sydney's road network, we aim to identify relationships between vehicle movement patterns and urban socio-economic indicators. This study is motivated by the need to provide urban planners, policymakers, and transportation engineers with a more comprehensive understanding of the factors influencing HV and LV movements in urban areas. The significance of this research extends beyond the immediate case study. By developing a machine learning-based regression model that links socio-demographic factors and other proxies to vehicle type splits, we create a framework that can be generalized to other regions or cities. Our approach addresses the scarcity of observed link-level decomposed flow data and offers a scalable solution for urban transportation analysis on a global scale, moving beyond traditional survey-based methods to provide a more accurate and transferable understanding of urban transportation patterns and their relationship to economic activities.

Major contributions of our study are as follows:

1. We develop a novel multi-stage methodology that integrates advanced network modeling, optimization techniques, and machine learning to disaggregate traffic flows.
2. We provide a comprehensive analysis of the factors influencing HV proportions in urban areas, offering a nuanced view of the complex dynamics governing urban freight movement.
3. We establish a generalizable framework for predicting vehicle type distributions across entire urban networks, enabling more accurate forecasting and scenario analysis for urban planning in diverse contexts.

By bridging the gap between aggregated traffic data and disaggregated vehicle flows, this research aims to enhance our understanding of urban transportation dynamics and provide practical tools for evidence-based decision making. The insights gained from this study have the potential to

make a noticeable difference towards improving urban mobility efficiency, reducing congestion, and supporting more sustainable urban development practices across various urban environments.

The remainder of this paper is organized as follows: Section 2 describes different data sources used. Section 3 describes detailed methodology. Section 4 presents results and discusses their implications for urban planning and transportation policy. Finally, Section 5 concludes with a summary of our findings and suggestions for future research directions.

DATA

The study leverages three comprehensive datasets that capture various aspects of the urban activity, and nightlight intensity in Sydney, Australia. These datasets provide a rich and diverse set of information, enabling a thorough analysis of the relationship between urban characteristics and the composition of traffic flows.

The first dataset, Points of Interest (POI), is sourced from SafeGraph, a leading provider of geospatial data. SafeGraph is renowned for its comprehensive and accurate POI data, which is widely used in academic research [25–27], government analyses, and business applications. The company employs rigorous data collection and validation processes, including both automated and manual checks, to ensure data quality. The POI data for the year 2023 describes the distribution of various amenities and services associated with POIs in the Sydney region. This information is crucial for understanding the spatial patterns of urban activities and their potential influence on traffic composition.

The visible infrared imaging radiometer suite (VIIRS) day/night band (DNB) nighttime lights data, maintained and provided by the Earth Observation Group[1], is increasingly recognized as the gold standard for economic studies utilizing night-lights [28–30]. This dataset captures the intensity of artificial lighting during nighttime hours, which has been demonstrated to correlate strongly with economic activity and urbanization levels [31]. For this study, we employed the Annual VNL V2 [32] average radiance composite data for 2023. This dataset is part of the VIIRS nighttime light time series, which provides cloud-free average radiance values from nighttime VIIRS DNB data. The VNL V2 product offers several advantages over its predecessors, including improved filtering of aurora, stray light, and other ephemeral lights. We selected the 2023 data to ensure our analysis reflects the most recent available information on nighttime light patterns in our study area.

Lastly, the study incorporates traffic volume data from Transport for New South Wales for the year 2023. The Traffic Volume Viewer provided by the transport authority splits the traffic flow for specific links in the road network into HV and LV, offering valuable insights into their respective proportions and counts [33]. This data source has been utilized in published research studies, demonstrating its reliability and relevance in traffic analysis [34].

Table 1 presents summary statistics for key variables in the dataset, encompassing urban features, population characteristics, and traffic-related metrics. The statistics reveal several noteworthy patterns across the studied areas. Among urban amenities, services and retail establishments appear to be the most prevalent, with mean counts of 709.50 and 296.83, respectively. The high standard deviations for these variables suggest considerable variation in their spatial distribution. Nightlight intensity, used as a proxy for economic activity, exhibits a mean value of 25.51

[1] https://eogdata.mines.edu/products/vnl/

with a standard deviation of 24.19, indicating diverse levels of urbanization or economic development across the study area. The population variable also shows substantial variation, with a mean of 27,659.27 and a standard deviation of 31,814.99, reflecting the heterogeneous distribution of population densities within Sydney.

TABLE 1 Descriptive Statistics of Key Variables

	POI_education	POI_employment	POI_food	POI_medical	POI_retail
mean	21.47	36.07	198.20	184.22	296.83
std	25.61	62.15	221.02	230.21	331.23
min	0.00	0.00	2.00	1.00	4.00
25%	5.00	6.25	53.00	36.00	84.75
50%	11.00	17.50	119.50	101.00	210.00
75%	29.00	36.00	245.75	215.25	345.75
max	161.00	621.00	1455.00	1157.00	2139.00
	POI_services	POI_transit	Nightlight_intensity	Population	Area
mean	709.50	49.60	25.51	27659.27	20.07
std	863.77	53.07	24.19	31814.99	19.43
min	10.00	1.00	0.42	14.72	0.89
25%	224.25	16.00	7.55	7506.12	6.20
50%	403.50	31.00	20.68	16118.90	6.07
75%	857.50	55.75	33.80	33588.74	43.42
max	6868.00	322.00	127.02	157582.34	43.46

METHODOLOGY

This study presents a novel multi-stage approach to estimate the proportions of HVs and LVs in OD matrices for transportation network analysis. Our methodology integrates advanced network modeling techniques, optimization algorithms, and machine learning to provide a comprehensive solution for disaggregating vehicle types in OD flows. The approach consists of three stages: OD matrix estimation, optimization-based proportion estimation, and machine learning-based generalization.

OD Matrix Estimation

The first stage of our methodology focuses on estimating the OD matrix for the study area using the RapidEx [1]. RapidEx is a powerful network modeling software that leverages OpenStreetMap data to extract road network information that has been successfully applied to analyze similar road networks [35]. To manage computational complexity while maintaining essential traffic corridors, we focused on motorways, trunk roads, and primary roads. This network simplification strategy allows for efficient processing without compromising the integrity of major HV and LV flow patterns. Travel time data for each link was collected using the Google Maps API for a representative weekday morning peak period (June 6, 2024, 7:00-9:00 AM). This time window was selected to capture peak traffic conditions. Additionally, we gathered population data from WorldPOP [36] and categorized POI information to inform the demand estimation process.

Figure 1 illustrates the road network of Greater Sydney, color-coded as motorways (dark blue), trunk roads (cyan), and primary roads (brown). Secondary and tertiary roads are omitted, as the selected network captures the majority of HV and LV flows. The overlay shows zones (158) based on the Uber H3 hexagonal hierarchical spatial index. These zones start at resolution 6 (approximately 36 km^2 per hexagon) and are further subdivided up to resolution 8 (approximately 0.9 km^2 per hexagon) based on POI density.

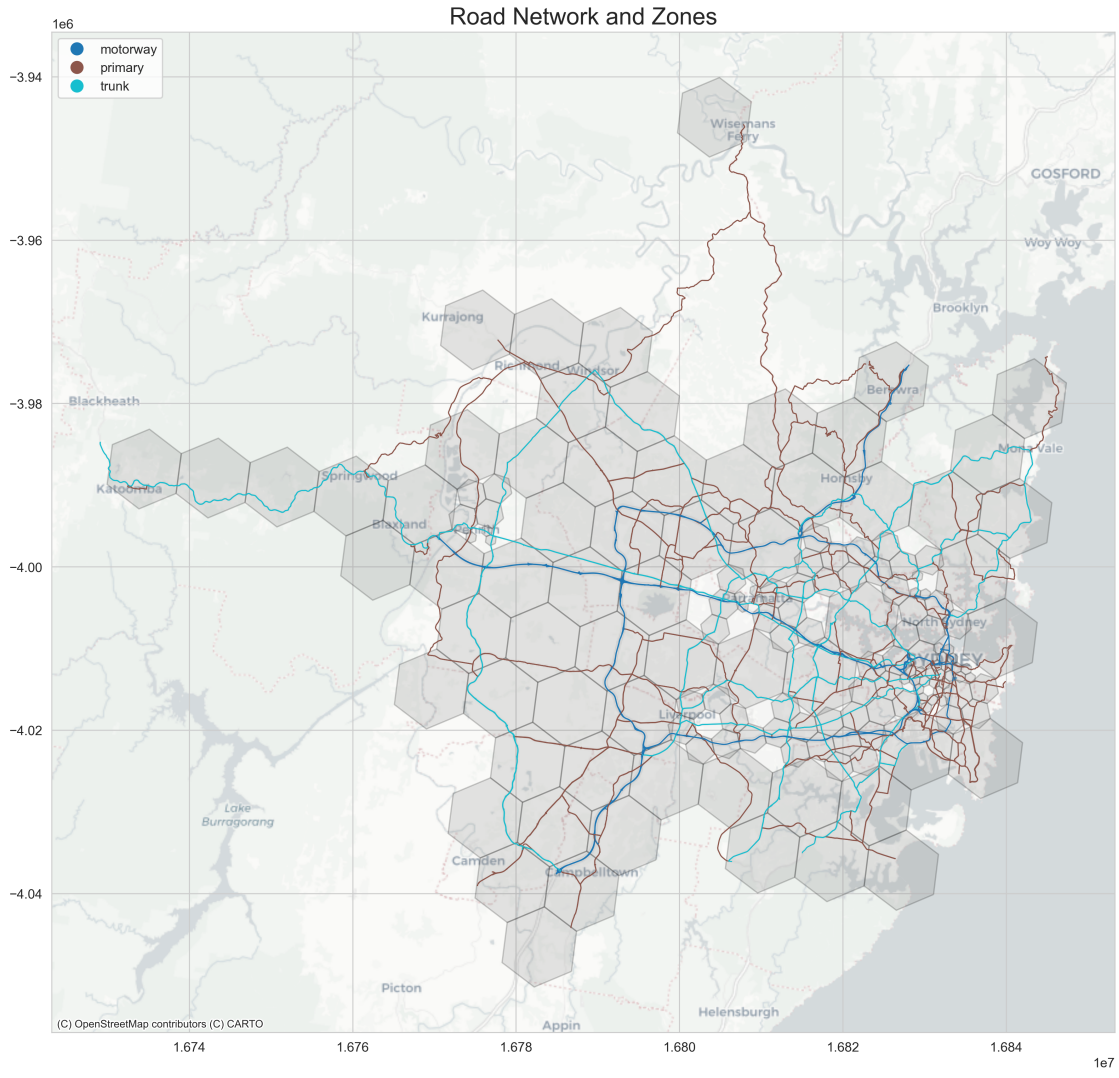

FIGURE 1 Road Network and Zones of Greater Sydney

The RapidEx tool employs a bi-level approach for OD matrix estimation. In the upper level, OD demand is initially estimated using a line search for total demand ranging from 5,000 to 1,500,000 vehicles, utilizing population and POI data proportions to distribute demand among OD pairs. The lower level consists of a traffic assignment problem, determining link-level flows and travel times based on the current OD estimate. A genetic algorithm then operates at the upper level, adjusting the OD demand based on a fitness function that measures the match between estimated and observed traffic flows or travel times at the lower level. This bi-level problem is solved

iteratively until a specified convergence criterion is met. For clarity, it's important to note that the OD demand estimated by RapidEx is equivalent to Passenger Car Units (PCU), a standardized measure of traffic flow accounting for various vehicle characteristics. For a more comprehensive understanding of the methodology and its implementation, readers are encouraged to refer to [1].

Optimization-Based Proportion Estimation

The second stage of our methodology focuses on estimating the proportions of HVs and LVs for each OD pair using an optimization-based approach. Since the OD demand estimated by RapidEx is in PCU and we have the OD contribution to each link (derived during the traffic assignment step), different proportions of HVs and LVs will not impact travel times as the PCU value remains constant. The number of HVs and LVs can vary while maintaining the same PCU demand, as multiple values satisfy the PCU for different HV and LV combinations.

The optimization model considers all OD pairs and observed flow [33] at a few links to identify the optimal and unique proportions. By leveraging the OD matrix from the previous stage and the contribution of each OD pair to link flows, we formulated a quadratic programming optimization model to determine HV and LV proportions at the OD level. Since demand is constant and the value of travel time will not be affected by these proportions, we eliminated the need to solve the traffic assignment problem at every iteration, thus making the optimization process extremely fast.

The optimization problem is defined as follows:

Objective Function:

$$\min \sum_{l \in obs} \left(F^H_{l,obs} - \sum_{(i,j) \in OD} ODL_{i,j,l} \times p^H_{i,j} \right)^2 + \left(F^L_{l,obs} - \sum_{(i,j) \in OD} ODL_{i,j,l} \times p^L_{i,j} \right)^2 \quad (1)$$

Subject to:

$$p^H_{i,j} + p^L_{i,j} = 1 \forall (i,j) \in OD \quad (2)$$

$$p^H_{i,j} \in [0,1] \forall (i,j) \in OD \quad (3)$$

$$p^L_{i,j} \in [0,1] \forall (i,j) \in OD \quad (4)$$

Where $F^H_{l,obs}$ and $F^L_{l,obs}$ are the observed flows for HVs and LVs on link l, respectively. $ODL_{i,j,l}$ represents the demand on link l contributed by OD pair (i,j), and p^H_{ij} and p^L_{ij} are the proportions of HVs and LVs for OD pair (i,j).

To solve this optimization problem, we employed Gurobi 11, a commercial optimization solver known for its efficiency in handling large-scale quadratic programming problems. To ensure a high-quality solution, a relative optimality gap of 1e-6 is set as the stopping criterion.

The output of this stage is a set of HV and LV proportions for each OD pair in the study area. These proportions provide a detailed understanding of the composition of traffic flows between zones, enabling more accurate modeling and analysis of HV and LV movements in the network.

Machine Learning-Based Generalization

The final stage of our methodology focuses on developing a machine learning model to generalize the relationship between various socio-economic, land-use, and demographic factors and the proportions of HVs in OD flows. While the previous stage provides exact proportions at the OD pair level for the study area, this stage aims to build a model that can be applied to other regions or time periods where detailed data may not be available. We proposed an XGBoost regression model [37] to capture the complex relationships between these factors and the proportions of HVs. The proportion of LVs can be easily derived as the complement of the HV proportion (i.e., 1 - HV proportion).

To train the XGBoost model, we engineered a comprehensive set of features that capture the factors influencing HV and LV proportions. These features included POI data and its sub-categories (employment centers, educational institutions, transit hubs, retail establishments, medical facilities, and service related), which serve as proxies for land-use patterns and economic activity; nightlight intensity data, an additional proxy for economic activity [38], capturing the intensity of human settlements and industrial areas; population data, representing the distribution of residential areas and potential trip generation/attraction zones; and zonal area, accounting for the size and spatial extent of each zone. These features were compiled for both origin and destination zones of each OD pair, providing a rich set of explanatory variables for the XGBoost model.

To improve model quality and reduce noise in the dataset, we applied several preprocessing steps. OD pairs with zero demand and those with an average annual daily traffic (AADT) for HVs less than 1 were excluded. All features underwent z-score normalization to prevent bias from high-magnitude variables. The dataset was partitioned into an 80% training set and a 20% testing set.

To optimize the performance of the XGBoost model, we conducted an extensive hyperparameter tuning process using grid search. The grid search explored a vast space of 468,750 parameter combinations, including key parameters such as maximum depth, minimum child weight, subsample, column sample by tree, learning rate, alpha (L1 regularization), lambda (L2 regularization), and the number of estimators. This exhaustive search allowed us to identify the best combination of hyperparameters that maximize the model's predictive accuracy and generalization ability.

The trained XGBoost model was then evaluated on the testing set to assess its performance in predicting HV proportions for unseen OD pairs. To gain further insights into the model's behavior and interpretability, we conducted residual analysis to check for heteroskedasticity and bias, performed feature importance analysis to identify the most influential factors in determining HV proportions, and created partial dependence plots to visualize the relationship between key features and the target variable.

The machine learning-based generalization stage offers several benefits for extending the HV proportion estimates to other regions or time periods. The XGBoost model captures the complex non-linear relationships between the input features and the target variable, allowing for accurate predictions even in the absence of detailed flow data. Furthermore, the model's ability to handle large datasets and its robustness to outliers and noise make it suitable for application in diverse urban contexts.

The output of this stage is a fully trained XGBoost model that can predict HV proportions for any OD pair, given the corresponding input features. This model serves as a powerful tool for transportation planners and policymakers, enabling them to assess the impact of different scenarios

on the distribution of HVs in the network and make informed decisions regarding infrastructure investments, traffic management strategies, and environmental policies.

RESULTS & DISCUSSION

The efficacy of our multi-stage approach is demonstrated through rigorous validation and performance metrics. The OD matrix estimation showed high accuracy, with link flows within 4% error on links with available count data, and travel times within 10% of observed values for 85% of the links. This level of precision in the initial stage provides a solid foundation for subsequent analyses.

Our XGBoost model, optimized through extensive hyperparameter tuning, achieved robust performance metrics. The optimal configuration included 1200 estimators, a maximum depth of 5, minimum child weight of 4, subsample and column sample by tree both at 0.8, learning rate of 0.06, and regularization parameters alpha and lambda at 0.3 and 9, respectively. This configuration was determined using five-fold cross-validation, with mean squared error (MSE) as the primary selection metric, ensuring a balance between predictive accuracy and model generalizability.

The final model demonstrated strong predictive capability, achieving a training R^2 of 0.931, a test R^2 of 0.637, and a cross-validation R^2 of 0.566 (standard deviation: 0.015). The Mean Absolute Error (MAE) of 0.137 further underscores the model's accuracy. These performance metrics indicate that our model captures a significant portion of the variance in HV proportions across the network, providing reliable insights into the factors influencing urban freight movement patterns.

Our analysis reveals significant insights into the factors influencing HV proportions in the origin-destination matrices. The machine learning model proposed demonstrates a robust ability to capture complex patterns in transportation flows, as evidenced by the feature importance analysis and partial dependence plots. These findings provide a in-depth understanding of the relationships between urban characteristics and HV traffic, offering valuable information for urban planning and transportation policy decisions.

The feature importance analysis, presented in Figure 2, identified key drivers of HV proportions. The top five most influential features are nightlight_intensity_Destination (0.0912), services_Destination (0.0791), Area_Destination (0.0783), food_Destination (0.0681), and Population_Destination (0.0648). This hierarchy of features underscores the critical role of destination characteristics in determining HV traffic patterns. The presence of services and area-related features at both origins and destinations highlights the model's ability to capture the bi-directional nature of freight transportation. Notably, the inclusion of nightlight intensity as a significant factor demonstrates the model's capacity to incorporate proxy indicators of economic activity, enhancing its predictive power.

The partial dependence plots, shown in Figure 3, provide valuable insights into the relationships between specific features and HV proportions. Nightlight_intensity_Destination demonstrates a complex, non-linear relationship with HV proportion. It shows an initial increase, followed by a decline, and then another rise. This pattern might reflect the varying nature of nighttime activities in different urban contexts, with some levels of nightlight intensity corresponding to industrial areas that generate more HV traffic, while others might indicate residential areas with less freight movement. The services_Destination plot shows an initial sharp increase in HV proportion as service-related activities increase, followed by a more gradual rise and eventual plateau. This suggests that areas with even a moderate level of service activities generate significantly more HV

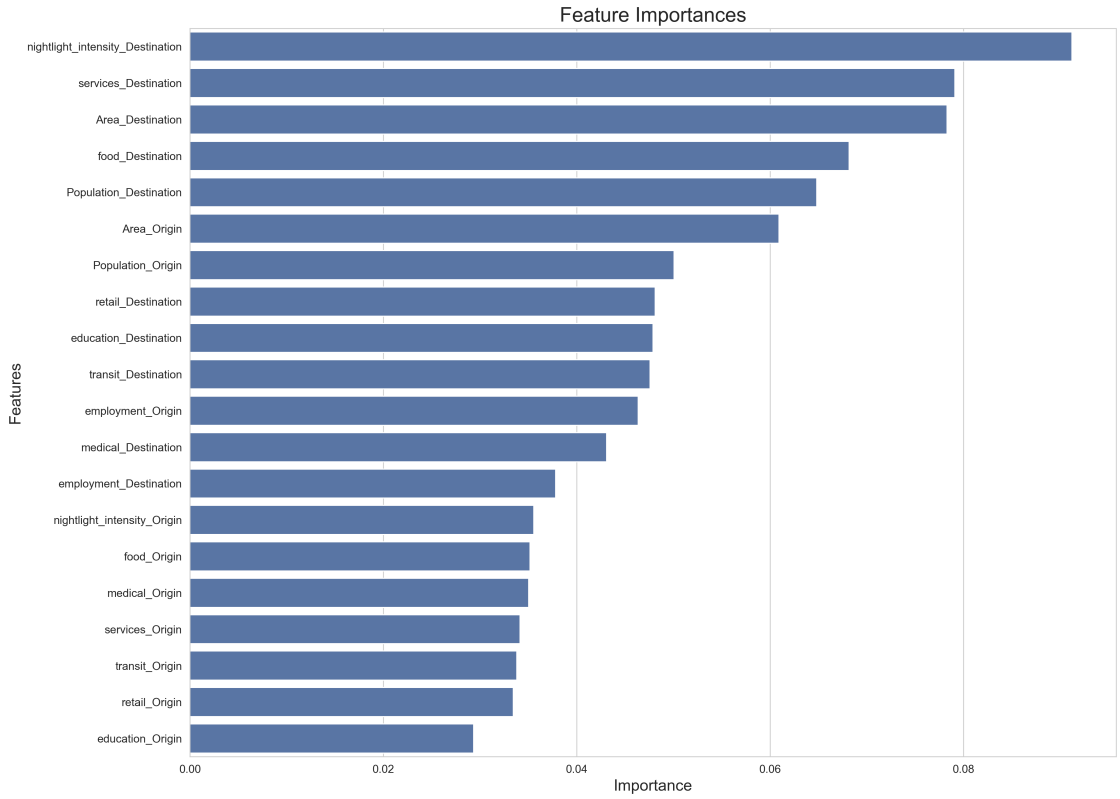

FIGURE 2 Feature Importances

traffic compared to areas with very low service presence. However, beyond a certain threshold, additional service destinations do not substantially increase HV proportions. This pattern aligns with the intuitive understanding that service-oriented areas attract freight movement, but there may be a saturation point in terms of the logistics needs of these areas. The Area_Destination plot exhibits a step-like pattern, indicating discrete jumps in HV proportion as the destination area increases. This could reflect zoning or land-use policies that cluster industrial or logistics activities in specific area ranges, leading to abrupt increases in HV traffic at certain thresholds. The food_Destination plot reveals a sharp initial increase in HV proportion, followed by a gradual decline. This could indicate that areas with moderate food-related destinations attract more HV traffic, possibly for restocking, while areas with very high concentrations might rely more on frequent, smaller deliveries using light commercial vehicles. Population_Destination plot displays a fluctuating pattern, indicating a complex relationship between population and HV proportion. This could reflect the interplay between residential areas and the commercial/industrial zones that generate HV traffic, with certain population densities corresponding to mixed-use areas that attract more freight movement. Finally, the Area_Origin shows a simpler step function, suggesting that the size of the origin area has a more straightforward impact on HV proportion, with larger areas generally associated with higher HV traffic. This could reflect the tendency of larger origin areas to house more diverse economic activities that generate freight movement. These nuanced relationships highlight the complex interplay of urban features in determining HV traffic patterns, underscoring the value of our machine learning approach in capturing these intricate dynamics.

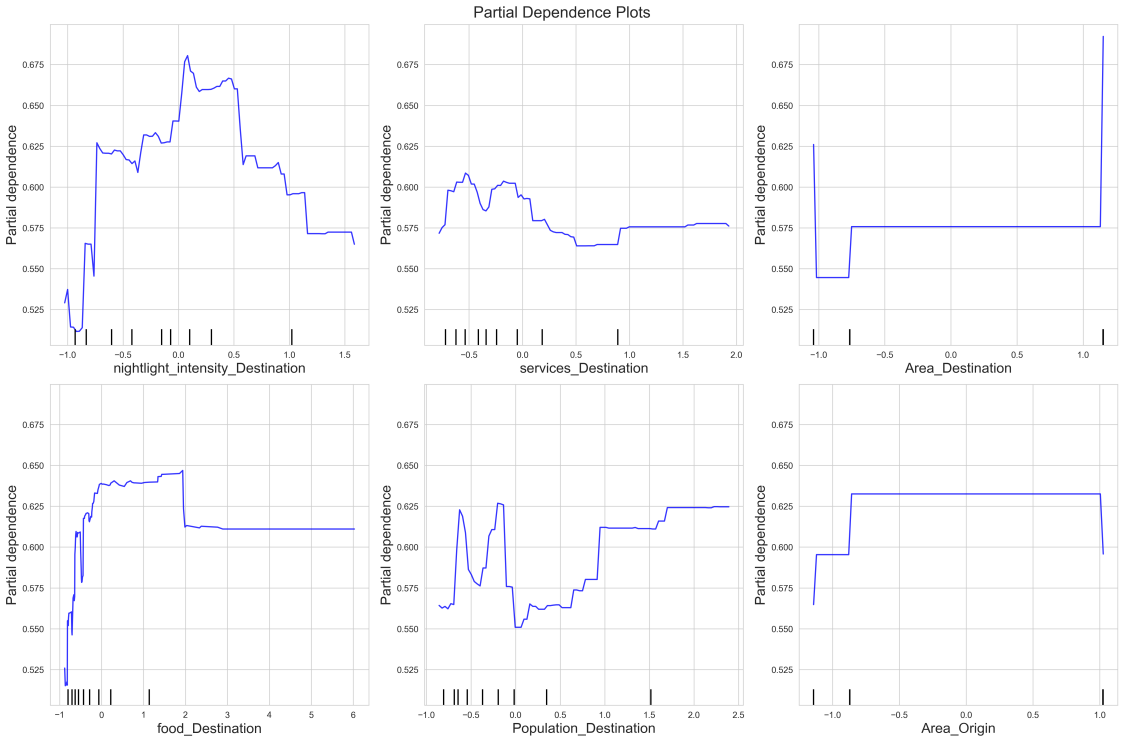

FIGURE 3 Partial Dependence Plots

The model's performance is further validated by the Q-Q plot presented in Figure 4. This plot compares the distribution of the model's residuals to a normal distribution, demonstrating a close alignment with the diagonal line. It indicates that the residuals closely follow a normal distribution, which is a desirable characteristic for regression models and supports the reliability of our predictions.

The residual plot, shown in Figure 5, provides additional insight into the model's performance. While the residuals are generally distributed around the zero line, there is a noticeable pattern of increasing spread as the predicted values increase. This pattern suggests the presence of heteroskedasticity in the model. The observed heteroskedasticity in our model, while noteworthy, should be interpreted within the context of real-world transportation patterns. It is important to recognize that large HV proportions are relatively rare in urban transportation networks, with LVs typically dominating traffic flows in most areas. This reality of transportation systems has significant implications for our model's performance and interpretation.

The increasing variability in residuals at higher predicted values likely corresponds to these less common, high-HV proportion scenarios. Given the rarity of such cases, the increased variability in predictions for these instances is less likely to significantly impact the model's overall utility for transportation planning and policy-making. The majority of predictions, which fall in the more common lower and moderate HV proportion ranges, demonstrate more consistent accuracy. Furthermore, this pattern in the residuals provides valuable insights into the nature of HV traffic prediction. The greater variability for high-HV proportion scenarios may reflect the more complex and diverse factors influencing these unusual cases. For instance, areas with exceptionally high HV proportions might be subject to unique local conditions which introduce additional complexity

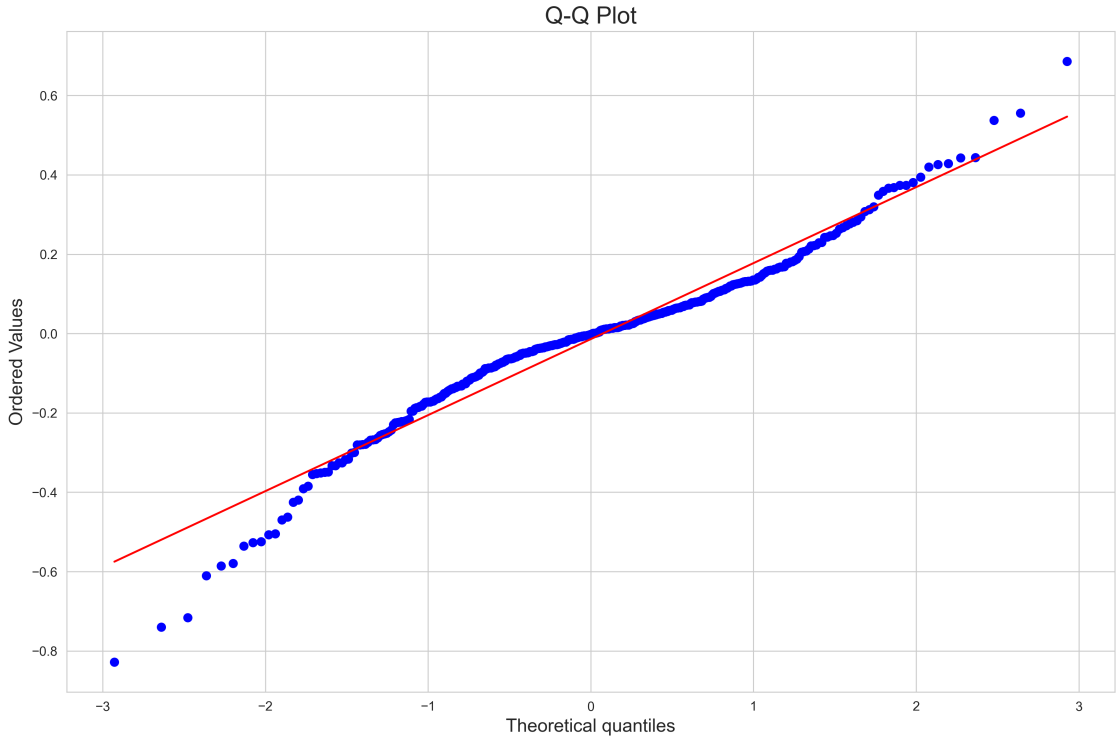

FIGURE 4 Q-Q Plot

not captured by our general set of predictors.

Despite this limitation, the model's strong performance in predicting the more common scenarios of lower to moderate HV proportions remains significant. The mean of residuals (-0.0158) being close to zero indicates minimal overall bias in the predictions, further supporting the model's reliability for typical urban transportation contexts. While our model exhibits some heteroskedasticity, particularly for the rare high-HV proportion scenarios, it successfully captures key relationships between urban features and HV proportions for the majority of more common cases. This characteristic aligns well with the practical needs of urban transportation planning, where accurate predictions for typical scenarios are most crucial. The model's performance, combined with its ability to highlight areas requiring specialized attention, provides a robust foundation for informed decision-making in urban freight management and transportation policy.

CONCLUSION

This study introduces an innovative approach to decomposing urban traffic flows into light vehicle (LV) and heavy vehicle (HV) categories, addressing a critical gap in transportation network analysis. Our three-stage methodology, combining OD matrix estimation, quadratic programming optimization, and XGBoost regression, provides a robust framework for estimating HV and LV proportions across Greater Sydney's road network. By leveraging diverse data sources, we have proposed a comprehensive model that achieves a test R^2 of 0.637, demonstrating its effectiveness in capturing complex urban mobility patterns.

Our analysis reveals intricate, non-linear relationships between urban features and HV proportions. The top five influential features identified are nightlight_intensity_Destination, ser-

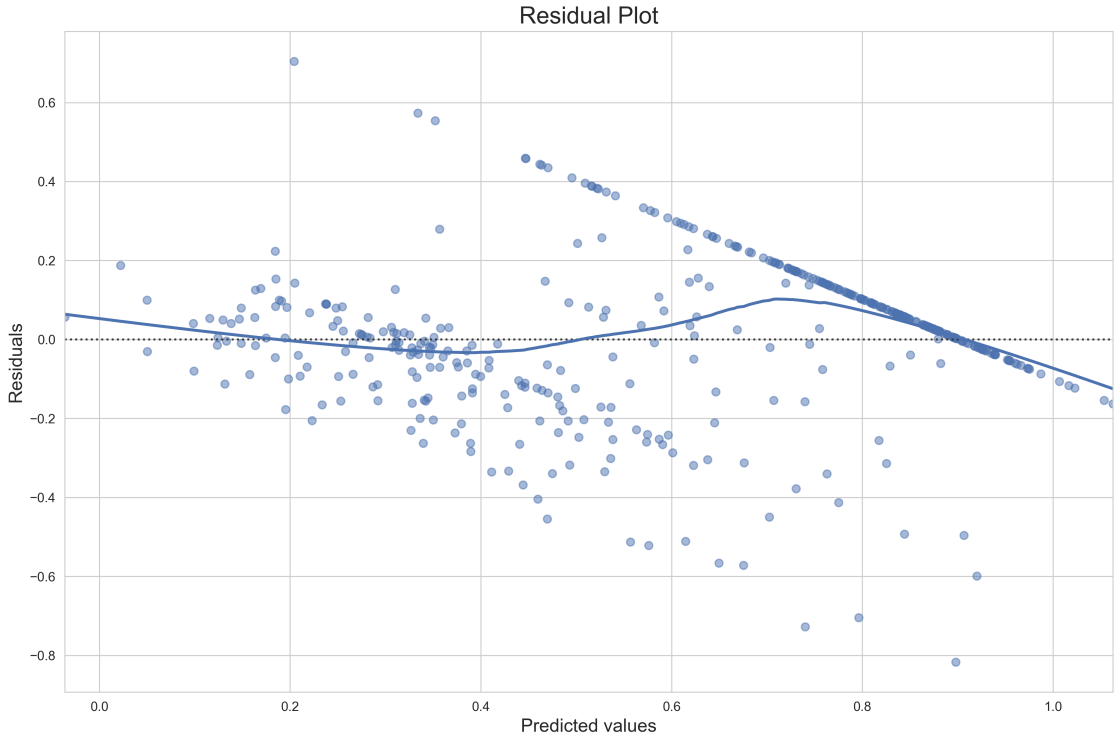

FIGURE 5 Residual Plot

vices_Destination, Area_Destination, food_Destination, and Population_Destination. Notably, nightlight intensity at the destination shows a complex pattern with initial increase, decline, and subsequent rise, reflecting varying nighttime activities across urban contexts. Services_Destination exhibits a sharp initial increase followed by a plateau, suggesting a saturation point in HV traffic generation. Area_Destination displays a step-like pattern, potentially indicating the impact of zoning policies on HV traffic. These nuanced relationships underscore the complex interplay of urban features in determining HV traffic patterns and highlight the value of our machine learning approach in capturing these intricate dynamics.

Expanding this framework to diverse global urban contexts would test its generalizability and potentially yield insights for creating more equitable and sustainable transportation systems. By incorporating socio-economic factors and developing policy recommendations, this approach could become a powerful tool for comprehensive urban planning. It would bridge the gap between transportation efficiency and social equity in our evolving smart cities, enabling the application of Mobility as a Resource (MaaR) concepts. This enhanced model could support decision-makers in optimizing urban mobility solutions that are not only efficient but also socially inclusive and environmentally sustainable, aligning with the broader goals of smart city initiatives and sustainable urban development.

Future research could address the limitations identified in this study, particularly focusing on improving predictions for high-HV proportion scenarios. Developing specialized models or incorporating additional features specific to areas with high HV traffic could enhance the framework's accuracy across all traffic conditions. This might involve integrating more detailed data on industrial activities, zoning regulations, or specific infrastructure designed for HVs. Addition-

ally, extending the model to capture dynamic scenarios by incorporating time-dependent variables could provide insights into daily or seasonal variations in HV proportions.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: PF, SRK, STW; data collection: PF, SRK; analysis and interpretation of results: PF, SRK, DYL, STW; draft manuscript preparation: PF, SRK, DYL, STW. All authors reviewed the results and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

The authors extend their gratitude to Anna Sotnikova for her assistance in matching observed flows to the exact links, which significantly contributed to the quality of this research.

REFERENCES

1. Waller, S. T., S. Chand, A. Zlojutro, D. Nair, C. Niu, J. Wang, X. Zhang, and V. V. Dixit, Rapidex: a novel tool to estimate origin–destination trips using pervasive traffic data. *Sustainability*, Vol. 13, No. 20, 2021, p. 11171.
2. Evgenikos, P., G. Yannis, K. Folla, R. Bauer, K. Machata, and C. Brandstaetter, Characteristics and causes of heavy goods vehicles and buses accidents in Europe. *Transportation research procedia*, Vol. 14, 2016, pp. 2158–2167.
3. Dablanc, L., City distribution, a key element of the urban economy: guidelines for practitioners. In *City distribution and urban freight transport*, Edward Elgar Publishing, 2011.
4. Taniguchi, E., R. G. Thompson, and T. Yamada, Recent trends and innovations in modelling city logistics. *Procedia-Social and Behavioral Sciences*, Vol. 125, 2014, pp. 4–14.
5. Figliozzi, M. A., The impacts of congestion on commercial vehicle tour characteristics and costs. *Transportation research part E: logistics and transportation review*, Vol. 46, No. 4, 2010, pp. 496–506.
6. Hartgen, D. T., M. G. Fields, A. L. Layzell, and E. S. Jose, How employers view traffic congestion: Results of National Survey. *Transportation research record*, Vol. 2319, No. 1, 2012, pp. 56–66.
7. Song, Y., X. Wang, G. Wright, D. Thatcher, P. Wu, and P. Felix, Traffic volume prediction with segment-based regression kriging and its implementation in assessing the impact of heavy vehicles. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 20, No. 1, 2018, pp. 232–243.
8. Stopher, P. R. and S. P. Greaves, Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice*, Vol. 41, No. 5, 2007, pp. 367–381.
9. Caspersen, E., An explorative approach to freight trip attraction in an industrial urban area. *City Logistics 3: Towards Sustainable and Liveable Cities*, 2018, pp. 249–268.
10. Sánchez-Díaz, I., J. Holguín-Veras, and X. Wang, An exploratory analysis of spatial effects on freight trip attraction. *Transportation*, Vol. 43, 2016, pp. 177–196.
11. Holguín-Veras, J., M. Jaller, L. Destro, X. Ban, C. Lawson, and H. S. Levinson, Freight generation, freight trip generation, and perils of using constant trip rates. *Transportation Research Record*, Vol. 2224, No. 1, 2011, pp. 68–81.
12. Pani, A. and P. K. Sahu, Planning, designing and conducting establishment-based freight

surveys: A synthesis of the literature, case-study examples and recommendations for best practices in future surveys. *Transport Policy*, Vol. 78, 2019, pp. 58–75.

13. Middela, M. S. and G. Ramadurai, Spatial Seemingly Unrelated Regression Models for Freight Trip Generation by Vehicle Type: Application to the Chennai Metropolitan Area in India. *Transportation Research Record*, Vol. 2676, No. 4, 2022, pp. 380–392.
14. Giuliano, G., S. Kang, Q. Yuan, N. Hutson, and M. T. Center, *The freight landscape: using secondary data sources to describe metropolitan freight flows*, 2015.
15. Kang, S., Q. Yuan, N. Hutson, G. Giuliano, and M. T. Center, *The freight landscape: using secondary data sources to describe metropolitan freight flows.*, 2010.
16. Alho, A. R. and J. d. A. e Silva, Analyzing the relation between land-use/urban freight operations and the need for dedicated infrastructure/enforcement—Application to the city of Lisbon. *Research in Transportation Business & Management*, Vol. 11, 2014, pp. 85–97.
17. Mellander, C., J. Lobo, K. Stolarick, and Z. Matheson, Night-time light data: A good proxy measure for economic activity? *PloS one*, Vol. 10, No. 10, 2015, p. e0139779.
18. Yan, G., L. Zou, and Y. Liu, The Spatial Pattern and Influencing Factors of China's Night-time Economy Utilizing POI and Remote Sensing Data. *Applied Sciences*, Vol. 14, No. 1, 2024.
19. Jiang, S., A. Alves, F. Rodrigues, J. Ferreira Jr, and F. C. Pereira, Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, Vol. 53, 2015, pp. 36–46.
20. Yuan, J., Y. Zheng, and X. Xie, Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 186–194.
21. Gao, S., K. Janowicz, and H. Couclelis, Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, Vol. 21, No. 3, 2017, pp. 446–467.
22. Zhang, Y. and A. Haghani, A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 308–324.
23. Xu, J., M. Saleh, and M. Hatzopoulou, A machine learning approach capturing the effects of driving behaviour and driver characteristics on trip-level emissions. *Atmospheric Environment*, Vol. 224, 2020, p. 117311.
24. Zafar, N. and I. Ul Haq, Traffic congestion prediction based on Estimated Time of Arrival. *PloS one*, Vol. 15, No. 12, 2020, p. e0238200.
25. Jiao, J., M. Bhat, and A. Azimian, Measuring travel behavior in Houston, Texas with mobility data during the 2020 COVID-19 outbreak. *Transportation letters*, Vol. 13, No. 5-6, 2021, pp. 461–472.
26. Li, Z., H. Ning, F. Jing, and M. N. Lessani, Understanding the bias of mobile location data across spatial scales and over time: a comprehensive analysis of SafeGraph data in the United States. *Plos one*, Vol. 19, No. 1, 2024, p. e0294430.
27. Prestby, T., J. App, Y. Kang, and S. Gao, Understanding neighborhood isolation through spatial interaction network analysis using location big data. *Environment and Planning A: Economy and Space*, Vol. 52, No. 6, 2020, pp. 1027–1031.
28. Bennett, M. M. and L. C. Smith, Advances in using multitemporal night-time lights satellite imagery to detect, estimate, and monitor socioeconomic dynamics. *Remote Sensing of Environment*, Vol. 192, 2017, pp. 176–197.

29. Gibson, J., S. Olivia, G. Boe-Gibson, and C. Li, Which night lights data should we use in economics, and where? *Journal of Development Economics*, Vol. 149, 2021, p. 102602.
30. Price, N. and P. M. Atkinson, Global GDP Prediction With Night-Lights and Transfer Learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 15, 2022, pp. 7128–7138.
31. Chen, X. and W. D. Nordhaus, Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, Vol. 108, No. 21, 2011, pp. 8589–8594.
32. Elvidge, C. D., M. Zhizhin, T. Ghosh, F.-C. Hsu, and J. Taneja, Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sensing*, Vol. 13, No. 5, 2021, p. 922.
33. New South Wales Roads and Maritime Services, *Traffic Volume Viewer*. `http://www.rms.nsw.gov.au/about/corporate-publications/statistics/traffic-volumes/aadt-map/index.html#?z=6`, 2023, [Online; accessed 6-June-2024].
34. Infante, W. and J. Ma, Clustering and Previous Visit Dependency Technique for Electric Vehicle Station Visits. In *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, 2018, pp. 1–5.
35. Waller, S. T., M. Qurashi, A. Sotnikova, L. Karva, and S. Chand, Analyzing and modeling network travel patterns during the Ukraine invasion using crowd-sourced pervasive traffic data. *Transportation research record*, Vol. 2677, No. 10, 2023, pp. 491–507.
36. Bondarenko, M., D. Kerr, A. Sorichetta, and A. J. Tatem, *Census/projection-disaggregated gridded population datasets for 189 countries in 2020 using Built-Settlement Growth Model (BSGM) outputs*. `https://www.worldpop.org/`, 2020.
37. Chen, T. and C. Guestrin, Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
38. Elvidge, C. D., K. E. Baugh, S. J. Anderson, P. C. Sutton, and T. Ghosh, The Night Light Development Index (NLDI): a spatially explicit measure of human development from satellite data. *Social Geography*, Vol. 7, No. 1, 2012, pp. 23–35.