Data Science for Transportation (2025) 7:27
https://doi.org/10.1007/s42421-025-00143-3

RESEARCH

Bus Stop Typology Reveals Crash Risk Environments

Tolu Oke[1] · Alexandra Pate[2] · Francis Tainter[1] · Jimi Oke[1] · Michael Knodler[1]

Received: 1 November 2025 / Revised: 12 November 2025 / Accepted: 12 November 2025

Abstract

Bus stops are critical connectors in the public transit network, yet in Massachusetts, 46% of crashes involving vulnerable road users (VRUs) occur near bus stops. Traditional analysis methods overlook the complex patterns around bus stops that contribute to crashes. To address this, we developed a framework that groups bus stops into safety typologies in order to predict crash risk in a more targeted manner and guide interventions. First, we fused multi-source data for 1773 stops to obtain 313 crash-relevant features. Then we applied explanatory factor analysis to extract 13 underlying components of crash risk based on these features. Using these factors, we fitted a Gaussian mixture model, which resulted in 13 distinct bus stop types. Finally, we trained extreme gradient boosting models to predict crash risk for each type. Interpreting the models via SHapley Additive exPlanations (SHAP values), we identified crash risk characteristics, assessed their impact, and proposed targeted countermeasures to inform effective safety interventions. Our findings confirm that bus stop safety is inherently context-dependent, necessitating typology-informed solutions rather than a one-size-fits-all approach. The novel framework presented here can serve as a foundation for further research on bus stop safety, as well as provide planners and decision makers with actionable guidance for targeted safety improvements.

Keywords Pedestrian safety · Bus stop safety · Crash risk · Machine learning · Factor analysis · Gaussian mixture models · Extreme gradient boosting models · Transportation safety

Introduction

Public transit is essential for enhancing urban mobility and enabling access to jobs, housing, education, healthcare, and other opportunities. Buses proved to be a crucial and resilient component of the public transit system, with the smallest ridership decline and fastest rebound of all modes impacted by the COVID-19 pandemic (APTA 2024). Bus stops serve as critical transit infrastructure, providing convenient and safe locations for passenger boarding, alighting, and waiting. However, their design and operational characteristics can create complex pedestrian–vehicle safety challenges. Studies show that bus stops are associated with a higher frequency of traffic conflicts for vulnerable road users (VRUs), particularly pedestrians and cyclists (Pulugurtha and Penkey 2010; Ukkusuri et al. 2012; Quistberg et al. 2015; Kraidi and Evdorides 2020; Su and Sze 2022; Ashraf et al. 2022; Rahman et al. 2022). This crash concentration results from multiple interacting elements, including increased pedestrian volumes, bus passenger boardings and alightings, sight obstructions created by stopped buses, time–pressure associated with transit use, vehicular traffic flow, and surrounding bus stop and road conditions.

Traffic crashes resulted in nearly 40,000 fatalities in the U.S. in 2021, a daily average exceeding 120 individuals, and have since risen to their highest levels in over a decade (NHTSA 2023). Following the national trend, Massachusetts also experienced record-breaking traffic fatalities, with 409

✉ Tolu Oke
toke@umass.edu

Alexandra Pate
apate@umass.edu

Francis Tainter
ftainter@umass.edu

Jimi Oke
jimi@umass.edu

Michael Knodler
mknodler@umass.edu

[1] Department of Civil and Environmental Engineering, University of Massachusetts Amherst, 130 Natural Resources Rd, Amherst, MA 01003, USA

[2] Department of Mathematics, Spelman College, 350 Spelman Ln SW, Atlanta, GA 30314, USA

deaths in 2022 (Impact 2024). VRUs are disproportionately impacted. Despite representing only 10% of trips, pedestrians and bicyclists account for a growing 20% of traffic fatalities. News Release (2023). In Massachusetts, pedestrian fatalities reached a historic high of 98 in 2022, representing almost a quarter of all traffic deaths. To fully understand the scope of the issue, which extends beyond fatalities, Fig. 1A displays the trend for all crashes (fatal and non-fatal) involving a VRU in Massachusetts from 2016 to 2024. Following a pre-pandemic decline from a peak of nearly 4,000 crashes in 2016, the numbers dropped sharply in 2020 due to COVID-19 but have since rebounded steadily to pre-pandemic levels (3300–3450) from 2021 through 2024.

Although crashes seldom occur precisely at bus stops, the pedestrian traffic they generate concentrates risk in the immediate vicinity. Defining the crash catchment area can be challenging. An overly narrow buffer fails to capture crashes associated with pedestrian movement around the stop, while an excessively large buffer over-attributes crashes unrelated to the transit environment. As shown in Fig. 1B, the proportion of VRU crashes increases with the buffer size, from 23% within 100 feet to nearly 60% within 500 feet. Despite this clear trend, the specific

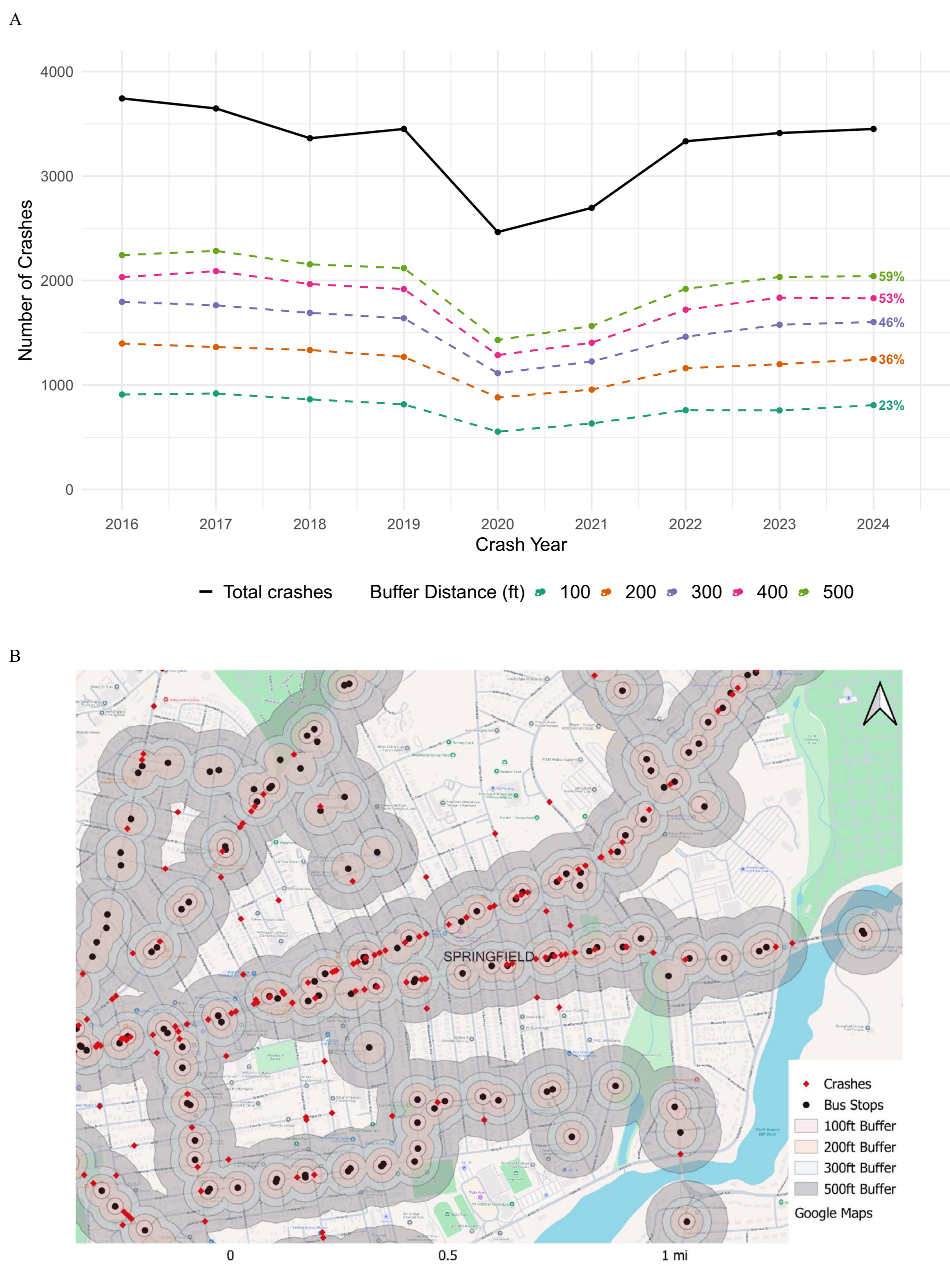


Fig. 1 Crashes Near Bus Stops in Massachusetts: **A** Crashes Near Bus Stops by Year. The solid line shows total crashes, and the dashed lines show crashes within 100–500 ft of a bus stop. Percentages show the proportion of total crashes within each buffer distance in 2024. **B** Crashes Near Bus Stops with 100–500 ft buffers

role of buses and bus stops in crash outcomes, remains unclear.

Traditional safety analyses often group bus stops by administrative jurisdiction or geographic location, treating them as independent entities. This approach overlooks the critical operational and environmental characteristics, such as roadway geometry, traffic volume, pedestrian infrastructure, and adjacent land use that are more directly linked to crash propensity and severity than administrative boundaries. Grouping stops by risk-related characteristics, rather than jurisdictional boundaries, enables transportation planners to prioritize high-risk locations more effectively. For example, stops in high-speed corridors with poor lighting demand different safety measures than those in dense commercial zones with heavy pedestrian activity. Therefore, a critical gap exists in identifying a framework for categorizing bus stops based on the characteristics that directly influence their safety performance. Closing this gap is essential for understanding high-risk bus stop profiles and developing effective mitigation strategies to enhance the safety of vulnerable road users.

This study addresses the gap by introducing a cluster-based approach to bus stop safety analysis. The objective is to group stops based on the physical, operational, and transit features that influence crash occurrence and severity. By identifying the underlying latent factors and clusters of stops with similar risk profiles, the analysis can uncover underlying patterns to improve the accuracy of crash prediction. Further, agencies can replace one-size-fits-all solutions with targeted countermeasures, that maximize the impact of safety investments and promote equitable improvements across agency boundaries.

This is the first study that contributes to the growing body of transportation safety literature by: (1) developing a machine learning (ML) methodology for bus stop clustering based on diverse latent factors previously unexplored in combination, (2) predicting crashes by cluster, and (3) providing cluster-informed safety interventions for improving pedestrian safety across different stop typologies. To our knowledge, no prior data-driven typology of bus stops relating to crashes exists.

Further, the application of machine learning (ML) models in pedestrian crash analysis introduces significant novelty compared to traditional statistical approaches by enabling the discovery of complex, nonlinear, and interaction effects among diverse risk characteristics that were previously difficult to capture. Their ability to flexibly integrate diverse data improves predictive accuracy, while interpretable machine learning (IML) techniques, translate these complex models into interpretable and actionable insights for safety planning.

Literature Review

Bus Stop Safety and Crash Risk Characteristics

Bus stops attract high volumes of pedestrian activity, which can significantly increase pedestrian crash risk. Roads with transit service experienced up to six times more pedestrian crashes, linked to two to three times higher pedestrian volumes (Pulugurtha and Penkey 2010). Other research has reinforced this spatial relationship (Hess et al. 2004; Quistberg et al. 2015; Ulak et al. 2021; Zegeer et al. 2010), with one study quantifying a 2.9% rise in pedestrian collisions for every additional bus stop in lower-income Florida communities (Dumbaugh et al. 2024). Although a high volume of pedestrian crashes occurs near bus stops, one study indicates that only about 5% involve bus passengers (Pessaro et al. 2017), with the vast majority involving non-transit users, suggesting that the primary risk is proximity to the transit infrastructure, rather than the population of transit users.

Research has extensively examined how roadway characteristics, the built environment, and bus stop configuration influence pedestrian safety. A key finding is that high-risk locations are consistently associated with characteristics like retail presence, high traffic volumes (AADT), and a greater number of traffic lanes (Hess et al. 2004; Anis et al. 2025). The configuration of bus stops themselves is critical with curbside stops being linked to higher injury rates than layby (or pull-out) stops, often due to their proximity to intersections (Phillips et al. 2021). Furthermore, pedestrians near bus stops face elevated risks from high traffic speeds, multiple lanes, roadside hazards, and a lack of amenities like sidewalks, crosswalks, and lighting (Yendra et al. 2024; Anis et al. 2025). The surrounding land use also plays a significant role, with increased crash frequencies near supermarkets, schools, hospitals, and religious facilities (Ulak et al. 2021). Interestingly, crash severity can vary by location, with mid-block crashes often being more severe than those at intersections due to a sparsity of crossing treatments (Truong and Somenahalli 2011; Phillips et al. 2021; Rewalt et al. 2025). However, the existing research has not identified which specific types of bus stops are associated with higher crash risk or the specific combinations of bus stop features that lead to adverse safety outcomes.

Machine Learning in Bus Stop Classification and Crash Prediction

Recent advances in data availability and computational power have enabled widespread adoption of machine learning methodologies in crash analysis (Iranitalab and Khattak 2017; Agarwal 2011; Silva et al. 2020; Wen et al. 2021; Santos et al. 2022; Chakraborty et al. 2023; Formosa et al. 2023; Ul Arifeen et al. 2023; Ali et al. 2024; Budzyński et al. 2024; Rafe et al. 2024; Zermane et al. 2024; Butt and Shafique 2025; Chen et al. 2025; Xue et al. 2025). Dimensionality reduction techniques such as principal component analysis (PCA), explanatory factor analysis (EFA), and uniform manifold approximation and projection (UMAP) help manage large datasets to extract salient features for analysis (Karlaftis and Golias 2002; Ye et al. 2016; Xu et al. 2019; Furno et al. 2024; Jin et al. 2024). Clustering techniques allow for a proactive grouping of locations by latent safety characteristics, successfully identifying high-risk zones based on environmental and design features (Ye et al. 2016; Rahman et al. 2019; Silva et al. 2020; Abdalazeem and Oke 2025; Brustad et al. 2025; Han and Oke 2025). For predictive tasks, ensemble methods like random forest and extreme gradient boosting (XGBoost) models have demonstrated superior performance, accurately forecasting injury severity and identifying key contributing characteristics better than classical logistic or regression models (Agrawal et al. 2019; Santos et al. 2022; Ali et al. 2024; Ahmed and Gayah 2025; Abdalazeem and Oke 2025; Maiti and Leclercq 2025; Prajapati and Guler 2025). Neural networks also contribute by modeling complex spatiotemporal patterns, with studies showing that driver and vehicle characteristics can outweigh roadway variables in determining crash severity (Bao et al. 2019; Zhang et al. 2020). While these methodological frameworks are highly relevant, the application of machine learning has been focused on vehicle crash severity, with just few studies dedicated to vehicle-pedestrian collisions (Ul Arifeen et al. 2023; Weng et al. 2023; Wang et al. 2024). This leaves the specific context of pedestrian safety analysis at bus stops largely unexamined.

Research Gap and Contributions

Despite extensive research on pedestrian crash characteristics, significant knowledge gaps remain regarding transit bus stop safety. While some studies have demonstrated correlations between bus stops and pedestrian crashes, only limited research has specifically examined crashes occurring near transit bus stops and the specific factors contributing to elevated crash rates in these transit environments (Hess et al. 2004; Ulak et al. 2021; Dumbaugh et al. 2024; Roll and McNeil 2021; Craig et al. 2019; Lin et al. 2018; Miranda-Moreno et al. 2011). In addition, there is limited understanding of the optimal integration of infrastructure improvements and mitigation strategies for pedestrian crashes in transit environments (Hess et al. 2004; Ulak et al. 2021).

This study addresses these critical gaps through a comprehensive analysis of bus stop safety in Massachusetts. We examine the relationship between roadway and transit bus stop infrastructure, bus operations, and nearby pedestrian crashes to provide empirical evidence for safer roadway and transit infrastructure. Our research pursues three primary objectives: (1) to develop a methodology for bus stop clustering by identifying latent risk characteristics that contribute to pedestrian safety incidents, (2) to apply advanced machine learning methods to predict pedestrian crash risk at bus stop by types, and (3) to develop type-specific safety interventions for improving pedestrian safety. Our methodological approach integrates machine learning (ML) models, including Factor Analysis, Gaussian Mixture Models, and XGBoost prediction to create a robust analytical framework capable of handling complex transit safety data. This framework is applied to a comprehensive dataset that fuses multiple sources, including GTFS data, crash records, demographics, road inventory, and vehicle census, encompassing over 300 initial variables that capture diverse potential characteristics. Importantly, our analysis directly translates into actionable cluster-type-specific intervention, bridging the gap between analytical findings and practical implementation for transportation planners.

Methodology

We employ a multi-dimensional analytical framework that extracts patterns of significance from pedestrian crashes near transit bus stop, using diverse predictors, including transit schedules, traffic volumes, roadway features, built environment characteristics, and demographic and socioeconomic context. The methodological framework shown in Fig. 2 integrates machine learning (ML) models, including explanatory factor analysis (EFA) for uncovering latent factors, Gaussian mixture models (GMM) for bus stop clustering, extreme gradient boosting (XGBoost) models for crash prediction, and SHapley Additive exPlanations (SHAP) plots for model interpretation, to create a robust analytical framework capable of handling complex transit safety data. These ML models advance pedestrian safety analysis by uncovering complex, non-linear characteristics and hidden spatiotemporal patterns, difficult to capture with traditional models. By flexibly integrating diverse data, these models achieve high predictive accuracy. More importantly, interpretable ML (IML) techniques like SHAP make these predictions accessible, providing planners with clear, actionable insights.

Fig. 2 Research framework

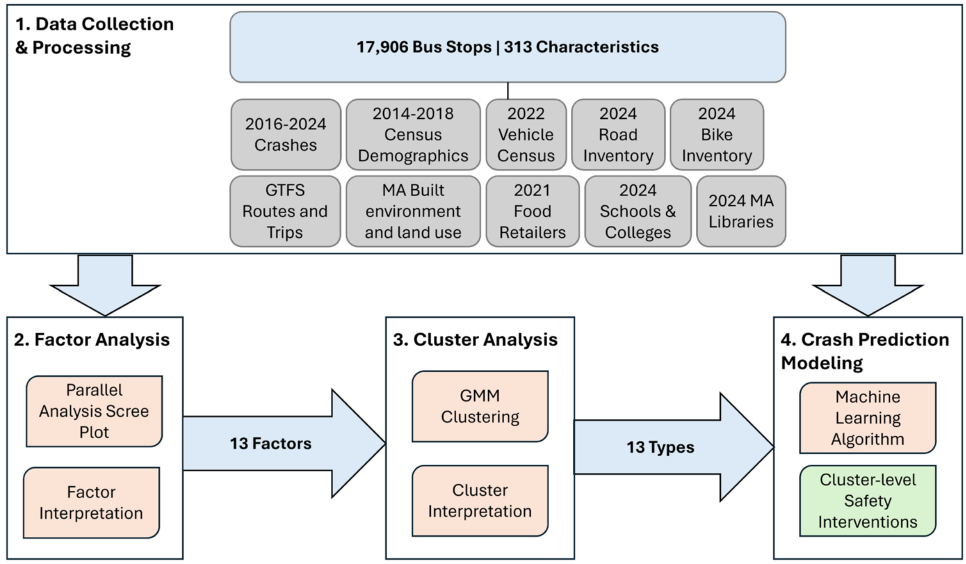


Table 1 Massachusetts is served by the Massachusetts Bay Transportation Authority (MBTA) and 15 distinct Regional Transit Authorities (RTAs)

RTA	Name	Source
BRTA	Berkshire Regional Transit Authority	Berkshire Regional Transit Authority (2025)
BAT	Brockton Area Transit Authority	BAT (2025)
CATA	Cape Ann Transportation Authority	Cape (2025)
CCRTA	Cape Cod Regional Transit Authority	Home (2025)
FRTA	Franklin Regional Transit Authority	FRTA (2025)
GATRA	Greater Attleboro Taunton Regional Transit Authority	GATRA (2023)
LRTA	Lowell Regional Transit Authority	Lowell Regional Transit Authority (2025)
MART	Montachusett Regional Transit Authority	MART (2025)
MBTA	Massachusetts Bay Transportation Authority	mbta.us (2025)
MVRTA	Merrimack Valley Regional Transit Authority	MeVa (2025)
MWRTA	MetroWest Regional Transit Authority	MWRTA (2025)
NRTA	Nantucket Regional Transit Authority	The Wave (2025)
PVTA	Pioneer Valley Transit Authority	PVTA (2024)
SRTA	Southeastern Regional Transit Authority	SRTA (2025)
VTA	Vineyard Transit Authority	Vineyard Transit Authority (2025)
WRTA	Worcester Regional Transit Authority	WRTA (2025)

Data Collection

Data from multiple sources were integrated to create a comprehensive database of bus stop characteristics. The Massachusetts Bay Transportation Authority (MBTA) and fifteen Regional Transit Authorities (RTAs) provide fixed-route bus service to 17,906 stops across urban, suburban, and rural areas (Table 1, Fig. 3). By leveraging the varied transit landscape from the 17,906 bus stops in Massachusetts, from high-frequency urban corridors to lower-density suburban and rural routes, the research findings are expected to be broadly applicable in other operational and geographic environments.

Pedestrian and bicyclist crash data for 2016–2024 were sourced from the MassDOT Impact Crash Data Portal (Impact 2024). The bus stop locations were then merged with the crash locations to identify the crash frequency and severity occurring near the bus stops. While bus stops themselves are rarely the site of collisions, a significant share of crashes occur nearby, reflecting the pedestrian activity around typical bus stop environments. To determine a meaningful crash catchment area for bus stops, we analyzed crashes within increasing distances from bus stops. As

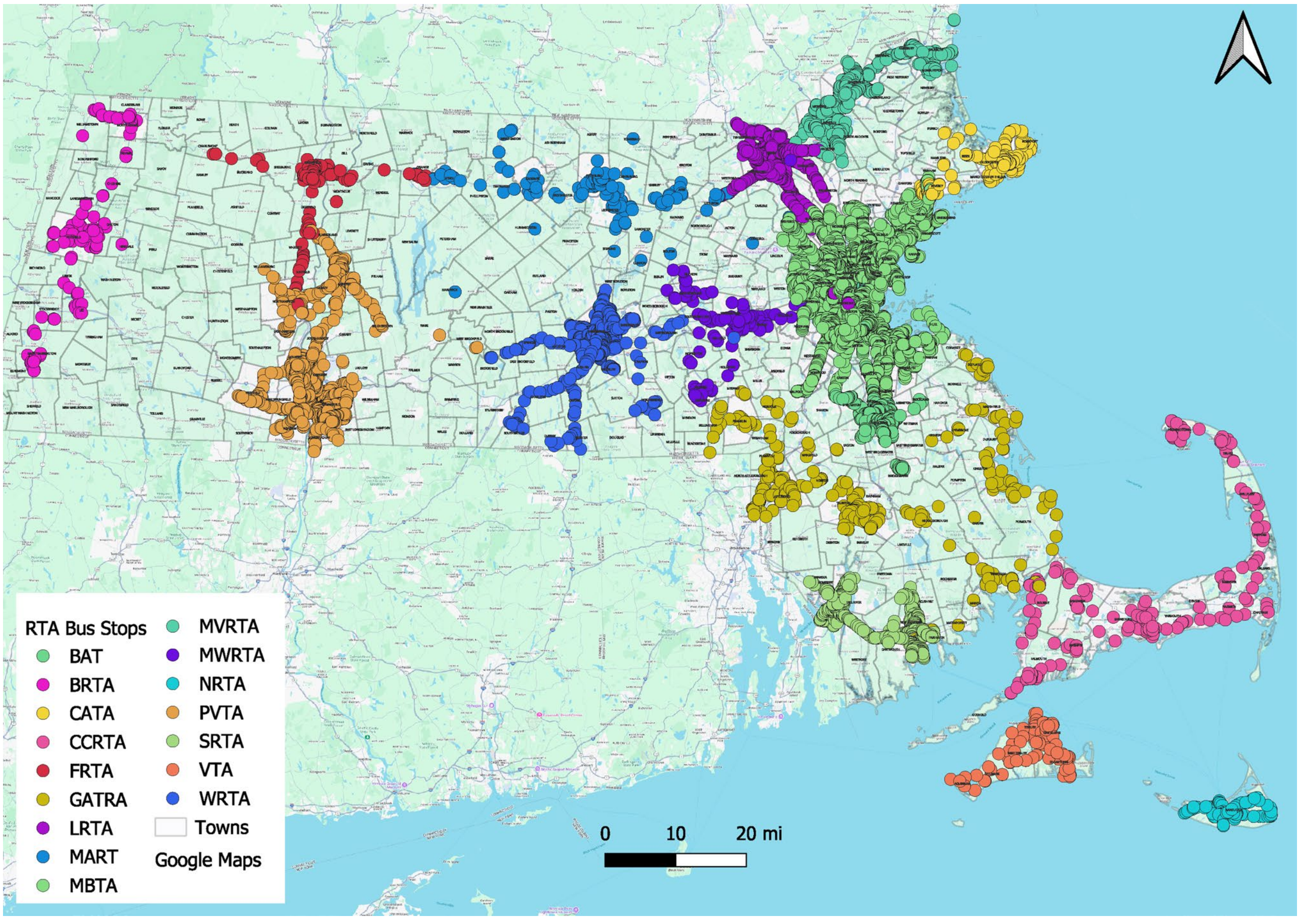


Fig. 3 Location of Massachusetts bus stops by regional transit agency

expected, crashes within buffers increase with buffer size. For example, with a 20-foot buffer, only 3% of stops have a nearby crash, covering just 2% of all crashes. A 100-foot radius captures 23% of all crashes affecting 31.4% of stops, expanding to 300 feet captures just over half of all bus stops and 42% of total crashes, and 500 feet captures 67% of stops and 60% of crashes. Since transit user activity, such as walking, crossing, and waiting, typically occurs within a few hundred feet, a buffer of 300 feet is optimal to effectively capture crashes linked to transit while excluding unrelated incidents. Figure 4 shows the location of the crashes in Massachusetts and the crashes aggregated by Census block groups. To calculate the average annual crash rates for each bus stop, the total number of pedestrian crashes within a 300-foot buffer was divided by the 8-year study period.

Additional features were incorporated based on crash literature: roadway characteristics from the 2022 MassDOT Road Inventory (MassGIS Data 2025a); demographics and socioeconomics from the American Community Survey (U.S. Census Bureau 2025); vehicle data from the Massachusetts Vehicle Census (Massachusetts Vehicle Census 2025); bicycle facilities from the 2024 Bicycle Facility Inventory (Bike Inventory 2025); and proximity to and density of schools (Massachusetts Public School Districts 2025; MassGIS Data 2025b), libraries (MassGIS Data 2025c), and food retailers (Grocery Store Locations in Massachusetts 2025) near bus stops.

Data cleaning involved validating variables against GoogleMaps, selecting the most accurate sources, and substituting COVID-era (2020–2021) traffic counts with data from other years. Variables with zero variance or high correlation were excluded. Additionally, duplicate stops and those with extensive missing data were removed. Numerical variables were standardized (mean-centered and scaled by standard deviation), and categorical variables were converted to binary dummy variables. The final dataset contained 17,736 bus stops with 313 features. Figure 5 shows histograms for key bus stop characteristics.

Factor Analysis

We conducted an exploratory factor analysis (Thurstone 1931; Fabrigar and Wegener 2012) to reduce the dimensionality and extract low-dimensional latent components from

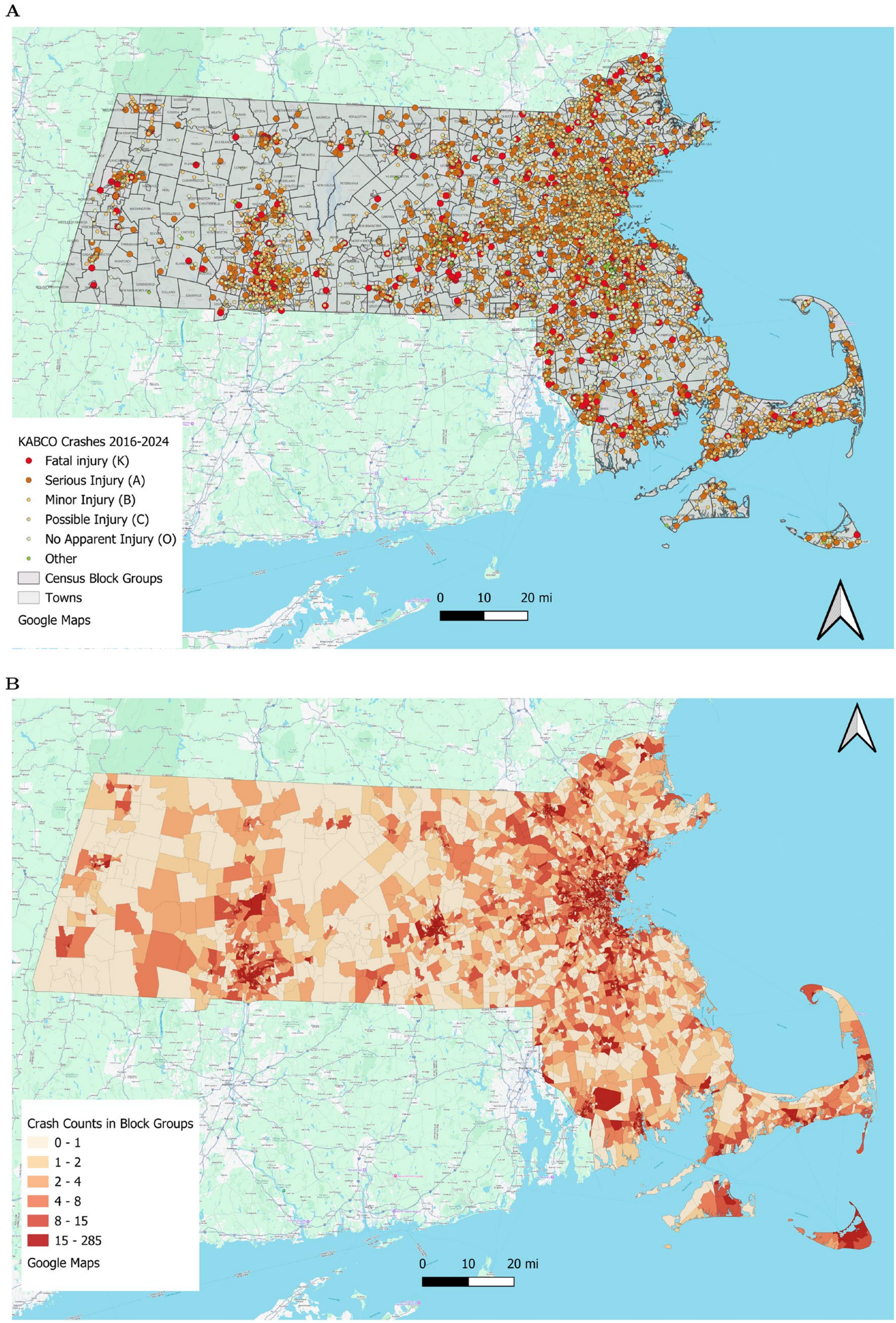


Fig. 4 2016–2024 crashes: A Location of crashes. B Crashes aggregated by Census block group

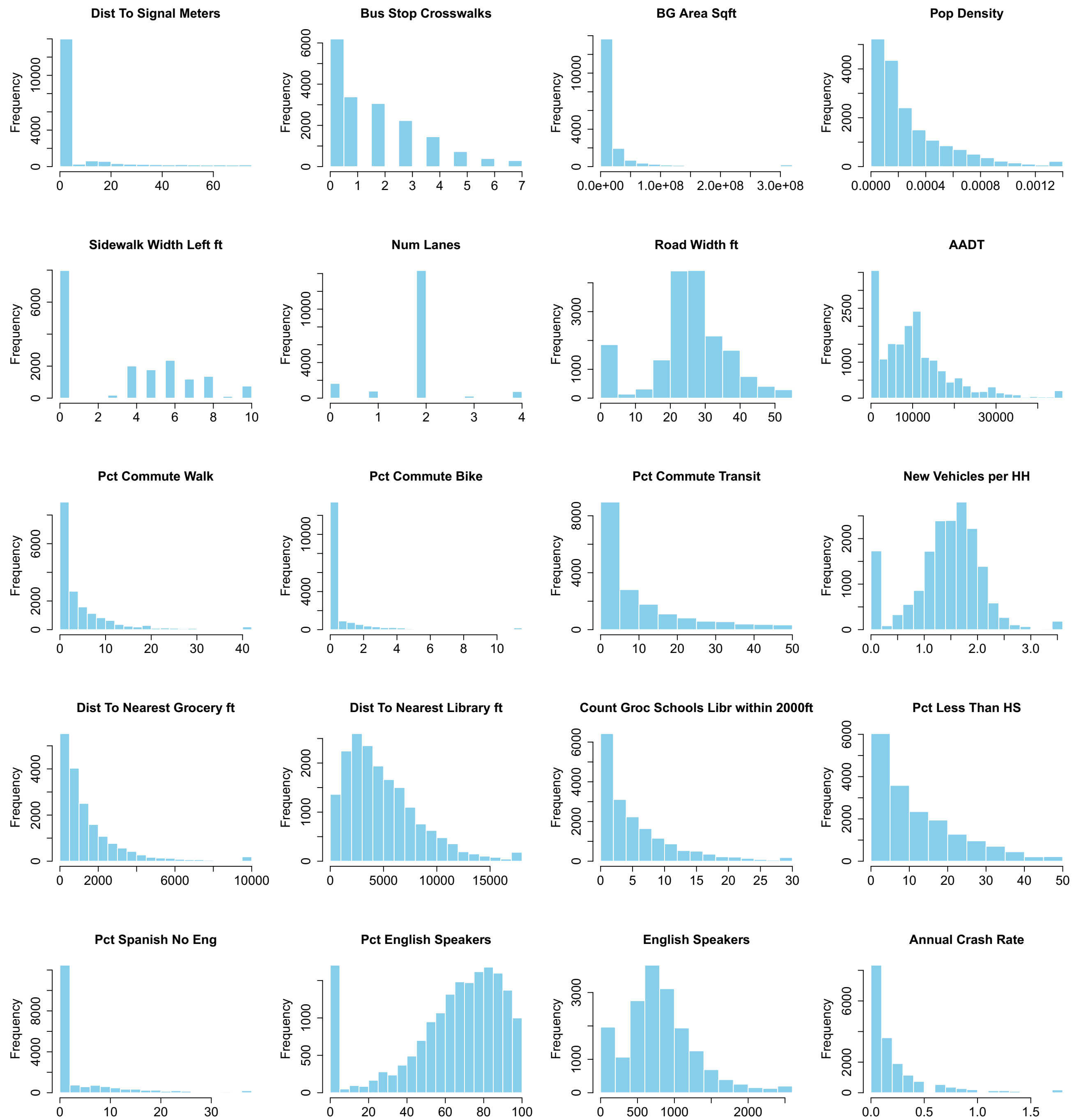


Fig. 5 Histograms of some key bus stop characteristics

the dataset. Factor analysis transforms the original variables into a set of uncorrelated factors that capture the maximum variance in the dataset, effectively reducing noise and multicollinearity while maintaining interpretability.

The factor analysis model expresses observed variables as a linear combination of underlying latent factors plus variance for each variable:

$$\boldsymbol{X} = \boldsymbol{L}\boldsymbol{F} + \epsilon \tag{1}$$

where $\boldsymbol{X}$ is an $N \times D$ matrix of N bus stops (17,736) on D observed variables (313), $\boldsymbol{L}$ is the $D \times L$ factor loading matrix, showing how strongly each observed variable loads on each factor, $\boldsymbol{F}$ is an $L \times N$ matrix of the L latent factors (with $L < D$), and ϵ represents the variances.

To improve interpretability, the loading matrix is rotated by multiplying by an orthogonal or oblique rotation matrix $\boldsymbol{T}$ (where $\boldsymbol{T}^{\intercal}\boldsymbol{T} = \boldsymbol{I}$). The rotated model becomes:

$$\boldsymbol{X} = (\boldsymbol{LT})(\boldsymbol{T}^{-1}\boldsymbol{F}) + \epsilon = \boldsymbol{L}^{*}\boldsymbol{F}^{*} + \epsilon \tag{2}$$

The rotated components $\boldsymbol{L}^{*}$ and $\boldsymbol{F}^{*}$ provide clearer factor structure without changing the fundamental data relationship. We selected oblique rotation (oblimin) for its superior balance of low error and interpretability. Since variables are naturally interrelated, e.g., areas with high transit access often have more traffic and denser development, oblique rotation better captures these correlations (Kieffer 1998).

The selection of the optimal number of factors was guided by a combination of heuristic methods and statistical fitness metrics. The Kaiser criterion (eigenvalues > 1.0) (Kaiser 1960) and the scree plot's elbow point (Cattell 1966) both suggested a 13-factor solution (Fig. 6). To validate this, we computed a range of model fitness metrics for 1 to 20 factor solutions (Table 2). The results showed that beyond thirteen factors, the improvement in model fit diminished. Specifically, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) trends flattened, and the Root Mean Square Error of Approximation (RMSEA) and Tucker–Lewis Index (TLI) showed minimal further improvement. Consequently, the 13-factor solution was selected, as it provides the optimal balance of statistical fitness, parsimony, and interpretability. This represents a substantial simplification of the original 313 bus stop characteristics into 13 core latent factors, representing the most important underlying patterns.

Cluster Analysis

Cluster analysis is an unsupervised learning approach that groups observations based on similarity measures to identify natural groupings within complex datasets (Jain et al. 1999; Fraley and Raftery 2002; Everitt et al. 2011). In the context of bus stop safety analysis, clustering can reveal underlying patterns that enable the identification of stop types that share similar geographic and operational characteristics and possibly similar risk profiles.

The thirteen extracted factors were used as input variables for the cluster analysis, ensuring that the clustering process focused on the most informative dimensions of bus stop variation. We used the Gaussian Mixture Model (GMM) (Dempster et al. 1977), which is a probabilistic clustering technique that assumes data points are generated from a mixture of several Gaussian distributions with unknown parameters. Unlike K-Means, which assigns each point to a single cluster, GMM assigns probabilities to each data point for belonging to different clusters. The probability of a data point $\boldsymbol{x}$ is:

$$\hat{p}(\boldsymbol{x}) = \sum_{k=1}^{K} \boldsymbol{\pi}_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3}$$

where $\boldsymbol{K}$ is the number of clusters, and $\boldsymbol{\pi}_k$ is the mixing weight of the k-th Gaussian component (with the probability of selecting that cluster, $\sum \boldsymbol{\pi}_k = \mathbf{1}$). GMM assumes that the data is generated from a weighted sum of $\boldsymbol{K}$ Gaussian distributions, $\mathcal{N}(z \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, each specified by its mean ($\boldsymbol{\mu}_k$) and covariance matrix ($\boldsymbol{\Sigma}_k$). The optimal parameters are obtained using the Expectation–Maximization (EM) algorithm (Dempster et al. 1977).

We used a grid search to evaluate various numbers of clusters and covariance specifications (spherical, tied, diagonal, and full). We selected the optimal number of clusters

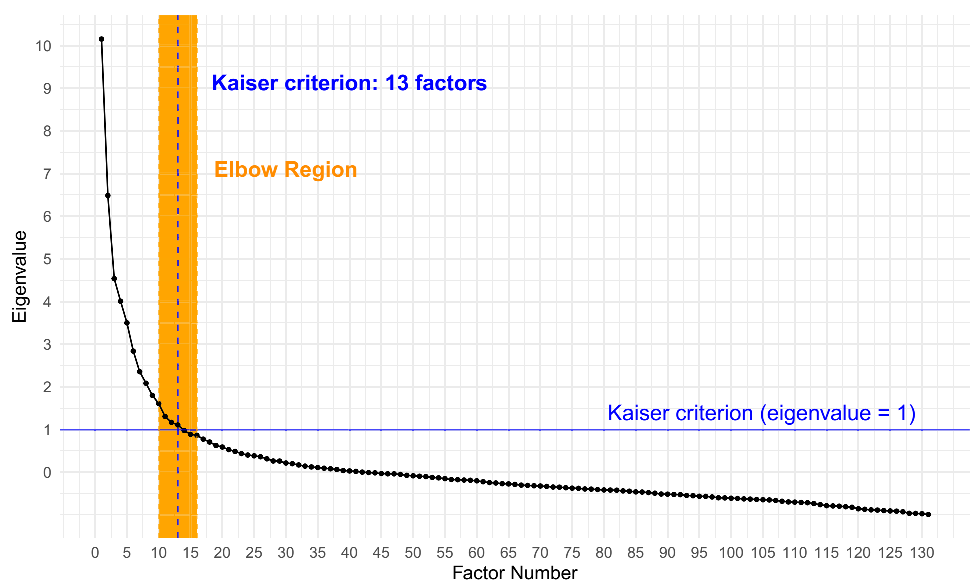


Fig. 6 Scree plot of eigenvalues of factors arranged in descending order of magnitude. The horizontal solid blue line represents the Kaiser criterion at 13 factors and the orange shaded area highlights the elbow region

Table 2 Factor analysis model fitness metrics (1–20 Factors)

Factors	χ^2 ($\times 10^5$)	df	p-value	BIC ($\times 10^5$)	AIC ($\times 10^5$)	TLI	RMSEA	RMSR
1	13.80	8384	0.00	12.98	13.97	0.089	0.096	0.095
2	12.74	8254	0.00	11.94	12.91	0.146	0.093	0.081
3	11.91	8125	0.00	11.11	12.07	0.189	0.091	0.074
4	10.97	7997	0.00	10.18	11.13	0.242	0.087	0.066
5	10.27	7870	0.00	9.50	10.42	0.279	0.085	0.059
6	9.66	7744	0.00	8.90	9.81	0.311	0.083	0.054
7	9.45	7619	0.00	8.70	9.60	0.315	0.083	0.053
8	8.96	7495	0.00	8.22	9.11	0.340	0.082	0.049
9	8.55	7372	0.00	7.82	8.69	0.360	0.081	0.046
10	8.26	7250	0.00	7.55	8.40	0.371	0.080	0.046
11	7.73	7129	0.00	7.03	7.87	0.402	0.078	0.043
12	7.33	7009	0.00	6.64	7.47	0.423	0.076	0.040
13	**7.07**	**6890**	**0.00**	**6.40**	**7.21**	**0.434**	**0.076**	**0.040**
14	6.96	6772	0.00	6.29	7.09	0.433	0.076	0.039
15	6.68	6655	0.00	6.02	6.81	0.447	0.075	0.039
16	6.24	6539	0.00	5.60	6.37	0.474	0.073	0.036
17	5.94	6424	0.00	5.31	6.07	0.490	0.072	0.034
18	5.83	6310	0.00	5.21	5.95	0.491	0.072	0.033
19	5.54	6197	0.00	4.93	5.66	0.507	0.070	0.034
20	5.36	6085	0.00	4.76	5.48	0.515	0.070	0.033

χ^2 Chi-square test statistic, *df* degrees of freedom, *AIC* Akaike Information Criterion, *BIC* Bayesian Information Criterion, *TLI* Tucker–Lewis Index, *RMSEA* Root Mean Square Error of Approximation, *RMSR* Root Mean Square Residual. Bolded row indicates the selected 13-factor solution. χ^2, BIC, and AIC values are multiplied by 10^5 (shown in header) for readability

K by identifying the model with the highest log-likelihood score and Bayesian information criterion (BIC), and the lowest Akaike information criterion (AIC). The covariance specification names are three-letter codes that describe constraints on the covariance structure, specifically regarding the variances (first letter), covariances (second letter), and the orientation/shape (third letter) of the covariance matrices. The letter E (Equal) means the parameter is constrained to be the same across all groups, V (Variable), the parameter is allowed to vary across groups, and I (Identity), the parameter is constrained to form an identity matrix. The average silhouette scores were also computed to further inform the optimal cluster selection. As shown in Fig. 7, the 13-cluster, unconstrained model (*VVV*) was identified as optimal. This model, which allows variances, covariances, and orientations to vary freely across clusters, revealed distinct profiles that yielded the best fit for the data. We further computed the silhouette metric to assess cluster quality, which reinforced the choice of 13 clusters (Fig. 8).

Predictive Crash Modeling

Machine learning algorithms such as Random Forests (Breiman 2001) and Extreme Gradient Boosting (XGBoost) (Chen and Guestrin 2016) are particularly well-suited for crash prediction modeling because they are capable of handling complex, nonlinear relationships within data, while remaining relatively robust to outliers. By constructing multiple decision trees, the ensemble nature of these models creates a robust predictor. However, their approaches differ; while Random Forest builds trees in parallel and aggregates their predictions, XGBoost builds trees sequentially, with each new tree correcting the errors of the previous ones. This ensemble approach reduces the risk of overfitting inherent in single decision trees and provides a foundation for quantifying prediction uncertainty.

To determine the most relevant risk characteristics and enable type-specific safety interventions, we developed XGBoost models to predict crash rates for each of the 13 bus stop types from the 313 bus stop characteristics observed. These characteristics included bus stop design (e.g., configuration, shelter, lighting), transit operational characteristics (e.g., bus frequency, number of routes), sociodemographic variables (e.g., age, gender, income and education distributions), road design and infrastructure (e.g., crossing widths, speed limits, crosswalks), and traffic and land use characteristics (e.g., traffic volumes, vehicle types, proximity to major destinations, and the percentage of commuters that walk, bike, or take public transit to work). First, we partitioned the entire dataset (17,736 bus stops) into global training and test

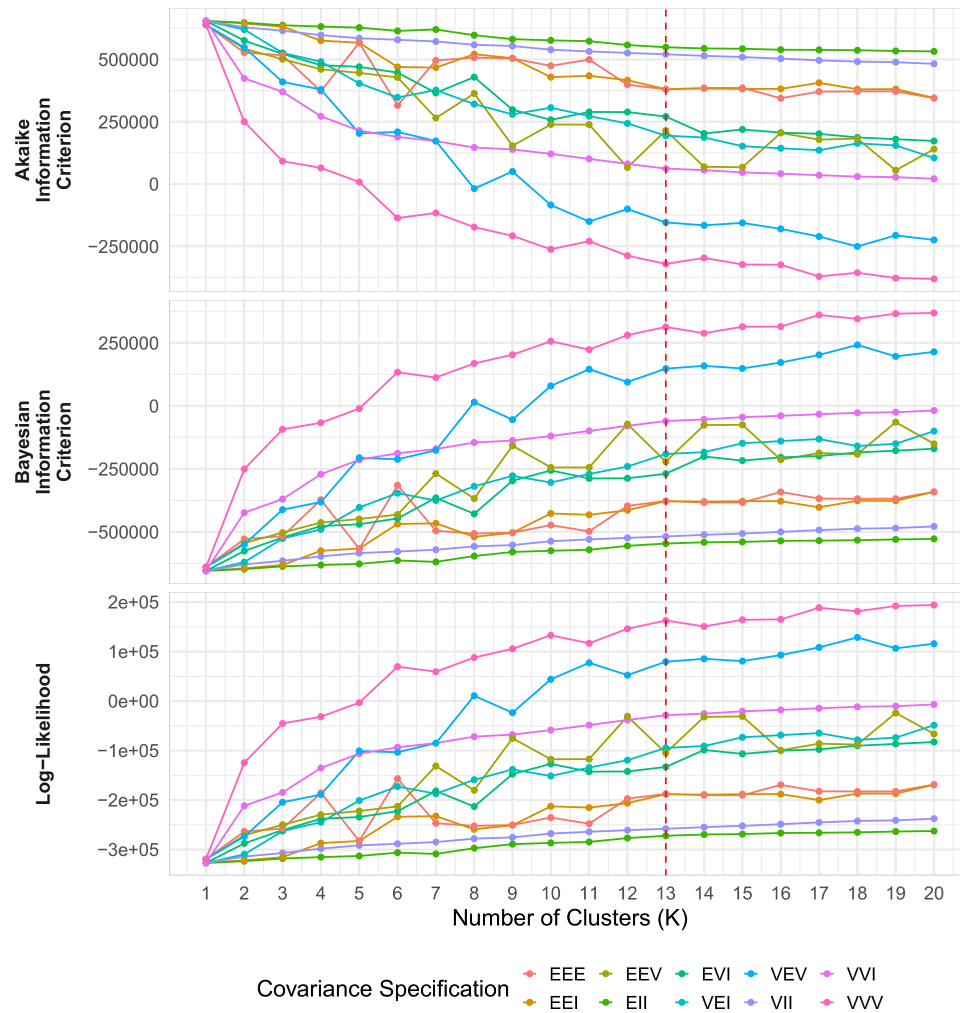


Fig. 7 Plots of AIC, BIC, and log-likelihood for various cluster numbers. Optimal number of clusters = 13 with Covariance Specification = VVV

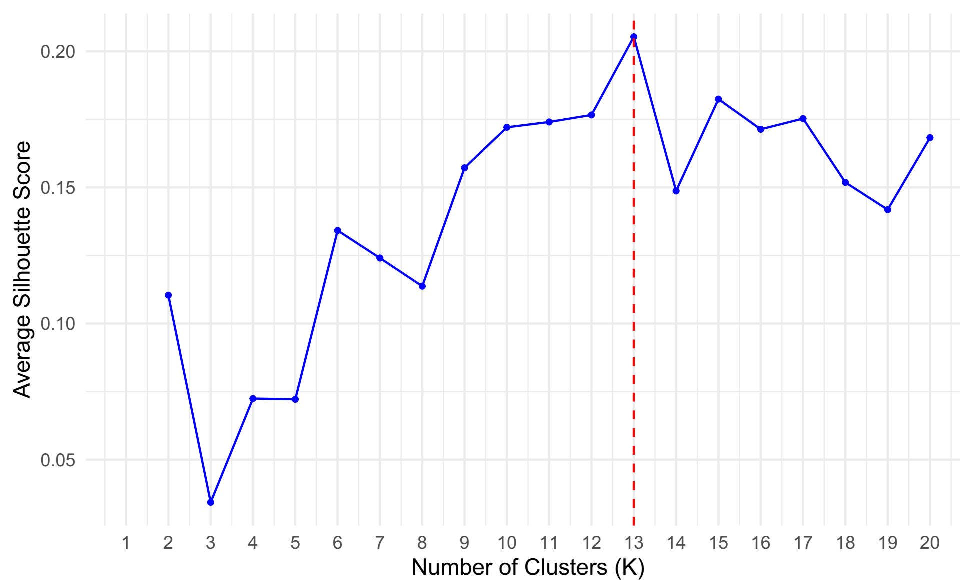


Fig. 8 Silhouette score with optimal cluster number = 13

Table 3 Grid search space for cluster-specific crash prediction models

Parameter	Description	Values
α	L_1 regularization term	$\{0, 0.01\}$
λ	L_2 regularization term	$\{1, 5\}$
ρ	Learning rate	$\{0.01, 0.1, 0.3\}$
$d_{\max}$	Maximum depth	[3, 10]
n_{est}	Number of estimators	$\{100, 200, 300, 400, 500\}$

partitions (using a 70:30 ratio). Then to fit each type-specific model, we used observations from the training and test sets, respectively, that belonged to each type. Hyperparameters for each model were optimized via a grid search with 3-fold cross-validation. The search space is shown in Table 3, while the optimal hyperparameters for each model are given in Table 4. In training all models, we used the mean squared logarithmic error (MSLE) as the objective, given its better performance with small positive values Hodson et al. (2021).

We trained the final type-specific models using the corresponding optimal hyperparameters. We evaluated model performance via R^2, mean absolute error (MAE) and mean squared logarithmic error (MSLE). Table 5 shows the predictive model performance summary across the 13 clusters. Cluster 4 achieved the strongest performance with an R^2_{test} of 0.73 using 285 samples, while Cluster 12 showed the weakest explanatory power ($R^2_{\text{test}} = 0.21$) despite having a comparable sample size of 199. The cluster sample sizes ranged from about 100 to 500 observations. The models generally showed moderate to good predictive capability, with 11 of the 13 clusters achieving R^2_{test} values above 0.50, suggesting that the clustering strategy effectively identified homogeneous subgroups with distinct patterns.

Table 4 Optimal hyperparameters for each cluster model, and corresponding cross-validation scores (per-fold, average and standard deviation) based on the mean squared logarithmic error (MSLE) objective

Cluster	α	λ	ρ	$d_{\max}$	n_{est}	Mean squared logarithmic error (MSLE)				
						Fold 1	Fold 2	Fold 3	Mean	SD
1	0.01	5	0.1	6	500	0.02137	0.02023	0.02538	0.02233	0.00221
2	0	1	0.1	5	200	0.01442	0.01237	0.01139	0.01273	0.00126
3	0	5	0.1	6	300	0.01978	0.02139	0.01974	0.02030	0.00077
4	0	5	0.3	4	200	0.03477	0.02533	0.02511	0.02840	0.00450
5	0	5	0.01	6	300	0.02819	0.03693	0.04211	0.03574	0.00574
6	0.01	5	0.1	5	500	0.02159	0.02950	0.02964	0.02691	0.00376
7	0	1	0.1	3	500	0.01691	0.02035	0.01621	0.01782	0.00181
8	0	5	0.1	3	300	0.02840	0.01895	0.02288	0.02341	0.00388
9	0	5	0.01	6	400	0.03487	0.04156	0.02561	0.03401	0.00654
10	0	5	0.1	5	500	0.02054	0.02412	0.02041	0.02169	0.00172
11	0	5	0.3	4	100	0.01621	0.03497	0.02246	0.02455	0.00780
12	0.01	1	0.1	4	200	0.02922	0.02410	0.02617	0.02650	0.00210
13	0	5	0.1	3	500	0.03810	0.03739	0.03527	0.03692	0.00120

Table 5 Performance of crash prediction models by cluster, showing R^2, mean absolute error (MAE) and mean squared logarithmic error (MSLE) for training and test sets

Cluster	Training				Test			
	Samples	R^2	MAE	MSLE	Samples	R^2	MAE	MSLE
1	1680	0.9997	0.0033	0.0000	530	0.524	0.123	0.021
2	853	0.9972	0.0109	0.0002	226	0.583	0.084	0.013
3	1105	0.9980	0.0079	0.0001	393	0.579	0.098	0.013
4	905	0.9994	0.0075	0.0001	285	0.733	0.126	0.019
5	762	0.8637	0.0926	0.0079	266	0.509	0.150	0.026
6	2000	0.9965	0.0147	0.0003	795	0.541	0.146	0.026
7	1273	0.9741	0.0284	0.0011	430	0.531	0.101	0.016
8	869	0.9706	0.0433	0.0022	309	0.560	0.118	0.020
9	337	0.9280	0.0477	0.0022	117	0.522	0.215	0.046
10	1632	0.9974	0.0113	0.0002	531	0.548	0.136	0.023
11	437	0.9990	0.0059	0.0001	149	0.304	0.135	0.022
12	638	0.9877	0.0291	0.0011	199	0.211	0.206	0.051
13	613	0.9930	0.0306	0.0009	204	0.636	0.180	0.032

We then used SHapley Additive exPlanations (SHAP values) (Lundberg and Lee 2017) to interpret the relevance and impact of the bus stop characteristics on crash rates in each type. The SHAP value, ϕ_d^l, indicates the magnitude and direction of each independent feature, $\boldsymbol{x}_d$ on crash outcomes, and is given by:

$$\hat{f} = \phi_0 + \sum_{d=1}^{M} \phi_d \tag{4}$$

where $$\phi_d = \sum_{\mathcal{S} \subseteq \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_D\} \setminus \{\boldsymbol{x}_d\}} \frac{|\mathcal{S}|!\,(|D| - |\mathcal{S}| - 1)!}{|D|!} \left[\hat{f}_{\mathcal{S} \cup \{d\}}(\boldsymbol{x}_{\mathcal{S} \cup \{d\}}) - \hat{f}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}) \right] \tag{5}$$

where ϕ_d represents the marginal contribution of feature $\boldsymbol{x}_d$ across all possible subsets $\mathcal{S}$ of input features. The magnitude of ϕ_d represents the strength or importance of the feature's contribution to the prediction, while the sign indicates the direction of the effect, positive SHAP values increase the predicted outcome, and negative values decrease it.

Results

Underlying Bus Stop Factors

The factor analysis reduced the dimensionality of the dataset from 313 observed bus stop characteristics to 13 latent factors, selected based on scree plot inspection and the Kaiser criterion eigenvalue threshold. Factor loadings quantify the magnitude and direction of each characteristics' contribution to a factor. Figure 9 shows the defining characteristics for each of the 13 factors, with the highest absolute loadings, along with an interpretive factor label derived from their common attributes.

Bus stops can be comprehensively characterized by assessing three key dimensions, the immediate physical environment, the broader transportation network integration and connectivity, and the surrounding sociodemographic

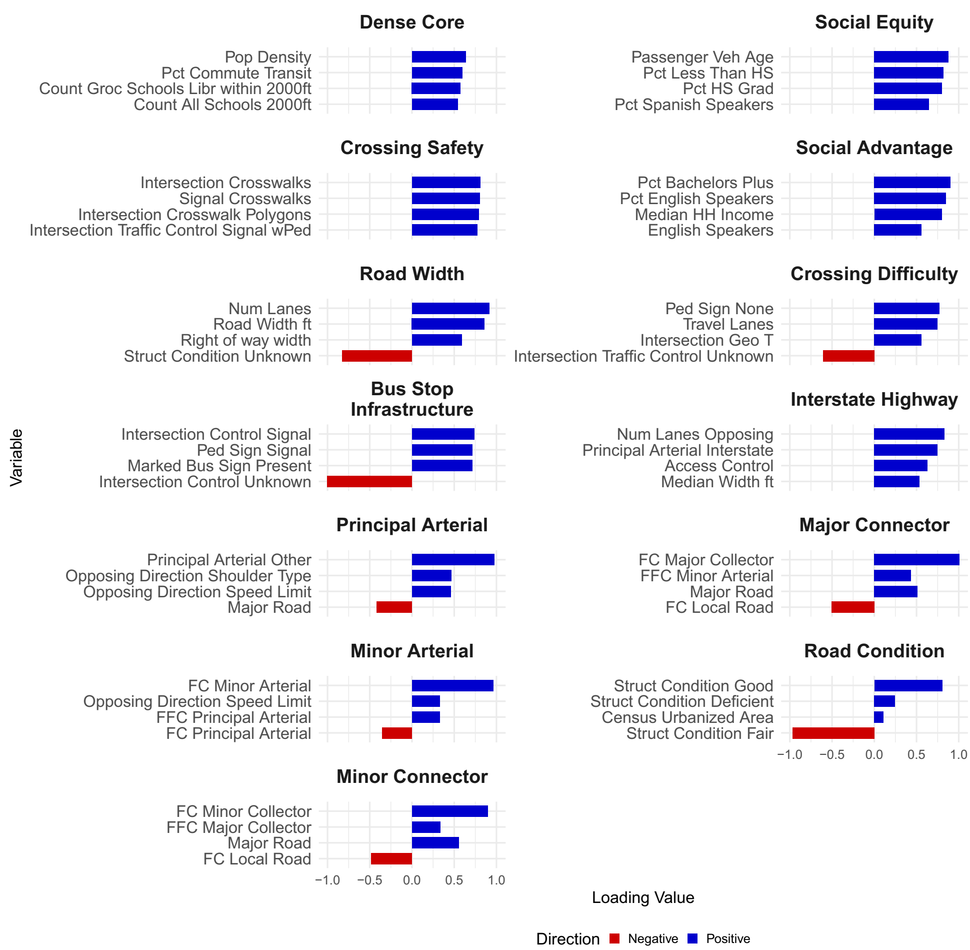


Fig. 9 Top variables per factor by loading strength. Positive loadings are represented in blue, while negative loading are in red

Table 6 Description of factors

Name	Key characteristics
Dense Core	These locations exhibit high residential density, high public transit use, and proximity to grocery stores, schools, and libraries signifying urban access
Social Equity	Locations with low educational attainment, high proportion of Spanish speakers, and older vehicles indicating lower socioeconomic status
Crossing Safety	Locations with crosswalks at bus stops, intersections, and signals, as well as pedestrian signals highlighting safe intersection crossings
Social Advantage	Bus stops with high educational attainment, high proportion of English speakers, and high income indicating higher socioeconomic status
Road Width	Locations characterized by wide roads and high capacity that can accommodate high traffic volumes
Crossing Difficulty	Indicates locations that lack pedestrian signs with wide roads and many lanes to cross, creating a difficult crossing environment for pedestrians
Stop Infrastructure	Highlights locations with traffic signals, pedestrian signals, bus signs, and crosswalks, highlighting infrastructure, safety, and visibility of pedestrians
Interstate Highway	Indicates locations designed for fast, long-distance travel with multiple opposing lanes, wide medians, and full access control
Principal Arterial	Strongly marks designated principal arterial routes with exposure to high-speed and high-volume oncoming and opposing direction traffic
Major Collector	Rural major collector or urban minor arterial locations serving as mid-level traffic connectors gathering traffic from local roads and connecting to major arterials or highways
Minor Arterial	Rural minor arterials or urban principal arterials characterized by high-speed, major traffic flow
Road Condition	Highlights built-up urban core areas with varying road conditions
Minor Connector	Standard collector rural and urban roads connecting local neighborhood streets to arterial systems with access to commercial areas

and land-use context. The immediate physical environment focuses on the presence and condition of roadway and bus stop infrastructure and amenities, including *bus stop infrastructure*, *road condition*, *road width*, *crossing difficulty* or *crossing safety*. The network integration and connectivity evaluate the stop's location and accessibility within the broader transportation system, including stops on *interstate highways*, *principal or minor arterials*, or *major or minor connectors*. Lastly, the sociodemographic and land-use context describes the characteristics of the community and area the stop serves such as *core density*, *social advantage or social equity*, based on indicators such as income levels, educational levels, car ownership, and english proficiency. Together, these factors provide a comprehensive characterization of bus stops for further analysis. A summary of all extracted factors and their interpretations is provided in Table 6.

Bus Stop Types

The cluster analysis identified 13 bus stop types, each exhibiting distinct patterns across the 13 underlying factors characterizing the bus stop's physical environment, road network connectivity, and sociodemographic and land-use context. In this study, the terms *cluster* and *type* are used interchangeably; however, for the presentation of results, the terminology *type* or *typology* will be adopted. Each bus stop type is characterized by examining the central tendency (mean and median) and variability (standard deviation) of the factor scores within each group. To facilitate interpretation, Figs. 10 and 11 provide complementary box-plot visualizations, illustrating the composition of each cluster by its factors, and conversely, how each factor manifests across the different clusters.

Typology Profiles

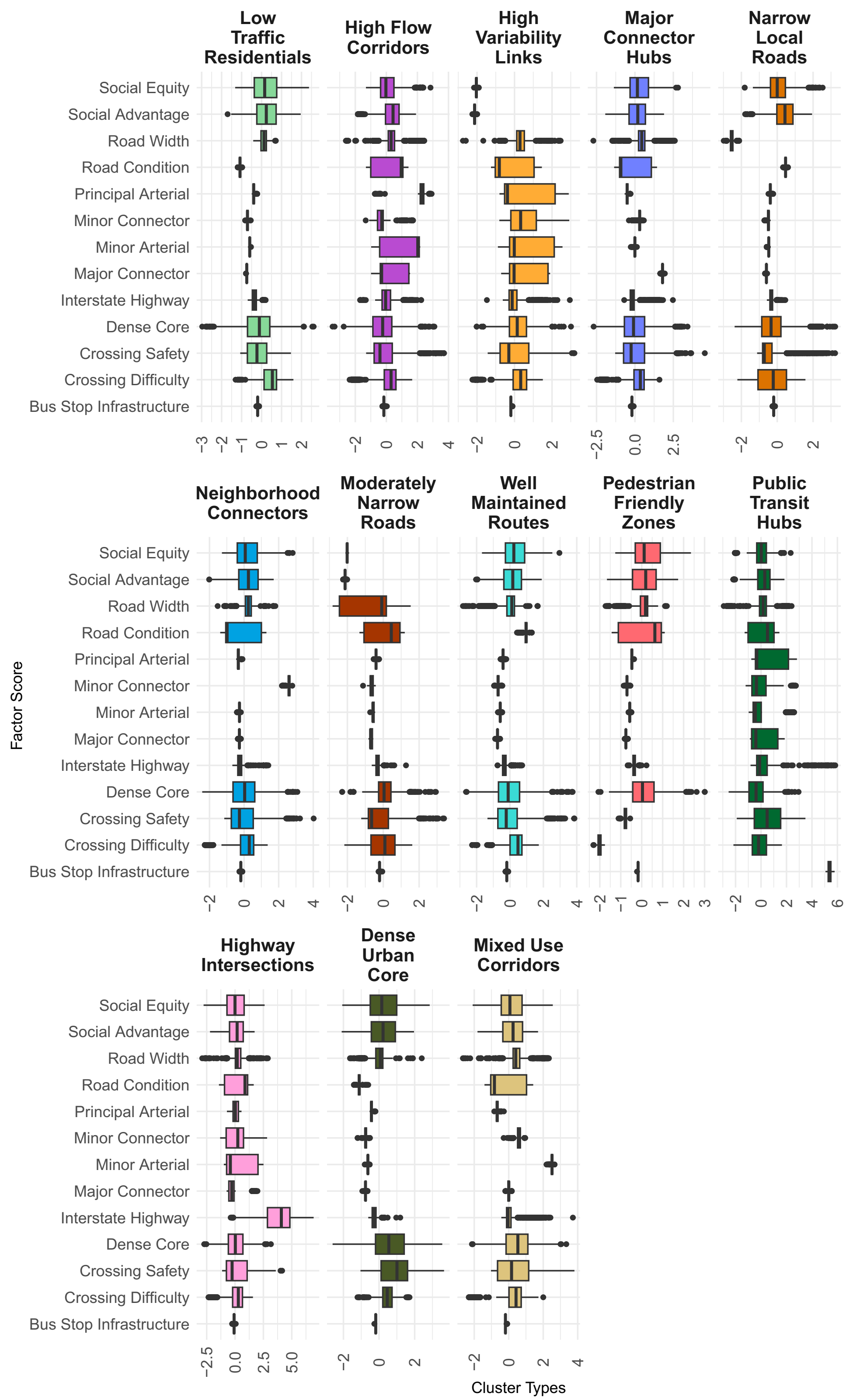
Type 1: Low-Traffic Residential

This type represents areas with limited bus stop infrastructure. It comprises 1871 bus stops (11% of the total) characterized by strong negatives in road condition, moderate negatives in bus stop infrastructure, and strong positives in crossing difficulty and road width. These patterns suggest relatively well-maintained streets that are primarily residential and carry low traffic volumes but lack crosswalks and other pedestrian infrastructure.

Type 2: High-Flow Corridors

This type includes 2,321 (13%) bus stops on major transportation corridors, dominated by principal and minor arterials. They are characterized by negatives on bus stop infrastructure and crossing safety, and positives on crossing difficulty with high variability in road conditions and road widths. These areas likely experience very high traffic

Fig. 10 Box plots of types by factor

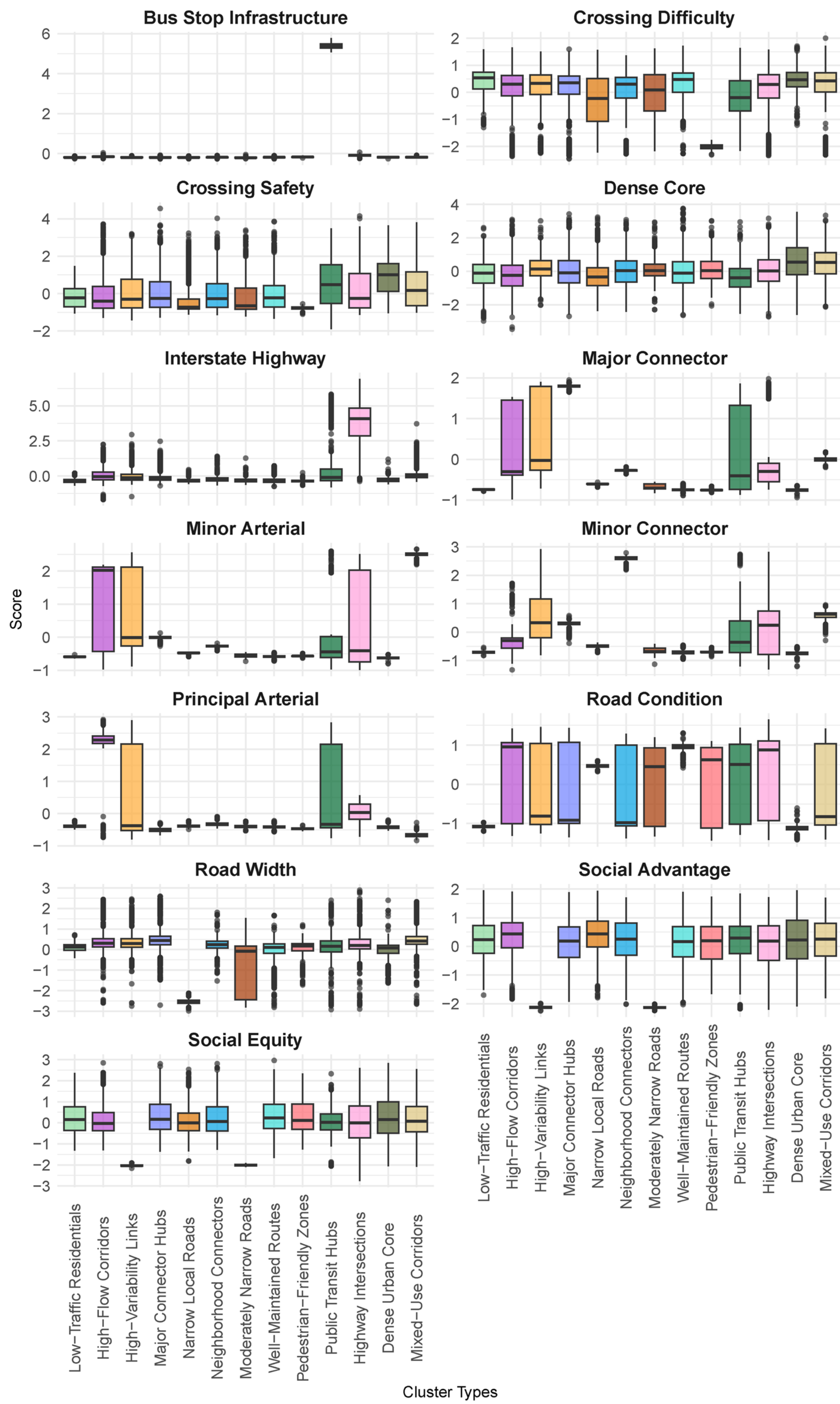


volumes and cut through a variety of neighborhoods with limited infrastructure.

Type 3: High Variability Links

High Variability Links includes 724 (4%) observations with highly variable infrastructure characteristics and

Fig. 11 Box plots of factors by type



quality, and extremely low scores on social equity and social advantage. The stops are strongly associated with Minor Connectors and Minor Arterials, indicating that they serve a key functional role in the road network, carrying significant traffic. However, the infrastructure does not match the need, with low scores in bus stop infrastructure. This cluster exhibits the most infrastructure variability, pointing to an inconsistent and potentially neglected physical environment.

Type 4: Major Connector Hubs

Major Connector Hubs comprises 2997 (17%) bus stops along major (mid-level) traffic connectors that collect traffic from local roads and connect to major arterials or highways, with somewhat poor road conditions.

Type 5: Narrow Local Roads

This type is characterized by extremely narrow roads with consistently good road conditions across 1296 (7%) bus stops. It shows low scores on all road type factors (no major roads) and positive scores on social advantage and a long positive tail on crossing safety. These narrow, well-maintained local streets show the most consistent road characteristic profile across all factors.

Type 6: Neighborhood Connectors

This type represents minor (neighborhood) connector-dominated areas with 1,433 (8%) bus stops, featuring mixed social conditions, and moderate overall infrastructure. These areas show highly consistent secondary connectivity specialization, representative of a typical residential area that is well-connected internally but not a major thoroughfare.

Type 7: Moderately Narrow Roads

Moderately Narrow Roads with 844 (5%) observations, shows some similarity to Type 3 with extremely low scores on social equity and social advantage, but limited infrastructure variability and characteristically narrow roads with good (positive) conditions, suggesting areas are on the periphery.

Type 8: Well-Maintained Routes

This type represents areas with consistently excellent road conditions with low scores on traffic-heavy road types across 2,416 (14%) bus stops. These stops are likely in stable, desirable neighborhoods with good infrastructure.

Type 9: Pedestrian-Friendly Zone

Pedestrian-Friendly Zone includes 805 (5%) bus stops in good condition and extremely low scores on crossing difficulty. It has low traffic (negative scores on road types) and average social factors, which suggests it could be a modern planned development or a closed-off campus.

Type 10: Public Transit Hub

This type comprises transit-oriented areas with 586 (3%) bus stops, featuring extremely high bus stop infrastructure with somewhat better crossing safety and average scores across all other factors. This cluster represents central downtown or major transfer stations where bus routes converge.

Type 11: Interstate Highway Intersection

This type represents interstate highway-dominated corridors with 691 (4%) bus stops, with positive scores on road width and road conditions. These corridors are highway-focused, likely serving as access routes to off-highway areas such as big box stores, with unpredictable local infrastructure.

Type 12: Dense Urban Core

Dense Urban Core includes 792 (4%) bus stops in dense urban core areas. This cluster represents a dense urban center where walking is common and safer crossing infrastructure exists. However, heavy road usage makes maintenance challenging, resulting in poor road conditions.

Type 13: Mixed-Use Arterial Corridor

This type features a very high minor arterial score with 960 (5%) bus stops. These areas are likely bustling corridors lined with businesses, apartments, and higher traffic. They have a high dense core score and wider road width, confirming their role as busy, commercial strips.

Distribution of Bus Stops by Type

The distribution of the 17,736 bus stops across the 13 types reveals a strategic balance between concentrating service on high-capacity corridors while ensuring equitable residential coverage (Fig. 12).

Major Connector Hubs, *Well-Maintained Routes*, and *High-Flow Corridors* represent the largest segments and together with *Dense Urban Core* account for almost 50% of bus stops. These clusters reflect the reality that bus stops are concentrated in densely populated areas and along roads designed for significant traffic volume and high capacity, where transit is most economical and useful.

Simultaneously, the system maintains extensive reach into residential neighborhoods, as evidenced by the substantial share of stops on *Low-Traffic Residential* (11%), *Narrow Local* (7%), and *Neighborhood Connector* (8%) roads. This 26% share confirms that the bus network is not limited to major roads but also serves community-level residential areas to provide essential first-last-mile connectivity.

The presence of smaller, specialized types like *Pedestrian-Friendly Zones* (5%) and *Public Transit Hubs* (3%) highlights tailored solutions in certain urban contexts. Other small clusters, such as *Highway Intersections*(4%), align with operational practicality and reflect their limited utility for bus operations. In general, the distribution of the cluster types reflects a strategic approach that concentrates service on major routes and dense urban centers while ensuring coverage across residential neighborhoods.

Spatial Distribution

Figure 13 maps the 13 distinct bus stop types, revealing clusters of similar stops that transcend agency boundaries. This geographic pattern suggests underlying commonalities in operational characteristics and potential risk profiles.

The major corridor networks, including Route 2, Interstate 91, Route 116, and Interstate 495, are classified as *High-Flow Major Corridors*, *High Variability Links*, and *Major Connector Hubs*. These corridors extend throughout eastern Massachusetts and into central Massachusetts,

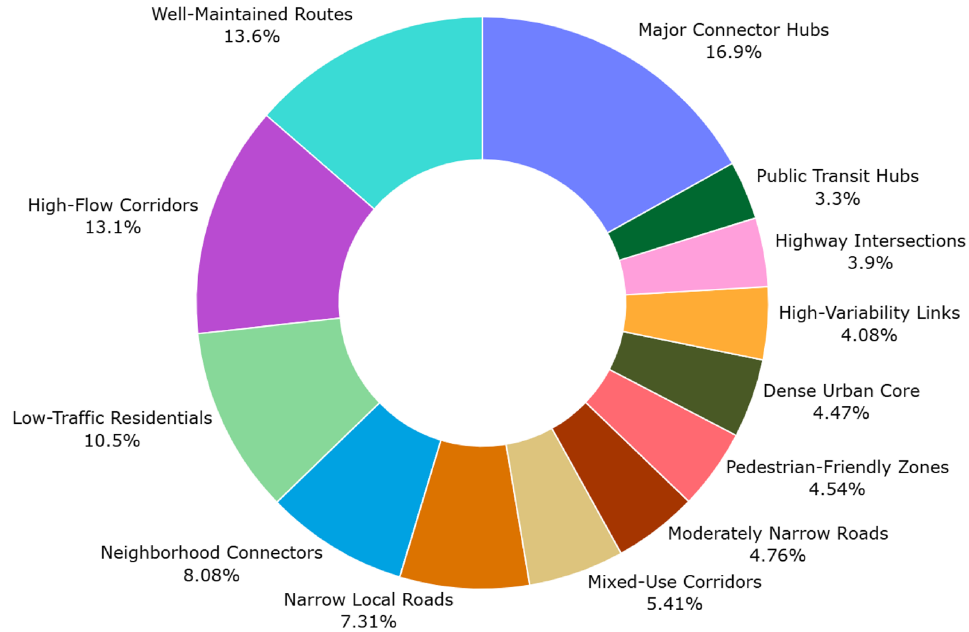


Fig. 12 Distribution of bus stops by type

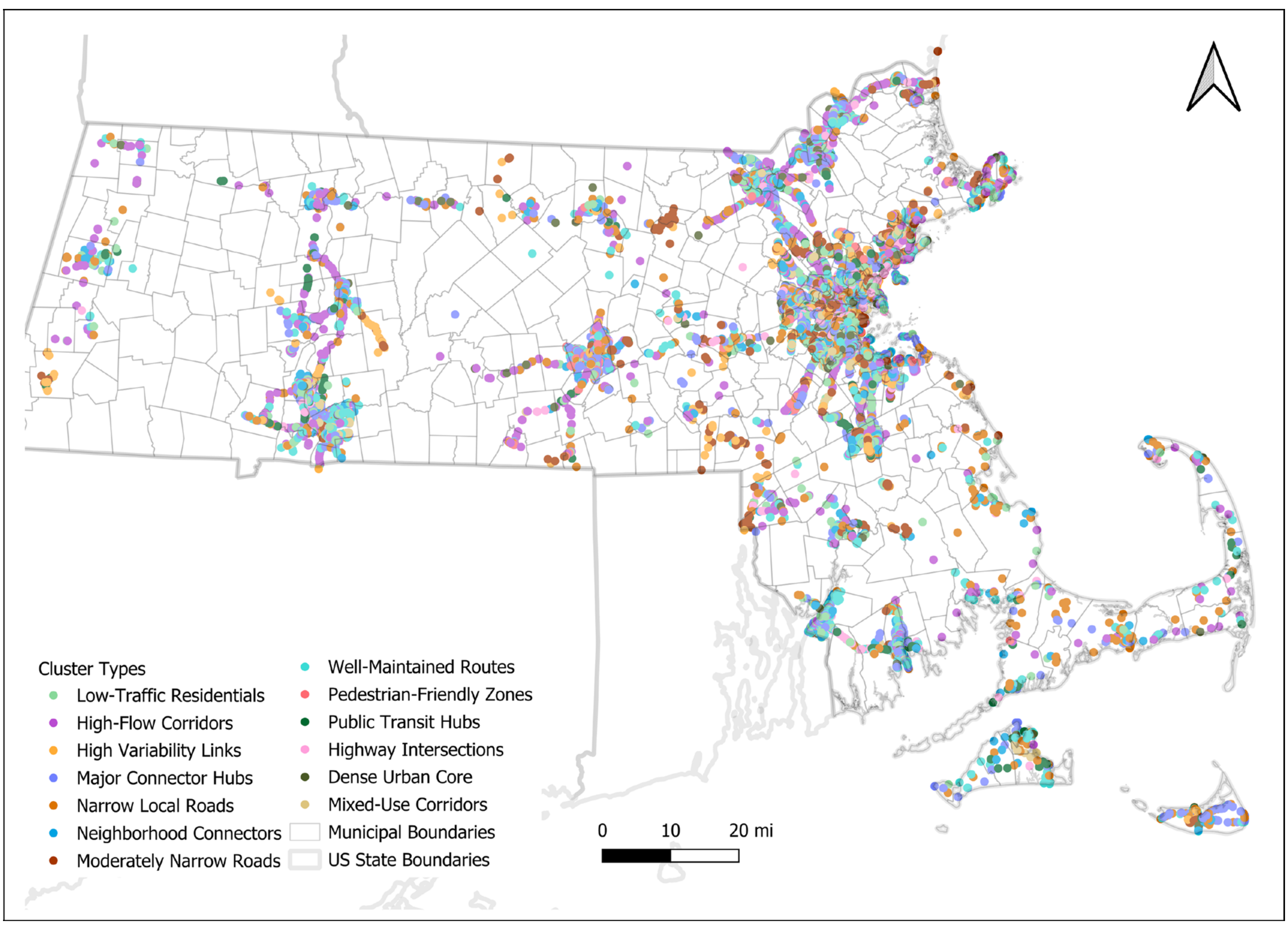


Fig. 13 Spatial distribution of bus stops by type

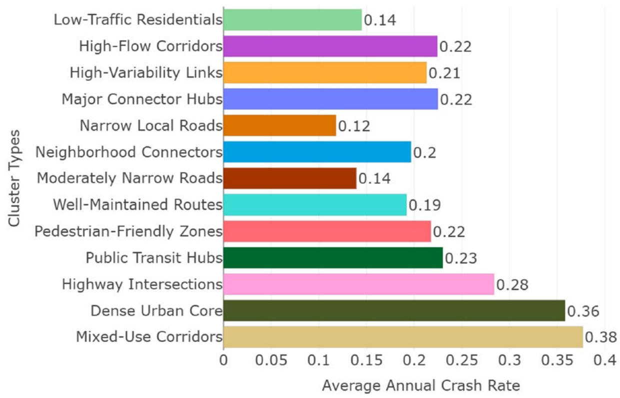


Fig. 14 Average annual crash rate by type

creating linear concentrations of high-connectivity and high-risk infrastructure.

Substantial areas of suburban and outer-ring communities show concentrations of *Low-Traffic Residentials, Narrow Local Roads*, and *Moderately Narrow Roads*, representing neighborhoods with low to moderate connectivity, and narrow road widths.

The map reveals that distinct types perform specific functional roles within regional transportation systems, with corresponding safety implications.

Type-Specific Safety Outcomes

Figure 14 shows the 13 types by their mean annual crash rate. Crash risk is most acute in dense, complex urban environments, and lowest in constrained, local-access networks.

The *Mixed-Use Arterial Corridors* and *Dense Urban Core* stand out with the highest mean crash rates of 0.38 and 0.36, respectively. These are complex environments characterized by high traffic volumes, multi-modal users, and mixed land uses, where competing demands for space create inherent conflict. Despite the presence of safety infrastructure, these areas demand the highest priority for targeted intervention, requiring coordinated investment in road redesign and advanced safety measures.

Notably, *Public Transit Hubs* (0.23), *Pedestrian-Friendly Zones* (0.22), and *Well-Maintained Routes* (0.19) also exhibit elevated crash rates. specialized transit, pedestrian infrastructure, and road quality do not, by themselves, guarantee safety. The high concentration of activity in these areas creates complex interactions that necessitate innovative design solutions beyond basic infrastructure provision.

High-Flow Corridors and *Highway Intersections* represent the high-speed, high-volume backbone of transit, serving regional and metropolitan functions. These corridors show elevated crash rates, of 0.22 and 0.28, respectively, reflecting the inherent risks of prioritizing efficiency when moving large traffic volumes.

Middle Tier Connectors, like the *Major Connector Hubs* and *Neighborhood Connectors* serve as intermediate links between major arterials and local neighborhoods. The slightly higher crash rate of *Major Connector Hubs* (0.22) compared to *Neighborhood Connectors* (0.2) suggests that reduced functional demands and lower speeds on the latter contribute to improved safety outcomes

In contrast, narrow local access networks like *Narrow Local Roads*, *Moderately Narrow Roads*, and *Low-Traffic Residentials*, demonstrate the lowest crash rates of 0.12, 0.14, and 0.14, respectively. These types form safe, neighborhood-scale networks where narrow roads and low traffic volumes naturally calm vehicle speeds, resulting in strong safety performance and providing essential community connectivity.

Typology-Informed Crash Risk Characteristics and Countermeasures

The XGBoost analysis and the SHAP plot identify the top five characteristics of bus stop crash risk for each type (Fig. 15). The most important features influencing crash outcomes relate to land use, activity density, and infrastructure. Appearing in all 13 clusters, the presence of major destinations like grocery stores, schools, and libraries within 2,000 feet of a bus stop and/or the distance to the nearest grocery store both have the strongest relationship with crash rates. This aligns with the positive associations observed for the proportion of walking and biking commuters, AADT, and population density, indicating that areas with higher pedestrian activity and greater land use intensity experience elevated crash rates. Rather than representing direct causal factors, these features likely serve as proxy indicators for exposure risk. Areas with more pedestrian generators naturally experience higher volumes of both VRU and vehicle traffic, thereby increasing opportunities for conflicts. From an infrastructure perspective, crosswalk prevalence, distance to traffic signals or intersections, and Arterial road indicators, characteristic of complex, high-capacity intersections, also emerge as positive predictors of crash risk. Together, these findings suggest that crash risk at bus stops is primarily driven by the confluence of increased activity levels and the infrastructure designed to manage them.

Demographic characteristics including the percentage of Asian speakers with limited English proficiency, Spanish speakers with limited English proficiency, and overall proportion of residents with limited English proficiency all show positive associations with crash rates. These patterns, along with the positive association observed for populations with less than high school education, suggest that crashes are more prevalent in areas serving racially diverse

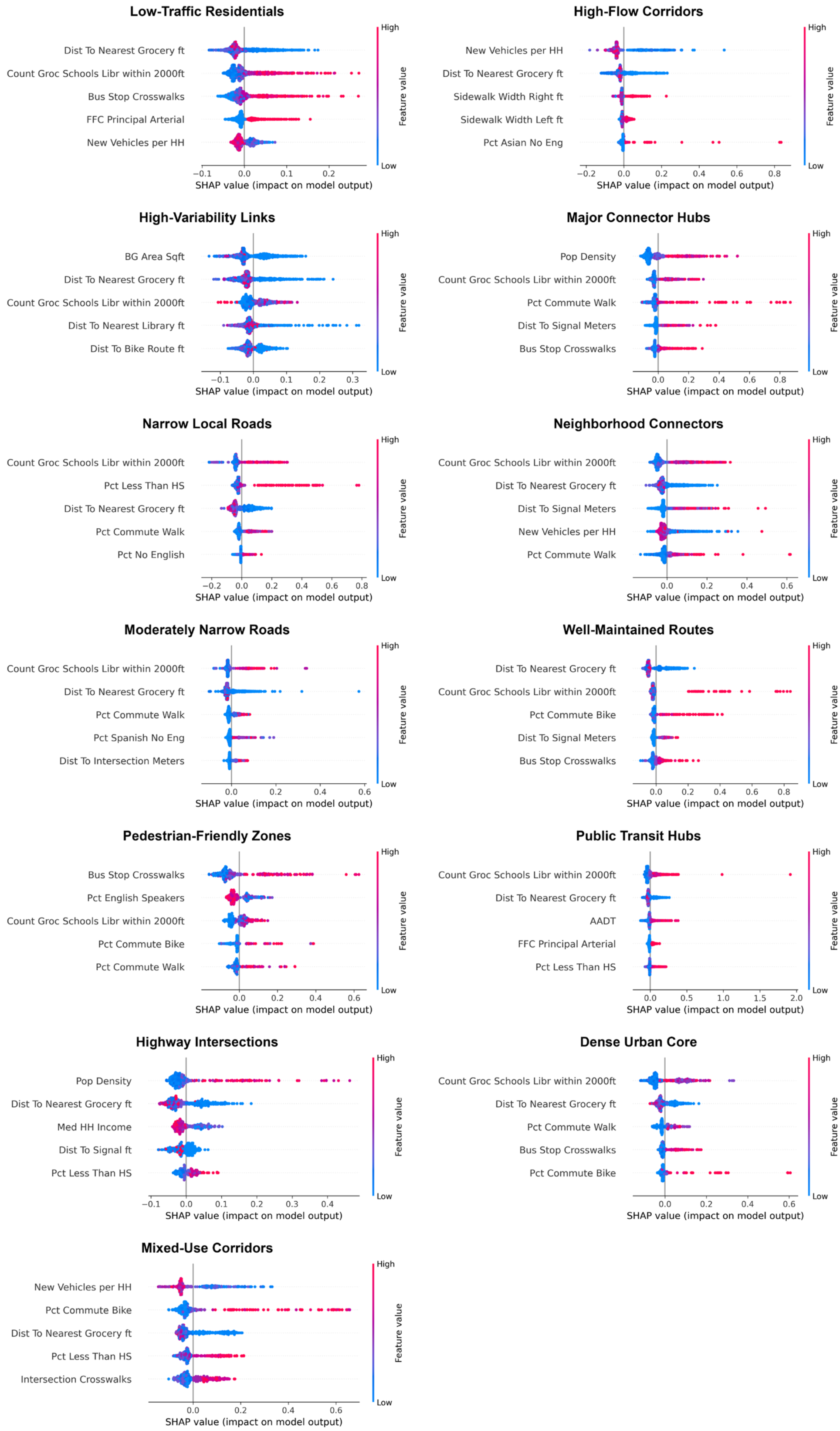


Fig. 15 Type-specific characteristics for crash rate prediction based on SHAP values. Each point represents a bus stop, colored by the feature's value (blue = low, red = high). The SHAP values indicate both the magnitude and direction of each feature's effect on predicted crash risk, positive values increase, and negative values decrease the predicted risk

and educationally disadvantaged populations. However, the underlying mechanisms require further investigation to determine whether these associations reflect infrastructure deficiencies, behavioral features, or differential exposure patterns.

While many of these crash risk characteristics are non-actionable predictor, that is, they are fixed geographic or demographic features, they can still inform the development of actionable mitigation strategies. The Federal Highway Administration's Proven Safety Countermeasures (PSC) initiative comprises 28 evidence-based strategies organized across four safety focus areas, speed management, roadway departure, intersections, and pedestrians/bicyclists, with several crosscutting countermeasures addressing multiple categories (FHWA 2025). These strategies were matched to the 13 types based on the top predictive risk characteristics identified through the SHAP analysis, ensuring that interventions directly address the specific crash risk in each cluster (Table 7).

For bus stops near grocery stores, pedestrian-focused countermeasures were prioritized, including crosswalk visibility enhancements (which can reduce pedestrian injury crashes by up to 40%), pedestrian hybrid beacons (55% reduction in pedestrian crashes), and leading pedestrian intervals (13% reduction in pedestrian-vehicle crashes). And for those where bus stop infrastructure appeared as a top feature, rectangular rapid flashing beacons (47% pedestrian crash reduction) and pedestrian refuge islands were recommended at the bus stop to address the concentrated pedestrian activity.

For bus stops with significant active transportation features, particularly those in *Pedestrian-Friendly Zones*, *Dense Urban Core* and *Mixed-Use Corridors*, separated bicycle lanes were prioritized given their 53% reduction in bicycle-vehicle crashes when vertical separation elements like flexible delineators are used, alongside road diets (19–47% total crash reduction) and complete streets designs.

For bus stop types characterized by high vehicle ownership and traffic volumes, such as *High-Flow Corridors* and *Public Transit Hubs*, speed management strategies were emphasized, including appropriate speed limits for all road users, which can achieve a 26% reduction in traffic fatalities with comprehensive speed management implementation, speed safety cameras, which can reduce crashes on urban principal arterials by up to 54%, use of traffic control devices like RRFBs and PHBs (47–55% pedestrian crash reduction), road diets and roadway reconfiguration (19–17% total crash reduction), dedicated left-turn and right-turn lanes (28–48% and 14–26% crash reductions respectively), and roundabouts (78–82% reduction in fatal and injury crashes).

Bus stop types that exhibited racial diversity, language barriers, and educational disadvantage as predictive features, received emphasis on roadside enhancements and walkway infrastructure (22–44% reduction in crashes) and multilingual safety campaigns, presuming that these communities may be systematically underserved in pedestrian infrastructure provision. The number of new vehicles per household was associated with lower crash rates. This negative association may reflect the presence of advanced safety technologies in newer vehicles, such as automatic emergency braking and pedestrian detection systems, or may serve as a proxy for neighborhood affluence and corresponding investments in safer street infrastructure.

The highest-risk clusters, *Mixed-Use Corridors*, *Dense Urban Core*, and *Highway Intersections*, would benefit from combining multiple PSCs for compounded crash reduction impacts. This includes crosswalk visibility enhancements, automated speed enforcement, separated active transportation infrastructure, road diets, use of traffic control devices, median barriers and refuge islands, and variable speed limits responsive to changing traffic and density conditions. These insights can also guide the strategic relocation of bus stops to safer locations.

Discussion

This discussion examines the key findings, their theoretical implications, and practical recommendations emerging from the analysis of 13 distinct transit bus stop types. We also discuss limitations and provide potential directions for future work.

The Infrastructure–Safety Paradox and the Narrow Road Advantage

Conventional road safety strategy holds that dedicated infrastructure, such as traffic signals, crosswalks, and pedestrian signals, directly enhances safety. However, this study reveals a counterintuitive infrastructure–safety paradox, the presence of such infrastructure is often associated with higher crash rates, while its absence in areas with minimal, constrained designs (like narrow roads) correlates with better safety performance. Bus stop types such as *Public Transit Hubs*, *Dense Urban Core*, and *Mixed-Use Corridors*, which score highly on the *Crossing Safety* and *Bus Stop Infrastructure* factors, exhibit elevated crash rates. Conversely, areas characterized by physical constraints, such as *Narrow Local Roads* and *Moderately Narrow Roads*, demonstrate the lowest crash rates.

This paradox reflects two fundamentally different safety philosophies: "safety through design" versus "safety through limitation." The former relies on the engineered approach of adding safety infrastructure to manage complex, high-risk environments, while the latter depends on using the road's physical geometry (like narrow widths) to naturally calm

Table 7 Pedestrian crash risk characteristics and potential countermeasures by cluster type

Cluster type	Risk characteristics	Potential countermeasures
Low-Traffic Residentials	Grocery distance Major destination proximity Bus stop crosswalks Principal arterial roads New vehicles per household	Crosswalk visibility enhancements near grocery20–25 mph speed limits near grocery Enhanced bus stop infrastructure with lighting Enhanced delineation on principal arterials Rectangular rapid flashing beacons (RRFBs)
High-Flow Corridors	New vehicles per household Grocery distance Right Sidewalk width Left Sidewalk width Asian-speaking %	Pedestrian hybrid beacons (PHBs) Leading pedestrian intervals (LPIs) Sidewalk network expansion and connectivity Curb extensions/bump-outs at key crossings Asian-language safety campaigns
High-Variability Links	Area square footage Grocery distance Major destination proximity Library distance Bike route distance	Variable speed limits responsive to density/time Leading pedestrian intervals (LPIs) Parking lot access management Pedestrian corridors to community facilities Bicycle safety education programs
Major Connector Hubs	Population density Major destination proximity Walking mode share Signal DistanceBus stop crosswalks	Enhanced crosswalk systems (RRFBs/PHBs) Pedestrian priority zones near community hubs Leading pedestrian intervals (LPIs) Dedicated Left and Right Turn Signals Crosswalk visibility enhancements
Narrow Local Roads	Major destination proximity < high school education % Grocery distance Walking mode share English language %	Enhanced crosswalk systems (RRFBs/PHBs) Community-based safety education partnerships Sidewalk improvements to grocery stores Enhanced non-motorized infrastructure Multilingual pedestrian safety campaigns
Neighborhood Connectors	Major destination proximity Grocery distance Nearest signal distance New vehicles per household Walking mode share	Access managementCoordinated signal systems in grocery areasOptimized signal timing for pedestrian phases Density-responsive traffic control measures Road diets and roadway reconfiguration
Moderately Narrow Roads	Major destination proximity Grocery distance Walking mode share Spanish-speaking % Intersection Distance	Enhanced crosswalk safety (bilingual signage) Walking safety programs Spanish-language safety campaigns Midblock crosswalk systems (RRFBs/PHBs)
Well-Maintained Routes	Grocery distance Major destination proximity Bicycle mode share Nearest signal distance Bus stop crosswalks	Pedestrian-priority corridors on grocery routes Signal timing for peak grocery shopping hours Separated bicycle lanes Extended crossing times Enhanced bus stop safety (shelters, lighting)
Pedestrian-Friendly Zones	Bus stop crosswalks English language %Major destination proximity Bicycle mode share Walking mode share	Pedestrian-activated signals at bus stops Multi-modal complete streets integration Protected bicycle lanes with vertical separation Bicycle safety education programs Exclusive pedestrian phases
Public Transit Hubs	Major destination proximity Grocery distance AADT traffic volumePrincipal arterial roads Less than HS education %	Raised medians with pedestrian refuge islands Exclusive pedestrian signal phases Speed safety cameras on high-AADT links Road diets and roadway reconfiguration Roadside enhancements in underserved areas
Highway Intersections	Population density Grocery distance Median Household Income Signal distance Less than HS education %	Dynamic speed limits responsive to density Road diets and roadway reconfiguration Grade-separated pedestrian crossings Rooundabouts Roadside enhancements in underserved areas
Dense Urban Core	Major destination proximity Grocery distance Walking mode share Bus stop crosswalks Bicycle mode share	Exclusive pedestrian phases Comprehensive sidewalk networks Multi-modal complete streets integrationEnhanced crosswalk systems (RRFBs/PHBs) Separated bicycle lanes
Mixed-Use Corridors	New vehicles per household Bicycle mode share Grocery distance Less than HS education % Intersection Crosswalks	Automated speed enforcement Protected bicycle lanes with grocery access Reduced lane widths for traffic calming Roadside enhancements in underserved areas Enhanced crosswalk systems (RRFBs/PHBs)

traffic by forcing behavioral change. Critically, formal safety infrastructure typically represents a response to pre-existing high-risk conditions characterized by substantial traffic volume, elevated speeds, and complex conflict points. The infrastructure itself thus serves as a proxy indicator for greater exposure risk rather than direct causal factors. In narrow road environments, physical constraints create "safety through inconvenience," where the built environment naturally suppresses traffic speed and volume, reducing both crash frequency and severity without requiring engineered interventions. Indeed, road diets have been associated with up to 47% reduction in crashes (FHWA 2025).

The pronounced protective effect of narrow road design offers a powerful insight for urban planners. In residential and low-volume settings, geometric constraints that force drivers to slow down may be a more effective safety strategy than adding supplemental infrastructure.

The Connectivity–Safety Trade-off

Enhanced network connectivity, activity, and infrastructure density are associated with elevated crash risk, revealing an inherent tension between transportation system efficiency and safety performance. High-connectivity types, such as *Major Corridors*, *Major Connectors*, *Interstate Corridors*, and *Minor Connectors*, all exhibit moderate to high crash rates. The presence of major pedestrian generators (grocery stores, schools, and libraries) near a bus stop demonstrate the strongest associations with crash rates. Serving as proxy indicators, these features signify heightened exposure risk.

The Connectivity–Safety trade-off presents a fundamental challenge for transportation planning, which conventionally seeks to maximize connectivity and activity density for economic efficiency while simultaneously minimizing safety risks. The finding suggests that these objectives may be structurally incompatible without sophisticated design interventions that address both dimensions concurrently.

The analysis further reveals a clear danger hierarchy across road functional classifications, with arterials emerging as the highest-risk infrastructure type. Arterials are characterized by moderate speeds that are dangerous to pedestrians and cyclists but insufficient separation for vehicle safety, combined with mixed land uses that generate diverse traffic patterns and conflict points. Advancing arterial safety will require not only speed management and activity separation but also collaboration between multiple disciplines and systems to ensure an integrated focus on safe roadways, safe vehicles, and safe road behaviors.

Limitations and Future Research

This study provides valuable insights into bus stop safety, but several limitations should be considered. First, the analysis focuses exclusively on Massachusetts and aggregates eight years of crash data (2016–2024) without modeling temporal trends, assuming static risk characteristics. In addition, bus stops that underwent safety improvements during the study period cannot be distinguished from those that remained unchanged, potentially obscuring the effectiveness of prior interventions. Relatedly, infrastructure characteristics were primarily drawn from the 2022 Road Inventory (U.S. Census Bureau 2025), representing a snapshot in time. Changes in road conditions, bus stop amenities, or nearby land uses that occurred after 2022 are not captured in the analysis. Second, crashes were associated with bus stops solely based on the 300-foot buffer. While the selected buffer radii are typical for exposure modeling, it was not empirically optimized for this study. Also, the analysis included all crashes within the bus stop buffers and did not differentiate between crashes directly caused by the bus stop and those that were merely spatially proximate. Finally, the analysis focuses on crash frequency, not severity, to ensure statistical power. Consequently, our findings may not fully represent the most serious safety outcomes, as the relationship between crash types and frequency may differ from that of severity.

Building on these limitations, future research should explore application to other geographic contexts, integrate temporal trends and analyses of crash severity, and employ normalization techniques that account for exposure such as traffic volume and pedestrian activity.

Conclusion

Bus stops represent critical nodes in urban transportation systems where the goals of mobility, accessibility, and safety intersect, and sometimes conflict. Every day, millions of people across the US rely on bus stops to access jobs, education, healthcare, and opportunities. The critical need to safeguard vulnerable road users (VRUs), who face a disproportionately high risk in traffic crashes, highlights the importance of designing safe interfaces between roadway, transit, and pedestrian or bike infrastructures.

This study introduces a data-informed framework for understanding and addressing bus stop safety through the integration of machine learning-based factor analysis, cluster analysis, crash prediction, and a comprehensive multi-source data fusion. By analyzing 17,736 bus stops across Massachusetts through 13 latent infrastructure and sociodemographic factors, we identified 13 distinct bus stop types that represent fundamentally different operational contexts and risk profiles. This cluster-based approach moves beyond traditional jurisdictional or simple geographic groupings of bus stops to reveal the underlying patterns that link bus stop characteristics with crash outcomes. Further, the framework is highly generalizable and readily deployable to other locations.

Based on the findings of previous studies, we believe that the insights gained from this study are applicable and transferable to other locations that have bus stops with typical characteristics similar to those of the types identified. Future work could validate this claim by applying the framework to elicit typologies from other states in the US (and even other countries) and quantifying their correspondence.

This typology of 13 distinct stop types, each with a unique risk profile, provides a framework for replacing generic safety measures with targeted solutions. The SHAP analysis identifies the specific characteristics driving risk, creating a direct, evidence-based link to targeted intervention strategies. Priority should be given to *Mixed-Use Arterial Corridors* and *Dense Urban Core* locations (collectively representing 9% of bus stops but exhibiting the highest crash rates). Risk in these areas is driven by multi-modal concentration, proximity to and density of major destinations, and crosswalk prevalence. Addressing these risk characteristics would require combining multiple Proven Safety Countermeasures to achieve compounded crash reduction impacts.

The infrastructure–safety paradox demonstrates that adding crosswalks, signals, and other formal treatments is necessary but insufficient in high-risk environments. Agencies must embrace complementary strategies including traffic volume reduction, speed management through self-enforcing design (narrow lanes, tight curb radii, vertical deflection), access management, and land use coordination. The success of low-infrastructure clusters (*Narrow Local Roads*) in achieving superior safety outcomes suggests that constraining traffic through design, rather than attempting to manage high volumes through control devices, may be the most effective long-term strategy.

Ultimately, creating safe transit environments requires collaboration between multiple disciplines and systems to ensure an integrated focus, including safe roadways, safe vehicles, and safe road behaviors. Given the national Vision Zero and Safe System initiatives to eliminate traffic fatalities and reduce serious injuries, understanding and addressing bus stop safety must be a central component of these efforts.

The path forward requires continued research, sustained investment, and willingness to challenge conventional approaches when evidence suggests better alternatives exist. This study represents one step in that ongoing journey toward transportation systems that enable mobility without compromising safety.

Acknowledgements The authors would like to thank the New England University Transportation Center (NEUTC) led at the University of Massachusetts, Amherst for their partial financial support of this research. Funding for the UTC Program is provided by the Office of Assistant Secretary for Research and Innovation (OST-R) of the United States Department of Transportation (USDOT). The recommendations of this study are those of the authors and do not represent the views of NEUTC or USDOT.

Author Contributions The authors confirm their contribution to the paper as follows: study conception and design: TO; data collection: TO, AP; algorithm development and coding: TO, AP, JO; analysis and interpretation of results: TO, AP, FT, JO, MK; draft manuscript preparation: TO; manuscript review: TO, AP, FT, JO, MK. All authors reviewed the results and approved the final version of the manuscript.

Funding This research was funded by the US Department of Transportation via the New England University Transportation Center (Grant Number: 69A35523483).

Data Availability Data and code used to generate the results in this paper are available upon reasonable request.

Declarations

Conflict of Interest The authors declare no conflict of interest.

Ethical Approval This study did not require ethical approval as it did not involve any human or animal subjects.

References

Abdalazeem M, Oke J (2025) Roadway crash typology of census tracts enables targeted interventions via interpretable machine learning. Data Sci Transport 7(2):14. https://doi.org/10.1007/s42421-025-00128-2

Agarwal NK (2011) Estimation of pedestrian safety at intersections using simulation and surrogate safety measures

Agrawal V, Chatterjee S, Mitra S (2019) Crash severity analysis through nonparametric machine learning methods. J East Asia Soc Transport Stud **13**:2614–2629. https://doi.org/10.11175/easts.13.2614

Ahmed T, Gayah VV (2025) A novel bi-level clustering optimization approach to balance treatment of crash data. Accid Anal Prev 219:108107. https://doi.org/10.1016/j.aap.2025.108107

Ali Y, Hussain F, Haque MM (2024) Advances, challenges, and future research needs in machine learning-based crash prediction models: a systematic review. Accid Anal Prev 194:107378. https://doi.org/10.1016/j.aap.2023.107378

Anis M, Geedipally SR, Lord D (2025) Pedestrian crash causation analysis near bus stops: Insights from random parameters Negative Binomial-Lindley model. Accid Anal Prev 220:108137. https://doi.org/10.1016/j.aap.2025.108137

APTA (2024) Ridership trends. https://transitapp.com/APTA. Accessed 06 March 2023

Ashraf MT, Dey K, Pyrialakou D (2022) Investigation of pedestrian and bicyclist safety in public transportation systems. J Transp Health 27:101529. https://doi.org/10.1016/j.jth.2022.101529

Bao J, Liu P, Ukkusuri SV (2019) A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accid Anal Prev 122:239–254. https://doi.org/10.1016/j.aap.2018.10.015

BAT—Brockton Area Transit. https://www.ridebat.com/. Accessed 29 Oct 2025

Berkshire Regional Transit Authority—Travel With Us. https://berkshirerta.gov/. Accessed 29 Oct 2025

Bike Inventory 2024. https://geodot-massdot.hub.arcgis.com/datasets/bike-inventory-2024. Accessed 24 Oct 2025

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Brustad HK, Midtbø JE, Tomba GS, Kivelä M, Gregersen FA, Alessandretti L (2025) Extracting commuters from automated road traffic counters: a Gaussian mixture approach. Data Sci Transport 7(2):9. https://doi.org/10.1007/s42421-025-00123-7

Budzyński A, Federowicz M, Jabłoński A, Hasan W, Gorszanów J, Nikitishyn T (2024) A machine learning approach for predicting road accidents. Saf Defense 10(2):60–70. https://doi.org/10.37105/sd.230

Butt MS, Shafique MA (2025) A literature review: AI models for road safety for prediction of crash frequency and severity. Discov Civ Eng 2(1):99. https://doi.org/10.1007/s44290-025-00255-3

Cape Ann Transportation Authority—Cape Ann Transportation Authority (2025). https://canntran.com/. Accessed 29 Oct 2025

Cattell RB (1966) The scree test for the number of factors. Multivar Behav Res 1(2):245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chakraborty M, Gates TJ, Sinha S (2023) Causal analysis and classification of traffic crash injury severity using machine learning algorithms. Data Sci Transport 5(2):12. https://doi.org/10.1007/s42421-023-00076-9

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–794. https://doi.org/10.1145/2939672.2939785. arXiv:1603.02754 [cs]. Accessed 28 Oct 2025

Chen J, Liu P, Wang S, Zheng N, Guo X (2025) Prediction and interpretation of crash severity using machine learning based on imbalanced traffic crash data. J Saf Res 93:185–199. https://doi.org/10.1016/j.jsr.2025.02.018

Craig CM, Morris NL, Van Houten R, Mayou D (2019) Pedestrian safety and driver yielding near public transit stops. Transp Res Rec 2673(1):514–523. https://doi.org/10.1177/0361198118822313

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc Ser B (Methodol) 39(1):1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dumbaugh E, Stiles J, Mitsova D, Saha D (2024) The most vulnerable user: examining the role of income, race, and the built environment on pedestrian injuries and deaths. Transp Res Rec 2678(2):743–752. https://doi.org/10.1177/03611981231175888

Everitt BS, Landau S, Leese M, Stahl D (2011) Cluster analysis, 5th edn. Wiley-Blackwell, New York. https://doi.org/10.1002/9780470977811. Accessed 11 July 2025

Fabrigar LR, Wegener DT (2012) Exploratory factor analysis. Exploratory factor analysis. Oxford University Press, New York, viii, 159

Formosa N, Quddus M, Man CK, Timmis A (2023) Appraising machine and deep learning techniques for traffic conflict prediction with class imbalance. Data Sci Transport 5(2):4. https://doi.org/10.1007/s42421-023-00067-w

Fraley C, Raftery AE (2002) Model-based clustering discriminant analysis, and density estimation. J Am Stat Assoc 97(458):611–631. https://doi.org/10.1198/016214502760047131

FRTA. https://www.frta.org/. Accessed 29 Oct 2025

Furno A, Zanella AF, Stanica R, Fiore M (2024) Spatial and temporal exploratory factor analysis of urban mobile data traffic. Data Sci Transport 6(1):4. https://doi.org/10.1007/s42421-024-00089-y

GATRA (2023) Greater Attleboro and Taunton Regional Transit Authority. https://www.gatra.org/. Accessed 29 Oct 2025

Grocery Store Locations in Massachusetts 2021. https://www.arcgis.com/apps/mapviewer/index.html?webmap=c07a77521b0c4f91b0a473cab37de927. Accessed 24 Oct 2025

Han Z, Oke J (2025) Novel typology of fatal crash behaviors via natural language processing. Data Sci Transport 7(3):22. https://doi.org/10.1007/s42421-025-00135-3

Hess P, Moudon A, Matlick J (2004) Pedestrian safety and transit corridors. J Public Transport **7**(2). https://doi.org/10.5038/2375-0901.7.2.5

Hodson T, Over T, Foks S (2021) Mean squared error deconstructed. J Adv Model Earth Syst 13. https://doi.org/10.1029/2021MS002681

Home—Cape Cod Regional Transit Authority (2025). https://capecodrta.org/. Accessed 29 Oct 2025

Impact M (2024) MassDOT: crash data portal. https://apps.impact.dot.state.ma.us/cdp/home. Accessed 13 Dec 2024

Iranitalab A, Khattak A (2017) Comparison of four statistical and machine learning methods for crash severity prediction. Accid Anal Prev 108:27–36. https://doi.org/10.1016/j.aap.2017.08.008

Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323. https://doi.org/10.1145/331499.331504

Jin L, Xu X, Wang Y, Lazar A, Sadabadi KF, Spurlock CA, Needell Z, Don DR, Amirgholy M, Asudegi M (2024) Macroscopic traffic modeling using probe vehicle data: a machine learning approach. Data Sci Transport 6(3):17. https://doi.org/10.1007/s42421-024-00102-4

Kaiser HF (1960) The application of electronic computers to factor analysis. Educ Psychol Meas 20:141–151. https://doi.org/10.1177/001316446002000116

Karlaftis MG, Golias I (2002) Effects of road geometry and traffic volumes on rural roadway accident rates. Accid Anal Prev 34(3):357–365. https://doi.org/10.1016/s0001-4575(01)00033-1

Kieffer KM (1998) Orthogonal versus oblique factor rotation: a review of the literature regarding the pros and cons. Technical report. ERIC Number: ED427031. https://eric.ed.gov/?id=ED427031. Accessed 28 Oct 2025

Kraidi R, Evdorides H (2020) Pedestrian safety models for urban environments with high roadside activities. Saf Sci 130:104847. https://doi.org/10.1016/j.ssci.2020.104847

Lin L, Wang Q, Sadek AW (2018) Civil and environmental engineering: novel machine learning methods for accident data analysis. Technical report. /view/dot/40076. Accessed 25 Aug 2025

Lowell Regional Transit Authority. https://lrta.com/. Accessed 29 Oct 2025

Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. arXiv:1705.07874. Accessed 29 Oct 2025

Maiti N, Leclercq L (2025) Estimating spatial mean speeds from local sensors: a machine-learning approach. Data Sci Transport 7(1):6. https://doi.org/10.1007/s42421-025-00120-w

MART. https://www.mrta.us/. Accessed 29 Oct 2025

Massachusetts Public School Districts (Feature Service) | MassGIS Data Hub. https://gis.data.mass.gov/maps/145c945f4fa744e8951c47b696c73758/about. Accessed 24 Oct 2025

Massachusetts Vehicle Census 2022. https://geodot.mass.gov/search?collection=dataset. Accessed 24 Oct 2025

MassGIS Data: Colleges and Universities | Mass.gov. https://www.mass.gov/info-details/massgis-data-colleges-and-universities Accessed 24 Oct 2025

MassGIS Data: Libraries | Mass.gov. https://www.mass.gov/info-details/massgis-data-libraries. Accessed 24 Oct 2025

MassGIS Data: MassGIS-MassDOT Roads | Mass.gov. https://www.mass.gov/info-details/massgis-data-massgis-massdot-roads. Accessed 24 Oct 2025

mbta.us—Under Construction. https://www.mbta.us/. Accessed 29 Oct 2025

MeVa—Merrimack Valley Transit. https://mevatransit.com/. Accessed 29 Oct 2025

Miranda-Moreno LF, Morency P, El-Geneidy AM (2011) The link between built environment, pedestrian activity and pedestrian-vehicle collision occurrence at signalized intersections. Accid Anal Prev 43(5):1624–1634. https://doi.org/10.1016/j.aap.2011.02.005

MWRTA. https://www.mwrta.com/. Accessed 29 Oct 2025

News Release (2023) U.S. traffic fatalities 25 percent higher than a decade before despite dropping each year since 2021... (2024). https://tripnet.org/reports/addressing-americas-traffic-safety-crisis-national-news-release-07-02-2024/. Accessed 30 Oct 2025

NHTSA (2023) Crash Stats.pdf. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813428. Accessed 30 July 2024

Pessaro B, Catalá M, Wang Z, Spicer M (2017) Impact of transit stop location on pedestrian safety. World Transit Research

Phillips RO, Hagen OH, Berge SH (2021) Bus stop design and traffic safety: an explorative analysis. Accid Anal Prev 153:105917. https://doi.org/10.1016/j.aap.2020.105917

Prajapati AK, Guler I (2025) Surrogate-modeling for pavement maintenance scheduling. Data Sci Transport 7(3):23. https://doi.org/10.1007/s42421-025-00134-4

Proven Safety Countermeasures | FHWA (2025). https://highways.dot.gov/safety/proven-safety-countermeasures. Accessed 29 Oct 2025

Pulugurtha SS, Penkey EN (2010) Assessing use of pedestrian crash data to identify unsafe transit service segments for safety improvements. Transp Res Rec 2198(1):93–102. https://doi.org/10.3141/2198-11

PVTA (2024). https://www.pvta.com/index.php. Accessed 01 Aug 2024

Quistberg DA, Koepsell TD, Johnston BD, Boyle LN, Miranda JJ, Ebel BE (2015) Bus stops and pedestrian–motor vehicle collisions in Lima Peru: a matched case–control study. Inj Prev 21(e1):15–22. https://doi.org/10.1136/injuryprev-2013-041023

Rafe A, Arman MA, Singleton PA (2024) A comparative study using generalized ordered Probit, stacking ensemble, and TabNet: application to determinants of pedestrian crash severity. Data Sci Transp 6(2):13. https://doi.org/10.1007/s42421-024-00098-x

Rahman MS, Mohiuddin H, Kafy A-A, Sheel PK, Di L (2019) Classification of cities in Bangladesh based on remote sensing derived spatial characteristics. J Urban Manag 8(2):206–224. https://doi.org/10.1016/j.jum.2018.12.001

Rahman M, Kockelman KM, Perrine KA (2022) Investigating risk factors associated with pedestrian crash occurrence and injury severity in Texas. Traffic Inj Prev 23(5):283–289. https://doi.org/10.1080/15389588.2022.2059474

Rewalt A, Brakewood C, Cherry C (2025) An analysis of pedestrian safety at bus stops using FARS data. J Saf Res 95:147–159. https://doi.org/10.1016/j.jsr.2025.09.002

Roll J, McNeil N (2021) Understanding pedestrian injuries and social equity. https://rosap.ntl.bts.gov. Accessed 03 Oct 2025

Santos K, Dias JP, Amado C (2022) A literature review of machine learning algorithms for crash injury severity prediction. J Saf Res 80:254–269. https://doi.org/10.1016/j.jsr.2021.12.007

Silva PB, Andrade M, Ferreira S (2020) Machine learning applied to road safety modeling: a systematic literature review. J Traffic Transport Eng (Engl Ed) 7(6):775–790. https://doi.org/10.1016/j.jtte.2020.07.004

SRTA—Southeastern Regional Transit Authority. https://www.srtabus.com/. Accessed 29 Oct 2025

Su J, Sze NN (2022) Safety of walking trips accessing to public transportation: a Bayesian spatial model in Hong Kong. Travel Behav Soc 29:125–135. https://doi.org/10.1016/j.tbs.2022.06.003. (**Accessed 2025-10-16**)

The Wave | Nantucket Regional Transit Authority. https://www.nrtawave.com/. Accessed 29 Oct 2025

Thurstone LL (1931) Multiple factor analysis. Psychol Rev 38(5):406–427. https://doi.org/10.1037/h0069792. (**Place: US Publisher: Psychological Review Company**)

Truong L, Somenahalli S (2011) Using GIS to identify pedestrian-vehicle crash hot spots and unsafe bus stops. J Public Transport 14. https://doi.org/10.5038/2375-0901.14.1.6

Ukkusuri S, Miranda-Moreno L, Ramadurai G, Isa-Tavarez J (2012) The role of built environment on pedestrian crash frequency. Saf Sci **50**. https://doi.org/10.1016/j.ssci.2011.09.012

Ul Arifeen S, Ali M, Macioszek E (2023) Analysis of vehicle pedestrian crash severity using advanced machine learning techniques. Arch Transp 68:91–116. https://doi.org/10.61089/aot2023.ttb8p367

Ulak MB, Kocatepe A, Yazici A, Ozguven EE, Kumar A (2021) A stop safety index to address pedestrian safety around bus stops. Saf Sci 133:105017. https://doi.org/10.1016/j.ssci.2020.105017. (**Accessed 2024-01-18**)

U.S. Census Bureau (2025) American Community Survey 2014–2018 5-Year Data Release. Section: Government. https://www.census.gov/newsroom/press-kits/2019/acs-5-year.html. Accessed 24 Oct 2025

Vineyard Transit Authority. https://www.vineyardtransit.com/ Accessed 29 Oct 2025

Wang Y, Sun W, Ricord S, Souza CM (2024) Assessing pedestrian safety on roads through machine learning approaches for state highways in Washington State. https://rosap.ntl.bts.gov. Accessed 31 Oct 2025

Wen X, Xie Y, Jiang L, Pu Z, Ge T (2021) Applications of machine learning methods in traffic crash severity modelling: current status and future directions. Transp Rev 41(6):855–879. https://doi.org/10.1080/01441647.2021.1954108

Weng Y, Das S, Paal SG (2023) Applying few-shot learning in classifying pedestrian crash typing. Transp Res Rec 2677(8):563–572. https://doi.org/10.1177/03611981231157393

WRTA—Worcester Regional Transit Authority. https://therta.com/. Accessed 29 Oct 2025

Xu X, Liang T, Zhu J, Zheng D, Sun T (2019) Review of classical dimensionality reduction and sample selection methods for large-scale data processing. Neurocomputing 328:5–15. https://doi.org/10.1016/j.neucom.2018.02.100

Xue J, Tan R, Ma J, Ukkusuri SV (2025) Data science in transportation networks with graph neural networks: a review and outlook. Data Sci Transport 7(2):10. https://doi.org/10.1007/s42421-025-00124-6

Ye Z, Wang C, Yu Y, Shi X, Wang W (2016) Modeling level-of-safety for bus stops in China. Traffic Inj Prev 17(6):656–661. https://doi.org/10.1080/15389588.2015.1133905

Yendra D, Haworth N, Watson-Brown N (2024) A comparison of factors influencing the safety of pedestrians accessing bus stops in countries of differing income levels. Accid Anal Prev 207:107725. https://doi.org/10.1016/j.aap.2024.107725

Zegeer C, Hunter W, Staplin L, Bents F, Huey R, Barlow J (2010) Safer vulnerable road users: pedestrians bicyclists, motorcyclists, and older users. Federal Highway Administration Office of Safety

Zermane H, Zermane A, Mohd Tohir MZ (2024) Machine learning techniques for fatal accident prediction. ACC J 30:24–49

Zhang C, He J, Wang Y, Yan X, Zhang C, Chen Y, Liu Z, Zhou B (2020) A crash severity prediction method based on improved neural network and factor analysis. Discret Dyn Nat Soc 2020(1):4013185. https://doi.org/10.1155/2020/4013185.