

Beyond aggregation: a multi-scale, interpretable machine learning assessment of built environment effects on heavy-truck crash severity in Fortaleza, Brazil

Irfan Ullah¹, Carlos Pinheiro^{2*}, Muhammad Asim Ayaz¹, Bruno Brito³ and Quan Yuan^{1*}

¹ The Key Laboratory of Road and Traffic Engineering, Ministry of Education, College of Transportation, Tongji University, Shanghai 201804, China

² Federal Institute of Education, Science and Technology of Ceara (IFCE), Complexo Industrial e Portuário do Pecém Highway, CE-422, CEP 62670-000, Sao Gonçalo do Amarante, CE, Brazil

³ Federal University of Ceara (UFC), Centro De Tecnologia - Petran, Campus do Pici | Bloco 703, CEP 60440-900 Fortaleza, CE, Brazil

* Correspondence: carlos.pinheiro@ifce.edu.br (Pinheiro C); quanyuan@tongji.edu.cn (Yuan Q)

Abstract

This study investigates the spatial determinants of freight-related injury crashes in Fortaleza, Brazil, focusing on three outcomes: Possible Injury Crashes (PIC), Evident Injury Crashes (EIC), and Fatal Injury Crashes (FIC). A rich urban dataset incorporating land use, freight infrastructure, road design, and traffic characteristics was analyzed across three spatial aggregations: hexagons, aggregated neighborhoods, and census tracts. Two complementary modeling techniques were employed: Zero-Inflated Poisson (ZIP) models were used to account for excessive zeros in crash counts, while Extreme Gradient Boosting (XGBoost) provided a high-performing machine learning approach capable of capturing nonlinear relationships. XGBoost results show that the density of traffic signals, density of freeway lane mileage, and density of truck volumes are the dominant factors influencing freight-related injury risk across spatial scales. ZIP results highlighted the significant impact of freight proximity (e.g., distance to ports, airports, and intermodal terminals), infrastructure density (e.g., traffic signals, freeway lanes), and land-use intensity (e.g., job density and retail zones) on crash occurrence and severity. To interpret the complex patterns captured by XGBoost, Shapley Additive Explanations (SHAP) analysis was used. SHAP dependence and interaction plots revealed threshold effects and spatial interactions such as increased EIC and FIC risk in areas with high truck exposure and dense signal or freeway infrastructure. Notably, PIC risk was highest close to the airports, especially where primary roads and poles coexisted. The study demonstrates that both linear and nonlinear modeling frameworks are necessary to uncover the layered nature of freight-induced crash risks. Policy implications include optimizing traffic control density, enhancing safety in freight corridors, and prioritizing pedestrian protection near logistics hubs. By integrating statistical and explainable machine learning methods, the study offers actionable insights for data-driven urban safety and mobility planning.

Keywords: Injury severity, Freight truck crashes, XGBoost, SHAP, Zero-Inflated Poisson, MAUP

Citation: Ullah I, Pinheiro C, Ayaz MA, Brito B, Yuan Q. 2026. Beyond aggregation: a multi-scale, interpretable machine learning assessment of built environment effects on heavy-truck crash severity in Fortaleza, Brazil. *Digital Transportation and Safety* 5(1): 65–83 <https://doi.org/10.48130/dts-0026-0006>

Introduction

The global supply-chain network has matured into a tightly knit web that moves raw materials and finished goods among suppliers, distribution hubs, and consumers spread across cities, regions, and countries. This logistical sophistication has significantly expanded the role of road freight, placing heavy trucks at the center of just-in-time distribution systems and next-day e-commerce delivery. Their expanding footprint on public roads, especially inside dense metropolitan areas, has come at a cost: between 2009 and 2019, fatal crashes involving heavy trucks rose by 47%, and injury-producing collisions by 52% worldwide^[1,2]. Crafting effective countermeasures demands a clear picture of the immediate sources of risk and of the broader environments that amplify or attenuate those hazards.

Crash occurrence is governed by mechanisms operating at two complementary scales. At the micro level, vehicle size, driver workload, braking performance, and cargo stability directly influence the likelihood and severity of a collision. At the macro level, the spatial distribution of truck movements interacts with the built environment (e.g., land-use mix, road-network density, and traffic flows) to determine where crashes cluster^[3,4]. Although broad literature links

these environmental attributes to truck safety, most studies aggregate data within a single zoning framework (for example, census tracts or administrative districts) and implicitly assume that results are invariant to scale. That assumption is rarely justified.

Spatial aggregation itself can alter descriptive statistics, correlation structures, and model coefficients, a phenomenon known as the Modifiable Areal Unit Problem (MAUP). When point-level events are fit inside arbitrary zones, changing zone size or boundaries may reverse coefficient signs, inflate type-I error rates, or mask local hot spots^[5,6]. Recent work in transport geography confirms that crash models calibrated at alternative spatial resolutions often yield divergent conclusions^[7], yet quantitative assessments of MAUP for heavy-truck safety remain scarce. Omitting this sensitivity could lead analysts to misidentify priority corridors and recommend infrastructure upgrades that fail to target the most hazardous locations.

Methodological innovation offers a path forward. ZIP regression separates two latent processes: a zone's propensity to record any crash and the expected count once that propensity is non-zero, thereby accommodating the abundance of zero-crash areas common to truck datasets^[4]. In parallel, gradient-boosted decision trees such as Extreme Gradient Boosting (XGBoost) can flexibly capture non-linearities and high-order interactions among built-environment, traffic, and operational predictors. When interpreted with

Shapley Additive Explanations (SHAP), XGBoost models retain transparency, allowing each variable's marginal contribution to be quantified^[8,9]. While both approaches have gained traction in general road-safety research, their joint use across multiple spatial resolutions has yet to be exploited for freight-truck crashes. This gap is practically consequential because freight-safety investments and enforcement strategies are typically planned using zone-based evidence, yet inferred risk factors and hotspot priorities can shift when the same crashes are aggregated at different spatial resolutions. Without scale-sensitive diagnostics, agencies risk mis-ranking contributing factors, targeting suboptimal locations, and implementing countermeasures that address downstream symptoms rather than upstream exposure and infrastructure conditions.

Geographical coverage represents a second shortcoming. Nearly 90% of global road traffic fatalities occur in low and middle-income nations, yet empirical investigations of crashes still concentrate on high-income settings^[10]. Rapid urbanization, heterogeneous road networks, and limited enforcement capacity in the Global South may produce risk mechanisms that differ markedly from those documented elsewhere. Whether built-environment correlates established in wealthier contexts remain valid under such conditions is an open and policy-critical question^[11]. Besides, freight statistics reinforce the contrast: trucks carry roughly 44% of US ton-miles but about 65% in Brazil and more than 80% across sub-Saharan Africa^[12,13].

This study tackles these intertwined gaps by analyzing how built-environment factors influence the severity of truck-involved crashes in Fortaleza, Brazil, one of the country's largest metropolitan areas and a representative freight hub in the Global South^[14]. Using geo-referenced police crash records for 2011–2025, linked to high-resolution land use, traffic, and network data, we pursue three objectives: (1) Employ statistical modelling by estimating ZIP regressions to pinpoint the most influential environmental and operational factors associated with crash severity, distinguishing outcomes as possible injury, evident injury, and fatality. (2) Undertake a machine learning validation step in which XGBoost models, interpreted with SHAP, are used to uncover non-linear relationships and interaction effects, thereby corroborating and extending the insights obtained from the ZIP estimations. (3) Test scale sensitivity by replicating both modeling approaches across three spatial aggregations, fine-grained hexagonal units, aggregated neighborhoods, and census tracts to examine how coefficient magnitudes, variable's importance rankings, and predictive performance vary with spatial aggregation. By triangulating results across methods and spatial schemes, the study delivers scale-robust, data-driven guidance for engineers and planners charged with designing safer logistics corridors. More broadly, our integrative framework shows how combining spatial econometrics with interpretable machine learning can overcome long-standing analytical pitfalls and generate actionable evidence for freight safety policy in rapidly urbanizing regions.

The remainder of this paper is organized as follows: the literature review section reviews prior research on MAUP aspects associated with truck safety and built-environment metrics. The study area section describes the data and study design, and data sources. The methodology section introduces the analytical approaches employed, including XGBoost, ZIP, and SHAP. The results section presents empirical findings across the three zoning schemes. Finally, the discussion and conclusions section synthesizes the key insights and is followed by policy implications aimed at enhancing heavy-truck safety in Fortaleza and similar logistics corridors.

Literature review

The MAUP in crash analysis

Recent advances in spatial econometrics underline how strongly statistical inference depends on the zoning system used to aggregate crash data^[15]. When point records are compressed into arbitrary areas, their variance structure and correlation pattern shift the essence of the MAUP^[16]. Neglecting this bias can hide spatial heterogeneity or even reverse the sign of built-environment coefficients, leading to mis-targeted countermeasures^[5,17]. Empirical evidence is compelling: Briz-Redón et al.^[6] found that replacing census tracts with regular hexagons in Valencia inverted the significance of several land-use predictors, while Zhai et al.^[18] showed that Bayesian multiscale Poisson lognormal models, which share random effects across block groups and tracts, lowered the Deviance Information Criterion and kept priority corridors stable across tiers of government. Synthesizing dozens of such cases, Xu et al.^[19] urge analysts to (1) work with the most disaggregated data available, (2) model spatial non-stationarity, (3) delineate optimized Traffic Safety Analysis Zones, and (4) report sensitivity diagnostics before any funds are allocated. Policy salience emerges whenever scale shifts alter where and why interventions should occur. Li & Chen^[7], for instance, showed that widening the buffer around Maryland transport hubs from 300 to 800 m raised the crash risk variance explained from 71% to 84%, prompting the authors to advocate buffer-specific density limits and design standards. Yet almost all MAUP-conscious studies examine mixed or passenger traffic. Dong et al.^[20] have already demonstrated that the determinants of heavy truck crashes diverge from those governing passenger-car collisions, implying that aggregation bias could distort freight diagnostics even more severely.

Built environment determinants of heavy truck crashes

Research that isolates how urban form shapes truck safety is still scant and largely United States-centered. Findings cluster around four domains: (1) freight-generator exposure: distance to airports, seaports or intermodal yards and the density of freight generation poles; (2) demographic and land use mix population and job densities and the proportion of industrial, residential and retail land; (3) road infrastructure controls densities of traffic signals and roundabouts; and (4) network structure and exposure lane-mile densities of freeways, primary and secondary roads plus truck flow^[2,4,21].

Applying a spatial Durbin model to 3,923 Los Angeles census tracts, Yang et al.^[21] show that proximity to intermodal facilities and a high density of warehouses markedly increase truck crash frequencies. Industrial land is likewise a strong positive driver, whereas residential and retail shares are neutral or protective. Network factors reinforce these patterns: traffic-signal density and the lane-mile densities of freeways, primary, and secondary roads remain significant, and truck volume density adds an exposure-based risk component. A companion study, Yang et al.^[4], covering 3,724 truck-injury crashes from 2014–2018 in the same tract system, corroborates that shorter distances to the airport, seaport, and intermodal facilities and dense freight trip generation clusters raise crash counts, while higher population density and a larger residential share reduce them. Roundabout and expressway-lane densities elevate risk, denser primary and secondary road grids dissipate it, and truck volume density again emerges as a strong positive predictor.

Built environment and heavy-truck crash severity

Adding a satellite perspective, Yu et al.^[2] merge Sentinel-2 imagery with GIS and socio-economic layers for 3,677 tracts; gradient-boosting models confirm that freight-facility concentration and truck-density exposure rank among the top crash predictors, whereas a larger residential share is consistently protective. They also find that freeway lane and roundabout densities contribute positively, while a finer arterial mesh spreads truck traffic more safely. In New York City, Wei et al.^[22] aggregate 2008–2012 crashes to traffic analysis zones and, using a spatial generalized ordered-probit, show that higher hourly truck volumes and dense retail service employment raise the odds of severe or fatal outcomes, whereas signalized intersections mitigate severity.

XGBoost + SHAP for built environment safety modelling

A recent wave of studies has paired XGBoost with SHAP to model crash outcomes because the technique marries high predictive accuracy with transparent inference. XGBoost captures the non-linear, high-order interactions that characterize urban-form variables, routinely outperforming Poisson, negative-binomial, and ordered-probit baselines, while SHAP decomposes each prediction into additive contributions that planners can audit at whichever spatial scale they manage^[4,8]. Together they answer a long-standing dilemma in traffic safety analytics: how to model complex built environment effects without sacrificing interpretability^[23,24].

In the above-mentioned studies, XGBoost-SHAP models calibrated at the census tract level show that road network aspects and demographics, and land use aspects are related to freight-truck crash counts, whereas higher shares of residential land mitigate risk^[4]. On Chicago expressways, a real-time accident-detection system achieves 99% accuracy, with SHAP pinpointing abrupt speed drops in dense urban segments as the dominant trigger and illustrating how proximity to the central business district interacts with residential density to heighten risk^[8].

Outside the United States, the XGBoost-SHAP pipeline continues to prove useful, though evidence is sparse. In the Republic of Korea, segment-scale XGBoost models coupled with SHAP show that high road network centrality inside dense mixed-use zones is a strong, non-linear driver of pedestrian fatalities and that the marginal effect of population density varies sharply by location heterogeneity that a global log-linear specification would have obscured^[25]. For Wuhan, China, an XGBoost model fed with streetscape variables extracted from panoramic street-view images reveals that larger visible road surfaces, higher pedestrian counts, and intense commercial frontage around intersections jointly raise pedestrian crash risk; SHAP visualizations clarify how each image-derived element pushes an intersection toward or away from the danger threshold^[24].

Studies of injury severity reach similar conclusions. A nationwide New Zealand investigation applies XGBoost-SHAP and shows that road class and multi-vehicle involvements are the most influential predictors of serious or fatal injury^[23]. In the United Kingdom, cyclist-injury research uses XGBoost-SHAP to demonstrate that overtaking maneuvers on high-speed arterials and the absence of cycle facilities are the chief drivers of fatal outcomes, with SHAP clarifying how these factors interact with traffic volume and lighting^[26].

Evidence from developing country corridors is rarer but equally instructive. On Pakistan's N-5 highway, a suite of boosting ensembles XGBoost among them interpreted with SHAP identifies undivided two-lane night-time segments and specific collision types as the principal fatality drivers, insights that conventional logit models had masked^[27].

Across all cases, three advantages recur: XGBoost substantially outperforms parametric baselines; SHAP consistently elevates built-environment variables, freight generation intensity, freeway-lane and roundabout densities, street centrality into the upper importance tier, and exposes threshold or saturation effects. These local SHAP explanations allow corridor-specific interventions rather than blanket programs^[4,24]. Yet, to date, no study has applied an interpretable XGBoost-SHAP workflow to heavy-truck crashes in a Global South metropolis or investigated whether SHAP variables' importance rankings remain stable when the analysis is replicated across alternative spatial aggregations, the very essence of the MAUP. The present research fills that dual void by integrating ZIP and XGBoost-SHAP across two spatial resolutions for Fortaleza, Brazil, thus providing the first scale-robust, machine learning assessment of freight-truck safety in a low-income metropolitan context. In this study, XGBoost and SHAP are employed in direct alignment with the research objectives to complement the ZIP analysis by uncovering non-linear relationships, interaction effects, and scale-sensitive patterns in built-environment risk factors. Specifically, these methods support the validation and extension of ZIP-based findings and enable assessment of whether dominant predictors of crash severity remain stable or shift across spatial aggregations, thereby directly addressing the MAUP.

Research gaps and study contributions

Overall, the literature reviewed highlights that existing research on heavy-truck safety is constrained by limited attention to MAUP-related scale effects, a predominantly United States-based evidence base, and a lack of interpretable machine-learning approaches capable of uncovering non-linear severity dynamics in rapidly urbanizing Global South environments. Motivated by these limitations, this study examines heavy-truck crashes in Fortaleza, Brazil, one of the country's largest metropolitan areas and a key freight hub, to generate contextually grounded and methodologically robust evidence. The present study addresses these intertwined gaps in three ways. First, we estimate ZIP models to identify statistically robust built-environment and operational predictors of truck-involved crash severity, distinguishing between possible injury, evident injury, and fatality outcomes. Second, we implement an interpretable XGBoost-SHAP workflow to uncover non-linearities and interaction effects that complement and extend the ZIP findings. Third, we evaluate scale sensitivity by replicating both modelling approaches across three spatial aggregations: fine-grained hexagons, administrative neighborhoods, and census tracts to assess whether coefficient magnitudes, variable-importance rankings, and predictive performance remain stable across zoning systems. By triangulating insights across modelling techniques and spatial schemes, this study delivers the first scale-robust, interpretable machine-learning assessment of heavy-truck crash severity in a Global South metropolitan context, offering engineers and planners actionable guidance for designing safer logistics corridors.

Study area

Fortaleza, the capital of Ceará in Brazil's northeast, houses about 2.5 million inhabitants inside a dense 312 km² municipal core and anchors a metropolitan region exceeding four million residents^[28]. Wholesale distribution and consumer services dominate local output, sustained by two seaports and an international airport that

feed maritime and air freight into a spoke-and-hub road system^[29]. Industrial plants, bulk warehouses, and logistics depots have gravitated to low-cost suburban parcels flanking the expressways, while residential and commercial growth has filled the gaps between them, creating a patchwork in which heavy-truck flows thread through neighborhoods of high housing, retail, and service density^[30].

Crash records reveal the safety toll of this juxtaposition: 4,358 truck-involved collisions were logged between 2011 and 2025, including 243 fatal and 2,012 injury crashes^[31]. Fortaleza's compact form, rapid urbanization, and intense freight exposure therefore provide an apt test bed for examining how land-use mix, road network density, and the siting of freight facilities shape heavy-truck crash severity. They also allow a direct assessment of the MAUP, because the same events can be aggregated along very different spatial footprints.

To prove that sensitivity, crashes, and built-environment indicators are summarized at three nested basic spatial units (BSUs) (Fig. 1). First, the study tessellates the municipality into 1,040 equal-area hexagons of 0.32 km² each. A regular hexagonal lattice avoids the uneven sizes and shapes of administrative zones and, according to freight practitioners, reduces boundary effects more effectively than square grids^[32,33]. Second, results are aggregated to the 121 officially recognized neighborhoods (mean area ≈ 2.6 km²), a scale familiar to planners and community organizations. Finally, crashes are summarized by 4,422 census tracts (mean area ≈ 0.07 km²), the spatial unit most commonly used in studies that relate urban form to safety outcomes^[6,18,21]. The conceptual framework illustrated in Fig. 2 focuses on various factors influencing traffic crash outcomes. These three spatial scales are used to examine MAUP by testing whether crash and built environment relationships change with spatial aggregation. Comparing model outputs across hexagons, neighborhoods, and census tracts allows us to assess the sensitivity of inferred risk factors to spatial resolution. It categorizes variables into four main groups: freight generators, road infrastructure, demographic profile & land uses, and road network. These categories include factors such as the distance to key transportation hubs (e.g., airport, seaport), road infrastructure elements (e.g., traffic signals, roundabouts), demographic aspects (e.g., population density, land use types), and road network characteristics (e.g., lane mileage, truck volumes). The central Response Features include possible injury crashes (PIC), evident injury crashes (EIC), and fatal injury crashes (FIC), highlighting the impact of these variables on different levels of crash severity.

Methodology

Zero-inflated Poisson (ZIP) regression models

The ZIP model, proposed by Lambert in 1992^[34], presents an alternative approach used for analyzing data characterized by a high frequency of zero values in the response feature. Such a high frequency of zeros in the data may violate the supposition of equal mean and variance inherent in the Poisson distribution.

For each observation of the dependent variable in the ZIP model, Y follows the distribution described in Eq. (1).

$$P(Y = y) = \begin{cases} \varphi + (1 - \varphi)e^{-\lambda}, & y = 0 \\ \frac{(1 - \varphi)e^{-\lambda}\lambda^y}{y!}, & y > 0 \end{cases} \quad (1)$$

where, $y = (0, 1, 2, \dots)$, λ represents the Poisson mean, and $\varphi \in [0,1]$ stands for the mixing probability parameter which accounts for the excess values of zeros. When $\varphi = 0$, Eq. (1) simplifies to the standard Poisson distribution in Eq. (2).

$$P(Y = y | \varphi = 0, \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, 2, \dots \quad (2)$$

XGBoost

XGBoost, a key decision tree-based optimized distributed gradient boosting algorithm, was developed by Chen & Guestrin^[35]. The mathematical expression used for XGBoost is shown as:

$$\hat{y}_n = \varphi(x_n) = \sum_{k=1}^K f_k(x_n) \quad (3)$$

where, \hat{y}_n is the predicted value for the n -th data point. x_n is a feature vector for the n -th data point. The $f_k(x_n)$ output of the k -th tree applied to x_n . The model prediction function is the sum of the predictions from all K trees. To reduce the model complexity, the penalty term θ is set as:

$$\theta(f_k) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (4)$$

where, γ controls the number of leaves, T controls the leaf weight ω , setting $\theta(f_k)$ simplifies the algorithm generated by the model and avoids the overfitting. The model reduces the objective function via an iterative technique. The model's objective function at the j -th iteration is minimized by f_j adding the term:

$$L_j = \sum_{n=1}^n l(y_n, \hat{y}_n^{(j-1)} + f_j(x_n)) + \theta(f_k) \quad (5)$$

SHAP

Although the stacked ensemble model performs exceptionally, the increase in complexity may reduce the interpretability of the

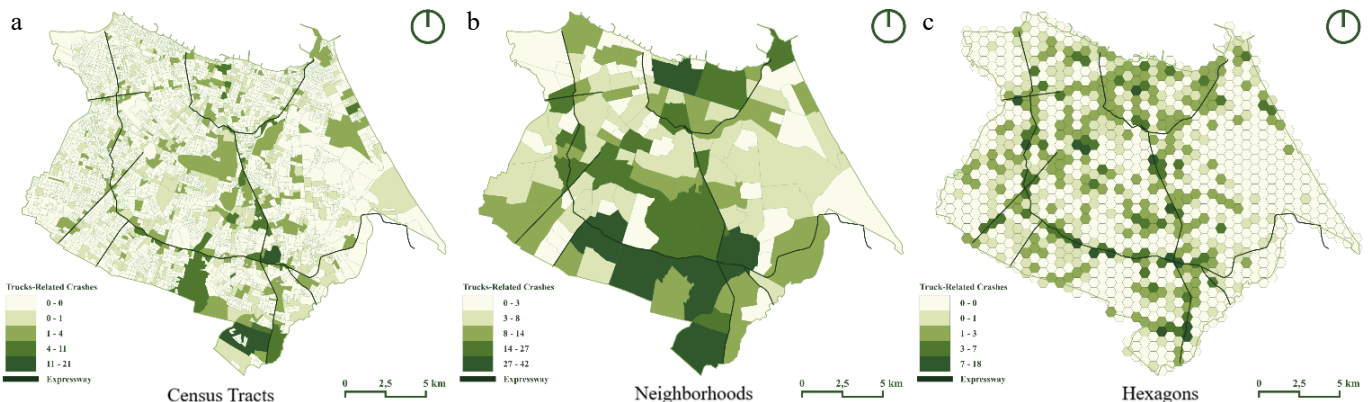


Fig. 1 The selected study area. Source: Authors' own elaboration based on study area boundary, road network, and crash data processed in GIS.

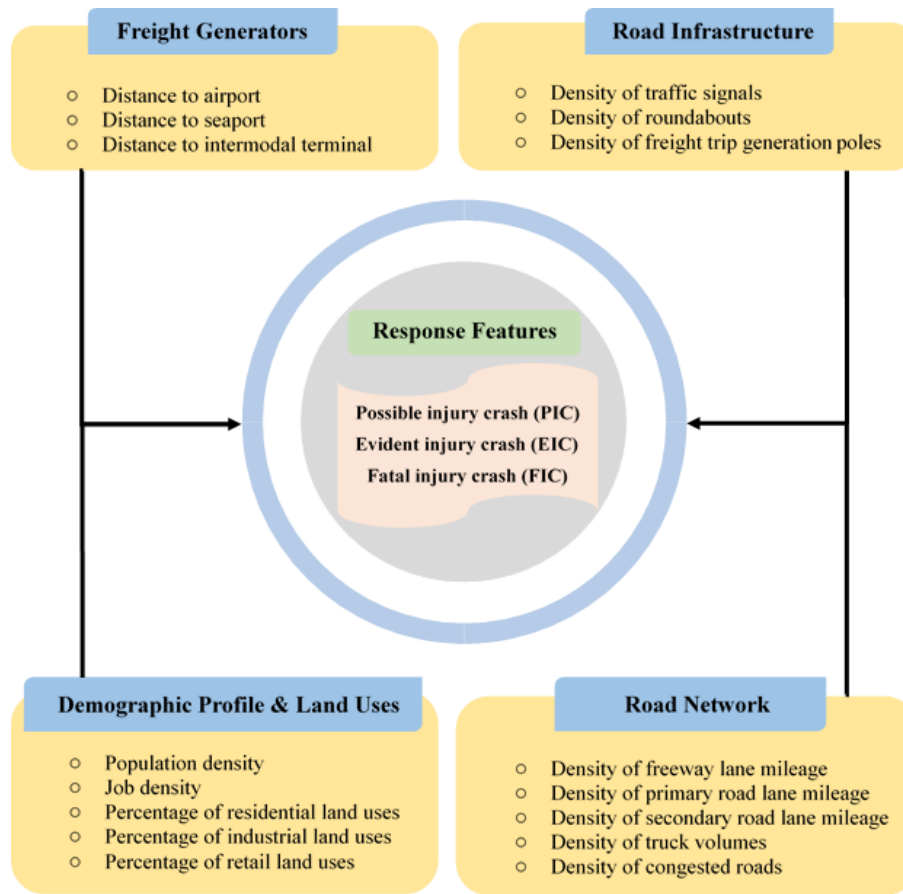


Fig. 2 Conceptual framework of the variables.

model. To address the complexity of interpreting feature contributions to stop duration predictions, this study employs SHAP analysis^[36].

Suppose D represents the complete set of input features used to predict freight trucks-related crashes, and let $S \subseteq D$ be any subset of these predictors. The model output $v(S)$ corresponds to the predicted crashes when only features in S are used. For a given feature i (such as freight generators), its contribution is computed as $(v(S \cup \{i\}) - v(S))$, indicating how much more accurate the crash prediction becomes when i is added to the subset S . By averaging this marginal contribution over all such subsets $S \subseteq D \setminus \{i\}$, the Shapley ϕ_i provides a principled measure of feature importance. This allows us to interpret the impact of each factor on crashes transparently and rigorously, enhancing model interpretability and aiding decision-making for freight truck-related crashes.

$$\phi_i(v) = \sum_{S \subseteq D \setminus \{i\}} \frac{|S|!(|D| - |S| - 1)!}{|D|!} (v(S \cup \{i\}) - v(S)) \quad (6)$$

The major principle of SHAP is that it treats each feature as a contributor, calculating SHAP values for each feature to summarize its contribution and ultimately obtain the model's final prediction.

Results

Descriptive statistics

Table 1 shows the descriptive statistics for equivalent hexagonal, neighborhood, and census tract units of Fortaleza city, revealing

substantial variation in truck-injury outcomes and built environment characteristics. On average, each hexagon records 0.68 PIC, 2.44 EIC, and 0.17 FIC, with maximum values reaching as high as 50.27 for EIC and 9.43 for both PIC and FIC, indicating the presence of significant crash hotspots. In terms of freight accessibility, the mean distances to intermodal terminals, airports, and seaports are 4.76, 9.38, and 13.46 km, respectively, reflecting varying exposure to freight corridors. The population and job densities average 7,338 persons per km² and 2,352 jobs per km², with some hexagons exceeding 30,000, indicating major urban activity centers. Residential land use dominates the landscape, comprising 64% on average, while industrial and retail uses are less prevalent but locally concentrated. Traffic signal and roundabout densities average 3.32 and 0.07 per km², respectively, while freight trip generation poles appear with a mean density of 4.24 per km². The supply of road infrastructure is characterized by 0.52 km/km² of freeway lanes, 1.37 km/km² of primary roads, and 1.59 km/km² of secondary roads. Additionally, about 5% of road segments are congested, and truck intensity averages 1.18 trucks per km of road, with some hexagons experiencing values as high as 59.56, highlighting stark heterogeneity in freight exposure across the city.

The neighborhood-level descriptive statistics indicate higher average injury counts and more aggregated land-use and infrastructure characteristics compared to hexagonal units. From Table 1, on average, neighborhoods report 0.88 PIC, 2.97 EIC, and 0.20 FIC, with maximum EIC counts reaching 12.67, suggesting notable injury clustering even at broader spatial scales. Freight infrastructure accessibility remains relatively similar, with average distances to

Table 1. Descriptive statistics.

Response variables	Hexagonal-based analysis				Neighborhood-based analysis				Census tracts-based analysis			
	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.
PIC	0.68	1.65	0.00	9.43	0.88	1.01	0.00	6.29	0.73	4.83	0	127.4
EIC	2.44	4.26	0.00	50.27	2.97	2.55	0.00	12.67	2.63	10.09	0	186.17
FIC	0.17	0.81	0.00	9.43	0.2	0.42	0.00	2.96	0.23	3.02	0	96.77
Predictors												
Freight generators												
Distance to airport (km)	9.38	3.36	0.00	18.37	8.46	2.8	1.48	15.48	8.61	2.74	0	17.32
Distance to seaport (km)	13.46	5.13	0.00	23.63	12.64	4.9	0.36	21.83	13.33	5.07	0	23.4
Distance to intermodal terminal (km)	4.76	2.24	0.09	10.89	4.36	2.18	0.25	9.74	4.23	2.25	0.06	10.27
Demographics & land uses												
Population density (persons/km ²)	7,338.51	5,696.26	0.00	29,802.4	16,813.09	8,190.26	737.16	38,948.43	17,056.7	17,351.8	0	629,596.41
Job density (jobs/km ²)	2,352.61	11,661.82	0.00	301,277.1	2,953.44	5,575.33	1.98	34,015.25	298.01	6,566.02	0	293,390.21
% age of residential land uses	64%	32%	0%	100%	76%	10%	30%	95%	69%	33%	0	100%
% age of industrial land uses	1%	7%	0%	100%	1%	3%	0%	18%	1%	6%	0	100%
% age of retail land uses	43%	28%	0%	100%	51%	15%	22%	85%	39%	31%	0	100%
Road infrastructure												
Density of traffic signals (per km ²)	3.32	6.59	0.00	59.7	4.63	5.17	0.00	27.79	4	12.57	0	255.36
Density of roundabouts (per km ²)	0.07	0.45	0.00	3.14	0.08	0.29	0.00	2.21	0.04	0.89	0	40.19
Density of freight trip generation poles (per km ²)	4.24	8.32	0.00	103.69	5.45	4.98	0.00	32.9	4.03	12.63	0	152.74
Road network												
Density of freeway lane mileage (km/km ²)	0.52	1.49	0.00	14.64	0.65	0.99	0.00	5.14	0.79	26.58	0	1,615.21
Density of primary road lane mileage (km/km ²)	1.37	2.09	0.00	11.77	1.98	1.84	0.00	9.43	1.54	3.43	0	32.75
Density of secondary road lane mileage (km/km ²)	1.59	1.86	0.00	9.22	2.10	1.53	0.00	7.83	2.16	3.89	0	56.57
Density of truck volumes (vehicle/km ²)	1.18	3.72	0.00	59.56	1,032.56	796.99	288.08	6,653.61	1.47	9.31	0	589.1
Density of congested roads (km/km ²)	0.05	0.10	0	1.47	0.06	0.04	0.00	0.19	0.06	0.19	0	6.22

intermodal terminals (4.36 km), airports (8.46 km), and seaports (12.64 km), albeit with slightly reduced variation. Population density across neighborhoods averages 16,813 persons per km², while job density averages 2,953 jobs per km², again indicating dense and economically active urban areas. Residential land use dominates with an average of 76%, whereas industrial and retail land uses remain low on average (1% and 51%, respectively), but with pockets of concentration. Traffic signals are more frequent (mean = 4.63 per km²), as are roundabouts (0.08 per km²) and freight trip generation poles (5.45 per km²). Road infrastructure supply also increases, with average densities of 0.65 km/km² for freeway lanes, 1.98 km/km² for primary roads, and 2.10 km/km² for secondary roads. Congested segments make up 6% of road networks on average. Notably, the density of truck volume intensity rises dramatically at this spatial scale, averaging about 1,033 trucks per km of road and reaching a maximum of 6,654, reflecting concentrated freight flows through key logistic corridors at the neighborhood level.

The descriptive statistics at the census tract level reveal substantial variability across crash occurrences, infrastructure, and land use characteristics. On average, PICs are low (mean 0.73), but some tracts report up to 127 crashes, while EIC and FIC average 2.63 and 0.23, respectively, with notable extremes. Infrastructure elements such as traffic signals, roundabouts, and FTGP pole density also vary, reflecting differences in traffic control and urban design. Population density exhibits wide disparities, with some tracts reaching over 629,000 persons per km², alongside highly concentrated job density in specific zones. Residential land use dominates (69% on average), while industrial and retail uses are less prevalent but can fully

occupy certain tracts. Road network characteristics, including freeway, primary, and secondary road densities, differ significantly, indicating the heterogeneity of transport infrastructure. Congestion levels also range widely, with some tracts exhibiting extremely high congestion percentages. Overall, this variability suggests diverse urban configurations, demographic profiles, and transport conditions across census tracts, all of which contribute differently to crash risks and severities.

Model comparison

In Tables 2–4, spatial modeling results for PIC, EIC, and FIC demonstrate that the identification and importance of risk factors are strongly influenced by both the spatial unit and the modeling approach. Table 2 summarizes the hexagonal-level analysis, which reveals critical spatial, demographic, and infrastructural factors influencing freight-related injury crashes in Fortaleza. For PIC, the ZIP model shows that distance to the airport, seaport, and intermodal terminal significantly affects crash likelihood; maintaining a greater distance to the airport and intermodal terminal reduces risk, while closer proximity to the seaport also decreases it. Additionally, the density of traffic signals and the percentage of congested roads are significant contributors. In contrast, the population density, percentage of residential and industrial land uses, density of roundabouts, density of primary road lane mileage, and freight trip generation poles show no significant influence. For EIC, similar trends are observed: shorter distances to freight terminals, especially distance to intermodal terminal and distance to airport, are associated with higher crash counts. The density of freight trip generation poles has a notable impact, while the density of traffic signals and the density

Built environment and heavy-truck crash severity

of freeway lanes remain highly significant. The percentage of congested roads again shows a strong positive effect. FIC is primarily associated with demographics and land use variables. The percentage of retail land uses and job density positively influence

FIC risk, while population density also plays a notable role. Interestingly, the percentage of congested roads shows a strong negative coefficient, suggesting potential risk-reducing effects possibly due to slower traffic in congested areas. The ZIP results emphasize linear

Table 2. Model comparison based on hexagonal-level analysis.

Category	Variable	PIC				EIC				FIC			
		ZIP		XGBoost		ZIP		XGBoost		ZIP		XGBoost	
		Co.eff	S.E	RI (%)	Rank	Co.eff	S.E	RI (%)	Rank	Co.eff	S.E	RI (%)	Rank
Freight generators													
	Distance to airport	-0.081***	0.02	4.23	9	-0.044***	0.009	2.91	13	0.007	0.038	4.82	12
	Distance to seaport	0.031*	0.013	3.76	13	0.035***	0.006	4.02	8	0.011	0.025	6.15	5
	Distance to intermodal terminal	-0.055*	0.025	3.33	15	-0.046***	0.011	2.56	16	-0.014	0.055	6.05	6
Demographics and land uses													
	Population density	0.045	0.079	3.69	14	-0.024	0.041	3.4	12	0.492**	0.157	3.87	16
	Job density	0.056**	0.021	4.19	10	0.006	0.009	2.76	15	0.103*	0.044	7.19	4
	%age of residential land uses	0.194	0.303	4.18	11	0.012	0.139	3.53	11	0.409	0.496	4.18	14
	%age of industrial land uses	0.888	1.847	3.93	12	0.691	0.386	2.9	14	0.017	4.833	5.19	9
	%age of retail land uses	0.811**	0.265	4.47	8	0.12	0.125	5.49	7	1.412**	0.535	4.76	13
Road infrastructure													
	Density of traffic signals	0.010*	0.005	19.89	1	0.014***	0.003	14.3	2	0.004	0.026	5.98	8
	Density of roundabouts	-0.009	0.076	2.56	16	0.041	0.03	5.65	6	0.295	0.156	8.21	2
	Density of freight trip generation poles	-0.024	0.043	10.45	2	0.070**	0.022	8.71	4	0.104	0.11	5.09	10
Road network													
	Density of freeway lane mileage	0.079***	0.02	8.07	4	0.089***	0.011	16.96	1	0.018	0.068	15.66	1
	Density of primary road lane mileage	0.037	0.021	7.12	5	0.012	0.013	3.75	9	-0.058	0.083	4.04	15
	Density of secondary road lane mileage	0.062*	0.026	4.6	7	-0.001	0.014	3.72	10	-0.003	0.068	6	7
	Density of truck volumes	-0.002	0.012	8.91	3	-0.001	0.006	12.93	3	0.009	0.038	7.96	3
	Density of congested roads	1.336**	0.46	6.62	6	1.441***	0.323	6.38	5	-1.424	3.126	4.87	11
	PSEUDO-R ²	0.042		0.167		0.074		0.349		0.044		-0.024	
	RMSE	1.559		1.561		3.923		3.608		0.803		0.784	
	MAE	0.955		0.861		2.562		2.152		0.300		0.223	

*, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 3. Model comparison based on neighborhood-level analysis.

Category	Variable	PIC				EIC				FIC			
		ZIP		XGBoost		ZIP		XGBoost		ZIP		XGBoost	
		Co.eff	S.E	RI (%)	Rank	Co.eff	S.E	RI (%)	Rank	Co.eff	S.E	RI (%)	Rank
Freight generators													
	Distance to airport	-0.139**	0.053	3.03	13	-0.091**	0.028	3.8	12	-0.1	0.103	5.98	7
	Distance to seaport	-0.009	0.036	2.79	15	-0.011	0.019	6.4	7	0.023	0.073	5.37	9
	Distance to intermodal terminal	-0.111	0.075	2.8	14	-0.076*	0.037	1.98	16	-0.175	0.136	6.63	6
Demographics and land uses													
	Population density	0.089	0.295	3.66	11	0.233	0.162	3.73	13	0.384	0.547	4.14	16
	Job density	0.118	0.130	6.19	4	-0.219***	0.057	3.56	14	-0.289	0.273	4.39	14
	% of residential land uses	0.126	1.478	3.07	12	0.059	0.791	4.44	9	-0.329	3.133	7.2	4
	% of industrial land uses	-0.019	5.259	2.64	16	-0.017	2.374	8.07	5	0.112	8.479	12.8	1
	% of retail land uses	-0.271	1	4.93	8	-0.097	0.526	7.42	6	0.015	1.954	5.44	8
Road infrastructure													
	Density of traffic signals	0.076*	0.037	24.62	1	0.041	0.025	8.43	4	0.049	0.092	4.94	12
	Density of roundabouts	0.084	0.345	5.89	5	0.036	0.213	3.45	15	-0.035	0.952	4.18	15
	Density of freight trip generation poles	-0.048	0.204	5.6	7	0.365**	0.133	11.55	3	0.537	0.423	8.43	2
Road network													
	Density of freeway lane mileage	0.101	0.103	6.89	3	0.140*	0.055	12.24	1	0.212	0.209	7.06	5
	Density of primary road lane mileage	-0.115	0.091	3.82	10	-0.075	0.056	5.33	8	-0.376	0.236	4.92	13
	Density of secondary road lane mileage	-0.185	0.103	5.73	6	-0.128*	0.060	3.81	11	-0.382	0.248	4.97	11
	Density of truck volumes	0	0	4.71	9	0	0	11.77	2	0	0	8.35	3
	Density of congested roads	0.077	4.802	13.64	2	-0.012	2.434	4.01	10	-0.025	9.565	5.21	10
	PSEUDO-R ²	0.153		0.080		0.122		0.106		0.099		-0.386	
	RMSE	0.818		0.920		2.059		2.336		0.388		0.411	
	MAE	0.558		0.628		1.529		1.630		0.234		0.230	

*, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 4. Model comparison based on census tracts-level analysis.

Category	Variable	PIC				EIC				FIC			
		ZIP		XGBoost		ZIP		XGBoost		ZIP		XGBoost	
		Co.eff	S.E	RI (%)	Rank	Co.eff	S.E	RI (%)	Rank	Co.eff	S.E	RI (%)	Rank
Freight generators													
	Distance to airport	-0.085	nan	7.22	5	-0.008	nan	5.45	11	0.012	nan	5.74	12
	Distance to seaport	-0.019	nan	6.36	7	-0.003	nan	6.93	7	0.038	nan	6.92	6
	Distance to intermodal terminal	-0.074	nan	5.23	11	-0.055	nan	6.17	9	-0.032	nan	10.18	2
Demographics and land uses													
	Population density	0.625	nan	5.66	9	0.381	nan	5.57	10	-0.022	nan	6.09	9
	Job density	0.032	nan	2.49	14	-0.002	nan	2.54	14	-0.011	nan		16
	% of residential land uses	-0.469	nan	6.07	8	-0.198	nan	4.88	13	-0.018	nan	9.12	4
	% of industrial land uses	1.176	nan	3.75	12	1.408	nan	5	12	0.001	nan	3.51	13
	% of retail land uses	-0.912	nan	9.76	3	0.008	nan	7.8	5	0.011	nan	6.53	7
Road infrastructure													
	Density of traffic signals	0.004	nan	14.93	1	0.004	nan	11.83	1	0.075	nan	5.87	11
	Density of roundabouts	-0.081	nan	1.22	16	-0.109	nan	1.69	15	0.001	nan	0.73	15
	Density of freight trip generation poles	-0.086	nan	3.25	13	-0.073	nan	7.12	6	0.028	nan	9.38	3
Road network													
	Density of freeway lane mileage	0.004	nan	1.48	15	0.006	nan	1.54	16	-0.014	nan	0.95	14
	Density of primary road lane mileage	0.017	nan	8.52	4	0.017	nan	8.12	3	0.052	nan	7.62	5
	Density of secondary road lane mileage	0.004	nan	5.5	10	0.018	nan	7.97	4	-0.01	nan	6.04	10
	Density of truck volumes	0.1	nan	7.19	6	0.051	nan	11.09	2	0.24	nan	15.12	1
	Density of congested roads	1.176	nan	11.36	2	-0.107	nan	6.31	8	0.001	nan	6.21	8
	PSEUDO-R ²	0.302		0.028		0.192		0.055		-2.392		-0.047	
	RMSE			4.659				9.742				2.825	
	MAE			1.18				3.951				0.319	

and interpretable relationships, while XGBoost rankings complement them by uncovering complex non-linear effects, with the density of traffic signals ranked 1 (most important predictor) for PIC, the density of freeway lane mileage ranked 1 for EIC, and FIC. This dual-method approach highlights the value of integrated modeling in identifying key drivers of urban freight crash severity.

Furthermore, the neighborhood-level analysis uncovers significant spatial and infrastructural factors that influence freight-truck-related injury severity in the city of Fortaleza. Results of the ZIP model from Table 3 show that distance to freight generators, particularly distance to airports, has a significant negative effect on crash frequency for both PIC and EIC, respectively. This indicates increased crash risk near airports. Similarly, the distance to intermodal terminals is also negatively associated with EIC and FIC, further supporting the role of freight proximity in injury severity. Among demographic variables, job density is strongly and negatively associated with EIC, implying lower crash risks in employment-intensive zones. Road infrastructure plays a crucial role, with the density of traffic signals being positively and significantly associated with PIC and maintaining high importance across models (Rank 1 in XGBoost for PIC). The density of freight trip generation poles has a significant positive impact on EIC and ranks highly for FIC in XGBoost (Rank 2). The density of freeway lane mileage emerges as a major contributor to EIC in ZIP; Rank 1, XGBoost, pointing to increased exposure to freight traffic on major routes. Notably, the density of secondary road lane mileage shows a consistent negative effect across EIC, indicating that more traffic corridors may reduce injury severity. In terms of performance, ZIP outperforms XGBoost with lower RMSE and MAE values, but XGBoost captures non-linear relationships such as the importance of density of truck volumes and density of freeway, which are not significant in ZIP but rank high in XGBoost for FIC (Ranks 3 and 5). Collectively, both models underscore the importance of congestion, freight activity, employment density, and road design in explaining spatial variation in injury severity from truck-related crashes.

In the Census Tracts-based analysis, as presented in Table 4, the ZIP model produced NaN values for all of the variables, indicating estimation limitations due to data sparsity. However, the XGBoost model successfully provided a complete ranking of variable importance for each injury type. For PIC, the top three predictors were density of traffic signals, density of congested roads, and % of retail land uses. For EIC, the most influential variables identified were the density of freeway lanes, the density of traffic signals, and the density of freight trip generation poles. In the case of FIC, the key predictors were truck volume density per road length, freeway lane density, and secondary road lane density. These results highlight the strength of machine learning approaches like XGBoost in capturing non-linear relationships and ranking predictors effectively, even when traditional models like ZIP struggle to produce stable estimates at the census tract level.

The comparative analysis between hexagonal, neighborhood, and census tract-level models reveals that hexagon-level modeling offers the most statistically robust insights into spatial crash determinants. The hexagonal ZIP model identifies a broader set of significant variables across all three injury severities, with clearer effect sizes (e.g., distance to airport for PIC: $\beta = -0.081^{***}$) and superior model performance metrics (lower RMSE and MAE). Neighborhood models, while capturing broader urban patterns, show reduced statistical significance and weaker model fit, with higher RMSE and fewer significant predictors. Similarly, the census tract-level models faced even greater limitations, with the ZIP model producing NaN values for several variables, indicating instability or insufficient variability within tracts.

It is important to note that the XGBoost models across all spatial units provided consistent variable rankings, but the hexagonal models again outperformed in capturing non-linear interactions, with variables like truck volumes ranking highly for EIC and FIC, despite ZIP insignificance.

SHAP analysis

The most important variables from hexagonal-based analysis were considered through SHAP analysis to understand the key drivers of EIC, FIC, and PIC, respectively. For EIC, as shown in Fig. 3a, the SHAP feature influence plot reveals that the density of traffic signals is the most influential predictor, with the highest mean SHAP value (+0.33), indicating that hexagons with more signalized intersections are strongly associated with a greater likelihood of EICs. This likely reflects the increased risk of vehicle conflict at controlled intersections, particularly where trucks interact with other road

users. The density of truck volumes per km² of road ranks second in importance (+0.28), underscoring that freight exposure is a critical factor in injury severity. Retail land use percentage also contributes significantly (+0.23), suggesting that commercial activity zones with high pedestrian and vehicle mixing are hotspots for evident injuries. Additional important predictors include density of freeway lane mileage, density of freight trip generation poles, and the percentage of congested roads, all of which point to infrastructure complexity and traffic flow stress as key contributors. Variables such as population density, density of primary roads, and distance to

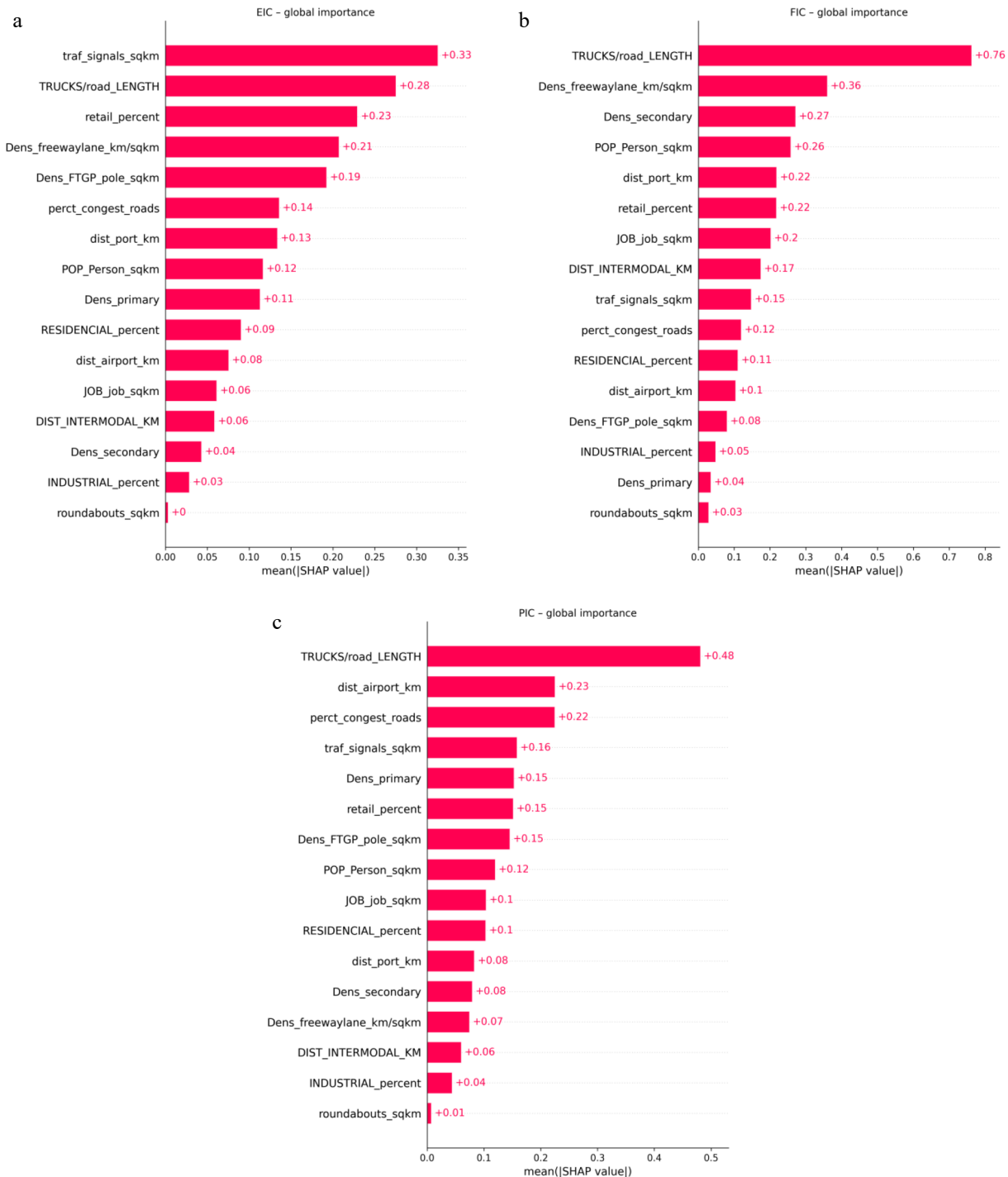


Fig. 3 Mean absolute SHAP value.

seaport show moderate effects, while others like percentage of industrial land use, density of secondary roads, and especially density of roundabouts have minimal influence on EICs. Overall, the SHAP results emphasize that EICs are primarily driven by areas with high traffic complexity, intense freight activity, and urban commercial concentration.

From Fig. 3b, while considering the key drivers of FIC, the analysis reveals that the density of truck volumes per km² of road is by far the most influential predictor, with a mean SHAP value of +0.76, highlighting that areas with intense freight exposure are significantly more prone to fatal outcomes. This is followed by the density of freeway (+0.36) and density of secondary road lane mileage (+0.27), suggesting that both high-speed arterial corridors and extensive secondary road networks contribute to fatal crash risk, possibly due to higher vehicle speeds and volume. Population density, distance to seaports, and percentage of retail land use also show notable influence, indicating that densely populated retail land and strategically located freight areas experience elevated fatality risks. Additionally, job density and distance to intermodal terminals reflect the impact of urban economic activity and freight logistics on FIC patterns. While the density of traffic signals, percentage of congested roads, and percentage of residential land use have a moderate influence, features like percentage of industrial land use, density of primary road lane mileage, and density of roundabouts appear to have limited explanatory power. Overall, the SHAP results suggest that fatal truck-related crashes are primarily driven by freight intensity, major road infrastructure, and urban activity centers, underscoring the need for targeted safety interventions in high-speed, freight-dominant corridors.

And finally, Fig. 3c shows that the SHAP feature influence analysis of PIC indicates that the density of truck volumes per km² of road is the most influential factor, with a mean SHAP value of +0.48, confirming that higher freight intensity strongly increases the likelihood of minor injury crashes. This is followed by distance to the airport (+0.23) and percentage of congested roads (+0.22), suggesting that areas closer to major freight hubs and those with higher traffic congestion are more prone to minor truck-related injuries. In addition to the density of traffic signal (+0.16), the density of primary road, percentage of retail land use, and density of freight trip generation poles all show similar influence levels (each +0.15), indicating that intersection complexity, freight generation hubs, and commercial activity contribute significantly to potential injury occurrences. Moderately important variables include population density, job density, and percentage of residential land use, which reflect the impact of traffic infrastructure and urban activity intensity. In contrast, variables like distance to seaports, density of secondary roads, and density of freeways have weaker contributions. The least influential predictors for PIC include the percentage of industrial land use and the density of roundabouts, with near-zero SHAP values, implying minimal effect. Overall, the SHAP results suggest that PICs are driven largely by freight volume, urban access points, and infrastructure-induced traffic complexity, reinforcing the need for localized safety measures in mixed-use, high-traffic zones.

From Fig. 4a, the SHAP summary plot for EIC at the hexagonal level reveals how both the magnitude and direction of key predictors influence model output. Notably, higher values of density of traffic signals, density of truck volumes per km², and percentage of retail land use consistently push SHAP values to the right, indicating a strong positive contribution to the likelihood of EICs. This suggests that intersections, freight volume, and commercial zones are critical risk areas. Similarly, the greater the density of freeway lane mileage,

the density of freight trip generation poles, and the percentage of congested roads also exhibit positive SHAP values at high feature levels, reinforcing the role of traffic intensity and infrastructure complexity in elevating injury risk. Population density and job density show moderate positive impacts, while distance-related variables, such as distance to seaports and distance to airports, reveal that shorter distances are associated with increased EIC risk. In contrast, features like percentage of industrial land use, density of secondary road lane mileage, and density of roundabouts have SHAP values clustered near zero, indicating minimal influence on EIC predictions. Overall, the SHAP plot highlights that evident injuries are driven by a combination of high traffic density, freight exposure, and complex urban infrastructure.

Similarly, for FIC shown in Fig. 4b, the most influential variable is the density of truck volumes per km² of road, where higher values (in red) strongly increase the likelihood of fatal crashes, as indicated by the SHAP values skewing to the right. Similarly, the density of freeway and the density of secondary road lane mileage show that high values are associated with a greater risk of fatal injuries, reflecting the dangers posed by fast-moving, freight-intensive road environments. Population density also contributes positively to FIC prediction, especially in high-density hexagons, likely due to increased exposure and interaction with freight traffic. Interestingly, distance to seaports and distance to intermodal terminals (i.e., lower values in blue) also lead to increased FIC risk, emphasizing the spatial concentration of fatal crashes near freight hubs. Urban activity indicators such as job density and percentage of retail land use are associated with elevated FIC predictions at higher values. In contrast, features like density of traffic signals, percentage of congested roads, and percentage of residential land use show a more mixed or moderate effect, while percentage of industrial land use, density of primary road lane mileage, and density of roundabouts exhibit very limited influence, with SHAP values centered around zero. Overall, this plot highlights that FICs are largely driven by intense freight exposure, proximity to logistic infrastructure, and high-capacity roads, reinforcing the need for safety interventions in areas where freight intensity overlaps with urban density.

Figure 4c represents the key driving factors in the case of PIC, where the density of truck volumes per km² of road is the most dominant factor, with higher values (red dots) strongly associated with an increase in PIC likelihood, as shown by positive SHAP values. Distance to airport (lower values in blue) also has a positive impact on PIC predictions, indicating that hexagons closer to major freight nodes are more prone to minor injuries. Additionally, a higher percentage of congested roads, density of traffic signals, and density of primary road lane mileage all contribute positively to the model's prediction, highlighting the role of traffic complexity and flow intensity in low-severity injury occurrence. Percentage of retail land use, density of freight trip generation poles, population density, and job density also influence PIC risk, with higher values generally increasing the output, although with slightly less impact. In contrast, predictors such as distance to ports, density of freeway lane mileage, percentage of industrial land use, and density of roundabouts show weaker and more neutral SHAP values, indicating limited influence on PICs at this resolution. Overall, the plot emphasizes that minor truck-related injuries are more likely in densely populated, commercially active, and congested areas with substantial truck presence, where infrastructure complexity intersects with freight movement.

While considering the SHAP waterfall plots, Fig. 5a provides a local explanation for hexagon 648, where the predicted risk of EIC is significantly higher than average, rising from a baseline value of 0.929 to a final prediction of 3.76. This increase is primarily driven by

Built environment and heavy-truck crash severity

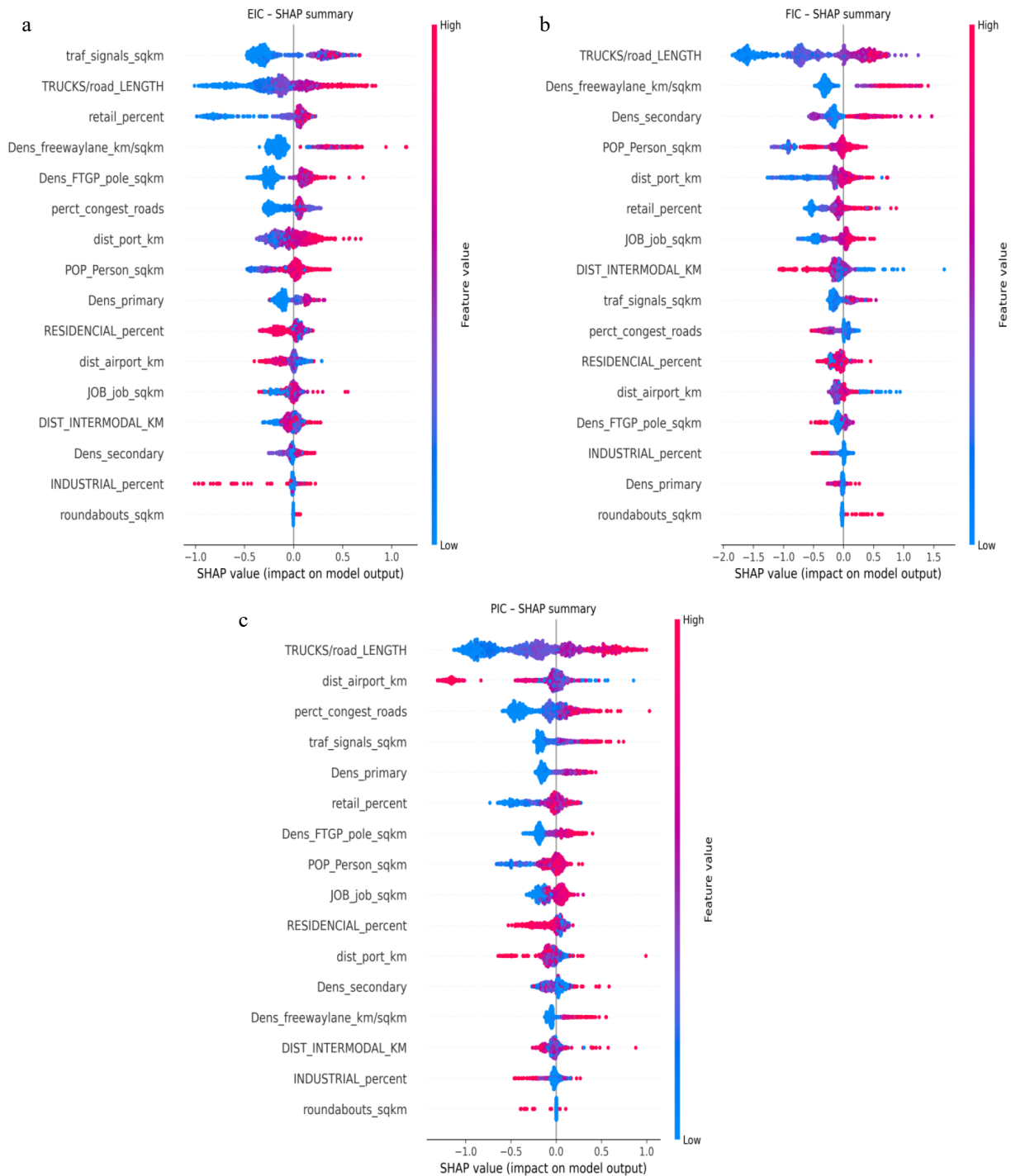


Fig. 4 Global SHAP summary plot.

several key features. The most influential contributor is the high density of freight trip generation poles (+0.71), followed by a substantial presence of percentage of freeway road lane mileage (+0.53) and high density of truck volumes per km² (+0.48), all of which reflect intense freight movement and complex road infrastructure. Additional factors such as the density of traffic signals, moderate distance to seaport, and a noticeable percentage of congested roads also push the prediction upward. Land-use characteristics like percentage of residential land uses and percentage of retail land uses, as well as job density, percentage of industrial land uses, contribute positively, but to a lesser degree. In contrast, the

density of primary road lane mileage slightly offsets the prediction (-0.08), possibly due to its traffic-distributing effect, while other features, such as density of roundabouts, population density, and distance to the airport, have a negligible influence. Overall, the plot illustrates that hexagon 648 exhibits a combination of freight exposure, traffic complexity, and mixed land use, which collectively elevate its likelihood of experiencing evident truck-related injuries.

Figure 5b also represents the SHAP waterfall plot, providing a local explanation for hexagon 317, where the predicted value for FIC is 1.96, significantly higher than the dataset's average baseline prediction of -1.558. This increase is driven by several high-impact

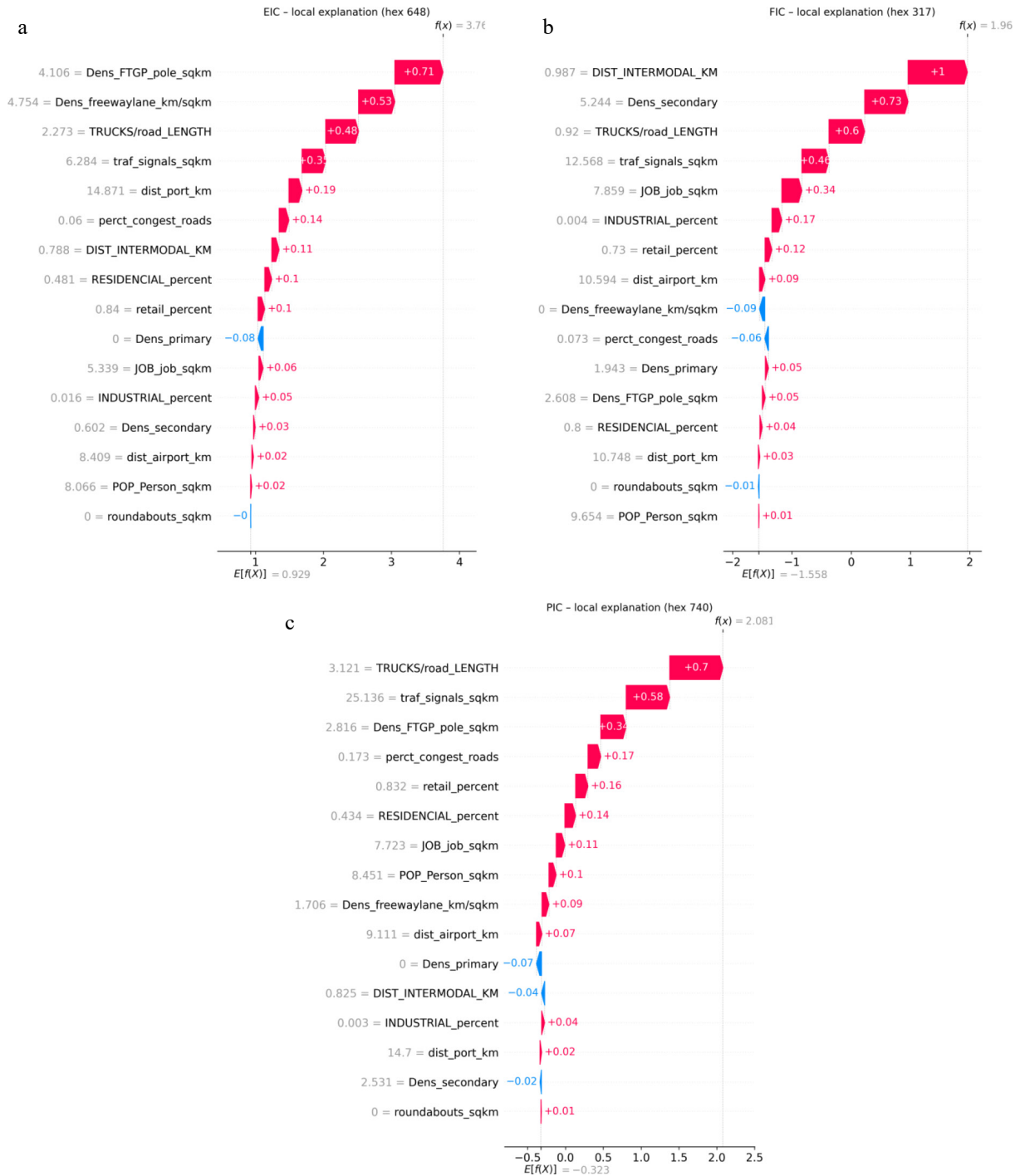


Fig. 5 SHAP waterfall plot.

features. The most influential is the short distance to the intermodal terminal (0.987 km), which contributes +1.00 to the prediction, indicating that proximity to freight hubs substantially raises fatal crash risk. This is followed by a high density of secondary road lane mileage (5.244 km/km²) and density of truck volumes per km² (0.92), contributing +0.73 and +0.60, respectively, further reinforcing the role of freight intensity and complex urban road networks. Additional contributors include a dense network or density of traffic signals (12.568 per km²) and high job density (7.859 jobs per km²), which add +0.46 and +0.34, likely reflecting high interaction points and mobility demand. Land-use indicators like percentage of industrial

land uses, percentage of retail land uses, and percentage of residential land uses also contribute positively, albeit at smaller magnitudes. Interestingly, some features reduce the risk prediction, including zero density of freeway (-0.09) and low percentage of congested roads (-0.06), which suggests the absence of high-speed truck corridors or traffic pressure may mitigate fatality risk in this area. Overall, this local SHAP explanation indicates that hexagon 317 is at elevated risk for fatal truck-related injuries due to its proximity to freight terminals, dense urban and employment activity, and complex secondary road infrastructure, with only limited mitigating effects from road congestion or highway infrastructure.

Built environment and heavy-truck crash severity

Finally, the SHAP waterfall plot for hexagon 740 is shown in Fig. 5c, where the predicted value for PIC is 2.081, considerably higher than the model's average baseline prediction of -0.323 . This elevated prediction is largely driven by a combination of freight intensity and urban infrastructure factors. The most influential contributor is the density of truck volumes per km^2 of road (3.121), which adds $+0.70$ to the prediction, indicating strong freight exposure in this area. This is followed by a very high density of traffic signals (25.136 per km^2), contributing $+0.58$, reflecting complex intersections that increase the likelihood of minor crashes. Other positive contributors include density of freight trip generation poles, percentage of congested roads, percentage of retail land uses and percentage of residential land uses, job density, and population density, each incrementally increasing the prediction by $+0.1$ to $+0.34$. These reflect a mix of commercial activity, urban intensity, and signal infrastructure, which are known to elevate low-severity crash risks in busy zones. In contrast, a few features modestly reduce the prediction. Notably, the density of primary road lane mileage (0) contributes -0.07 , suggesting that the absence of such roads may limit through traffic and slightly dampen crash risk. Distance to intermodal terminals (0.825 km) also reduces the prediction slightly (-0.04), perhaps due to limited direct exposure or shielding by other infrastructure. Similarly, secondary road lane mileage and distance to seaport have negligible to slightly negative impacts. Overall, this explanation shows that hexagon 740's high PIC prediction is driven by intense truck activity, complex road controls, and dense, active land use, hallmarks of a busy urban-freight interface where minor crashes are more likely.

SHAP dependence plots

Additionally, the SHAP dependence plots for all three injury types, PIC, EIC, and FIC, are presented to visualize how the top three most influential variables affect model predictions. These plots reveal the nature and direction of each variable's contribution to injury risk, highlighting potential thresholds or non-linear effects. For EIC, the top three predictors' density of traffic signals, density of truck volumes per km^2 of road, and percentage of retail land use show a strong positive influence on crash risk, with sharper increases in SHAP values at higher feature levels, particularly for density of traffic signals and freight exposure. Similarly, for FIC, key variables such as density of truck volumes per km^2 of road, density of freeway, and density of secondary road lane mileage exhibit steep positive associations, indicating that high-speed freight corridors and road complexity significantly elevate fatal crash risks. In the case of PIC, the dominant predictors include density of truck volumes, distance to airport, and percentage of congested roads, all of which demonstrate non-linear relationships where risk escalates beyond certain exposure thresholds. The remaining dependence plots for additional variables are provided in Supplementary Fig. S1 for further examination. These insights not only confirm the critical role of freight intensity and infrastructure but also guide targeted urban interventions by revealing how specific variables impact different types of truck-related injury outcomes.

The SHAP dependence plot shows that a higher density of traffic signals is strongly associated with increased risk of EIC. From Fig. 6a, while low densities (0–5 signals per km^2) have mixed effects, SHAP values become consistently positive beyond this range, indicating a clear upward trend. This suggests that hexagons with dense signalized intersections are more prone to EICs, likely due to increased vehicle interaction points. Similarly, Fig. 6b shows that the density of truck volumes per km^2 of road is a strong positive contributor to EIC.

As truck intensity increases, SHAP values sharply rise, especially between 0 and 5, indicating higher EIC risk in areas with more freight activity. Beyond a certain threshold, the effect plateaus, suggesting a saturation point where additional density of truck volumes adds limited extra risk. The percentage of retail land use has a non-linear relationship with EIC, shown in Fig. 6c. At very low retail presence (below 0.2), the SHAP values are highly variable, ranging from strongly negative to positive, suggesting inconsistent effects. However, once retail coverage exceeds almost 0.2, SHAP values stabilize above zero, indicating that higher retail concentration consistently increases EIC risk. This pattern reflects how commercial zones, likely involving more pedestrian and delivery activity, contribute to injury likelihood.

To discuss the key driving factors in the prediction of FIC, the SHAP dependence plot shown in Fig. 6d shows that the density of truck volumes per km^2 of road is a strong and consistent driver of FIC. As truck exposure increases, the SHAP values sharply rise, especially between 0 and 5, indicating a significant increase in predicted fatal crash risk. Beyond this point, the SHAP values plateau, suggesting that high truck density areas have already reached a threshold where further increases contribute marginally. This reinforces the critical role of freight intensity in fatal injury outcomes. Again, Fig. 6e shows that the density of freeway lane mileage is the secondary key representative positively associated with FIC. At very low densities (near zero), SHAP values are often negative or neutral, but as the density of freeway lane mileage increases beyond $1 \text{ km}/\text{km}^2$, SHAP values rise sharply, indicating a higher predicted risk of fatal crashes. The effect peaks and stabilizes at higher densities, suggesting that areas with more extensive freeway infrastructure, likely accommodating high-speed freight movement, are strongly linked to fatal injury outcomes.

Figure 6f represents the relationship between the density of secondary road lane mileage and FIC. At lower densities (below $3 \text{ km}/\text{km}^2$), SHAP values are predominantly negative, indicating a protective effect or lower predicted fatal crash risk. However, as density increases beyond this threshold, SHAP values rise sharply, peaking around $4\text{--}5 \text{ km}/\text{km}^2$, suggesting that very dense secondary road networks are associated with a higher risk of fatal injuries. This could reflect increased vehicle interactions or routing complexity in these areas. The pattern flattens or slightly declines at extreme densities, hinting at a saturation effect.

The SHAP dependence plot for the density of truck volumes per km^2 of road about PIC reveals a strong positive and non-linear effect. It is obvious from Fig. 6g that at very low truck road lengths, the SHAP values are predominantly negative, indicating a reduced likelihood of PIC. However, as the density of truck volumes increases, SHAP values rise sharply, suggesting that areas with higher truck volume density contribute more significantly to the predicted risk of possible injury crashes. This pattern plateaus after a certain point, indicating a threshold beyond which additional density of truck volumes does not increase risk proportionally.

The next plot from Fig. 6h represents the distance to the seaport about PIC, suggesting a subtle but nonlinear relationship. When the distance to a seaport is relatively low (under 20 km), the SHAP values fluctuate around zero, implying minimal influence on PIC risk. However, once the distance exceeds approximately 20 km, SHAP values turn more negative, indicating that locations farther from seaports are associated with a lower predicted risk of possible injury crashes. This could reflect reduced truck activity or freight movement in areas distant from major logistics hubs like ports.

At last, the percentage of congested roads concerning PIC shows a clear positive association, as shown in Fig. 6i. As the percentage

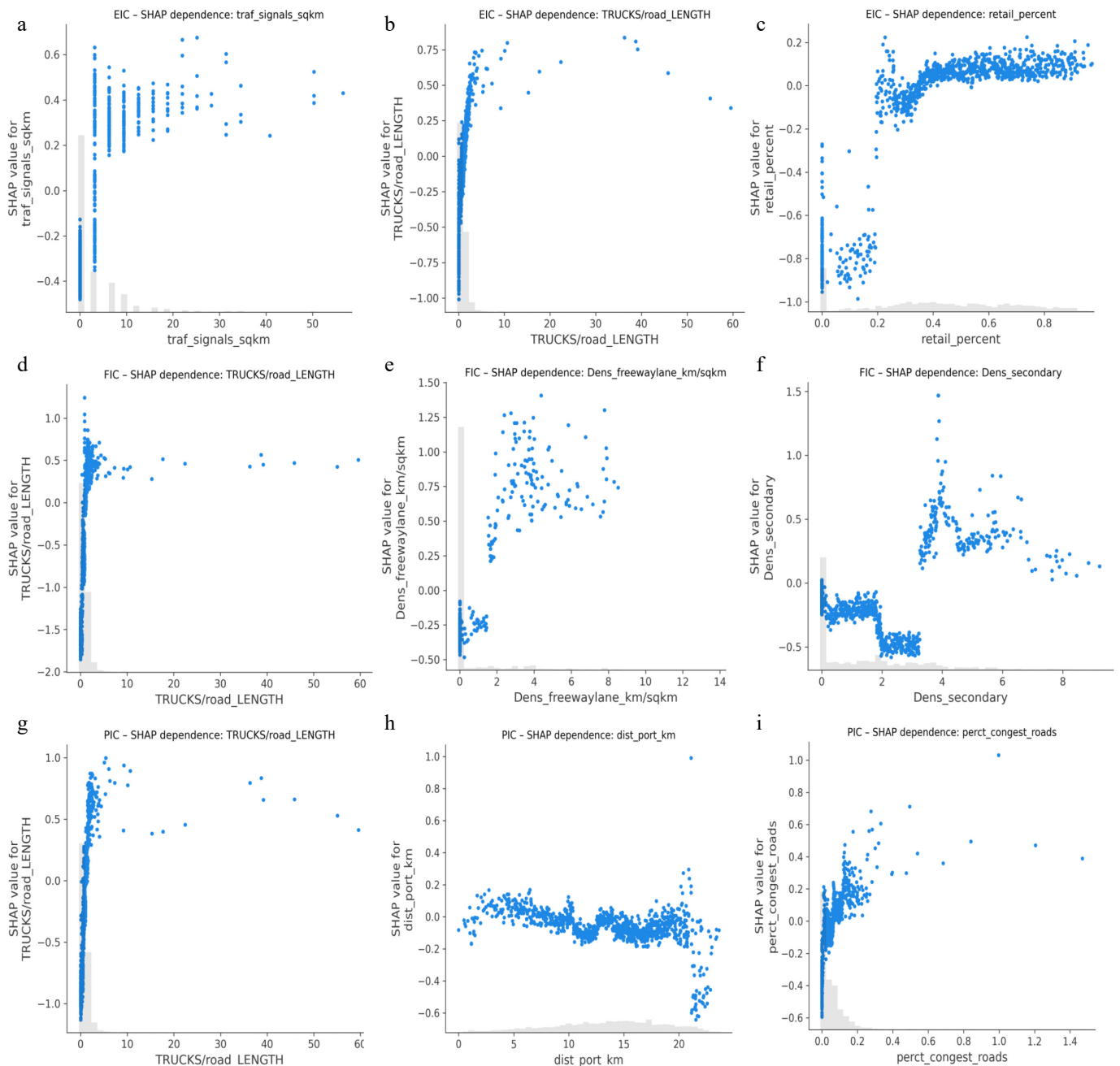


Fig. 6 SHAP dependence plot.

of congested roads increases, particularly beyond the 0.1 (10%) threshold, the SHAP values also increase, indicating a higher contribution to predicted injury crash risk. This suggests that areas with more frequent traffic congestion are more likely to experience possible injury crashes, likely due to increased interaction between vehicles, abrupt stops, and reduced maneuverability.

SHAP Interaction plot

The SHAP interaction analysis identifies the top three most influential variables for each injury severity type, presented to visualize how these predictors affect model predictions, while the remaining interaction plots are provided in [Supplementary Fig. S2](#). The interaction SHAP plot between the density of traffic signals per km² and the density of freeway lane mileage for EIC highlights a nuanced

relationship, shown in [Fig. 7a](#). As the density of traffic signals per km² increases, its SHAP value initially rises, indicating a stronger positive contribution to EIC risk, particularly when the density of freeway lane mileage is low (blue points). However, at higher traffic signal densities, the SHAP value tends to decrease, especially when combined with higher freeway lane densities (indicated by the color gradient turning pink). This suggests that the effect of traffic signals on injury risk is moderated by freeway infrastructure in denser freeway areas; the impact of traffic signals on EIC becomes less pronounced or even negligible.

From [Fig. 7b](#), the interaction SHAP plot between traffic signals per km² and density of freight trip generation poles for EIC illustrates how the presence of roadside infrastructure may shape crash risk. Initially, as the density of traffic signals increases, the SHAP values

Built environment and heavy-truck crash severity

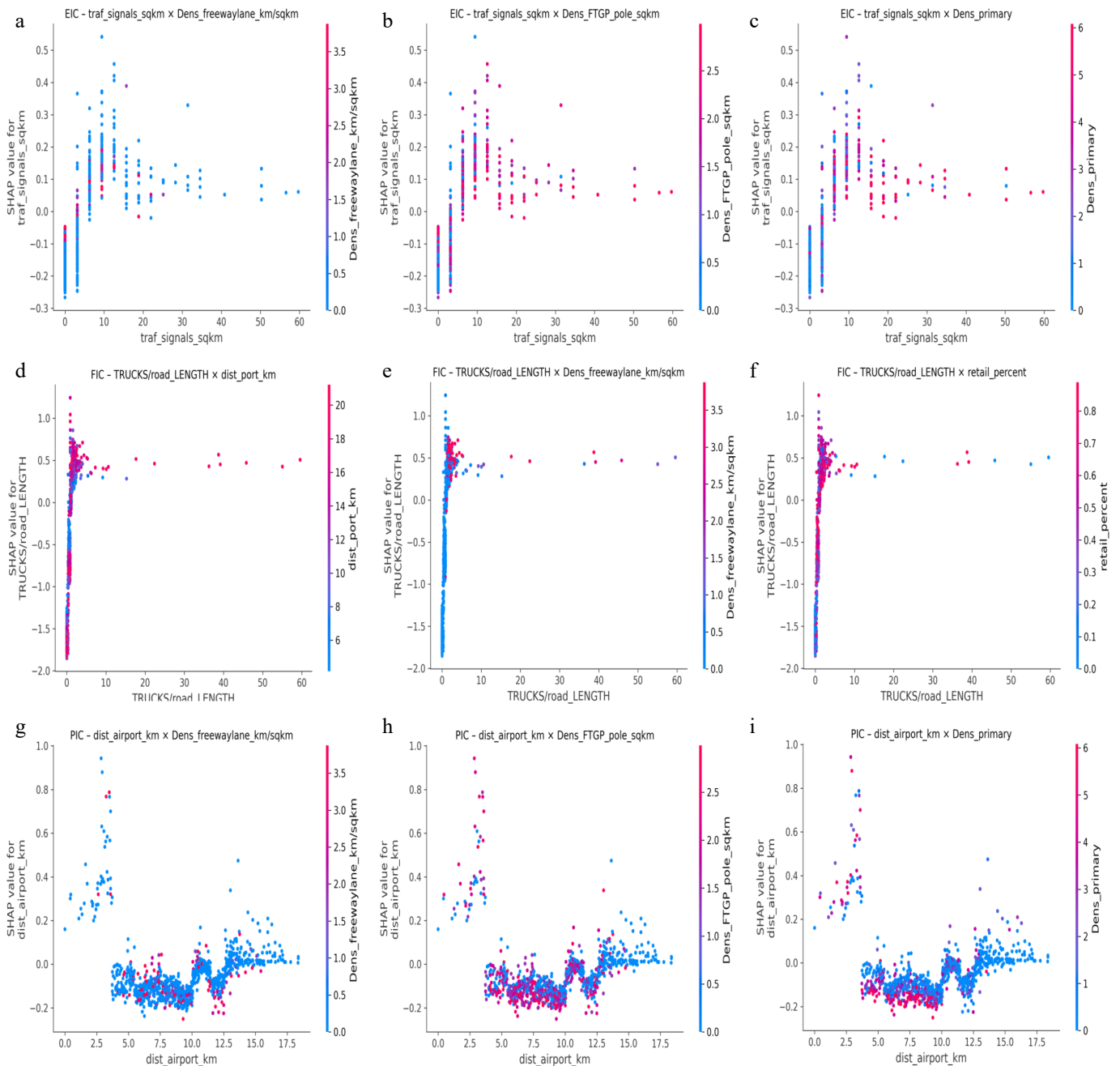


Fig. 7 SHAP interaction plot.

also increase, indicating a rising risk contribution, but this effect plateaus or slightly declines beyond 15 signals per km². Interestingly, the interaction color gradient shows that a higher density of freight trip generation poles tends to dampen the positive impact of traffic signals on injury risk, implying that well-distributed poles may serve as safety elements in high-signal areas. And the SHAP interaction plot for the density of traffic signals per km² for EIC in Fig. 7c reveals a strong non-linear relationship. Initially, an increase in the density of traffic signals leads to a rise in SHAP value, indicating elevated crash risk contributions. However, at higher densities of traffic signals (above 15 signals per km²), this influence stabilizes or slightly declines. Notably, the effect is more pronounced in regions with moderate density of primary road lane mileage, suggesting that primary roads amplify the safety relevance of signal placement,

while in areas with very high density of primary road lane mileage (red tones), the marginal safety contribution of additional signals diminishes.

For FIC, starting from Fig. 7d, the interaction plot reveals how the ratio of density of truck volumes per km² of road interacts with the distance to port in influencing crash risk. Higher density of truck volumes per square road length is generally associated with increased SHAP values, implying a stronger contribution to fatal crash likelihood. Interestingly, this effect is more pronounced in areas located farther from ports (depicted in red tones), suggesting that remote areas with high truck activity may face elevated fatal crash risks, possibly due to fewer safety controls or infrastructure limitations. In contrast, regions closer to ports (blue tones) show a more moderate SHAP contribution, indicating slightly better

mitigation despite truck presence. Figure 7e illustrates how the effect of the density of truck volumes per km² of road on FIC is influenced by the density of the freeway. The contribution of truck density to FIC is generally high, especially when the density of the freeway is low (blue). However, as the density of freeway lane mileage increases (shifting toward red), the SHAP values slightly moderate, suggesting that greater freeway infrastructure may help buffer the negative impact of truck concentration. This interaction implies that regions with more developed freeway networks can potentially mitigate fatal crash risks associated with heavy truck presence.

In addition, the SHAP interaction plot from Fig. 7f shows how the density of truck volumes per km² of road affects FIC, and how this effect varies with the percentage of retail land uses. The plot indicates that areas with higher densities of truck volumes per km² of road generally have higher SHAP values, meaning a greater positive contribution to FIC risk. Importantly, this relationship remains consistently high across varying percentages of retail land uses, suggesting retail presence does not significantly mitigate the safety risk posed by high truck concentrations. Dense retail areas (pink or red) show similarly elevated SHAP values, implying potential compound risks in commercial zones with heavy density of truck volumes. For PIC, the SHAP interaction plot shows that the relationship between distance from the airport and the densities of freeway lane mileage is moderated. Figure 7g suggests that being closer to airports (0–3 km) is associated with a higher SHAP value, indicating increased PIC risk. This effect is particularly pronounced in areas with low density of freeway lane mileage (shown in blue), possibly due to a higher percentage of congested roads or complex road structures near airports. As the distance increases beyond 7 km, SHAP values generally decrease and stabilize, indicating a lower contribution to PIC. Overall, proximity to airports appears to elevate PIC risk, especially in low freeway-density environments. Figure 7h shows the effect of distance from the airport on PIC, conditioned on the utility density of freight trip generation poles. The plot shows a clear trend: shorter distances to the airport (0–3 km) are associated with higher SHAP values, indicating a greater contribution to PIC risk. This effect is more pronounced in areas with higher FTGP pole densities (represented by pink or red dots), possibly reflecting complex road infrastructure or obstructions in such zones. As distance increases (> 5 km), SHAP values drop and stabilize, especially where FTGP pole density is low, suggesting reduced PIC risk farther from airports and in less obstructed environments.

Finally, the SHAP interaction plot shown in Fig. 7i illustrates how distance from the airport influences PIC, depending on the density of primary road lane mileage. Closer distance to airports (< 5 km) is associated with higher SHAP values, indicating a greater positive contribution to PIC risk. Within that short distance, a higher density of primary road lane mileage (pink or red dots) amplifies the effect, suggesting that denser major roads near airports increase crash risk, likely due to higher traffic volumes and complex road environments. Beyond 5 to 6 km, SHAP values decrease and stabilize, regardless of road density, showing that distance buffers the risk.

Discussion and conclusions

This study investigated the spatial determinants of freight-related injury crashes in Fortaleza, Brazil, focusing on three injury outcome PIC, EIC, and FIC, and comparing model outputs across three spatial aggregations: hexagons, neighborhoods, and census tracts. Using a rich and multidimensional urban dataset, key variables such as

infrastructure density (e.g., traffic signal and freeway lane densities), land use (e.g., percentage of retail zones), and proximity to logistical nodes (e.g., distance to seaports and airports) were analyzed using both statistical and machine learning techniques.

The hexagon-level ZIP model offered the most statistically robust insights. It identified several critical predictors of crash frequency and severity, including proximity to airports, seaports, and intermodal terminals, as well as job density and infrastructure elements like freeway and signal density. Notably, the negative association between airport distance and PIC aligns with prior urban freight safety literature that highlights elevated crash risks near logistics hubs^[37]. An initially counterintuitive finding that increased distance from ports correlates with higher crash likelihood was also observed, consistent with prior studies on port-induced traffic dispersal patterns^[4].

Retail land use emerged as a strong positive predictor, supporting the broader evidence that commercial zones are associated with increased traffic conflict and injury risk^[38]. Similarly, higher job density was significantly linked to increased crash frequency, reinforcing prior findings that employment centers contribute to transportation system stress and crash risk^[39,40]. Roadway features such as freeway and secondary road lane densities were also positively associated with crash counts, underscoring the impact of infrastructure design and traffic intensity on safety outcomes^[40].

At the neighborhood scale, the relationships between traffic infrastructure and freight exposure become more pronounced. ZIP models indicate elevated crash risk in areas located near airports and intermodal terminals, aligning with prior findings that emphasize the influence of freight hubs on injury severity^[4,21]. In addition, higher densities of freeway lane mileage and traffic signals are consistently associated with increased crash frequency, further underscoring the role of major roadway infrastructure in shaping exposure and conflict points^[4,21]. Notably, the ZIP model also identifies greater secondary road density as a protective factor against both PIC and EIC, suggesting that more distributed road networks may help diffuse traffic flows and reduce injury risk^[4]. At the census tract level, the ZIP model exhibited limited capacity to identify statistically significant predictors across crash types, largely due to data sparsity and the high proportion of zero-count observations. These shortcomings are consistent with long-recognized limitations of fine-scale spatial units, where low crash frequencies and pronounced spatial fragmentation restrict the informational content necessary for stable model estimation^[41]. Although zero-inflated models are specifically designed to address excess-zero scenarios, they may still yield estimation failures such as non-estimable coefficients or undefined outputs (NaN values) when event density is insufficient or the data structure lacks variability. Similar issues have been documented in diverse high-quality studies: DeSantis et al.^[42] reported non-estimable coefficients in sparse biomedical datasets; Garcia & Suarez^[43] observed non-convergence of a ZINB model in the context of zero-dominated social science data; and Evans et al.^[44] documented missing model outputs under extreme zero-inflation. Collectively, these findings reinforce that, despite their theoretical suitability, zero-inflated models can underperform when applied to highly disaggregated units with limited data support, ultimately undermining inference and predictive utility.

To address these limitations and better interpret spatial crash patterns, the ZIP model was initially employed to distinguish structural zeros from sampling-related variability, reflecting its dual-process framework. However, as discussed above, the ZIP model itself struggles when faced with an extremely high number of

Built environment and heavy-truck crash severity

zero-count observations, especially in fine-scale spatial units, resulting in estimation instability and, in some cases, undefined outputs such as NaN values. This limitation further justifies the adoption of more flexible modeling approaches. In this context, the study incorporated the XGBoost algorithm, which substantially outperformed ZIP in terms of predictive accuracy and model fit. While XGBoost validated several key predictors previously identified by the ZIP model, it also uncovered additional influential variables such as the truck-to-road ratio, whose nonlinear contributions to crash risk were not captured by the parametric structure of the ZIP framework.

To address the presence of excessive zeros in the crash data, the ZIP model was employed to distinguish structural zeros from true count variability, enabling clearer interpretation of the spatial risk landscape. However, to capture more complex, nonlinear relationships, the study incorporated the XGBoost algorithm, which outperformed ZIP in predictive accuracy and model fit. XGBoost identified many of the same top predictors but also highlighted additional variables (e.g., truck-to-road ratios) whose nonlinear effects were not captured by ZIP. At the hexagon level, XGBoost achieved the best overall performance, consistently identifying freeway lane mileage^[4], traffic signal density, truck volume^[2,4], and freight trip generation poles^[4] as top predictors. This intermediate resolution captured both local variability and broader spatial patterns, allowing the model to detect nonlinear relationships often missed by parametric approaches. Similarly, at the neighborhood level, XGBoost maintained stable predictive performance, with traffic infrastructure and freight-related variables remaining central. Industrial and retail land uses gained relevance, particularly for fatal crashes, reflecting broader land use dynamics^[2]. However, reduced granularity limited the detection of micro-scale crash patterns. While at the census tract level, although performance was affected by data sparsity, XGBoost still outperformed ZIP and identified key predictors such as traffic signal density, congestion, and truck volume^[4]. Variability in feature importance across crash types was higher, reflecting instability due to fragmentation and zero dominance in the data. Overall, results varied notably with spatial scale. Hexagons provided the most balanced and detailed outputs, neighborhoods offered stability with coarser insights, and census tracts, despite high resolution, were more prone to instability. These differences underscore the sensitivity of machine learning results to spatial unit selection.

To interpret the XGBoost 'black-box' structure, SHAP analysis was employed. SHAP visualizations revealed non-monotonic and threshold-based relationships among variables. For instance, traffic signal density positively influenced EIC up to a saturation point, after which its impact diminished, suggesting congestion thresholds. The density of truck volumes was a dominant predictor for FIC, especially in areas near ports or along freeway corridors. For PIC, the risk peaked in zones close to airports, particularly where high densities of freeway lanes or freight poles coexisted. These interaction effects provide nuanced insights that traditional statistical models often overlook.

The inclusion of three spatial scales, hexagons, neighborhoods, and census tracts, allowed for a meaningful comparison of how spatial aggregation influences model performance. Hexagon-level analyses consistently provided more statistically significant and spatially specific insights. Neighborhood-level models were useful for identifying broader planning patterns but exhibited reduced model performance and fewer significant predictors. Census tract-level ZIP models often failed to converge (yielding NaN values), indicating data sparsity or collinearity issues at that aggregation level. However, XGBoost managed to produce stable rankings for the

census tracts, demonstrating its resilience to such limitations. By combining interpretable statistical models with explainable machine learning techniques, this study offers a transparent and effective methodology for urban crash risk assessment. It underscores the potential of integrating ZIP, XGBoost, and SHAP to support evidence-based, spatially adaptive, and risk-sensitive urban transport and logistics planning.

Both ZIP and XGBoost models indicate higher crash risks near airports and intermodal terminals, in areas characterized by elevated retail activity, job density, freeway lane mileage, traffic signal density, and congestion. Notably, freeway lane mileage and traffic signal density consistently emerge as leading predictors in XGBoost models, demonstrating robust relationships regardless of crash severity type.

At this scale, the associations between traffic infrastructure and freight exposure are more prevalent. ZIP models highlight increased crash risk related to proximity to airports and intermodal terminals, as well as higher densities of freeway lane mileage and traffic signals. Interestingly, the ZIP approach identifies greater secondary road density as protective against both PIC and EIC. Similarly, the XGBoost model consistently ranks freeway lane mileage, truck volume density, freight trip generation poles, and traffic signal density among the top predictors.

However, the analysis offers the highest spatial granularity, clearly pinpointing local determinants of crash risk. XGBoost models robustly identify traffic signal density, road congestion, retail land use, and truck volume as critical predictors for PIC and EIC.

Policy implications

From a public policy perspective, the comparative findings across hexagons, neighborhoods, and census tracts, as well as across ZIP and XGBoost models, reveal that a uniform, 'one-size-fits-all' intervention strategy would be ineffective for managing freight-related injury crashes. The hexagon-level models, particularly when paired with XGBoost and SHAP analysis, demonstrated superior performance in identifying key spatial risk factors such as high truck traffic density, proximity to freight hubs (airports, seaports, intermodal terminals), and infrastructure indicators like freeway lane and traffic signal density. These insights enable interventions such as retiming or reducing traffic signals in high-risk freight corridors, improving infrastructure near logistics hubs, or enhancing pedestrian safety around areas of concentrated retail or job density. For instance, findings showed that closer proximity to the airport significantly increased PIC and EIC risk, and truck traffic per road length strongly influenced FIC. These patterns underscore the need for targeted mitigation strategies in specific hotspots, rather than applying broad city-wide measures. Although not the most fine-grained unit, hexagons provided a strong spatial resolution for identifying high-risk corridors, supporting interventions such as reconfiguring traffic infrastructure or managing freight access near logistics clusters.

At the neighborhood level, although some significant predictors overlapped with hexagonal results (e.g., traffic signals, freeway lanes), the overall statistical robustness and model fit declined. Still, this scale is valuable for district-level policies such as land-use adjustments, zoning reforms, and corridor-level freight regulations that affect broader areas but do not require high geographic precision. On the other hand, census tract-level analysis revealed considerable estimation issues in ZIP models (e.g., NaN values), suggesting this scale may be too disaggregated or statistically unstable for precise crash prediction. However, XGBoost rankings at

the census level still offered useful insights into dominant predictors, highlighting the role of machine learning in complementing traditional methods. The neighborhood unit aligns with recommendations by Briz-Redón et al.^[6] and Xu et al.^[19], who stress the importance of selecting spatial units that balance interpretability and statistical robustness, particularly when informing urban-scale policy decisions.

At the census tract level, despite challenges related to data sparsity and estimation instability, particularly in ZIP models, the results offer important policy insights when interpreted with caution. The ability of XGBoost to identify consistent predictors at this fine-grained resolution suggests its utility for informing localized, equity-oriented interventions. When complemented by socio-demographic data, this scale is valuable for integrating traffic safety with broader agendas such as environmental justice and spatial equity in urban planning^[30]. Thus, while not optimal for predictive modeling with traditional methods, census tracts remain relevant for policy frameworks that require fine spatial targeting and socially inclusive planning.

Therefore, policymakers should adopt a multi-layered spatial strategy: using hexagonal units for precision interventions, neighborhood models for regional infrastructure and mobility planning, and census tracts for broader socio-economic profiling. This hybrid approach ensures that interventions are both spatially responsive and functionally scalable, aligning urban safety investments with the actual geography of risk exposure. It also provides a strong basis for data-driven decision-making, prioritizing high-impact zones and allocating resources efficiently to reduce the frequency and severity of freight-related crashes in urban environments.

Taken together, the findings advocate for a multi-scalar, data-informed policy strategy. Hexagonal units are well-suited for operational interventions in high-risk areas, neighborhood-level models support structural planning and freight zoning, and census tract data can enrich socio-spatial profiling when supplemented with robust methods like XGBoost. This layered approach reflects guidance from^[18], who emphasizes the importance of multiscale spatial modeling to mitigate MAUP effects and enhance the relevance of empirical evidence for policy formulation. Matching spatial resolution to policy scope enables more accurate risk targeting and resource allocation, improving the efficacy of crash mitigation strategies in freight-intensive urban areas.

Author contributions

The authors confirm their contributions to the paper as follows: study conception and design: Ullah I, Pinheiro C, Yuan Q; data collection: Pinheiro C, Brito B; analysis and interpretation of results: Ullah I, Ayaz MA, Yuan Q; draft manuscript preparation: Ullah I, Ayaz MA, Brito B, Yuan Q. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgments

This study was funded by the National Natural Science Foundation of China (Grant No. 52302394), and the Shanghai Municipality Science and Technology Commission (Grant No. 24692106600).

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

Received 4 December 2025; Revised 26 February 2026; Accepted 20 March 2026; Published online 31 March 2026

References

- [1] Bunn TL, Liford M, Turner M, Bush A. 2022. Driver injuries in heavy vs. light and medium truck local crashes, 2010–2019. *Journal of Safety Research* 83:26–34
- [2] Yu C, Hua W, Yang C, Fang S, Li Y, et al. 2024. From sky to road: incorporating the satellite imagery into analysis of freight truck-related crash factors. *Accident Analysis & Prevention* 200:107491
- [3] McDonald N, Yuan Q, Naumann R. 2019. Urban freight and road safety in the era of e-commerce. *Traffic Injury Prevention* 20:764–770
- [4] Yang C, Chen M, Yuan Q. 2021. The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: an exploratory analysis. *Accident Analysis & Prevention* 158:106153
- [5] Sexton D. 2008. Modifiable areal unit problem (MAUP). In *Wiley StatRef: Statistics Reference Online*, eds Balakrishnan N, Colton T, Everitt B, Piegorisch W. Hoboken, NJ: Wiley. doi: 10.1002/9781118445112.stat03806
- [6] Briz-Redón Á, Martínez-Ruiz F, Montes F. 2019. Investigation of the consequences of the modifiable areal unit problem in macroscopic traffic safety analysis: a case study accounting for scale and zoning. *Accident Analysis & Prevention* 132:105276
- [7] Li C, Chen L. 2025. Exploring the impact of built environment on crash risks at transportation hubs. *Accident Analysis & Prevention* 218:108079
- [8] Parsa AB, Movahedi A, Taghipour H, Derrible S, Mohammadian AK. 2020. Toward safer highways: application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention* 136:105405
- [9] Ziakopoulos A, Yannis G. 2020. A review of spatial approaches in road safety. *Accident Analysis & Prevention* 135:105323
- [10] Soliani RD, Argoud ARTT, Santiago F, Lopes AVB, Emekwuru N. 2024. Catastrophic causes of truck drivers' crashes on Brazilian highways: Mixed method analyses and crash prediction using machine learning. *Multimodal Transportation* 4:100173
- [11] Haghani M, Behnood A, Dixit V, Oviedo-Trespalacios O. 2022. Road safety research in the context of low- and middle-income countries: macro-scale literature analyses, trends, knowledge gaps and challenges. *Safety Science* 146:105513
- [12] U. S. Department of Transportation, Bureau of Transportation Statistics. 2023. *National transportation statistics: Table 1-50 – U. S. ton-miles of freight by mode, 1960–2023*. Washington, DC: BTS. Available at www.bts.gov
- [13] Economic Commission for Latin America and the Caribbean (ECLAC). 2024. *Freight transport and logistics statistics yearbook: Brazil profile 2020*. United Nations-ECLAC, Santiago, Chile. Available at <https://statistics.cepal.org/yearbook/2024/statistics.html?lang=en&theme=economic>
- [14] Lopes AS, Cavalcante CB, Vale DS, Loureiro CFG. 2020. Convergence of planning practices towards LUT integration: seeking evidences in a developing country. *Land Use Policy* 99:104842
- [15] Pinheiro CDP, Gonzalez-Feliu J, Bertoincini BV. 2025. Addressing spatial heterogeneity and MAUP in urban transport geography: a multi-scale analysis of accessibility and warehouse location. *Journal of Transport Geography* 123:104078
- [16] Openshaw S, Taylor PJ. 1979. A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In *Statistical Applications in Spatial Sciences*, ed. Wrigley N. London: Pion. pp. 127–144

- [17] Reda AK, Tavasszy L, Gebresenbet G, Ljungberg D. 2023. Modelling the effect of spatial determinants on freight (trip) attraction: a spatially autoregressive geographically weighted regression approach. *Research in Transportation Economics* 99:101296
- [18] Zhai X, Sze NN, Lee JJ, Xu P, Huang H. 2025. Multi-scale approaches to cope with scale effect issues in macroscopic crash analysis. *Accident Analysis & Prevention* 214:107971
- [19] Xu P, Huang H, Dong N. 2018. The modifiable areal unit problem in traffic safety: basic issue, potential solutions and future research. *Journal of Traffic and Transportation Engineering (English Edition)* 5:73–82
- [20] Dong C, Clarke DB, Richards SH, Huang B. 2014. Differences in passenger car and large truck involved crash frequencies at urban signalized intersections: an exploratory analysis. *Accident Analysis & Prevention* 62:87–94
- [21] Yang C, Chen M, Yuan Q. 2021. The geography of freight-related accidents in the era of E-commerce: evidence from the Los Angeles metropolitan area. *Journal of Transport Geography* 92:102989
- [22] Zou W, Wang X, Zhang D. 2017. Truck crash severity in New York City: an investigation of the spatial and the time of day effects. *Accident Analysis & Prevention* 99:249–261
- [23] Ahmed S, Hossain MA, Ray SK, Bhuiyan MMI, Sabuj SR. 2023. A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. *Transportation Research Interdisciplinary Perspectives* 19:100814
- [24] Yue H. 2024. Investigating the influence of streetscape environmental characteristics on pedestrian crashes at intersections using street view images and explainable machine learning. *Accident Analysis & Prevention* 205:107693
- [25] Chang I, Park H, Hong E, Lee J, Kwon N. 2022. Predicting effects of built environment on fatal pedestrian accidents at location-specific level: application of XGBoost and SHAP. *Accident Analysis & Prevention* 166:106545
- [26] Scarano A, Sadeghi M, Mauriello F, Riccardi MR, Aghabayk K, et al. 2025. Cyclist crash severity modeling: a hybrid approach of XGBoost-SHAP and random parameters logit with heterogeneity in means and variances. *Journal of Safety Research* 93:373–398
- [27] Dong S, Khattak A, Ullah I, Zhou J, Hussain A. 2022. Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with SHAPley Additive exPlanations. *International Journal of Environmental Research and Public Health* 19:2925
- [28] Instituto Brasileiro de Geografia e Estatística. 2022. *IBGE Cities database*. Brasília: IBGE. Available at: <https://cidades.ibge.gov.br/brasil/ce/Fortaleza/panorama>
- [29] Pinheiro CDP, Gonzalez-Feliu J, Bertoncini BV. 2025. A novel comprehensive spatial accessibility indicator to capture the latent nature of accessibility in logistic warehouses. *Research in Transportation Economics* 109:101517
- [30] Pinheiro CDP. 2024. *Evaluating accessibility and transportation justice in urban freight transport: implementation, spatial dynamics and policy implications*. PhD thesis. Universidade Federal do Ceará, Fortaleza. <https://repositorio.ufc.br/handle/riufc/81139>
- [31] Autarquia Municipal de Trânsito e Cidadania. 2025. *Traffic incidents recorded in the municipality of Fortaleza, Brazil*. Fortaleza: AMC. <https://dados.fortaleza.ce.gov.br/organization/amc>
- [32] Ducret R, Lemarié B, Roset A. 2016. Cluster analysis and spatial modeling for urban freight. Identifying homogeneous urban zones based on urban form and logistics characteristics. *Transportation Research Procedia* 12:301–313
- [33] Regal-Ludowieg A, Gonzalez-Feliu J, Rodríguez M. 2022. Delivery bay location and dimensioning for city logistics uses: an interactive modelling approach. In *Production and Operations Management*, eds Florez JV, de Brito R Junior, Leiras A, Alberto Paz Collado S, Alvarez MDG, et al. vol 391. Cham: Springer. pp. 475–481 doi: [10.1007/978-3-031-06862-1_35](https://doi.org/10.1007/978-3-031-06862-1_35)
- [34] Lambert D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34:1–14
- [35] Chen T, Guestrin C. 2016. XGBoost: a scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '16:785–794
- [36] Lundberg SM, Lee SI. 2017. A unified approach to interpreting model predictions. *arXiv* 00:1705.07874
- [37] Shi Z, Wang Y, Guo D, Jiao F, Zhang H, Sun F. 2025. The Urban Intersection Accident Detection Method Based on the GAN-XGBoost and Shapley Additive Explanations Hybrid Model. *Sustainability* 17:453
- [38] Laphrom W, Se C, Champahom T, Jomnonkwo S, Wipulanusatd W, et al. 2024. XGBoost-SHAP and unobserved heterogeneity modelling of temporal multivehicle truck-involved crash severity patterns. *Civil Engineering Journal* 10:1890–1908
- [39] Wang Y, Kockelman KM. 2013. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. *Accident Analysis & Prevention* 60:71–84
- [40] Ukkusuri S, Miranda-Moreno LF, Ramadurai G, Isa-Tavarez J. 2012. The role of built environment on pedestrian crash frequency. *Safety Science* 50:1141–1151
- [41] Lord D, Washington S, Ivan JN. 2007. Further notes on the application of zero-inflated models in highway safety. *Accident Analysis & Prevention* 39:53–57
- [42] DeSantis SM, Lazaridis C, Ji S, Spinale FG. 2014. Analyzing propensity matched zero-inflated count outcomes in observational studies. *Journal of Applied Statistics* 41:127–141
- [43] García J, Suárez MJ. 2023. The relevance of specification assumptions when analyzing the drivers of physical activity practice. *Economic Modelling* 119:106127
- [44] Evans TG, Castellino F, Dobczyk MK, Tucker G, Walley AM, et al. 2024. Assessment of CD8⁺ T-cell mediated immunity in an influenza A(H3N2) human challenge model in Belgium: a randomized phase 2 study. *Lancet Microbe* 5:645–654



Copyright: © 2026 by the author(s). Published by Maximum Academic Press, Fayetteville, GA. This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit <https://creativecommons.org/licenses/by/4.0/>.