scientific reports

Check for updates

OPEN

# Segment level safety analysis using lane-changing behavior and driving volatility features from connected vehicle trajectories

Lei Han✉ & Mohamed Abdel-Aty

**Frequent crashes on urban arterial segments pose significant safety and mobility concerns. While existing safety studies primarily focus on macro infrastructure and traffic features, ignoring the critical influences of micro-level risky driving behavior (e.g., lane-changing). To address such gaps, we developed a directional-level segment crash analysis method, leveraging the high-resolution Connected Vehicle (CV) data. A Constrained Gaussian Mixture Method was proposed to identify lane-changing behavior from raw CV trajectories. Micro-level driving behavior features were then extracted considering risky driving behavior, driving volatility, and aggressive speeding. A Bivariate hierarchical negative binomial model was employed to jointly estimate the heterogeneous impacts of driving behavior features on rear-end (RE) and sideswipe (SW) crashes. While a hierarchical zero-inflated Poisson model was utilized to identify significant contributors to speeding crashes. Empirical experiments at Hillsborough County highlight the critical role of risky driving behavior features for segment safety: (1) Segments with a high proportion of free-flow trajectories tend to experience fewer RE and SW crashes. (2) Driving fluctuation of stop-and-go vehicles is positively related to the frequency of RE crashes. (3) Risky right lane-changings coupled with hard accelerations are significantly associated with SW crashes. (4) Aggressive speeding behavior is highly related to speeding crashes.**

**Keywords** Arterial segment safety, Connected vehicle trajectory, Driving volatility, Lane-changing behavior, Sideswipe crashes, Rear-end crashes

Frequent crashes on urban arterial roads pose a significant safety concern, and result in substantial delays across the roadway network[1,2]. To address these safety challenges, extensive efforts have been dedicated to urban arterial crash frequency modeling and safety evaluation to seek effective countermeasures[1,3–6]. Within existing urban arterial segment crash analysis, researchers have identified various contributing factors (e.g., segment geometry design, surrounding access points, and traffic volume)[1,4,7], and have employed statistical and machine learning models to establish their impact on segment crash frequency[2,8,9].

However, previous segment crash studies mainly focus on the safety impact of macro-level features (e.g., segment length, speed limit) while ignore the key role of micro-level risky driving behavior, which is the leading cause of crashes[10,11]. For most arterial systems, traditional vehicle detectors are either absent or installed at limited locations, providing highly aggregated traffic information (e.g., vehicle counts and speeds)[12,13]. As a result, existing studies have largely relied on macroscopic traffic indicators (e.g., bidirectional Annual Average Daily Traffic) and infrastructure features (e.g., roadside access point count), failing to capture microscopic traffic dynamics and driving behavior characteristics specific to each direction.

With the applications of vehicle connection techniques, Connected Vehicle (CV) trajectory data has become increasingly available with high-resolution updates (within seconds) and extensive spatial coverage at both road and city levels. Leveraging such CV data, it is possible to extract enriched traffic information (e.g., traffic volume and speed) at a detailed directional level[13,14], enabling more precise crash analysis at the unidirectional segment level. Moreover, micro-level risky driving behavior (e.g., hard braking, lane-changing) can be extracted from CV trajectories to investigate their relationship with segment crashes[15,16].

This study aims to develop a directional-level segment crash frequency model integrating microscopic driving behavior features from emerging CV trajectory data. While some recent studies have extracted the micro-level hard acceleration and braking events into the modeling process[15–17], critical lane-changing behavior and driving

Department of Civil, Environmental & Construction Engineering, University of Central Florida, Orlando, FL 32816, USA. ✉email: le966091@ucf.edu

volatility at segments remain unexplored. For instance, risky lane-changing maneuvers can heighten lateral conflict risks, potentially causing sideswipe crashes[18,19]. Similarly, driving volatility at segments, characterized by variations in driving operations and speed profiles, can cause traffic flow turbulence to increase the rear-end crash risk[20]. Therefore, it is essential to capture these key features and quantify their impact on segment crashes. However, the lack of lane-level information and GPS location drifts poses challenges in extracting lane-changing behavior and driving volatility features from CV trajectories. Meanwhile, the impact of risky driving behavior on different crash types (e.g., rear-end and sideswipe) may exhibit significant heterogeneity due to their distinct crash mechanisms.

To address these challenges, this study makes the following contributions:

1) Incorporating microscopic traffic characteristics and driving behaviors features for segment level crash analysis.
2) Extracting lane-changing behavior and driving volatility features from CV trajectories to explore their relationships with roadway segment crashes.
3) Jointly estimating the heterogeneous impacts of risky driving behaviors on different types of crash frequencies (e.g., Rear-end (RE) and Sideswipe (SW)).

Following this section, Sect.Literature review presents the related literature review, followed by the details of the proposed methodology in Sect. Methodology. Section Data Preparation shows the data preparation and Sect. 5 illustrates the experiment results. Finally, the conclusion of this study is presented in Sect. Conclusion.

## Literature review

### Segment crash contributing factors

This study focuses on urban arterial roadways, where traffic is highly interconnected with other roadways and frequently interrupted by intersections. Based on existing studies[7,21,22], the crash-contributing factors for this type of roadways can be mainly categorized into three groups: road design features, traffic characteristics, and surrounding socio-economic conditions. (1) Several road geometric design features (e.g., segment length, number of lanes, curve ratio, and number of access points) have been found to be significantly correlated with segment crash frequency[4,7,8]. (2) For segment traffic characteristics, traffic-volume-related variables (e.g., Annual Average Daily Traffic) have demonstrated significant positive effects on crash frequency[4,7,23]. Additionally, higher ratio of trucks and posted speed limits has been linked to increased crash likelihood[24–26]. (3) Various socio-economic features near segments (e.g., population, poverty ratio, number of nearby bus stops, and commercial land use) have been found to show positive effect on segment crash frequency[4,27,28]. However, these factors only reflect the static infrastructure and aggregated traffic status at the bi-directional segment level, failing to capture detailed traffic characteristics specific to each road direction[12,29]. Moreover, micro-level risky driving behavior, that is more directly linked to crashes, are still not being considered in existing studies.

To overcome such limitations, recent studies have utilized CV data to capture micro-level traffic features and evaluate their impacts on crashes at the unidirectional segment level. For example, Wang et al. (2018)[2] extracted speed variation from taxi GPS data for each arterial segment direction and found that larger speed variation was significantly associated with increased crash frequency. Gupta et al., (2024)[15] identified three risky driving behaviors (i.e., hard acceleration, braking, and cornering) from Michigan CV data and found a strong positive correlation between the frequency of such risky driving events and crash frequency. These studies have highlighted key impacts of risky driving behavior on segment crashes. However, frequent lane-changing and driving volatility caused by aggressive driving are significant contributors to crashes on arterial segments. Despite their importance, these factors remain largely overlooked in previous studies. Therefore, how to capture crucial lane-changing behavior and driving volatility parameters from CV data and further analyze their impact on crashes still needs further investigation.

### Segment crash frequency modeling

Two primary types of models have been developed for segment crash frequency prediction: statistical and machine learning (ML) methods. Among statistical models, the Poisson family, the Negative Binomial family, and other count-data models have been widely employed[4,7,15,30]. For ML approaches, tree-based models (e.g., random forest, XGBoost, and LightGBM) and neural network models (e.g., multilayer perceptron (MLP) and convolutional neural network (CNN)) are two common modeling methods. While ML models often serve superior predictive ability, they earn a "black-box" designation because of the difficulty in unraveling how specific elements might influence predictions[12,31]. In contrast, statistical models provide strong mathematical interpretability of variable correlations, making them particularly valuable for the decision-making of safety countermeasures[32].

Within statistical methods, earlier studies often adopted a univariate modeling framework to study a single crash frequency variable or multiple crash frequency variables (such as crash frequency by crash type)[33,34]. However, these approaches are not appropriate for modeling multiple dependent variables for the same observational unit as they fail to account for common unobserved heterogeneity affecting the various dependent variables[35]. To address this issue, the multivariate framework have been proposed to account for the potential dependency across multiple dependent variables including multivariate Poisson, multivariate negative binomial, and multivariate Poisson lognormal models[32,36]. Existing safety studies have demonstrated that multivariate models can effectively capture the variable influences across multiple dependent variables and their unobserved heterogeneity[3,37,38]. Given the different crash mechanisms in different crash types (e.g., rear-end and sideswipe crashes), this study aims to develop a multivariate model to jointly estimate the heterogeneous impacts of risky driving behavior on different types of crash frequencies.

## Methodology

### Unidirectional segment analysis unit

In existing arterial segment crash modeling, the basic analysis unit remains limited to the bi-directional segments. Specifically, previous studies commonly define the research segment as the bi-directional roadway section between two intersections, excluding the extended intersection areas (typically 250 ft) (Fig. 1(a)). Features from both directions (e.g., west- and east-bound) are treated as a single entity. However, the road surrounding environment (e.g., access points, lane configurations) and traffic conditions (e.g., tidal traffic patterns) differ significantly between directions, thus leading to distinct crash patterns and their associated safety impacts. Traditional bi-directional analysis overlooked such heterogeneity, thus leading to potentially biased crash assessments[12]. Additionally, the conventional unit is insufficient for capturing complete vehicle trajectories as it does not include the two extended intersection areas.

To address these limitations, the traditional bi-directional segment units should be refined into unidirectional segments, as shown in Fig. 1(b). The segment is extended to encompass the entire roadway from the intersection exit approach to the next intersection's stop line. This ensures the full coverage of the vehicle trajectory travelling at the segment. More importantly, the roadway is further divided by direction into two basic segment units, allowing for a more detailed analysis of direction-specific macro road traffic characteristics and micro driving behaviors.

### Lane-Changing identification from CV trajectories

For traffic safety analysis, lane-changing behavior have been recognized as critical vehicle maneuvers that can potentially cause lateral conflicts and crash risk[18,39]. Especially for the high-volume arterials, risky lane-changing with hard braking or acceleration can increase driving volatility among surrounding vehicles and disrupt the whole segment's traffic flow, significantly raising the likelihood of crashes. Traditional aggregated data from fixed-location detectors lack the granularity to capture such detailed information, limiting the scope of this exploration[12]. The emergence of CV data offers an opportunity to extract lane-changing behavior from individual vehicle trajectories and assess their relationship with crashes[40].

However, raw CV data only provides basic vehicle locations, such as latitude and longitude, making it challenging to directly identify lane-change behavior. To address this issue, Fig. 2 provides the proposed framework for lane-changing behavior identification from CV trajectory data, which consists of three main steps:

- Step 1: Taking roadway GIS as road centerline, spatial matching is performed to map CV trajectories onto individual road segments. A road Cartesian projection is applied to convert CV's GPS coordinates into lateral and longitudinal distances to road centerline.
- Step 2: Based on the distribution of lateral distances of massive CV trajectories, a Gaussian mixture method (GMM) constrained with road features (lane number and width) is employed to identify the lane boundaries and assign lane IDs to CV trajectories.
- Step 3: Lane-changing behaviors are detected by identifying the differences in lane IDs along sequential CV trajectory with additional lateral distance restrictions to account for the potential CV positioning drift.

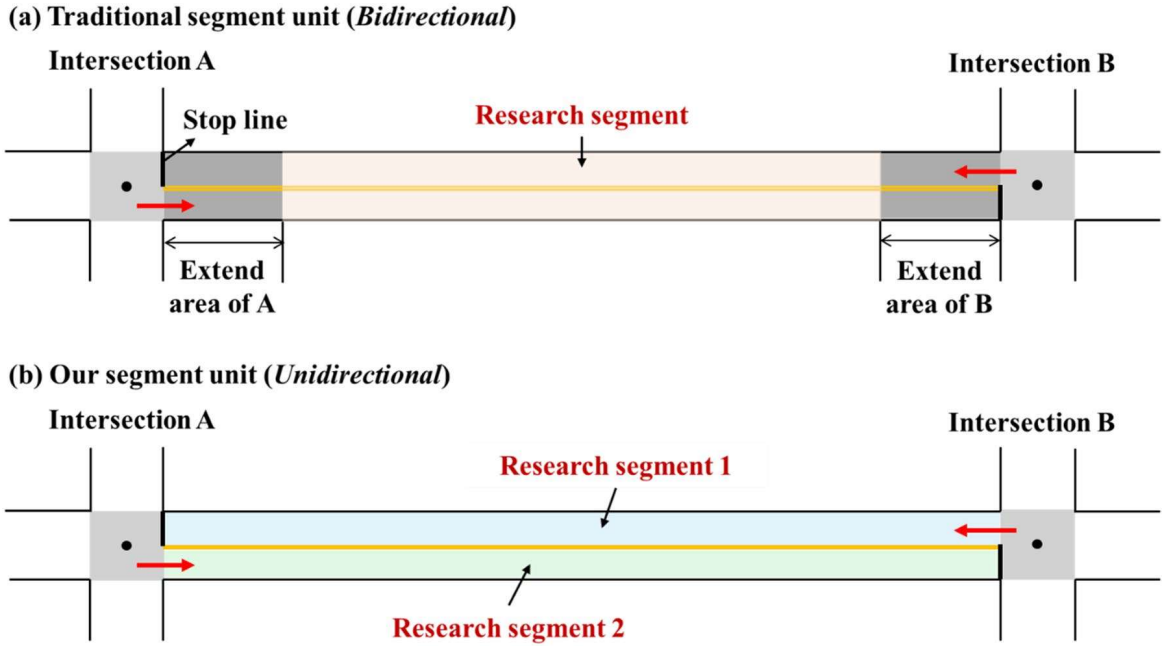


**Fig. 1**. Illustration of bi-directional and unidirectional segments.

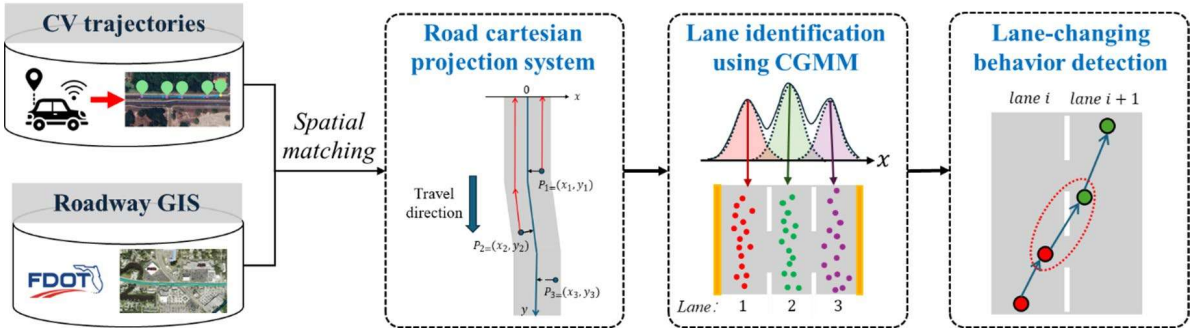


**Fig. 2**. The framework of lane-changing identification from CV Trajectories.

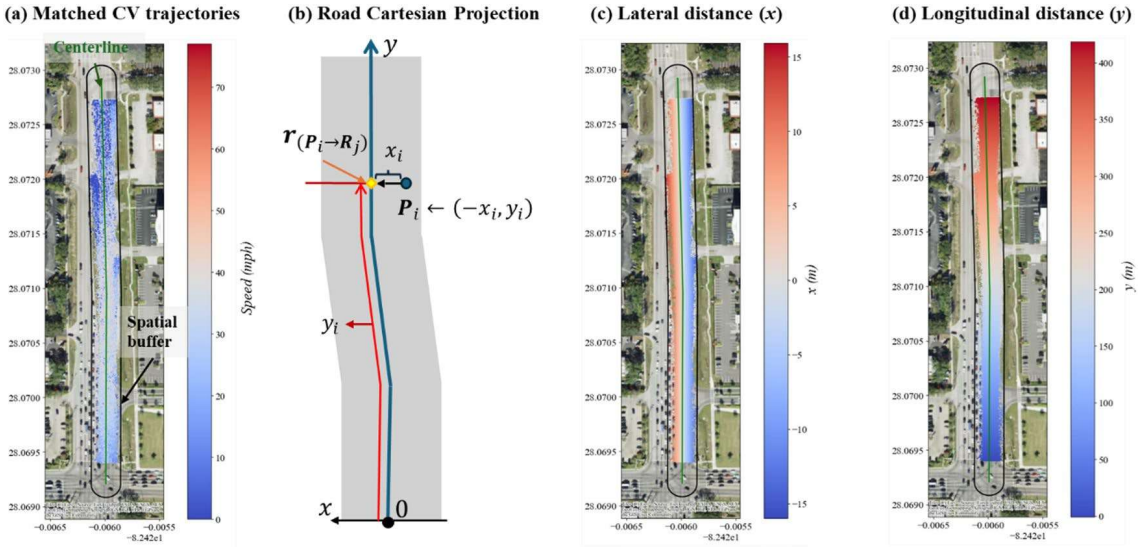


**Fig. 3**. The process of CV data spatial matching at RCP system. The maps were generated using ArcGIS 10.8 (Environmental Systems Research Institute, USA. https://www.esri.com/).

The implementation of these three steps of lane-changing identification framework is thoroughly described in the following subsections.

*CV data spatial matching at road cartesian projection system*
This step takes the CV data and road GIS as the inputs:

The CV trajectory can be represented as:

$$\boldsymbol{Tr}_v = \{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_{n_v}\}$$

where $n_v$ is the number of CV points of vehicle $v$, each CV point can be written as $\boldsymbol{P}_{i\in[1,n_v]} =< lon_i, lat_i, t_i, v_i, h_i >$, where $lon_i$ and $lat_i$ are the longitude and latitude in GPS coordinates, respectively. $t_i$, $v_i$, and $h_i$ represent the collecting timestamp, speed (mph), and vehicle heading (degree), respectively.

The road GIS is seen as the roadway central line as a polyline object:

$$\boldsymbol{R}_j = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{n_j}\}$$

where $n_j$ is the number of GIS points on segment $j$, each road GIS point denotes as $\boldsymbol{r}_{g\in[1,n_j]} =< long_g, lat_g >$ including longitude $long_g$ and latitude $lat_g$ in GPS coordinates.

To match the CV trajectories $\boldsymbol{Tr}_v$ with individual road segment $\boldsymbol{R}_j$, a spatial buffer is first utilized as shown in Fig. 3(a). Considering that arterial always has less than 6 lanes (including the left/right turning and merge lanes) per direction, the spatial buffer is set to (6/2) lane *4m/lane = 12 m on each side of the centerline. A heading verification is employed to filter out the CV trajectories in the opposite direction:

$$|h_i - \vartheta_j| \leq \phi_{max},\ \vartheta_j = arccos\frac{\boldsymbol{r}_1.\boldsymbol{r}_{n_j}}{|\boldsymbol{r}_1| * |\boldsymbol{r}_{n_j}|} \tag{1}$$

where $\vartheta_j$ is the calculated angle of road's travel direction. And $\phi_{max}$ is the angle difference threshold, which is set to 30 degrees in this study.

A Road Cartesian Projection (RCP) system is applied to convert CV's GPS coordinates into lateral and longitudinal distances related to road, as shown in Fig. 3(b):

$$x_i = Distance\left(\boldsymbol{P}_i, \boldsymbol{r}_{(\boldsymbol{P}_i \to \boldsymbol{R}_j)}\right),\ y_i = Distance\left(\boldsymbol{r}_{(\boldsymbol{P}_i \to \boldsymbol{R}_j)}, \boldsymbol{r}_1\right) \tag{2}$$

where $\boldsymbol{r}_{(\boldsymbol{P}_i \to \boldsymbol{R}_j)}$ is the point on $\boldsymbol{R}_j$ perpendicular to $\boldsymbol{P}_i$, $x_i$ measures the distance from $\boldsymbol{P}_i$ to $\boldsymbol{r}_{(\boldsymbol{P}_i \to \boldsymbol{R}_j)}$, representing the vehicle's lateral distance relative to the roadway centerline. It is noted that CV points on the left side of roadway centerline are assigned positive sign and vice versa to better distinct their lateral positions. $y_i$ is the distance from road start point $\boldsymbol{r}_1$ to $\boldsymbol{r}_{(\boldsymbol{P}_i \to \boldsymbol{R}_j)}$, representing the longitudinal distance of the vehicle along the roadway. Figure 3(c) and Fig. 3(d) illustrate the example of the lateral distance ( $x_i$) and longitudinal distance ( $y_i$) of CV trajectories, respectively.

*Lane identification based on constrained Gaussian mixture method (CGMM)*

In this step, a CGMM is introduced to identify lane boundaries based on the distribution of vehicles' lateral distance ( $x_i$). The underlying assumption is that CV trajectories tend to cluster near the center of each lane, with some spread due to inaccuracy of GPS and the natural movement of vehicles within a lane[41–43]. Accordingly, the lateral distances across the multi-lane roadway form a weighted sum of Gaussian distributions, i.e., the Gaussian mixture model (GMM):

$$p(x_i) = \sum_{k=1}^{K} w_k \frac{\exp\{-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}\}}{\sqrt{2\pi\,\sigma_k^2}} \tag{3}$$

where $K$ is the number of Gaussian components corresponding to the number of lanes. $w_1, w_2, \ldots, w_K$ are the weights of each component, which are positive and normalized: $w_{k\in[1,K]} > 0$ and $\sum_{k=1}^{K} w_k = 1$. $\mu_k$ and $\sigma_k$ represent the mean and variance of the $k$th Gaussian distributions, representing the centerline of each lane and the spread of vehicle positions within a lane, respectively (see Fig. 4(a)).

However, directly fitting this GMM may result in unrealistic parameter estimation without considering the actual road conditions. Based on the suggestions from related studies[42,44], two key constrains are introduced into the GMM:

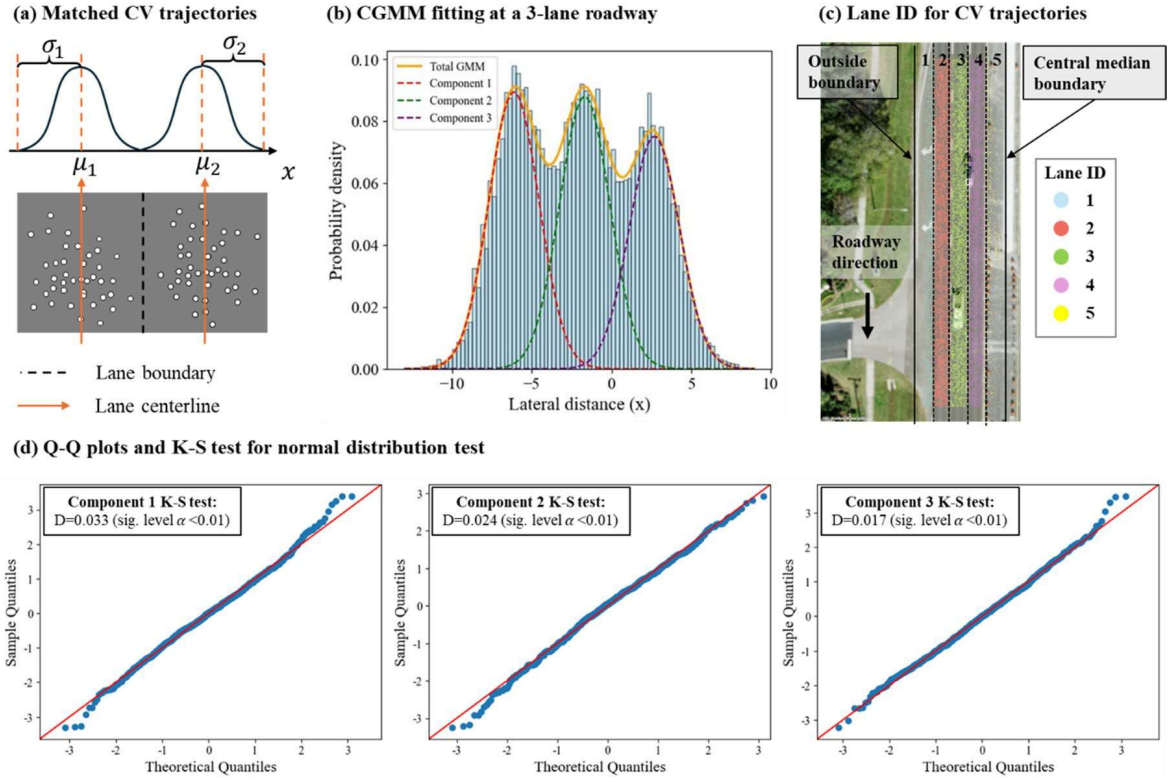


**Fig. 4**. The process of lane identification based on CGMM.

- The width of the lanes on the same roadway should be approximately the same. This observation can be translated into the constraint of $\mu_k$:

$$\mu_k = \mu + (k-1)\,\Delta\mu,\ k \in [1, K] \tag{4}$$

where $\Delta\mu$ is the change between two adjacent lane centerlines, which are the width of lane. $\mu$ is the value of either the leftmost or rightmost lane centerline.

- Given these equal-width lanes, vehicles are expected to drive within their lanes, making that the spread of trajectories for each lane remain approximately the same. Then all the Gaussian components can be assumed to share the same variance:

$$\sigma_k^2 = \sigma^2,\ k \in [1, K] \tag{5}$$

Based on these two constraints, GMM in (3) can be revised as Constrained GMM (CGMM):

$$p(x_i) = \sum_{k=1}^{K} w_k \frac{\exp\{-\frac{(x_i-\mu-(k-1)\Delta\mu)^2}{2\sigma^2}\}}{\sqrt{2\pi\sigma^2}} \tag{6}$$

To determine the CGMM, the Expectation-Maximization (EM) algorithm is commonly adopted to figure out the unknown parameter $\theta_k = (K, w_1, \ldots, w_k, \mu, \Delta\mu, \sigma)$. Given $\mathcal{D}$ as the observed data $x_{i\in[1,n]}$, the EM tries to minimize the following cost function:

$$-\frac{1}{n}\sum_{i=1}^{n}\log\left(p(x_i|\theta_k)\right) + \lambda * R(\mathcal{D}, \theta_k),\ R(\mathcal{D}, \theta_k) = \frac{logn}{2n} * (K+2) \tag{7}$$

where the first term is the negative meaning log-likelihood, while the second term $R(\mathcal{D}, \theta_k)$ is a Bayesian information criterion (BIC) regularization term to penalize complex models, and $\lambda$ >0 is a parameter to make a trade-off between model fitness and model complexity.

Figure 4(b) illustrates an example of the CGMM fitting results on a 3-lane arterial segment. Based on the estimated centerline values, Fig. 4 (c) presents the final lane IDs assigned to the raw CV trajectories. Lane ID 2, 3, and 4 correspond to the estimated component 1, 2, and 3 (in same color), respectively. While lane 1 and lane 5 are considered as outside lanes to account for the complexity of lane setting (e.g., right-turn and median-turn lane). Furthermore, the Q-Q plot and Kolmogorov-Smirnov (K-S) goodness of fit test were conducted to verify if the CV points within each lane is normally distributed or not. As visualized in Fig. 4 (d), the Q-Q plots show that the standardized sample quantiles align closely with the theoretical quantiles of the normal distribution. The K-S test results confirm that the estimated clusters follow normal distributions at significance level $\alpha \leq 0.01$. Across all segments, 80.4% of cases follow normal distribution at significance level $\alpha \leq 0.05$, and the remaining 19.6% are accepted at $\alpha \leq 0.10$. Therefore, these empirical findings support the normality assumption in the CGMM model that CV lateral distance ($x_i$) is well approximated by a normal distribution within each lane.

*Lane-changing detection considering CV positioning drifts*

Since that lane IDs $l_i$ are assigned to CV point $\boldsymbol{P}_{i\in[1,n_v]} =< lon_i, lat_i, t_i, s_i, h_i, l_i >$, detecting lane-changing behaviors becomes straightforward, that is to identify the changes in lane IDs between two consecutive timestamps:

$$LC_i = 1,\ if\ l_i \neq l_{i-1},\ i \in [1, n_v] \tag{8}$$

where $LC_i$ is the lane-changing indicator to be 1 when a lane-changing occurs. However, if vehicle points are near lane boundaries, small positioning drifts can lead to the false detection of lane changes. To address this issue, two additional strict rules are utilized:

$$\Delta t_{i-lc} = |t_{i+2} - t_{i-1}| < 15, \Delta x_{i-lc} = |x_{i+2} - x_{i-1}| > 3 \tag{9}$$

It means that a lane change should satisfy that the lateral position change $\Delta x_{i-lc}$ exceeds a threshold of 3 m and is completed in a short duration of 15s[45]. Note that the threshold of 3 m is smaller than the standard lane width (12ft = 3.65 m) to consider real-world lateral GPS fluctuation. By manually checking different thresholds, we have found that using the full 12ft threshold would miss approximately 10.5% of true lane changes. Conversely, applying a more lenient threshold (e.g., 2–2.5 m) would lead to a high false positive rate (around 20–30%). Therefore, to strike a balance between detection accuracy and false alarm, we adopted a threshold of 3 m for $\Delta x_{i-lc}$. Figure 5 provides an example of several lane-changing behaviors within a single CV trajectory. A total of 3 lane changes are observed: two right lane changes occurring in the middle part of the segment and one left lane change near the end of the segment, close to an intersection.

## Micro-level driving behavior features extraction

In this study, four kinds of micro-level driving behavior features from CV trajectories are extracted and then aggregated at the unidirectional segment level:

(1) *Traffic volume.*

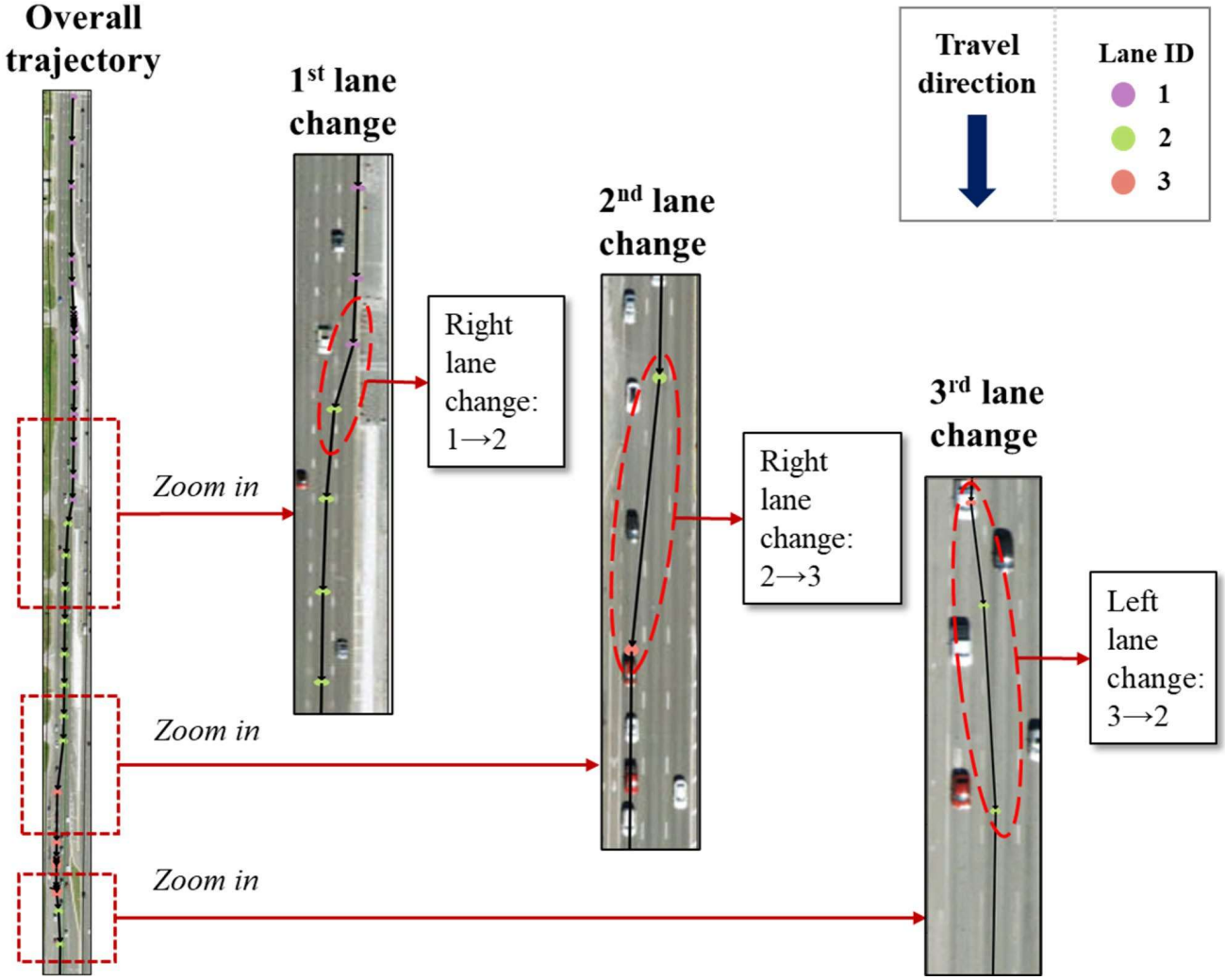


**Fig. 5**. The example of lane-changing behaviors within a CV trajectory.

The counts of vehicles are utilized to represent the traffic volume at each direction. Given the penetration rate of CV, the traffic volume of each unidirectional segment can be calculated:

$$Traffic\,volume_j = \#\ of\ vehicle\ at\ segment\ \boldsymbol{R}_j / p_{cv} \tag{10}$$

where $p_{cv}$ are set as 4.3% according to CV data company and the reverification using sampled detector data.

(2) *Counts of risky driving behavior*.

Risky driving behavior is recognized as extreme driving events to cause potential risk. For instance, if the linear acceleration exceeds a certain threshold, it is classified as "hard acceleration". Therefore, the linear acceleration of CV trajectories is first calculated:

$$a_i = \frac{\Delta v}{\Delta t} = \frac{v_i - v_{i-1}}{t_i - t_{i-1}} \tag{11}$$

A dynamic threshold approach[20,40] is adopted to identify the hard acceleration and braking. Specifically, different speed bins (i.e., 0-5mph, 5-10mph, …70-75mph) were defined, with each of them has its own upper bound calculated:

$$ACC\_threshold_m = \mu_m + 3 * \sigma_m \tag{12}$$

where $\mu_m$ represents the mean and $\sigma_m$ represents the standard deviation of acceleration within the respective speed bin $m$. Using this approach, hard acceleration and braking can be identified as shown in Fig. 6. It shows that the linear acceleration distribution varies across different speed bins, with two threshold curves: the positive curve identifies hard acceleration events, while the negative curve captures hard braking events.

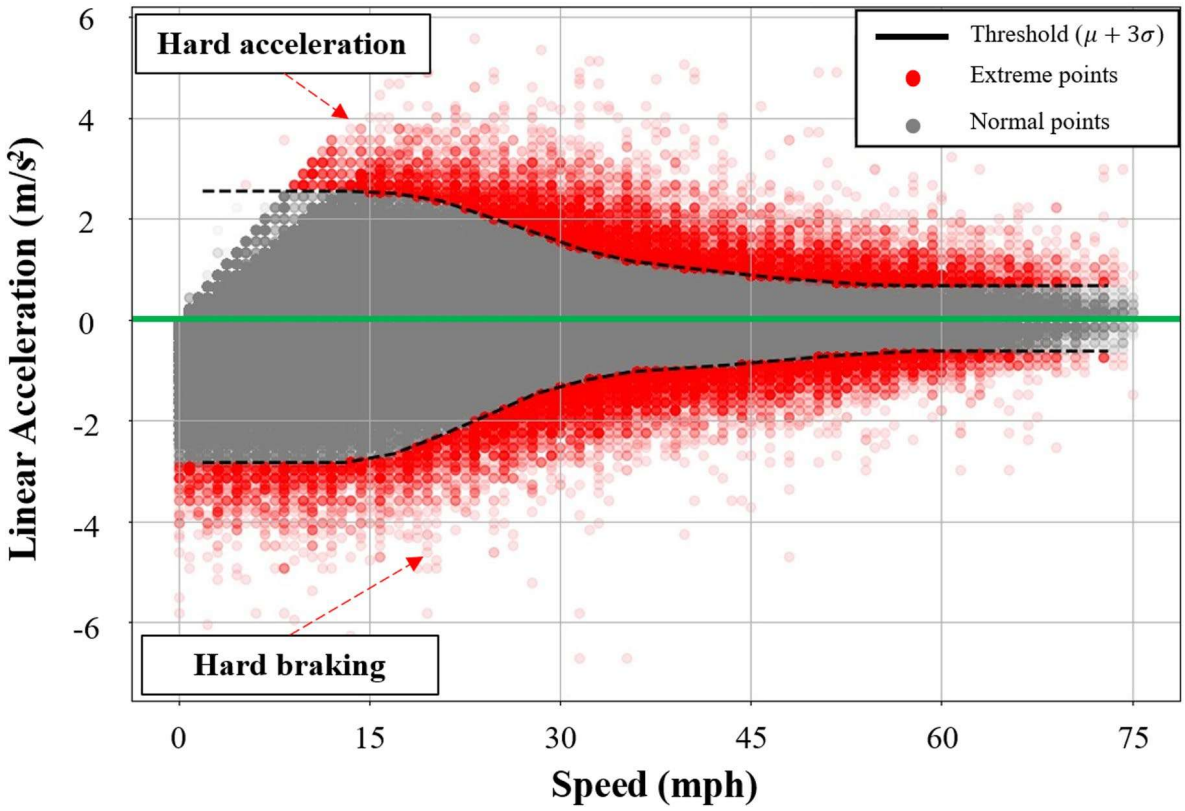


**Fig. 6**. Hard acceleration and braking identification with dynamic threshold.

| | Hard acceleration | Hard braking | Other |
|---|---|---|---|
| Left LC | Left LC with hard acceleration counts | Left LC with hard braking counts | Left LC counts |
| Right LC | Right LC with hard acceleration counts | Right LC with hard braking counts | Right LC counts |
| Non-LC | Hard acceleration counts | Hard braking counts | - |

**Table 1**. Counts of 8 risky driving behaviors (LC: lane-changing).

Moreover, drivers may perform lane-changing maneuvers during instances of hard braking or acceleration, which may further increase the risk of lateral conflicts (sideswipe) rather than longitudinal conflicts (rear-end). Considering their differing risk outcomes, a total of 8 risky driving behaviors are identified at the unidirectional segment level by combining two lane-changing types with hard events, as shown in Table 1.

(3) *Driving Volatility features.*

In this study, we also derived driving volatility measures for segment crash analysis. Driving volatility is a critical safety measure to reflect instantaneous driving decisions (such as variation in speed) when a vehicle is being driven at a specific roadway location. Several driving volatility metrics have been proposed and identified as leading indicator for understanding the occurrence of unsafe outcomes, such as conflicts or crashes[20,46,47]. Commonly, these metrics (e.g., speed standard deviation or coefficient of variation) are calculated at the individual vehicle level[14,46,48] and then aggregated at the segment or intersection level to represent overall safety level of the infrastructure[20,40].

For a segment between two adjacent intersections, vehicle trajectories are significantly affected by their arrival times relative to the downstream intersection signal timing to present significantly different patterns. Based on their speed profile, these trajectories can be mainly classified into three types as shown in Fig. 7: free-flow (FF), stop-and-go (SG), and interrupted slow (IS).

- The free-flow trajectory means that a vehicle passes through the segment freely without slowing down or stops, typically for those arriving at the intersection during the green time.
- The stop-and-go trajectory involves at least one stop due to segment traffic congestion or arriving at the red signal phase.
- The interrupted slow trajectory reflects vehicles that slow down significantly due to traffic congestion but continue to move forward without stopping.

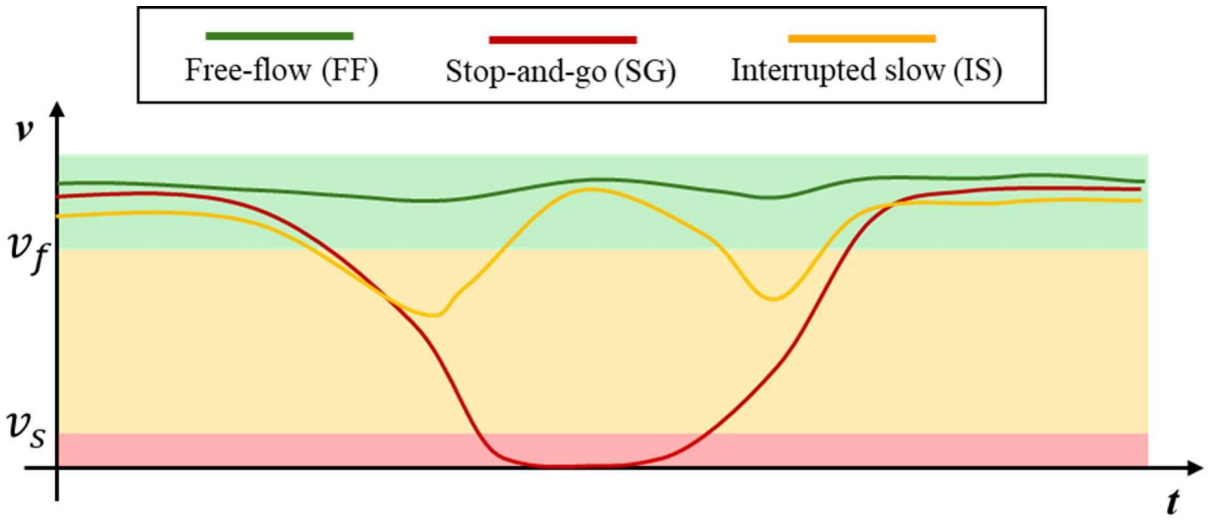


**Fig. 7**. Three patterns of CV trajectories within a segment.

| Features | Description | Formula* |
|---|---|---|
| Ratio | Ratio of each type of trajectories among all trajectories | $\frac{\#\ of\ c(\boldsymbol{Tr}_j)}{\#\ of\ \boldsymbol{Tr}_j} * 100$ |
| Speed standard deviation | Mean of the speed standard deviation of each trajectory $j$ ( $std_j$) | $\frac{1}{N} * \sum_{j=1}^{N} std_j, std_j = \sqrt{\frac{1}{n_j} \sum_{i=1}^{n_j} (v_i - \bar{v_i})^2}$ |
| Cumulative acceleration | Mean of the sum of acceleration values of each trajectory $j$ ( $s\,(a)_j$) | $\frac{1}{N} * \sum_{j=1}^{N} s\,(a)_j, s\,(a)_j = \sum_{i=1}^{n_j} \lvert a_i \rvert$ |

**Table 2**. Four driving volatility metrics. *: $N$ is the total number of trajectories on the segment, $n_j$ is the total number of trajectory points within trajectory $j$.

| Features | Description | Formula* |
|---|---|---|
| Speed difference | Mean of the difference between the max and min of speed of each trajectory $j$ ( $ds_j$) | $\frac{1}{N} * \sum_{j=1}^{N} ds_j,$ $ds_j = \max\left(v_{i\epsilon[1,n_j]}\right) - \min\left(v_{i\epsilon[1,n_j]}\right)$ |

However, existing studies typically aggregate all trajectories to calculate driving volatility metrics, ignoring these differences in trajectory patterns. To more precisely characterize driving volatility features at the segment level, CV trajectories are first divided into these three groups based on their minimum speed $v_{min} = \min\left\{v_{i\in\,[1,n_v]}\right\}$:

$$\begin{cases} c\,(\boldsymbol{Tr}_v) \triangleq FF,\ if\ v_{min} \geq v_f \\ c\,(\boldsymbol{Tr}_v) \triangleq IS,\ if\ v_s < v_{min} < v_f \\ c\,(\boldsymbol{Tr}_v) \triangleq SG,\ if v_{min} \leq v_s \end{cases} \tag{13}$$

where $v_f$ and $v_s$ are the free-flow and stop speed threshold, respectively. Based on existing studies[13,14], $v_s$ is chosen as 5 m/s, and $v_f$ is set as 80% of speed limit.

For each type of trajectory, four driving volatility metrics are calculated (see Table 2). Specifically, the speed standard deviation and cumulative acceleration of a trajectory represent the magnitude of speed fluctuation as the vehicle traverses the segment. The speed difference reflects the range of these fluctuation[20]. Therefore, a total of 3*4 = 12 driving volatility metrics can be extracted for each individual segment.

(4) *Speeding-related features.*

Speeding is another risky driving behavior that can prolong the braking distance and reduce the human reaction time which may led to increase potential crash risk[49,50]. In this study, a vehicle is identified as speeding if its speeds exceeding 10% above the speed limit ( $v_{speed}$) [50]. Furthermore, if a speeding vehicle has both hard acceleration and braking events as shown in Fig. 8, it is labeled as a risky speeding trajectory as they may indicate aggressive driving behavior characterized by rapid speed fluctuation and excessive speeding. Therefore, the count of speeding vehicles and risky speeding trajectories are utilized as two speeding-related features in this study.

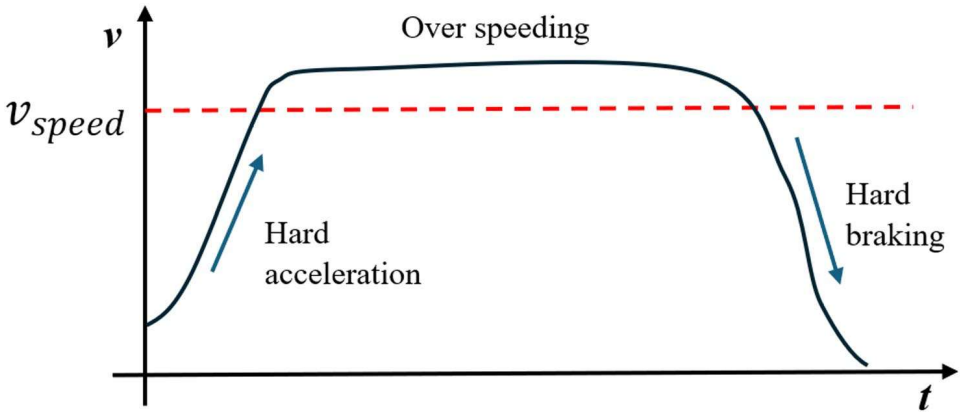


**Fig. 8**. The speed profile of risky speeding trajectory.

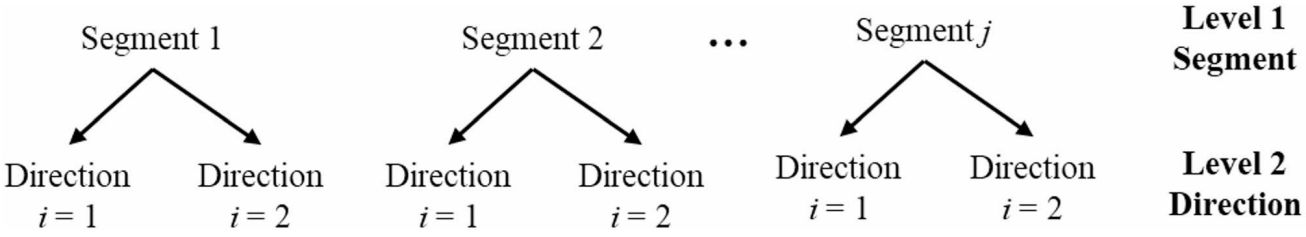


**Fig. 9**. Hierarchical structure of the data.

## Segment crash frequency modeling method

This study focuses on three types of crashes: rear-end (RE), sideswipe (SW), and speeding crashes. RE and SW crashes are two primary crash types occurring on segments due to the failure in driving interaction between vehicles at longitudinal and lateral movements[24,51]. Speeding crashes, another critical safety issue, occur more frequently on segments and always lead to more severe outcomes[52]. Given the strong correlation observed between RE and SW crashes, a Bivariate Random-Parameter Hierarchical Negative Binomial (BRPHNB) model is developed to jointly estimate the heterogeneous impact of risky driving behavior on their frequencies while accounting for potential shared features. Additionally, a Zero-Inflated Hierarchical Poisson (ZIHP) model is estimated to investigate the potential impact factors of speeding crashes.

*Bivariate random-parameter hierarchical negative binomial model*

Instead of the existing aggregated bi-directional segment crash analysis, this study considers each unidirectional segment as the basic modeling unit. Therefore, unidirectional segment-specific features, such as micro driving behavior metrics, are utilized for each direction. At the same time, some features, such as road length and speed limit, are shared between both directions of a segment. To handle such data structure, a two-level hierarchy structure is developed as shown in Fig. 9, where level one represents segment-level, and level two represents the single direction-level.

Accordingly, the Bivariate RPHNB model is specified as follows:

$$y_{ijk} \sim NB(\lambda_{ijk}, \alpha_k) \tag{14}$$

where $y_{ijk}$ is the crash frequencies at the direction $i$ = 1, 2 from segment $j$ = 1, 2, …, $m$ of crash type $k$ = 1 (SW), 2 (RE). $\lambda_{ijk}$ is the expected number of crashes at the direction $i$ from segment $j$ of crash type $k$. $\alpha_k$ is the dispersion parameter for crash type $k$ in negative binomial (NB) distribution to address the overdispersion of crash frequencies. To address the hierarchical data structure and unobserved heterogeneity, $\lambda_{ijk}$ can be expressed as:

$$\ln(\lambda_{ijk}) = Dir_{ijk} + Seg_{jk} + \epsilon_{ijk} \tag{15a}$$

$$Dir_{ijk} = \beta_{Dir,k} X_{Dir,ij} \tag{15b}$$

$$Seg_{jk} = \beta_{Seg,k} X_{Seg,j} + u_{jk},\ u_{jk} \sim N(0, \rho_k^2) \tag{15c}$$

$$\beta_{Seg/Dir,k} = b_{Seg/Dir,k} + \omega_{jk},\ \omega_{jk} \sim N\left(0, \sigma_k^2\right) \tag{15d}$$

$$\epsilon_{ijk} \sim \mathrm{MN}(0, \Sigma), \ \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1 \sigma_2 \rho_{12} \\ \sigma_1 \sigma_2 \rho_{21} & \sigma_2^2 \end{bmatrix} \tag{15e}$$

where $Dir_{ijk}$ and $Seg_{jk}$ are the unidirectional and segment level parameters, respectively. The former are determined by direction-level variables $X_{Dir,ij}$ and coefficients $\beta_{Dir,k}$. While the latter are derived from segment-level variables $X_{Seg,j}$, coefficients $\beta_{Seg,k}$, and random effects across segments $u_{jk}$, which have a mean of 0 and a variance of $\rho_k^2$. Noting that $\beta_{Seg/Dir,k}$ are setting as random parameters to capture the unobserved heterogeneity. $b_{Seg/Dir,k}$ are the mean of coefficient estimation and $\omega_{jk}$ is a randomly distributed term with a mean of 0 and a variance of $\sigma_k^2$. In addition, the error term $\epsilon_{ijk}$ is multivariate normally distributed with zero mean, variance $\sigma_k^2$, and correlation $\rho_{12} = \rho_{21}$ that represent the correlation between the error term of crash type SW and RE.

*Zero-inflated hierarchical Poisson model*

Given the high number of zero observations in speeding crash frequency, the ZIHP model for speeding crash analysis is developed:

$$P(y_{ij}) = \begin{cases} \pi_{ij} + (1 - \pi_{ij}) e^{-\lambda_{ij}}; y_{ij} = 0 \\ (1 - \pi_{ij}) \frac{e^{-\lambda_{ij}} * \lambda_{ij}^{y_{ij}}}{y_{ij}!}; \ y_{ij} > 0 \end{cases} \tag{16}$$

where $y_{ij}$ is the speeding crash frequencies at the direction $i$ from segment $j$. $\pi_{ij}$ is the probability that direction $i$ from segment $j$ will exist in the zero-crash state. $\lambda_{ij}$ is the expected number of speeding crashes at the direction $i$ from segment $j$. Similarly, a hierarchical model structure can be used for $\pi_{ij}$ and $\lambda_{ij}$:

$$\mathrm{logit}(\pi_{ij}) = \ln\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_{zero} Z_{Seg/Dir,ij} \tag{17a}$$

$$\ln(\lambda_{ij}) = \beta_{Dir} X_{Dir,ij} + \beta_{Seg} X_{Seg,j} + u_j, \ u_j \sim N(0, \rho_s^2) \tag{17b}$$

where $Z_{Seg/Dir,ij}$ and $\beta_{zero}$ are zero-inflated-related explanation variables and coefficients, respectively. While $u_j$ is a random effect parameter across segments, which have a mean of 0 and a variance of $\rho_s^2$.

*Model fitting and performance evaluation*

In this study, models were estimated in the Full Bayesian Framework using the Markov chain Monte Carlo (MCMC) simulation[53]. The R package "brms" was used to run three Markov chains for each parameter for total 20,000 iterations. The first 5,000 iterations were discarded as burn-in runs to exclude unstable iterations. To evaluate the model fit and performance, four measures were utilized in this study:

- Deviance Information Criteria (DIC) has been widely utilized as a local goodness of fit measure to compare the Bayesian models. Generally, a model with lower DIC is considered superior among the candidate models.
- Two global goodness of fit measures include the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE). The lower the values of MAE and MSE indicate a better the model in terms of prediction performance.
- Bayesian $R^2$ is used to evaluate the goodness-of-fit in the Bayesian framework[54]. Unlike classical $R^2$, Bayesian $R^2$ accounts for the uncertainty in parameter estimates by using the posterior distribution of the model. A greater Bayesian $R^2$ means a better model fitness.

## Data Preparation

In this study, eight arterials in Hillsborough County, Florida, were selected as they have been identified as top in the High Injury Network as shown in Fig. 10. Each arterial was divided into unidirectional segments (e.g., west- and east-bound) based on intersections, resulting in total 212 segments with a total length of 48.15*2 = 96.30 miles. More detailed information (e.g., speed limits and lane numbers) is shown in Table 3.

### Segment crash matching

The crash data, covering the three-year period from June 2021 to May 2024, were obtained from the Signal Four Analytics (S4A) system, which includes all crash records in Florida. For each crash, it has the detailed crash time, location, type, severity, number of vehicles involved, and other detailed information. To match these crashes with individual segments, the closet crashes to each segment were first detected. However, positional inaccuracies at crash locations lead to several false matches. To address this issue, the traveling direction of the crash vehicles was examined from the crash reports and then adjusted to align with the corresponding directional segments, as illustrated in Fig. 11(a). Finally, the frequencies of SW, RE, and speeding crashes of unidirectional segments can be obtained. Figure 11(b) and Fig. 11(c) show the spatial distribution of SW and RE crashes while the distribution of speeding crashes is not displayed due to space constraints.

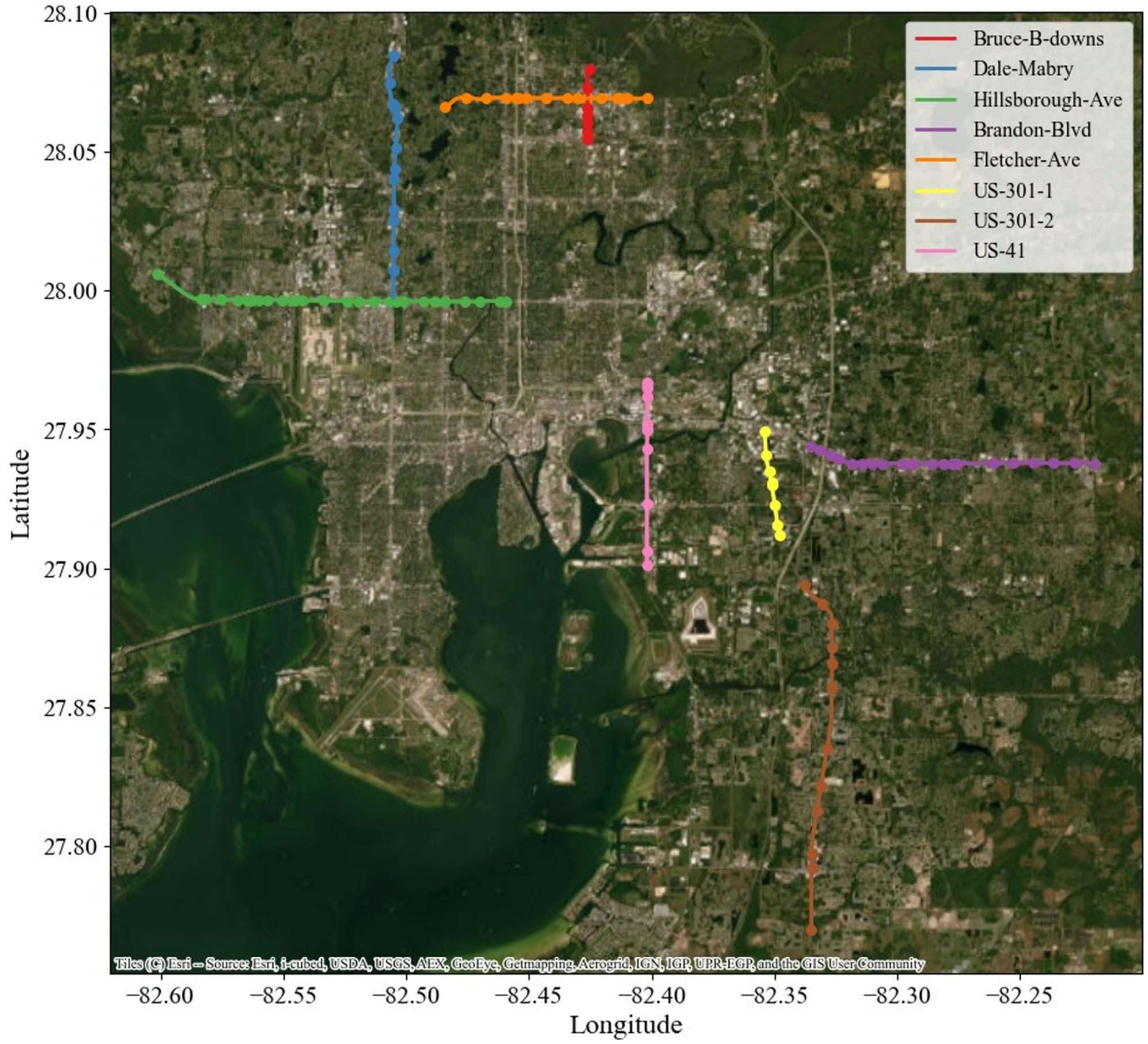


**Fig. 10**. Eight studied arterials in Hillsborough County, Florida. Generated using ArcGIS 10.8 (Environmental Systems Research Institute, USA. https://www.esri.com/).

| Name | Length (mile) | # of segments | Speed limit (mph) | Lane number* |
|---|---|---|---|---|
| Bruce B downs | 1.77 | 14 | 45 | 3 |
| Dale Mabry | 6.23 | 24 | 45 | 3,4 |
| Hillsborough Ave | 10.04 | 52 | 40,45,50 | 2,3 |
| Brandon Blvd | 8.12 | 38 | 45,50,55 | 2,3,4 |
| Fletcher Ave | 5.79 | 30 | 35,40,45 | 2 |
| US 301-1 | 2.65 | 14 | 50 | 2,3 |
| US 301-2 | 8.98 | 22 | 45,50,55 | 3 |
| US 41 | 4.57 | 18 | 40,45,50,55 | 2,3 |

**Table 3**. Eight studied arterial information. *: lane number per direction.

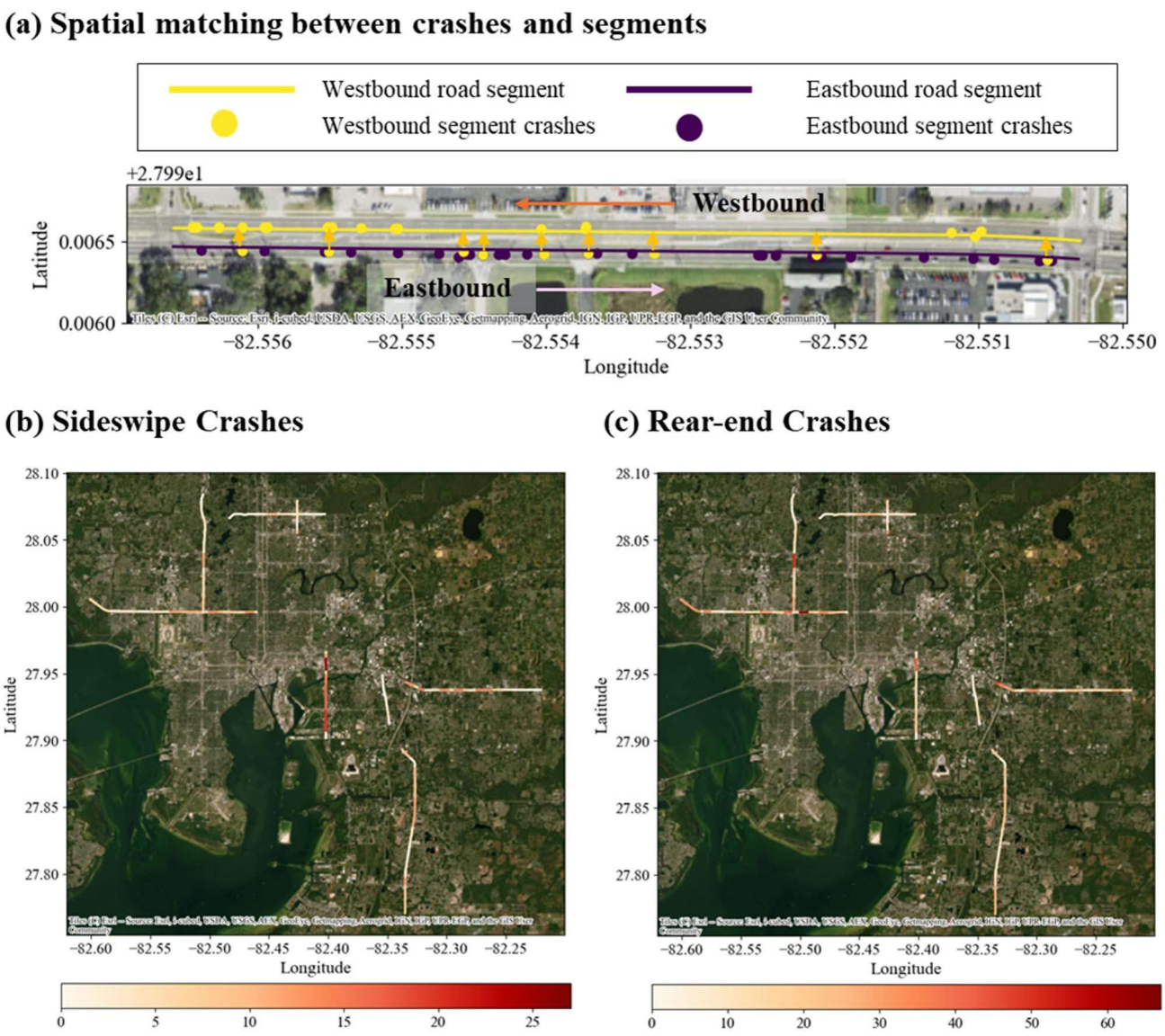


**Fig. 11**. Examples of segment crash matching. Generated using ArcGIS 10.8 (Environmental Systems Research Institute, USA. https://www.esri.com/).

| Parameters | Description | Unit |
|---|---|---|
| Journey Id | Unique identifier for a trip (from ignition start to end). | - |
| Capture time | 10-digitial UTC timestamp | s |
| Latitude | North-South position of the vehicle | - |
| Longitude | East-West position of the vehicle | - |
| Heading | The heading of the vehicle travel. (e.g., 0: North; 90: East) | ° |
| Speed | Speed of the vehicle at the instant the datapoint was captured | MPH |

**Table 4**. Description of the Raw CV data parameters.

## CV trajectory data

In this study, the CV data is provided by Streetlight company, which contains 3-second-interval vehicle trajectories from vehicle original equipment manufacturers (OEMs). As described in Table 4, the CV data includes journey Id, capture time, GPS location, heading, and speed. The data spans two distinct periods: Jan 3–13 and Jan 30 - Feb 8. On average, it includes over 4,692,975 CV trajectory points per day, derived from 154,997 journeys, providing full coverage of Hillsborough County, as shown in Fig. 12(a). As for CV market penetration, the data provider, Streetlight, claims an average penetration of 3–5%. Our independent validation with roadside detectors found

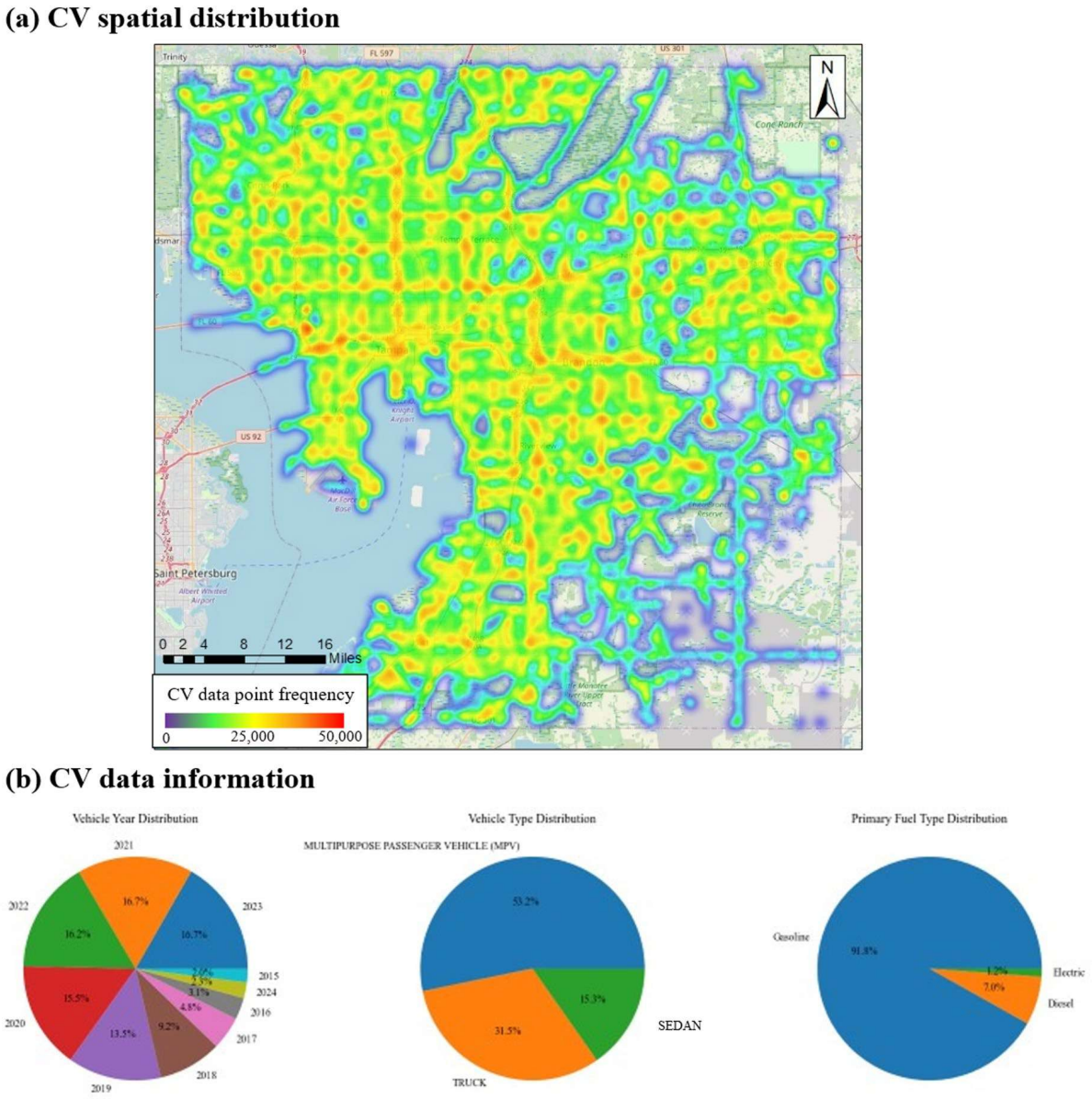


**Fig. 12**. CV data spatial distribution and information. Generated using ArcGIS 10.8 (Environmental Systems Research Institute, USA. https://www.esri.com/).

that the average market penetration is 4.17 ± 1.2%, which is a relatively high level among existing studies[15,16]. Such high market penetration supports the CV dataset to better capture driving behaviors in the studied areas. Moreover, given the CV dataset is composed of multiple types of personal, non-commercial vehicles with rich vehicle types, it can better represent the vehicles on the roadways to reduce sample bias[55]. Figure 12(b) presents vehicle year, vehicle type, and fuel type in the used CV data. The dataset covers vehicles produced within the last 10 years and encompasses a wide variety of vehicle styles. Over 53% are multipurpose passenger vehicles (e.g., SUV, Vans, and Minivans), 31.5% are large-size pickup trucks, and 15.3% are sedans. Regarding fuel type, 91.8% are on gasoline, 7% on diesel, and only 1.2% are electric vehicles (EVs). While driver identifiers are removed for data privacy, the dataset is expected to include drivers across genders and different age groups. Given the potential noise and errors in the raw CV trajectories, the outlier points are detected (e.g., speed >120mph) and a simple Gaussian filter is applied to smooth the raw data as suggested by existing studies[13,14]. Then, a total of 23 micro-level driving behavior features were extracted from the processed CV data using the methods in the Methodology section.

## Macro-level segment features

The macro-level segment features represent the static road design and traffic characteristics of segments. Referring to existing studies[8,56], ten macro-level segment features (see in Table 5) were identified and categorized into two types:

| Variable | Definition | Min | Max | Mean | STD |
|---|---|---|---|---|---|
| ***Segment-level features*** ( $X_{Seg}$ ): | | | | | |
| Log AADT | Log of Bi-directional AADT (pcu/day) | 9.61 | 11.23 | 10.84 | 0.25 |
| Log road length | Log of segment length (mile) | 3.97 | 7.34 | 5.90 | 0.71 |
| Speed limit | The post speed limit at segment (mph) | 35 | 55 | 45.90 | 4.61 |
| Median separation* | 1: Physical separation (e.g., concrete median); 0: Non-physical separation | | | | |
| Context classification* | C3R: Suburban residential; C3C: Suburban commercial; C4: Urban area | | | | |
| ***Unidirectional-level features*** ( $X_{Dir}$ ): | | | | | |
| Lane number* | Number of lanes in a single direction: 2, 3, 4 | | | | |
| Access point counts | Counts of access points, referring to entrance to sounding buildings (e.g., shopping mall) | 0 | 21 | 4.54 | 4.26 |
| Non-signal intersection counts | Counts of non-signal intersections connected with low-level roadways | 0 | 10 | 1.35 | 1.69 |
| Median turn counts | Counts of median opening allowing turns toward the opposite side of road | 0 | 10 | 1.29 | 1.45 |
| Bus station counts | Counts of bus stations on the segment | 0 | 4 | 0.96 | 1.10 |

**Table 5**. Summary of macro-level segment features. *: Categorical variables.

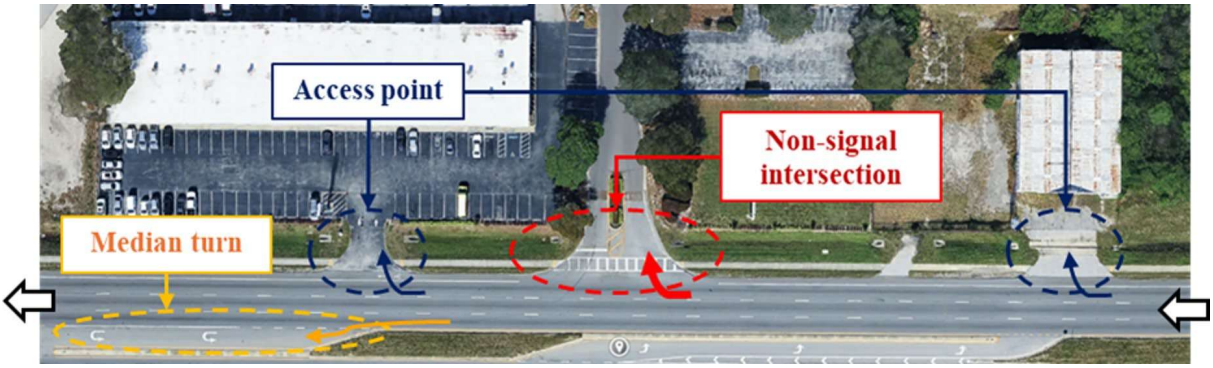


**Fig. 13**. Examples of access point, non-signal intersection and median turn. Generated using ArcGIS 10.8 (Environmental Systems Research Institute, USA. https://www.esri.com/).

- ***Segment-level features*** ( $X_{Seg}$ ): These features are shared by both directions of a segment to affect bidirectional traffic and crashes, including AADT, road length, speed limit, median separation type, and context classification.
- ***Unidirectional-level features*** ( $X_{Dir}$ ): These features are specific to each segment direction and have significant impacts on single-direction traffic safety. They encompass the lane number, counts of access points, non-signal intersections, median turns (as illustrations in Fig. 13), and bus station number.

## Results

### Lane-changing behavior analysis

Figure 14 presents an example of lane-changing (LC) behavior analysis for a unidirectional segment on Bruce-B-downs Blvd. Figure 14(a) shows the LC points along the roadway, with red points reflect right LCs and blue for left LCs. A clear spatial clustering pattern of LC points is observed as shown in Fig. 14(b) and (c). Specifically, hotpots of left LC points are clustered near the median turn lane, meaning that most vehicles trend to make left LCs to accesss the oppsite direction. In contrast, right LC hotpots are clustered near the right-turn lane at the intersection, indicating that vehicles commonly perform right LCs in preparation for a right turn at the intersection. Figure 14(d) also confirms these spatial trends: red LC points are concentrated at 80–220 m (median turn area), while right LC points are aggregated at 300–350 m range (right-turn lane area). Figure 14(e) shows the temporal differences of two LC types: left LCs mainly occur at midday (11:00–13:00), while right LCs exhibit pronounced morning (8:00–9:00) and evening peak (17:00–18:00).

### Segment crash frequency model results

*Rear-end and sideswipe crashes jointly model*

Before discussing the model results, the models' performance is compared in Table 6. Compared to the separate NB model (correlation coefficient between RE and SW crashes is set 0), the Bivariate NB model has better model fit (lower DIC) and improved predictive performance (lower MAE and RMSE, and higher Bayesian $R^2$). It highlights the necessity of jointly estimating RE and SW crashes to capture their correlation and shared variable impact. Furthermore, the Bivariate HNB model exhibits significant improvement over the Bivariate NB model, including a 7.8% reduction in DIC and over a 30% decrease in both MAE and RMSE for RE and SW crashes. These findings confirm that the introduction of hierarchical modeling structure helps to fit the heterogeneous

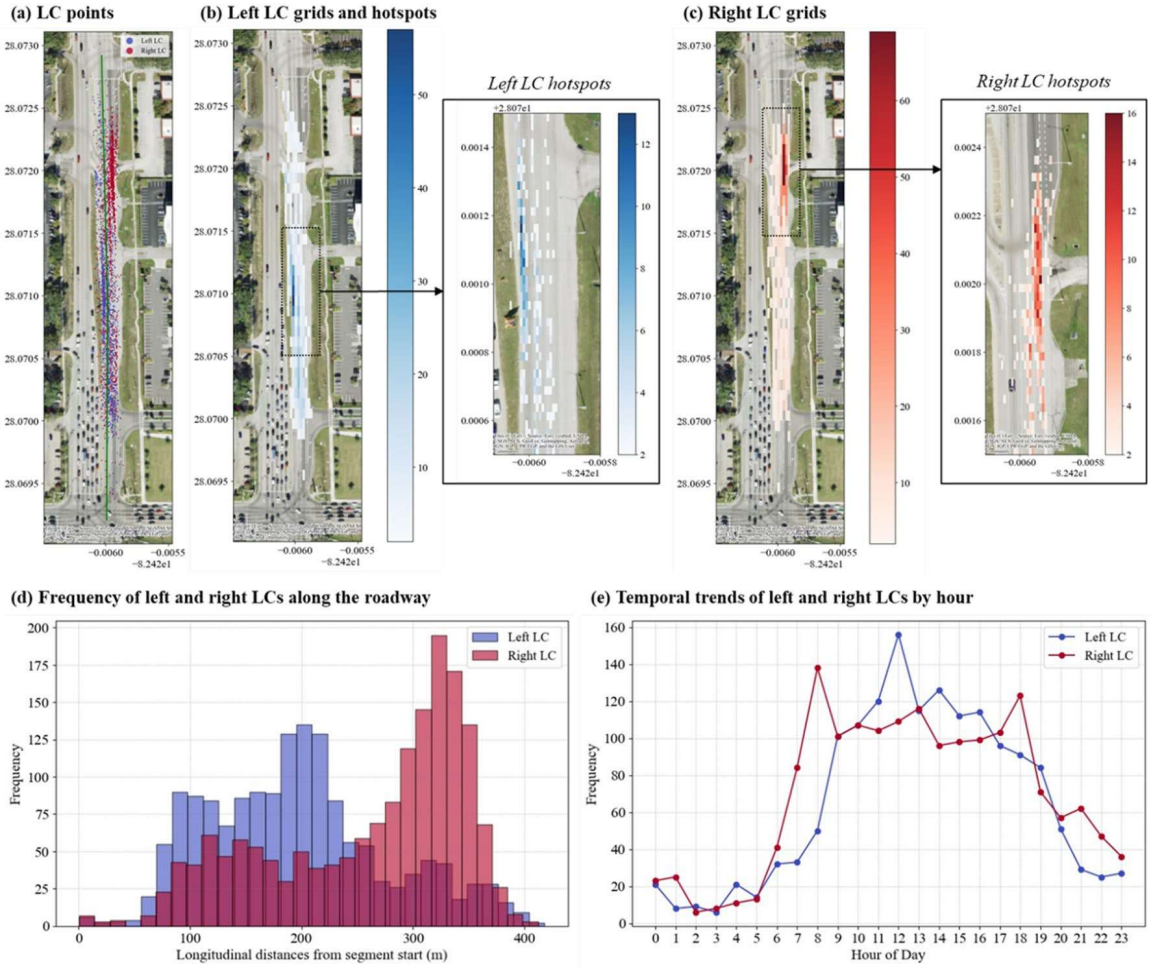


**Fig. 14**. Lane-changing behavior analysis for a unidirectional segment: Bruce-B-downs 0107. The map was generated using ArcGIS 10.8 (Environmental Systems Research Institute, USA. https://www.esri.com/).

| Models | | Separate negative binomial (NB) | Bivariate NB | Bivariate hierarchical NB (HNB) | Bivariate Random-Parameter HNB |
|---|---|---|---|---|---|
| DIC | | 4981.3 | 4945.7 | 4561.1 | 4434.4 |
| Rear-end (RE) crashes | MAE | 7.179 | 7.125 | 4.425 | 3.166 |
| | RMSE | 10.168 | 10.075 | 6.284 | 4.436 |
| | Bayesian $R^2$ | 0.412 | 0.427 | 0.619 | 0.701 |
| Sideswipe (SW) crashes | MAE | 2.837 | 2.658 | 1.730 | 1.546 |
| | RMSE | 4.095 | 3.796 | 2.273 | 2.041 |
| | Bayesian $R^2$ | 0.469 | 0.518 | 0.685 | 0.707 |

**Table 6**. Model performance comparison.

traffic and driving behavior patterns between two different directions. Finally, the Bivariate HNB model with random parameters (i.e., Bivariate RPHNB model) achieves the lowest DIC, reflecting the best model goodness-of-fit. It also demonstrates the lowest MAE and RMSE as well as the highest Bayesian $R^2$ for both RE and SW crashes. This suggests the presence of unobserved heterogeneity during the modeling process —specifically, the impact of certain variables is not fixed but vary among samples. Overall, these results confirm that the Bivariate RPHNB model is well suited for segment-level crash frequency modeling at detailed directional level, effectively capturing both the shared influences on RE and SW crashes and the potential heterogeneous effects.

Table 7 summarizes the estimated parameters of the Bivariate RPHNB model, highlighting 9 significant variables at the 90% Bayesian Credible Interval level. Among the macro-level segment features, C4 context classification (urban area), lane number of 3, and access point count are significantly related to both RE and SW crashes. In contrast, C3R context classification (suburban resident area) is significant only for SW crashes, while the count of median turning is significant only for RE crashes. For the micro-level driving behavior features, the natural log of traffic volume and the ratio of free-flow trajectories at segments are significant for both RE and

| Variables | Rear-end Crashes | | Sideswipe Crashes | |
|---|---|---|---|---|
| | Mean (S.D.)[1] | 90% BCI[2] | Mean (S.D.)[1] | 90% BCI[2] |
| Constant | -5.05 (1.13) | (-6.88, -3.23) | -6.00 (1.33) | (-8.21, -3.85) |
| ***Macro-level segment features*** | | | | |
| *C4 (base: C3C)* | 0.44 (0.13) | (0.22, 0.66) | 0.59 (0.14) | (0.37, 0.82) |
| *S.D. of C4* | *0.28 (0.13)* | *(0.04, 0.48)* | *0.17 (0.12)* | *(0.02, 0.41)* |
| C3R *(base: C3C)* | - | - | -0.51 (0.21) | (-0.87, -0.16) |
| *S.D. of C3R* | - | - | *0.25 (0.18)* | *(0.02, 0.58)* |
| Lane Number = 3 *(base: Lane Number = 2)* | 0.22 (0.12) | (0.02, 0.42) | 0.53 (0.13) | (0.32, 0.74) |
| Access point counts | 0.03 (0.01) | (0.01, 0.05) | 0.03 (0.01) | (0.01, 0.05) |
| Median turn counts | 0.11 (0.05) | (0.03, 0.18) | - | - |
| ***Micro-level Driving behavior features*** | | | | |
| Ln (Traffic Volume) | 0.68 (0.13) | (0.46, 0.89) | 0.63 (0.15) | (0.40, 0.87) |
| Ratio of free-flow (FF) trajectories | -1.68 (0.38) | (-2.32, -1.05) | -1.14 (0.39) | (-1.81, -0.51) |
| *S.D. of Ratio of FF trajectories* | *0.43 (0.28)* | *(0.04, 0.96)* | *0.32 (0.25)* | *(0.02, 0.81)* |
| Cumulative acceleration of stop-and-go (SG) trajectories | 0.06 (0.04) | (0.01, 0.13) | - | - |
| Right LC with hard acceleration counts | - | - | 0.23 (0.10) | (0.07, 0.39) |
| *S.D. of Right LC with hard acceleration counts* | - | - | *0.06 (0.04)* | *(0.01, 0.15)* |
| ***Model parameters*** | | | | |
| Segment level random effect $\rho_k$ | 0.43 (0.06) | (0.32, 0.53) | 0.39 (0.09) | (0.24, 0.52) |
| Dispersion parameter $\alpha_k$ | 8.78 (2.78) | (5.17, 13.86) | 9.90 (3.11) | (5.69, 15.58) |
| Correlation $\rho_{12} = \rho_{21}$ | 0.619 | | | |

**Table 7**. Estimated parameters of bayesian bivariate RPHNB modelt. 1:Standard deviation. 2: Bayesian Credible Interval.

| Features | Bidirectional | | Unidirectional | | |
|---|---|---|---|---|---|
| | C4 (base: C3C) | C3R (base: C3C) | Lane Number = 3 (base: Lane Number = 2) | Access point counts | Median turn counts |
| Rear-end crashes | +7.44 | - | +3.11 | +0.33 | +1.82 |
| Sideswipe crashes | +3.46 | -2.15 | +2.61 | +0.13 | - |

**Table 8**. Marginal effects of macro-level segment features.

SW crashes. However, the cumulative acceleration of stop-and-go (SG) trajectories is only significant for RE crashes while the counts of right LC involving hard acceleration is only significant for SW crashes. Additionally, the estimated correlation coefficient $\rho_{12} = \rho_{21}$ is estimated to be 0.619, indicating a relatively high correlation between the RE and SW crashes.

(1) *Macro-level segment features.*

Table 8 illustrates the marginal effects of macro-level segment features based on the Bivariate RPHNB model. For bidirectional features, context classification emerges as a significant contributor. Compared to the C3C (suburban commercial area), C4 segments (urban area) are associated with an estimated increase of 7.44 and 3.46 in RE and SW crash frequencies, respectively. It reveals that urban areas may suffer more safety issues due to their complexed traffic environment (e.g., mixed traffic involving bikes, pedestrians and dense road network)[6,57]. Conversely, SW crashes are estimated to decrease by 2.15 in the C3R segments (suburban residential area). As for unidirectional features, 3-lane segments are estimated to have an additional 3.11 RE crashes and 2.61 SW crashes compared to 2-lane segments. Similarly, each roadside access point is associated with an increase of about 0.33 RE crashes and 0.13 SW crashes. Previous studies have shown that access traffic, typically traveling at lower speeds, can significantly disrupt normal segment traffic flow to increase the risk of RE and SW crashes[4]. These findings highlight the importance of prioritizing safety management for multi-lane segments with high density of roadside access points. Interestingly, each additional median turn is estimated to increase RE crashes by 1.82. This may be because vehicles making median turns need to slow down in the left-most lane, making the following high-speed vehicles hard to brake in time to cause RE crashes.

(2) *Micro-level driving behavior features.*

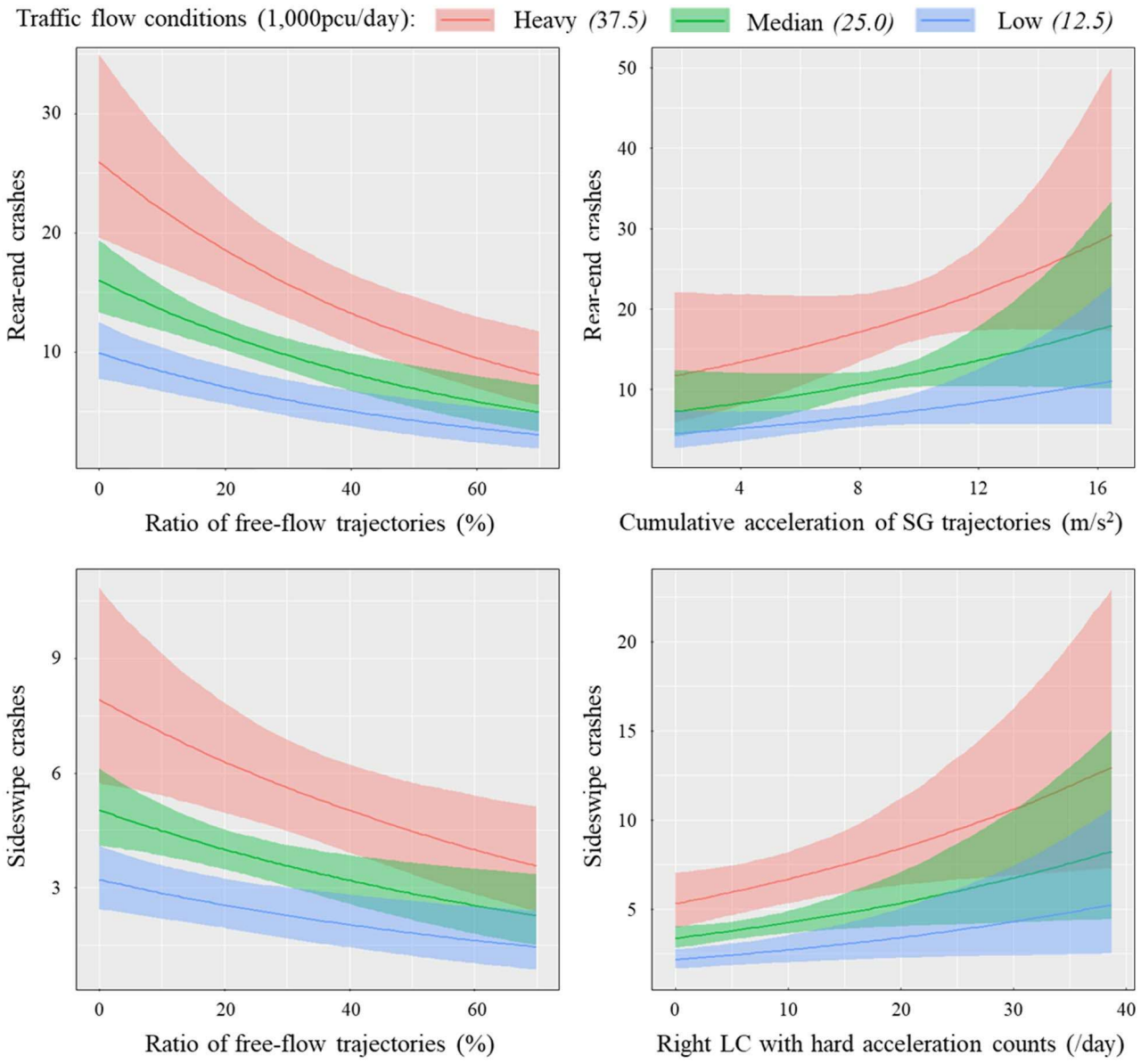


**Fig. 15**. Marginal effects of micro-level driving behavior features.

Figure 15 visualizes the marginal effects of three driving behavior features at different traffic flow conditions. Based on the results, the following conclusions can be drawn:

1) ***Ratio of free-flow trajectories***: This feature represents the proportion of trajectories that can freely pass the segment without slowdowns or stops. As it increases from 0 to over 60%, the estimated RE and SW crashes reduce sharply in different traffic situations. It indicates that if a segment has effective signal setting (e.g., coordinated green wave control) to allow more vehicles freely to pass through smoothly, it can also improve the traffic safety to reduce both RE crashes and SW crashes.
2) ***Driving volatility***: The cumulative acceleration of stop-and-go (SG) trajectories reflects the speed fluctuations of SG vehicles on a segment, where higher values indicate more frequent accelerations and braking of drivers. This metric exhibits a significant positive correlation with the frequency of RE crashes, with notably higher marginal effects under high-traffic conditions. These findings suggest that during heavy traffic periods, if a segment shows higher SG vehicles with frequent acceleration and braking, it has a higher probability of RE crashes and should be prioritized for traffic management.
3) ***Risky driving behaviors***: The count of right lane-changing with hard accelerations are found to significantly increase SW crash frequencies. It indicates that compared to left LC, right LC may pose a higher risk due to the typically larger blind spot on the right side[58]. Additionally, hard acceleration during a LC would significantly reduce the reaction time available to the following vehicle[39], therefore increase the likelihood of SW crashes. As traffic flow increases, the marginal effect of this measure also shows a notable

| Variables | ZIP | | Hierarchical ZIP | |
|---|---|---|---|---|
| | Mean (S.D.)[1] | 90% BCI[2] | Mean (S.D.)[1] | 90% BCI[2] |
| **Zero state ( $y_{ij} = 0$)** | | | | |
| Constant | 18.38 (4.08) | (12.48, 25.86) | 20.80 (5.14) | (13.38, 29.98) |
| Speed limit = 55mph (*base line: Speed limit = 45mph*) | 4.05 (1.21) | (2.27, 6.01) | 4.50 (1.94) | (2.21, 7.12) |
| Log road length | -3.14 (0.69) | (-4.41, -2.14) | -3.62 (0.88) | (-5.17, -2.35) |
| ***Poisson state*** ( $y_{ij} > 0$) | | | | |
| Constant | 6.30 (1.92) | (3.08, 9.40) | 5.67 (2.75) | (1.22, 10.16) |
| Access point counts | 0.07 (0.02) | (0.04, 0.09) | 0.06 (0.02) | (0.03, 0.09) |
| Ln (Traffic Volume) | -0.63 (0.20) | (-0.95, -0.30) | -0.59 (0.28) | (-1.04, -0.14) |
| Risky speeding trajectory counts | 0.14 (0.05) | (0.05, 0.22) | 0.17 (0.07) | (0.06, 0.29) |
| ***Model parameters*** | | | | |
| Segment level random effect $\rho_s$ | - | - | 0.50 (0.12) | (0.32, 0.71) |
| ***Model performance*** | | | | |
| DIC | 1095.2 | | **1002.3** | |
| MAE | 0.911 | | **0.716** | |
| RMSE | 1.249 | | **0.957** | |
| Bayesian $R^2$ | 0.265 | | **0.469** | |

**Table 9**. Estimated parameters and model performances of bayesian ZIP models. 1: Standard deviation. 2: Bayesian Credible Interval.

increase, indicating that the association between risky right lane changes and segment SW crashes is more pronounced under congested conditions.

*Speeding crashes model*

For speeding crash modeling, two zero-inflated Poisson (ZIP) models were estimated, with their parameters and performance metrics presented in Table 9. Compared to the traditional ZIP model, the hierarchical ZIP model demonstrates superior model goodness-of-fit. Specifically, the model DIC decreases from 1095.2 to 1002.3. The MAE and RMSE are reduced by 0.195 and 0.292, respectively, while the Bayesian $R^2$ improves from 0.265 to 0.469. Overall, these results indicate that for speeding crash modeling, the hierarchical model structure is still effective to capture the variations in traffic and driving behavior patterns between individual directions, therefore enhancing the model fit and predictive accuracy as the previous findings in RE and SW jointly modeling. It is worth noting that although the Random-Parameters hierarchical ZIP was also tested, no random parameters were found to be significant at the 90% Bayesian Credible Interval level.

Hierarchical ZIP model identifies 5 significant variables, including 2 in the zero-state and 3 at the Poisson state function. Based on the results, several conclusions can be summarized:

1) ***Zero-state variables***: A speed limit of 55mph is positively associated with the probability of a zero state. It suggests that compared to the baseline speed limit of 45 mph, segments with higher speed limits are less likely to experience speeding crashes, as the speed is already high, reducing the likelihood of vehicles exceeding the limit. In contrast, segment length shows a negative effect on the zero-state probability. In other words, longer segments trend to encourage faster driving, thereby increasing the likelihood of speeding crashes[50,52].
2) ***Poisson state variables***: The number of access points has a positive effect on increasing speeding crash frequency. This may be because vehicles entering a segment from roadside access points typically travel at lower speeds. High-speed vehicles travelling on the segment may be affected by lacking sufficient time to reduce their speed, therefore increasing the occurrence probability of speeding crashes. Conversely, the natural log of traffic volume shows a negative effect on speeding crash frequency. This is reasonable that segments with heavy traffic typically have slower traffic flow, making it difficult for vehicles to reach high speeds under such conditions. The count of risky speeding trajectories is the only driving behavior feature showing a positive effect on speeding crashes. As shown in Fig. 16, its marginal effects tend to increase as its value rises. This highlights that aggressive speeding behavior—characterized by drivers exceeding speed limits and engaging in harsh braking—significantly elevates the risk of speeding-related crashes.

*Typical traffic characteristics on Hillsborough and their safety impacts*

Considering the heterogeneous traffic conditions across regions (e.g., counties and states), we highlight two typical road and traffic characteristics in Hillsborough County that may differ from other locals. By comparing

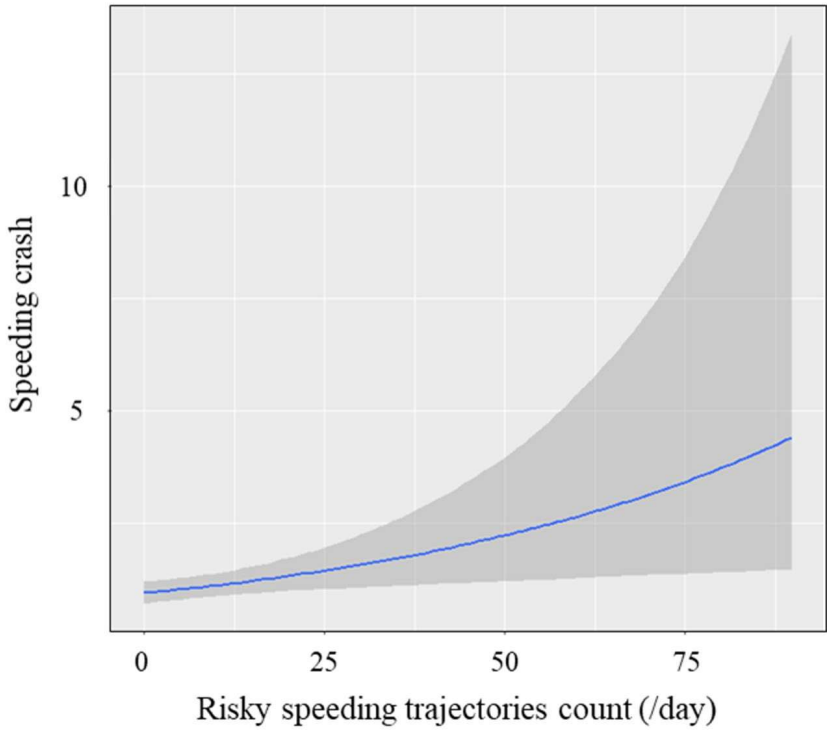


**Fig. 16**. Marginal effects of risky speeding trajectory counts on speeding crashes.

their associated crash and driving behavior patterns, these analyses provide a clearer context for our study and clarify its applicability:

(1) Compared with other counties, Hillsborough County is highly urbanized in Florida. As the urban core of the Tampa Bay metro, roughly 84% of population live in urban areas, producing a dense roadway network and heavy traffic activity along arterials. Reflecting the high level of urbanization, half of the study segments (106/212) are classified as urban corridors (C4 = 1), which typically experience more complex traffic interactions than suburban (C4 = 0) roads. Here, we compare the distributions of crashes, road length, and critical driving behaviors among the two groups. Figure 17(a)-(c) show that urban segments tend to have higher frequencies of RE, SW, and speeding crashes. Figure 17(d) shows that urban segments have much shorter segment lengths, indicating a dense network where vehicles may encounter lots of intersections and interrupted traffic. Figure 17(e–f) further indicate higher stop-and-go ratios and greater acceleration volatility on urban segments. These results reveal that the high urbanization environment in Hillsborough may intensify stop-and-go behavior and the associated acceleration volatility, thereby highlighting their impact on the segment crash frequency. Methods tailored to different contexts (e.g., long-distance, rural corridors) may therefore require different modeling and calibration, which warrants further investigation.

(2) Given Hillsborough's large population and its access-management context, main arterials are designed with multiple roadside access points as shown in Fig. 18 (a), which may substantially affect traffic operations and safety. To analyze these impacts, we classify the counts of access points into three levels: low (0–5), moderate (5–10), and high (>10), and compare the corresponding crash and risky driving behaviors as shown in Fig. 18(b)-(f). Figure 18(b)-(c) show clearly that both RE and SW crashes increase with access density. For instance, the average counts of RE crashes double from 10.5/year (0–5 group) to 22.3/year (>10 group). From the driving behavior perspective (Fig. 18(d)-(f)), higher frequencies of hard braking and risky left/right LC (i.e., LC with hard brake/accelerations) are observed on segments with moderate and high access count level. These findings indicate that segments with relatively more access points would introduce more cut-in traffic on Hillsborough arterials. Such impact on traffic flow is reflected in the elevated hard-braking and risky lane-changing indicators, which are positively associated with segment crashes in our estimated models. Similar conclusions are consistent with prior literature[4,7,8] on access management and safety, reinforcing our conclusions.

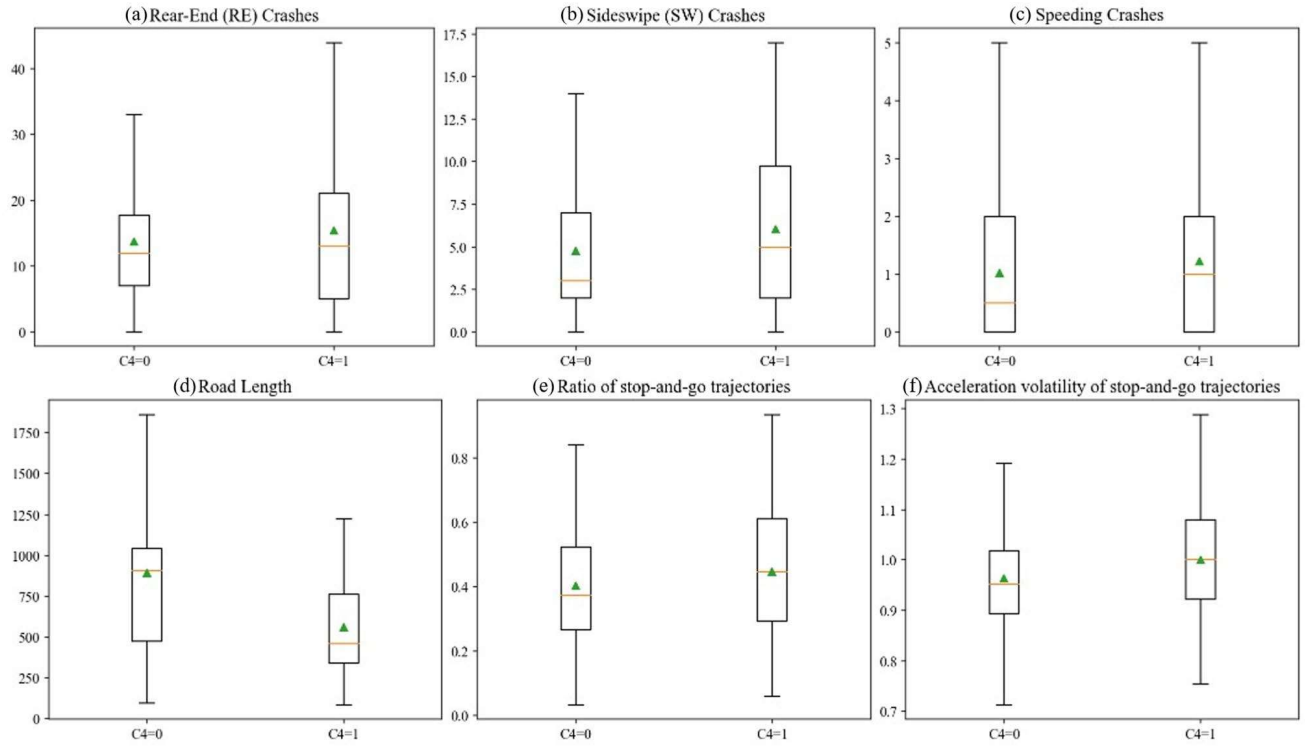


**Fig. 17**. Comparison of urban (C4 = 1) and suburban (C4 = 0) segments.

## Conclusion

This study focuses on the segment-level crash frequency modeling on urban arterial roads. Within existing studies, researchers have identified various crash contributing factors (e.g., segment geometry design and traffic volume) and explored their impacts on segment crash frequency. However, existing segment crash analyses are still being conducted at the bi-directional segment level, failing to capture the heterogeneous traffic characteristics specific to each direction —such as different access point configurations and traffic conditions—resulting in biased crash assessments. Meanwhile, due to data limitations, previous studies have focused on macro-level static infrastructure and highly aggregated traffic features, ignoring the critical influence of micro-level human driving behaviors on segment crashes.

To address these research gaps, we developed a directional-level segment crash frequency model using emerging CV data. Leveraging the advantage of CV trajectories that span the entire segment space, we extracted both macro-level segment traffic characteristics and micro-level driving behavior features for precise crash analysis at unidirectional segments. To be specific, a Constrained Gaussian Mixture Method (CGMM) was proposed to extract lane-level information from raw CV trajectories to further identify lane-changing behaviors. A total of 23 micro-level driving behavior features were quantified considering risky driving behavior, driving volatility, and risky speeding for each segment direction. Finally, a Bivariate RPHNB model was employed to jointly estimate the heterogeneous impacts of driving behavior features on RE and SW crash frequencies. While a hierarchical ZIP model was utilized to identify the significant contributors to segment speeding crashes.

High-resolution CV data at Hillsborough County were utilized for empirical experiments. Based on the results, the main findings of the study can be summarized as:

1) The proposed CGMM can effectively capture lane-level information from CV data to identify LC behaviors. At each unidirectional segment, a clear spatial clustering pattern of LC points can be observed, primarily concentrated before median turning lanes and intersection approaches.
2) Estimating correlations between RE and SW crashes and capturing heterogeneous traffic patterns in both directions, Bivariate RPHNB model demonstrates superior model fitness compared to traditional separate models, achieving a 11.0% reduction in DIC, as well as the Bayesian $R^2$ for both RE and SW crashes significant increase from 0.41 to 0.47 to 0.70 and 0.71, respectively.
3) Micro-level driving behavior features play a critical role for segment crash analysis. Segments with a high proportion of free-flow vehicles trend to experience fewer RE and SW crashes. Driving fluctuation of stop-and-go trajectories is positively related to the frequency of RE crashes, while risky right lane-changings involving hard accelerations are strongly associated with SW crashes on segments. Aggressive speeding behaviors are high related to a higher likelihood of speeding crashes.

This study highlights the benefits of using emerging CV data in arterial safety analysis. Even short-term CV data enables researchers to proactively identify hotspot segments without waiting for crashes to happen. These

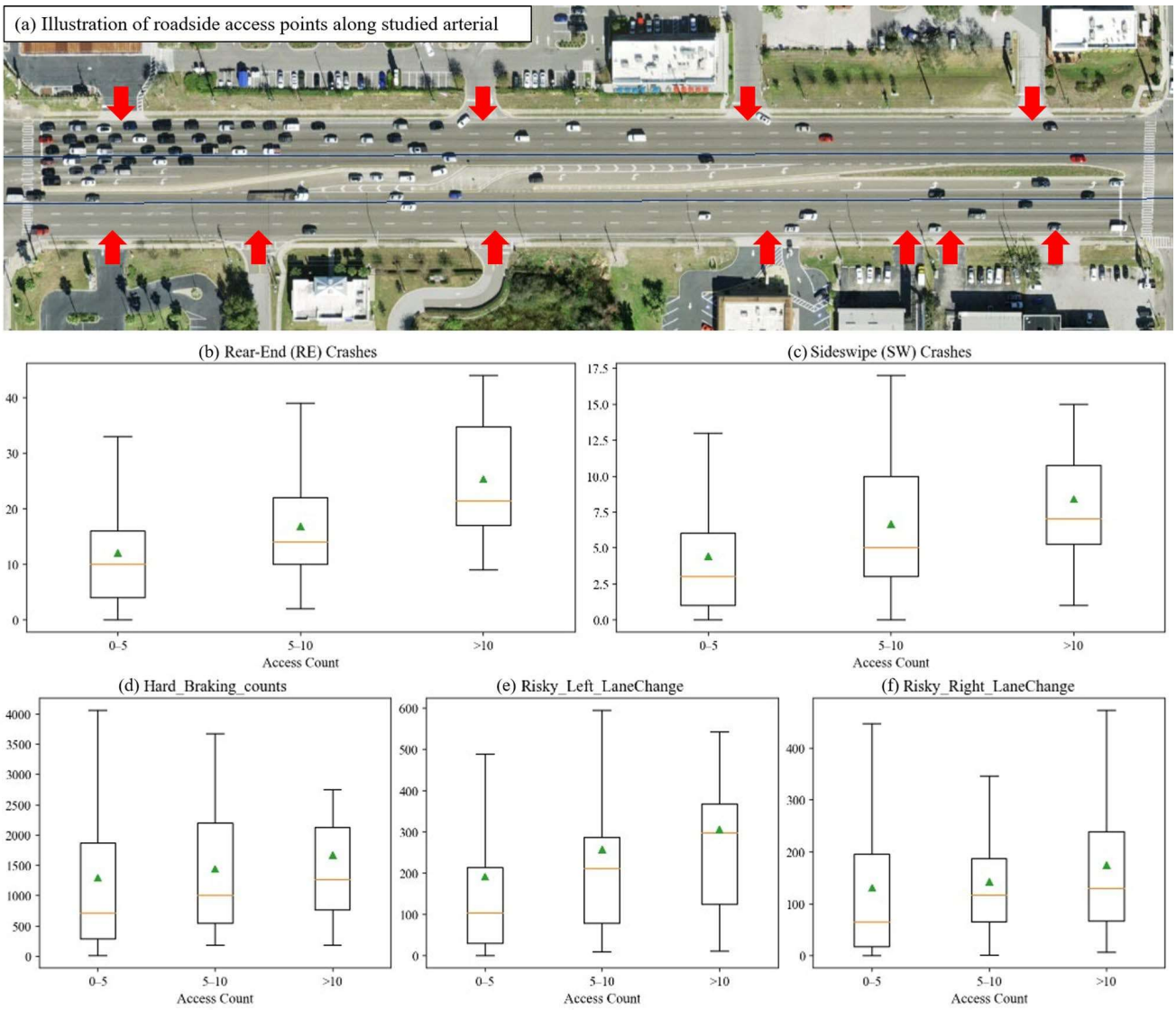


**Fig. 18**. Comparison of crashes and risky driving behaviors by access count groups. The map was generated using ArcGIS 10.8 (Environmental Systems Research Institute, USA. https://www.esri.com/).

hotspots may be those segments with few crash recordings but high risky driving events and high driving volatility[46]. On the other hand, utilizing V2I communication, proactive warnings could be generated at these high-risk segments to inform drivers about potential hazards, which could potentially enhance drivers' situational awareness and prevent crashes. Nonetheless, there are still a few limitations in the current study. Future research should consider testing the model in diverse geographic contexts and various roadway types to evaluate its generalizability. Second, the robustness of model prediction and the stability of model interpretations should be further evaluated under varying CV penetration rates.

## Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to the privacy of raw connected vehicle data, but the processed data is available from the corresponding author on reasonable request.

Received: 4 February 2025; Accepted: 2 December 2025
Published online: 09 December 2025

## References

1. Gu, Z., Bejleri, I. & Peng, B. Exploring characteristics and influencing factors of crash duration on urban arterials and collectors. *J. Transp. Saf. Secur.* **14**(9), 1470-1489 (2022).
2. Wang, X., Zhou, Q., Quddus, M., Fan, T. & Fang, S. Speed, speed variation and crash relationships for urban arterials. *Accid. Anal. Prev.* **113**, 236–243 (2018).

3. Alarifi, S. A., Abdel-Aty, M. & Lee, J. A bayesian multivariate hierarchical Spatial joint model for predicting crash counts by crash type at intersections and segments along corridors. *Accid. Anal. Prev.* **119**, 263–273 (2018).
4. Cai, Q., Abdel-Aty, M., Lee, J., Wang, L. & Wang, X. Developing a grouped random parameters multivariate Spatial model to explore zonal effects for segment and intersection crash modeling. *Anal. Methods Accid. Res.* **19**, 1–15 (2018).
5. Li, J. & Wang, X. Safety analysis of urban arterials at the meso level. *Accid. Anal. Prev.* **108**, 100–111 (2017).
6. Mahmoud, N., Abdel-Aty, M., Cai, Q. & Zheng, O. Vulnerable road users' crash hotspot identification on multi-lane arterial roads using estimated exposure and considering context classification. *Accid. Anal. Prev.* **159**, 106294 (2021).
7. Atumo, E. A., Li, H. & Jiang, X. Segment-Level Spatial heterogeneity of arterial crash frequency using locally weighted generalized linear models. *Transp. Res. Rec.* **2677**, 1637–1653 (2023).
8. Alarifi, S. A., Abdel-Aty, M. A., Lee, J. & Park, J. Crash modeling for intersections and segments along corridors: A bayesian multilevel joint model with random parameters. *Anal. Methods Accid. Res.* **16**, 48–59 (2017).
9. Li, P., Abdel-Aty, M. & Yuan, J. Real-time crash risk prediction on arterials based on LSTM-CNN. *Accid. Anal. Prev.* **135**, 105371 (2020).
10. Guo, M., Zhao, X., Yao, Y., Bi, C. & Su, Y. Application of risky driving behavior in crash detection and analysis. *Phys. Stat. Mech. Its Appl.* **591**, 126808 (2022).
11. Han, L., Yu, R., Wang, C. & Abdel-Aty, M. Transformer-based modeling of abnormal driving events for freeway crash risk evaluation. *Transp. Res. Part. C Emerg. Technol.* **165**, 104727 (2024).
12. Mannering, F., Bhat, C. R., Shankar, V. & Abdel-Aty, M. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Anal. Methods Accid. Res.* **25**, 100113 (2020).
13. Wang, X. et al. Traffic light optimization with low penetration rate vehicle trajectory data. *Nat. Commun.* **15**, 1306 (2024).
14. Han, L., & Abdel-Aty, M. Intersection crash analysis considering longitudinal and lateral risky driving behavior from connected vehicle data: A spatial machine learning approach. Accident Analysis & Prevention. 220, 108180 (2025).
15. Gupta, N. et al. Examining the relationship between connected vehicle driving event data and Police-Reported traffic crash data at the Segment- and event level. *Transp. Res. Rec.* **03611981241243329** https://doi.org/10.1177/03611981241243329 (2024).
16. Lee, T. & Rouphail, N. Enhanced crash frequency models using surrogate safety measures from connected vehicle fleet. *Transp. Res. Rec J. Transp. Res. Board.* **2678**, 463–478 (2024).
17. Han, L., Abdel-Aty, M., Yu, R., & Wang, C. LSTM + Transformer Real-Time Crash Risk Evaluation Using Traffic Flow and Risky Driving Behavior Data. IEEE Transactions on Intelligent Transportation Systems. (2024).
18. Chen, Z., Qin, X. & Shaon, M. R. R. Modeling lane-change-related crashes with lane-specific real-time traffic and weather data. *J. Intell. Transp. Syst.* **22**, 291–300 (2018).
19. Lee, C., Park, P. Y. & Abdel-Aty, M. Lane-by-Lane analysis of crash occurrence based on driver's Lane-Changing and Car-Following behavior. *J. Transp. Saf. Secur.* **3**, 108–122 (2011).
20. Wali, B., Khattak, A. J., Bozdogan, H. & Kamrani, M. How is driving volatility related to intersection safety? A bayesian heterogeneity-based analysis of instrumented vehicles data. *Transp. Res. Part. C Emerg. Technol.* **92**, 504–524 (2018).
21. Das, A., Abdel-Aty, M. & Pande, A. Genetic programming to investigate design parameters contributing to crash occurrence on urban arterials. *Transp. Res. Rec J. Transp. Res. Board.* **2147**, 25–32 (2010).
22. Fan, Y. et al. Comprehensive evaluation of signal-coordinated arterials on traffic safety. *Anal. Methods Accid. Res.* **21**, 32–43 (2019).
23. Zhang, Y., Xie, Y. & Li, L. Crash frequency analysis of different types of urban roadway segments using generalized additive model. *J. Saf. Res.* **43**, 107–114 (2012).
24. Das, A. & Abdel-Aty, M. A. A combined frequency–severity approach for the analysis of rear-end crashes on urban arterials. *Saf. Sci.* **49**, 1156–1163 (2011).
25. Das, S., Vierkant, V., Gonzalez, J. C., Kutela, B. & Sheykhfard, A. Bayesian network for motorcycle crash severity analysis. *Transp. Res. Rec.* **2677**, 51–63 (2023).
26. Mahmud, A. & Gayah, V. V. Estimation of crash type frequencies on individual collector roadway segments. *Accid. Anal. Prev.* **161**, 106345 (2021).
27. Vahedi Saheli, M. & Effati, M. Segment-Based count regression Geospatial modeling of the effect of roadside land uses on pedestrian crash frequency in rural roads. *Int. J. Intell. Transp. Syst. Res.* **19**, 347–365 (2021).
28. Xiao, D., Ding, H., Sze, N. N. & Zheng, N. Investigating built environment and traffic flow impact on crash frequency in urban road networks. *Accid. Anal. Prev.* **201**, 107561 (2024).
29. Yue, H. Investigating the influence of streetscape environmental characteristics on pedestrian crashes at intersections using street view images and explainable machine learning. *Accid. Anal. Prev.* **205**, 107693 (2024).
30. Stiles, J., Li, Y. & Miller, H. J. How does street space influence crash frequency? An analysis using segmented street view imagery. *Environ. Plan. B Urban Anal. City Sci.* **49**, 2467–2483 (2022).
31. Wu, Y. W. & Hsu, T. P. Mid-term prediction of at-fault crash driver frequency using fusion deep learning with city-level traffic violation data. *Accid. Anal. Prev.* **150**, 105910 (2021).
32. Bhowmik, T., Rahman, M., Yasmin, S. & Eluru, N. Exploring analytical, simulation-based, and hybrid model structures for multivariate crash frequency modeling. *Anal. Methods Accid. Res.* **31**, 100167 (2021).
33. Agbelie, B. R. D. K. A comparative empirical analysis of statistical models for evaluating highway segment crash frequency. *J. Traffic Transp. Eng. Engl. Ed.* **3**, 374–379 (2016).
34. Mousavi, S. M., Marzoughi, H., Parr, S. A., Wolshon, B. & Pande, A. A mixed crash frequency Estimation model for interrupted flow segments. 72–83 (2019). https://doi.org/10.1061/9780784482575.008
35. Mannering, F. L., Shankar, V. & Bhat, C. R. Unobserved heterogeneity and the statistical analysis of highway accident data. *Anal. Methods Accid. Res.* **11**, 1–16 (2016).
36. Haddad, A., Mondal, A., Eluru, N. & Bhat, C. R. A novel integrated approach to modeling and predicting crash frequency by crash event state. *Anal. Methods Accid. Res.* **41**, 100319 (2024).
37. Hosseinpour, M., Sahebi, S., Zamzuri, Z. H., Yahaya, A. S. & Ismail, N. Predicting crash frequency for multi-vehicle collision types using multivariate Poisson-lognormal Spatial model: A comparative analysis. *Accid. Anal. Prev.* **118**, 277–288 (2018).
38. Wang, C. et al. Random-parameter multivariate negative binomial regression for modeling impacts of contributing factors on the crash frequency by crash types. *Discrete Dyn. Nat. Soc.* **2020**, 6621752 (2020).
39. Arbis, D. & Dixit, V. V. Game theoretic model for lane changing: incorporating conflict risks. *Accid. Anal. Prev.* **125**, 158–164 (2019).
40. Gu, Y., Liu, D., Arvin, R., Khattak, A. J. & Han, L. D. Predicting intersection crash frequency using connected vehicle data: A framework for geographical random forest. *Accid. Anal. Prev.* **179**, 106880 (2023).
41. Chen, Y. & Krumm, J. Probabilistic modeling of traffic lanes from GPS traces. in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* 81–88 https://doi.org/10.1145/1869790.1869805 (ACM, San Jose California, 2010).
42. Shu, J. et al. Efficient Lane-Level map Building via Vehicle-Based crowdsourcing. *IEEE Trans. Intell. Transp. Syst.* **23**, 4049–4062 (2022).
43. Uduwaragoda, E. R. I. A. C. M., Perera, A. S. & Dias, S. A. D. Generating lane level road data from vehicle trajectories using Kernel Density Estimation. in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)* https://doi.org/10.1109/ITSC.2013.6728262 384–391 (2013).

44. Tang, L., Yang, X., Dong, Z. & Li, Q. C. L. R. I. C. Collecting Lane-Based road information via crowdsourcing. *IEEE Trans. Intell. Transp. Syst.* **17**, 2552–2562 (2016).
45. Li, G., Yang, Z., Pan, Y. & Ma, J. Analysing and modelling of discretionary lane change duration considering driver heterogeneity. *Transp. B Transp. Dyn.* **11**, 343–360 (2023).
46. Kamrani, M., Wali, B. & Khattak, A. J. Can data generated by connected vehicles enhance safety? Proactive approach to intersection safety management. *Transp. Res. Rec J. Transp. Res. Board.* **2659**, 80–90 (2017).
47. Khattak, Z. H., Smith, B. L., Park, H. & Fontaine, M. D. Cooperative lane control application for fully connected and automated vehicles at multilane freeways. *Transp. Res. Part. C Emerg. Technol.* **111**, 294–317 (2020).
48. Yu, R. et al. Exploring the Temporal associations between multi-type aberrant driving events and crash occurrence. *Accid. Anal. Prev.* **206**, 107698 (2024).
49. Høye, A. Speeding and impaired driving in fatal crashes—Results from in-depth investigations. *Traffic Inj Prev.* **21**, 425–430 (2020).
50. Job, R. S. & Brodie, C. Road safety evidence review: Understanding the role of speeding and speed in serious crash trauma: A case study of new Zealand. *J. Road. Saf.* **33**, 5–25 (2022).
51. Chen, C., Zhang, G., Yang, J., Milton, J. C. & Alcántara, A. Dely. An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. *Accid. Anal. Prev.* **90**, 95–107 (2016).
52. Abdel-Aty, M., Ugan, J. & Islam, Z. Exploring the influence of drivers' visual surroundings on speeding behavior. *Accid. Anal. Prev.* **198**, 107479 (2024).
53. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).
54. Gelman, A., Goodrich, B., Gabry, J. & Vehtari A. R-squared for bayesian regression models. *Am. Stat.* **73**, 307–309 (2019).
55. Zhang, S. & Abdel-Aty, M. Real-time crash potential prediction on freeways using connected vehicle data. *Anal. Methods Accid. Res.* **36**, 100239 (2022).
56. Wen, X., Xie, Y., Wu, L. & Jiang, L. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. *Accid. Anal. Prev.* **159**, 106261 (2021).
57. Al-Omari, M. M. A., Abdel-Aty, M. & Cai, Q. Crash analysis and development of safety performance functions for Florida roads in the framework of the context classification system. *J. Saf. Res.* **79**, 1–13 (2021).
58. Potts, I. B., Bauer, K. M., Torbic, D. J. & Ringert, J. F. Safety of channelized Right-Turn lanes for motor vehicles and pedestrians. *Transp. Res. Rec.* **2398**, 93–100 (2013).


## Author contributions

The authors confirm contribution to the paper as follows: study conception and design: L. H. and Dr. M. A.; data collection: L. H.; analysis and interpretation of results: L. H.; draft manuscript preparation: L. H. and Dr. M. A. All authors reviewed the results and approved the final version of the manuscript.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to L.H.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.