Accident Analysis and Prevention 198 (2024) 107479

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Exploring the influence of drivers' visual surroundings on speeding behavior

Mohamed Abdel-Aty, Jorge Ugan *, Zubayer Islam

Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816, USA

ARTICLE INFO

Keywords:
Dash camera
Probe vehicle data
Speeding behavior
Time-headway
Close car following, visual environment
Computer vision

ABSTRACT

Despite awareness campaigns and legal consequences, speeding is a significant cause of road accidents and fatalities globally. To combat this issue, understanding the impact of a driver's visual surroundings is crucial in designing roadways that discourage speeding. This study investigates the influence of visual surroundings on drivers in 15 US cities using 3,407,253 driver view images from Lytx, covering 4,264 miles of roadways. By segmenting and analyzing these images along with vehicle-related variables, the study examines factors affecting speeding behavior. After filtering the images, to ensure an accurate representation of the driver's view, 1,340,035 driver view images were used for analysis. Statistical models, including hurdle beta and bivariate probit models with random driver effects as well as Machine Learning's eXtreme Gradient Boosting (XGBoost), were employed to estimate speeding behavior. The results indicate that factors within the driver's visual environment, weather conditions, and driver heterogeneity significantly impact speeding. Speeding behavior also varies across geographic locations, even within the same city, suggesting a connection between local context and speeding. The study highlights the importance of the driver's environment, showing that more open spaces encourage speeding, while areas with trees and buildings are associated with reduced speeding. Notably, this research differs from previous studies by utilizing real-time data from dash cameras, providing a dynamic and accurate representation of the driver's visual surroundings. This approach enhances the reliability of the findings and empowers transportation engineers and planners to make informed decisions when designing roadways and implementing interventions to address effectively excessive speeding. In addition to examining speeding behavior, the study also analyzes time-headway, a key factor affecting safety and risky driver behavior, to explore its relationship with speeding. The findings offer valuable insights into the factors influencing speeding and the driver's visual environment. These insights can inform efforts to create environments that discourage speeding (and close car following) and ultimately reduce severe accidents caused by excessive speed (and tailgating).

1. Introduction

Speeding presents a grave threat to road safety in the United States and around the world, as evidenced by concerning statistics that highlight its severe consequences. In 2020, the National Highway Traffic Safety Administration (NHTSA) reported that speeding played a role in 29 % of all traffic fatalities, claiming the lives of over 11,258 individuals nationwide (NHTSA, 2020). These figures serve as a stark reminder of the pressing need to address the issues associated with speeding and emphasize the critical importance of comprehending the myriad factors contributing to this perilous behavior. By identifying and gaining a thorough understanding of the underlying causes, we can lay the groundwork for effective solutions and work towards creating safer roads for everyone. Table 1 reveals the top four states with the highest number of speed-related fatalities in the United States: Texas, California, South Carolina, and North Carolina. Additionally, Florida has been included due to its ranking as the third highest state in terms of total traffic fatalities. Exploring the reasons behind the significant number of fatalities in these states would provide valuable insights into the factors contributing to the overall high fatality rate in the country.

Speeding on roadways is influenced by factors such as the driver's visual environment, weather conditions, and location. However, previous studies have not fully captured important factors due to limited data availability. These studies have overlooked the real-time information

* Corresponding author.
E-mail addresses: m.aty@ucf.edu (M. Abdel-Aty), jorge.ugan@ucf.edu (J. Ugan), zubayer.islam@ucf.edu (Z. Islam).

https://doi.org/10.1016/j.aap.2024.107479
Received 19 September 2023; Received in revised form 29 November 2023; Accepted 14 January 2024
Available online 20 January 2024

that drivers experience, which is crucial for understanding their speeding behavior. To address this research gap, this paper aims to; (i) to identify regional variations in speeding patterns; (ii) assess the impact of road design and infrastructure on speeding; and (iii) identify potential interventions and countermeasures. These are achieved by developing group random effect hurdle beta, grouped random effect multinomial logit, and grouped random effect bivariate probit models. To understand speeding behavior, various variables were taken into account, including the speed limit, proportions of sky and buildings, the number of surrounding vehicles, weather conditions, and more, all derived from each driver's visual environment.

1.1. Literature review

Several studies have examined the factors influencing drivers' speeding behavior. Cai et al. (Cai et al., 2022) applied machine learning methods on google street view images to analyze the effects of speeding accidents. Bassani et al. (Bassani et al., 2014) focused on urban arterials and analyzed the 85th percentile speed. They investigated the influence of road attributes, such as the presence of shoulder, bus and taxi lanes, and sidewalks, on drivers' speed decisions. Bhowmik et al. (Bhowmik et al., 2019) and Cai et al. (Cai et al., 2021) developed multilevel ordered probit fractional split models to examine the effects of roadway attributes, traffic data, land use, socio-demographic characteristics, and environmental factors on speed proportions on various arterials. Eluru et al. (Eluru et al., 2013) studied the effects of different roadway geometric factors, including speed limit, number of lanes, lane width, and number of sidewalks, on speed for arterials. Ghasemzadeh and Ahmed (Ghasemzadeh and Ahmed, 2019) utilized naturalistic driving data to explore speeding behavior and developed a multilevel logistic model that captured regional heterogeneity in speeding behavior. Mahmoud et al. (Mahmoud et al., 2021; Mahmoud et al., 2022) developed a tobit models to understand factors contributing to operating speeds and analyze the difference between operating speed and target speed.

The study by Edquist et al. (Edquist et al., 2012) took a different approach, conducting a driving simulator study to examine the influence of visual complexity in the roadside environment on factors such as travel speed and reaction time. The study highlighted the significant role of visual complexity in affecting driver workload and performance. Similarly, Atombo et al. (Atombo et al., 2016) conducted a survey study that revealed the notable effects of the driving environment on speeding and overtaking violations. Furthermore, Marshall et al. (PE et al., 2018) developed statistical models to analyze the effects of tree density on accident frequency in urban areas, demonstrating that a higher density of trees can potentially reduce accidents. However, research exploring the relationship between drivers' visual environment and traffic safety remains limited, partly due to the challenges associated with obtaining data from the drivers' perspective.

While Google Street View has been utilized to understand the effects of the driver's visual environment on traffic safety, its suitability for examining speeding behavior is limited due to its inability to provide real-time data. Google Street View relies on static and outdated images of roadways, which may not accurately reflect the conditions drivers experience in real time. Moreover, it only provides a perspective from a single driver during a specific time when the vehicle passed through the road, thus limiting its representation of the diverse visual environments that different drivers encounter. Consequently, there can be disparities between the visual cues presented in Google Street View and the actual on-road environment that impacts driver behavior. This disparity underscores the importance of utilizing dash cameras to comprehend speeding incidents. Dash cameras, in contrast, offer a more precise representation of the driver's environment by capturing continuous and dynamic recordings of the road. They provide real-time information regarding traffic density, road conditions, signage, and the behaviors of other road users. By utilizing dash camera footage, researchers can analyze the contextual factors contributing to speeding incidents with more accuracy, leading to a better understanding of the driver's actual perspective. Several studies have examined the limitations of Google Street View and its impact on assessing road conditions and pedestrian safety.

Isola et al. (Isola, 2019) highlights the potential limitations of using Google Street View for evaluating environmental safety features at the sites of pedestrian-vehicle collisions. The study suggests that while Google Street View provides a visual representation of the road environment, it may not capture all relevant features or accurately reflect the conditions at the time of the incident. Outdated imagery and limited coverage can affect the accuracy of the assessment, potentially leading to incomplete or inaccurate conclusions about the safety features in the road environment. In contrast, Mooney et al. (Mooney, 2020) focused on the development and validation of a Google Street View pedestrian safety audit tool. This research recognized the potential of Google Street View in assessing pedestrian safety but also acknowledged certain limitations. The study found that Google Street View can be a valuable tool for conducting safety audits, but its effectiveness relies on the quality and recency of the imagery. Outdated imagery or limited coverage can hinder the tool's accuracy in identifying specific pedestrian safety features or capturing changes in the road environment. Furthermore, Rundle et al. (Rundle et al., 2011) compared the use of Google Street View with vi deotape recordings to examine speeding behavior. The researchers found that Google Street View images did not adequately capture the dynamic aspects of the road environment, such as changes in traffic flow or road construction, which could influence driver speed. In contrast, the analysis of vi deotape recordings provided a more detailed and accurate representation of the driving environment, allowing for a more nuanced understanding of speeding behaviors. The findings were based on a relatively small sample size, comprising only 38 block segments and utilizing 143 Google Street View images. Additionally, the study acknowledged the temporal variability of the Google Street View data as a constraint, highlighting its inherent instability over short periods of time. In conclusion, while Google Street View can be a valuable tool for understanding the road environment and assessing pedestrian safety, it is essential to recognize its limitations. Outdated imagery, limited coverage, and the lack of real-time information are factors that can affect its accuracy. Dash cameras, on the other hand, offer a more precise and detailed perspective, enabling researchers to capture the contextual factors and visual cues that influence speeding behavior more effectively. However, implementing cameras on numerous vehicles across a vast area in the US to achieve a significant sample size is both a

Table 1
States with the Highest Speeding-Related Traffic Fatalities, 2020 [1].

State	Total Traffic Fatalities	Speeding-Related Fatalities by Roadway Function Class				
		Total Speeding-Related Fatalities	Interstate	Arterials	Collectors	Locals
Texas	3874 (9.98 %)	1446 (37.33 %)	196 (13.55 %)	765 (52.90 %)	310 (21.44 %)	171 (11.83 %)
California	3847 (9.91 %)	1228 (31.92 %)	190 (15.47 %)	774 (63.03 %)	164 (13.36 %)	100 (8.14 %)
South Carolina	1064 (2.74 %)	494 (46.43 %)	59 (11.94 %)	363 (73.48 %)	26 (5.26 %)	46 (9.31 %)
North Carolina	1538 (3.96 %)	489 (31.79 %)	46 (9.41 %)	189 (38.65 %)	137 (28.02 %)	116 (23.72 %)
Florida	3331 (8.58 %)	285 (8.56 %)	20 (7.02 %)	147 (51.58 %)	44 (15.44 %)	40 (14.04 %)
National	38,824 (100.00 %)	11,258 (29.00 %)	1438 (12.77 %)	5741 (50.99 %)	2189 (19.44 %)	1704 (15.14 %)

time-consuming and expensive process.

Using speeding proportions as a measure can effectively reflect the overall level of speeding. To calculate speeding proportions, speed data along with the corresponding speed limit on the roadway is required. Traditionally, transportation agencies and state departments of transportation have relied on fixed sensors like loop detectors and cameras to monitor traffic. While these fixed sensors can provide relatively accurate traffic data, they come with high deployment and maintenance costs. Additionally, their geographical scalability is limited as a large number of sensors need to be installed to assess the traffic condition in an area (Young, 2007). Probe vehicle technology has emerged as a cost-effective alternative for traffic monitoring, and the coverage of probe data has significantly expanded (Ahsani et al., 2019). Numerous studies have been conducted to validate the accuracy and reliability of probe source data (Ahsani et al., 2019; Abdelraouf et al., 2022; Adu-Gyamfi et al., 2017; Hu et al., 2016), indicating notable improvements in the data quality of probe vehicles.

In summary, this study aims to make several contributions to the field of speeding analysis. First, it seeks to identify regional variations in speeding patterns by examining data collected from different states and cities. Second, it aims to assess the influence of road design and infrastructure on speeding by examining the correlation between the elements within the driver's view and instances of speeding. Third, it aims to investigate the interplay between time-headway and speeding, analyzing how the distance maintained between vehicles over time affects speeding behaviors and the associated safety implications. Fourth, this study also delves into real-time traffic information, a factor that has not been explored in studies relying on Google Street View or simulator-based approaches. By incorporating real-time data, including traffic congestion, weather, and dynamic road conditions, this study aims to provide a more comprehensive understanding of the contextual factors influencing speeding behavior. Lastly, this research aims to identify potential interventions and countermeasures by utilizing various statistical methods to analyze the data, thereby contributing to the development of effective strategies for mitigating speeding-related risks and improving the overall road safety.

The paper is organized as follows: The introduction discusses existing literature on speeding analysis. The data preparation section outlines the data used and the methods for extracting data from driver's view images. The methodology section presents the algorithms and models employed for object classification, depth estimation, and analyzing speeding behavior. The results and discussions section presents the modeling results of speeding proportion levels. The conclusion summarizes the study's findings. By following this structure, the study aims to contribute new insights to the field of speeding analysis and provide a comprehensive understanding of the factors influencing speeding behavior.

1.2. Data preparation

The dataset utilized in this paper was provided by Lytx, encompassing vehicle-specific data that showcases near-real-time road-view images captured by the Lytx Camera Network. These images originate from Lytx cameras installed on the vehicles and include pertinent information such as GPS location, heading, speed, and timestamp for each image. The Lytx data used in this study boasts an accuracy of over 95 %, covering both GPS and speed information. Data for each driver is updated every 30 s. The study period spanned from the 1st to the 17th of December 2022, encompassing a two-week duration. The dataset consisted of a total of 3,407,253 images from 15 major cities in 5 states, as depicted in Table 2, which provides a breakdown of the data used for this study.

To ensure the suitability of the data, a rigorous data preparation pipeline was implemented as depicted in Fig. 1. The pipeline involved applying specific criteria to filter the images. These criteria included ensuring that the images represented the driver's view, with the horizon positioned at the middle of the image. Additionally, the images had to possess a clear view without any obstructions and maintain an acceptable camera resolution that focused on the driver's perspective. The images used in the study were collected from 15 cities renowned for high-speed driving across five states: Florida, Texas, California, North Carolina, and South Carolina. The data collection specifically targeted interstates and highways within these cities. To provide a visual representation of the various city environments, Fig. 2 presents sample images from the dataset. These images highlight the differences among the visual environments of the different cities, showcasing the unique characteristics of each location.

Table 2
Breakdown of Images Collected by City and State Level.

State	City	Images	Total Roadway Length (miles)	Size (GB)	Date & Time
Texas	Dallas	263,129	464.74	4.14	12/1 – 12/17 (7AM – 5PM)
	San Antonio	271,011	522.67	4.17	
	Fort Worth	261,495	432.10	3.94	
California	Los Angeles	296,115	432.57	5.07	
	San Diego	285,491	374.23	4.58	
	San Jose	226,620	139.15	4.02	
South Carolina	Greenville	147,689	64.14	2.47	
	Spartanburg	107,179	45.92	1.82	
	Charleston	204,892	126.82	3.30	
North Carolina	Charlotte	238,287	313.61	3.81	
	Raleigh	217,369	175.40	3.71	
	Greensboro	230,172	680.65	3.73	
Florida	Miami	226,478	136.61	3.74	
	Orlando	211,982	235.12	3.66	
	Tampa	219,344	120.22	3.74	
Total		3,407,253	4263.96	55.90	

Furthermore, the database was augmented with additional features, including speed limit information, weather conditions, object detection, and segmentation. Weather data was collected using the Visual Crossing weather API, which leveraged GPS coordinates and timestamps to provide accurate and timely information on prevailing weather conditions. In the case of speed limit data, this information was obtained through GIS shapefiles provided by the Department of Transportation agencies of the relevant state or city. For each driver, GPS coordinates were supplied, and the API or GIS shapefiles, depending on the data source, were utilized to acquire corresponding speed limit data. This dual-data source approach ensured the inclusion of crucial environmental and contextual factors in the dataset, encompassing details about the drivers' visual surroundings, roadway design, and prevailing weather attributes. The detailed process of data extraction, including the specific methodologies employed for collecting weather and speed limit data, will be comprehensively explained in the methodology section. Table 3 presents a sample of the data, with each column showcasing the variables utilized to model speeding behavior for each drive's perspective. The dataset was meticulously aggregated on a per-driver basis, aiming to provide a comprehensive representation of driving scenarios. To enhance dataset accuracy, a manual filtering process was applied to scrutinize the original 3,407,253 Lytx images, resulting in a refined database comprising approximately 1,340,035 images over the two-week period.

The speeding proportion was calculated using the vehicle speed and the speed limit information as shown in Eq. (1).

$$y = \frac{speed - speedlimit}{speedlimit} \tag{1}$$

To obtain a visualization of the high-speeding areas in the study, we developed 15 maps that depict the driver's speed proportion using different colors. The relationship between the speeding proportions and color is depicted in Fig. 3. Figs. 4–8 display the speeding proportions of various roadway segments and hotspot locations. Each region in the study is highlighted with a specific color corresponding to the speeding

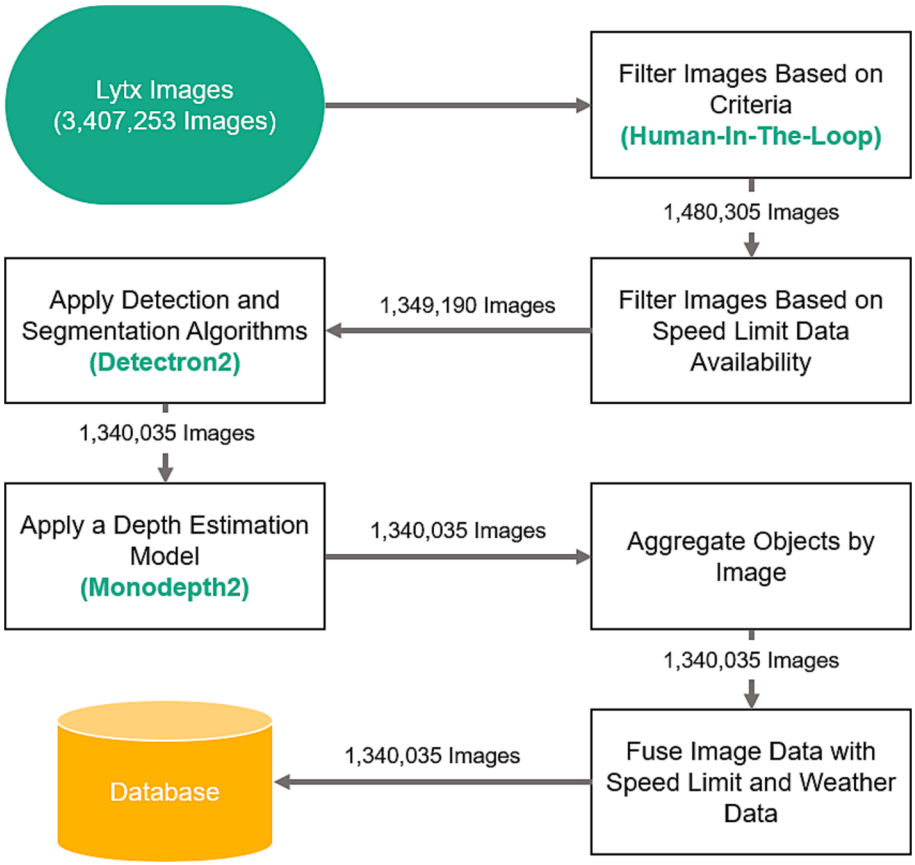

Fig. 1. Data Preparation Pipeline.

proportions above the speed limit. These figures allow us to identify regions and segments with high speeding proportions. By gaining insights into areas with a significant speeding issue, we can better comprehend the contributing factors associated with excessive speeding.

Table 4 presents the distribution of speeding proportions above the speed limit for each driver. In this analysis, any speed that was less than 0.05 over the speed limit was considered as non-speeding. The results indicate that the largest proportion of speeding above the speed limit falls within the range of 0.05 to 0.20, accounting for 75.52 % of the observations. As the proportion of speeding above the speed limit increases, there is a corresponding decrease in the frequency of drivers at those higher levels.

Table 5 presents the distribution of time headway for each driver that had a leading vehicle in front of them. Classification of the time-headways showed similar distribution with previous studies (Ayres et al., 2001). The results indicate that the largest proportion of time headway gap was from the long headway class within the range of 2 to 12 secs, accounting for 44.31 % of the observations with a leading vehicle. As the time headway decreases, there is a corresponding increase in the frequency of drivers at those higher levels.

Fig. 9 complements the analysis by providing a kernel density estimate, which illustrates the probability density function of the speeding proportions across different cities. The plots demonstrate that each city exhibits varying levels of speeding and frequencies. For instance, focusing on Florida, we observe that Miami, Tampa, and Orlando all display similar peak shapes, but the magnitude of each peak differs. It is noteworthy that despite Orlando having the lowest peak among the Florida cities, it exhibits a higher density for higher speeding proportions. When examining California and South Carolina, a similar trend emerges in all three cities within each state, making it challenging to differentiate the cities based solely on their speeding proportions. However, it is worth noting that the peaks in the South Carolina cities have a higher magnitude compared to those in the California cities. Overall, the analysis reveals distinct variations in speeding proportions and frequencies among cities, highlighting the unique characteristics of each location. Fig. 10 presents a probability density function depicting the distribution of time-headways across various cities. In each city, two distinct peaks can be observed for the time-headways. The majority of states exhibited variations in the distribution of time-headways across different cities. However, the differences among Texas drivers were not as easily discernible.

The descriptive statistics of the prepared feature data set and brief description is mentioned in Table 6. For each of the variables that were considered in the model there is the mean, standard deviation, minimum value, and maximum value.

Fig. 11 provides an overview of the variables derived from the drivers view, which were considered in the models. The calculations involved determining the proportions of sky, building, and road surface within the image, obtained by dividing the respective areas by the total image area. A similar approach was employed for determining the proportion of the closest vehicle based on the count of surrounding vehicles. In cases where a leading vehicle was present, the time-headway was calculated using the distance from the leading vehicle to the following vehicle. Vehicle time headway refers to the amount of time, measured in seconds, that passes between the arrival of the leading vehicle and the following vehicle at a specific observation point. It is a metric that indicates the distance or space between two consecutive vehicles.

2. Methodology

Identification and segmentation of objects and features of the images has been a growing field of interest in computer science. Deep learning

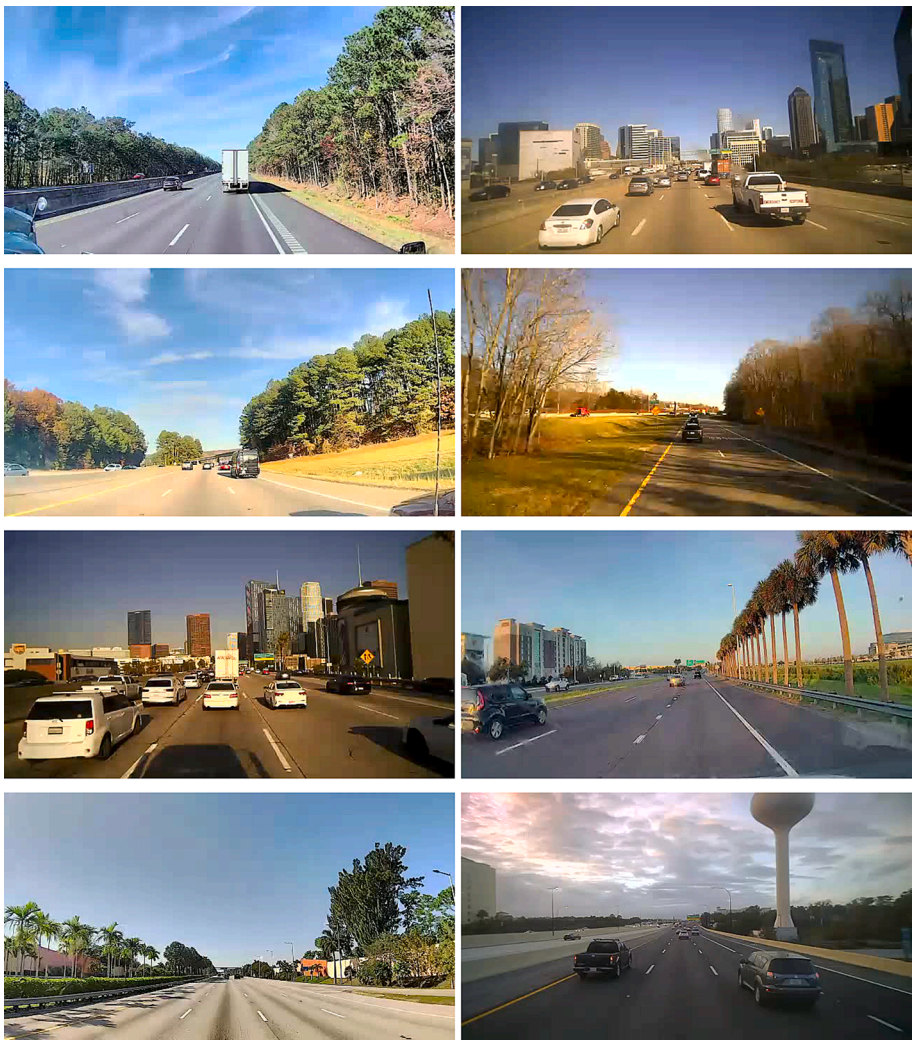

Fig. 2. Samples of images from the database, presented from left to right in the following order: Charleston, SC; Dallas, TX; Raleigh, NC; Greensboro, NC; Los Angeles, CA; Tampa, FL; Miami, FL; Orlando, FL.

has been heavily applied and developed for semantic segmentation from images. In this study, "Detectron2" from Facebook was used to cluster objects. Detectron2, starting with maskrcnn-benchmark (Washington et al., n.d.), is Facebook AI research next generation software system that implements state-of-the-art object detection algorithms by reaching 34.9 mask average precision (Wu et al., 2019). It is also suggested that the model using the Detectron2 framework could reach the state-of-the-art performance for labeling objects in drivers' view (Yu et al., 2021; Washington et al., n.d.). For example, Syed et al. (Washington et al., n.d.) found that the Detectron2 framework could have a pixel accuracy of around 90 % to detect pedestrians in different cloth and offer more stable detection results compared to other detection frameworks with impacts of the pixel area, occlusion rate, and distance. Yu et al. (Yu et al., 2021) developed models to classify risky driving scenes based on the Detecton2 framework, which could reach 96.4 % classification accuracy. As shown in Fig. 12, different objects in the environment such as roads, trees, sky, and buildings in the drivers' view could be labelled from the images. Based on the clustering results, we could know the object type by each pixel in the image. Then, the proportion of pixels by object type in the drivers' view could be calculated, such as the proportion of trees and the proportion of roads. Fig. 12, shows the percentage of each object (e.g., car, truck) displayed in the image. This percentage represents the confidence or probability associated with the detection of each object. A higher percentage indicates greater confidence in the detection result, suggesting a higher likelihood of the object's presence.

Meanwhile, the depth information could be obtained from the 2D images. Since the dash camera could be treated as a mono camera, a self-supervised monocular depth estimation method (monodepth2) proposed by Godard et al. (Washington et al., n.d.) was used to obtain the depth information. It was suggested that the depth estimation method could provide an absolute relative error of 0.115 for monocular depth estimation on the KITTI benchmark, achieving state-of-the-art depth estimation. The detection range by this method is from 0 to 80 m. Fig. 12 illustrates the depth information subtracted from the image.

After extracting the features using detection and segmentation algorithms, as well as depth estimation models, the next step involved fusing these features with speed limit and weather data. The speed limit information was utilized to calculate the proportion of speed, as mentioned earlier. This proportion of speed is an essential factor in understanding the degree of speeding behavior. Additionally, the weather data from Visual Crossing was incorporated into the dataset to analyze the impact of specific weather conditions such as precipitation and cloud cover on speeding and the proportion of speed. Visual Crossing is a comprehensive weather service that utilizes Application Programming Interfaces (APIs) to retrieve accurate and up-to-date weather data. Through its API integration, Visual Crossing establishes a connection with reliable weather data providers, granting users access to a vast range of meteorological information. Leveraging these APIs, Visual Crossing efficiently retrieves weather data such as temperature,

Table 3
Sample data extracted from the final aggregated dataset.

Latitude	Longitude	City	ID	Hour	Day	Speed_limit	Speed	Heading	Precip prob
37.26	−121.86	San Jose	1	11	16	65	68.97	269.90	0
28.54	−81.33	Orlando	2	14	12	65	57.21	269.12	0
32.67	−96.72	Dallas	3	18	7	70	50.91	245.00	100
35.89	−78.59	Raleigh	4	13	17	70	70.21	136.80	0
35.76	−78.60	Raleigh	5	10	7	65	67.97	107.26	100

cloudcover	windgust	humidity	visibility	windspeed	winddir	car	truck	Car prop	Truck prop
9.90	6.90	74.00	9.90	5.40	240.80	2	1	0.03	0.02
53.30	9.20	76.40	9.50	10.30	12.40	3	0	0.04	0.00
86.60	19.00	86.40	7.70	12.80	188.60	1	0	0.00	0.00
37.10	16.10	77.90	9.90	8.10	227.90	7	0	0.01	0.00
98.70	19.20	94.70	3.90	7.90	238.90	4	0	0.00	0.00

Road prop	Tree prop	Grass prop	Mountain prop	Barrier prop	Sky prop	Building prop	Prop speed	Close_veh angle	Close_veh prop
0.22	0.04	0.00	0.00	0.00	0.58	0.00	0.06	0.00	0.00
0.36	0.11	0.00	0.00	0.00	0.48	0.00	−0.12	21.09	0.01
0.37	0.00	0.10	0.00	0.00	0.45	0.03	−0.27	41.78	0.00
0.22	0.35	0.00	0.00	0.00	0.31	0.00	0.00	43.31	0.01
0.56	0.20	0.03	0.00	0.00	0.18	0.00	0.05	−15.15	0.00

Close_veh depth	Speeding	Headway	Peak	Dow	Weekday
0.00	1	1.62	0	4	1
5.78	0	0.45	0	0	1
9.46	0	2.20	1	2	1
4.06	1	1.59	0	5	0
19.29	1	1.65	1	2	1

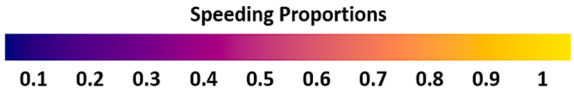

Fig. 3. Speeding Proportions Color Bar.

humidity, precipitation, wind speed, and more from various sources across the globe. By seamlessly tapping into these external APIs, Visual Crossing ensures that users receive real-time and historical weather data (Crossing, 2023). By including weather-related features, the study aims to explore how different weather conditions influence drivers' behavior and their tendency to speed. Therefore, the final dataset used for modeling purposes comprises a combination of features related to the drivers' visual environment (extracted from the detection, segmentation, and depth estimation models) and weather conditions. This comprehensive dataset provides valuable insights into the relationship between the drivers' surroundings, weather conditions, and their propensity for speeding. Fig. 13 illustrates how the dataset was employed for each model, showing the relationship between speeding behavior and time headway with the driver's environment.

2.1. Grouped random effect hurdle beta model

As previously mentioned, the proportions of speeding are expected to be continuous numbers within the range of (0, 1). However, it is important to consider that the proportion of speeding can also be negative or zero if drivers adhere to the speed limit or drive below it. Therefore, the proportions of speeding exhibit a mixed distribution, wherein a continuous nonnegative random variable is combined with a probability mass truncated at zero. One approach for modeling such mixed distributions is the hurdle model (Boucher and Santolino, 2010; Cai et al., 2016; Ma et al., 2016; Ma et al., 2015; Wu et al., 2018; Ugan et al., 2022), which introduces a threshold between positive and non-positive outcomes. Hurdle models have been successfully employed in previous studies related to traffic accidents. For instance, Boucher and

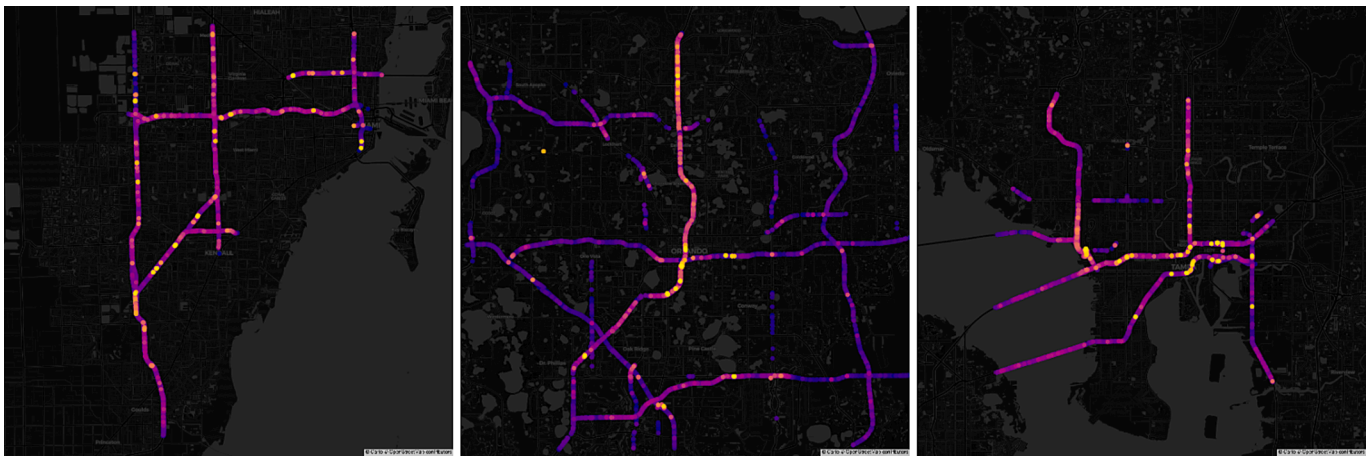

Fig. 4. Speeding Map of Regions in Florida, from left to right, Miami (45423 image coordinates), Orlando (41738 image coordinates), and Tampa (52035 image coordinates).

Fig. 5. Speeding Map of Regions in South Carolina, from left to right, Charleston (16452 image coordinates), Greenville (24459 image coordinates), and Spartanburg (9393 image coordinates).

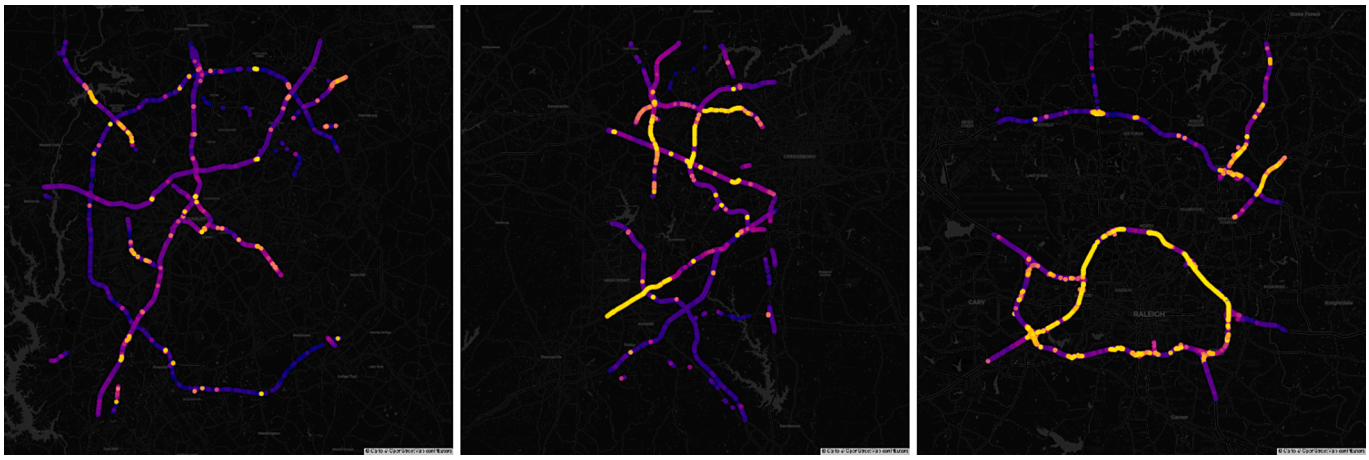

Fig. 6. Speeding Map of Regions in North Carolina, from left to right, Raleigh (43673 image coordinates), Greensboro (44095 image coordinates), and Charlotte (34008 image coordinates).

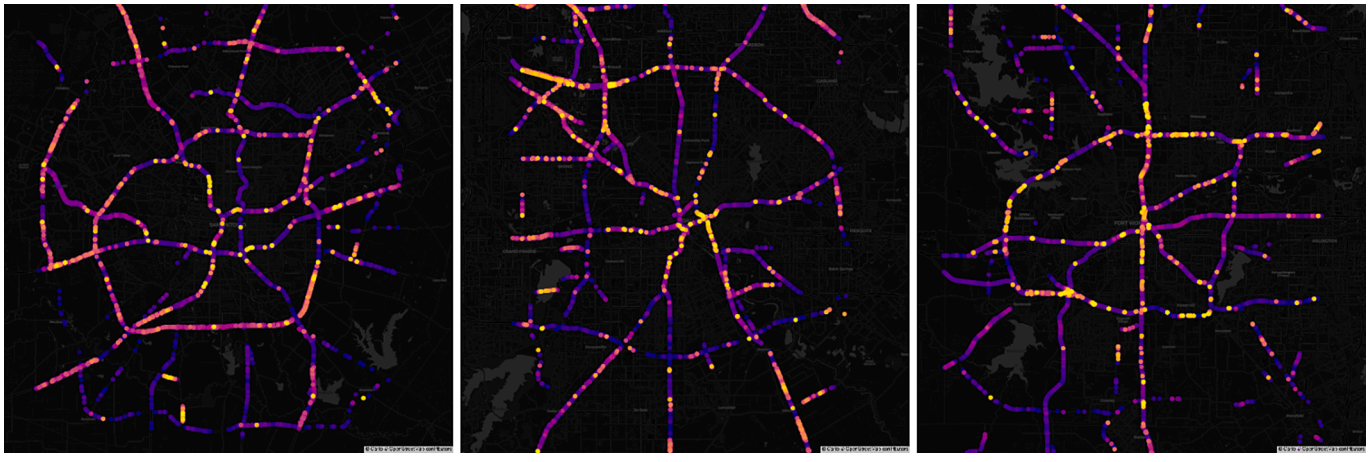

Fig. 7. Speeding Map of Regions in Texas, from left to right, San Antonio (44687 image coordinates), Dallas (35800 image coordinates), and Fort Worth (48398 image coordinates).

Santolino[28] utilized a hurdle model to analyze disability score data and highlighted that one advantage of these models is the ability to separately model the zero score process. Similarly, Ma et al. (2015) (Ma et al., 2015) and Ma et al. (2016) (Ma et al., 2016) proposed a hurdle regression framework to analyze accident rates of road segments and extended the modeling structure to handle equivalent property damage-only accident rates. Both studies consistently demonstrated the superior performance of hurdle models compared to the Tobit model for censored

Fig. 8. Speeding Map of Regions in California, from left to right, San Diego (21866 image coordinates), San Jose (13147 image coordinates), and Los Angeles (11108 image coordinates).

Table 4
Distribution of Speeding Proportions.

Interval for Speeding Proportion	Frequency (*f*)	Cumulative Frequency (*cf*)	Relative Frequency (%)
[0.05, 0.20)	205,499	205,499	75.52 %
[0.20, 0.40)	44,138	249,637	16.22 %
[0.40, 0.60)	12,140	261,777	4.46 %
[0.60, 0.80)	6763	268,540	2.48 %
[0.80, 1.00)	3567	272,107	1.31 %
Total	272,107	–	100.00 %

Table 5
Distribution of Driving Behavior from Time-Headway.

Interval for Time Headway (secs)	Classification	Frequency (*f*)	Cumulative Frequency (*cf*)	Relative Frequency (%)
(0, 1)	Short	81,500	81,500	21.59 %
[1, 2)	Typical	128,737	210,237	34.10 %
> 2	Long	167,285	377,522	44.31 %
Total		377,522	–	100.00 %

accident data. In the context of pedestrian and bicycle accidents, Cai et al. (Cai et al., 2016) applied hurdle negative binomial models based on traffic analysis zones. Moreover, Wu et al. (Wu et al., 2018) utilized hurdle beta models to examine the proportion of average speed reduction under different fog conditions, using real-time fog warning systems. These studies also indicated that the dual-state model outperforms the conventional single-state model. Collectively, these findings suggest that the hurdle model offers high flexibility in dealing with such mixed distributions. Therefore, we have adopted the hurdle modeling structure to analyze the proportion of speeding in this study.

The hurdle model, proposed by Mullahy (Mullahy, 1986), is composed of two components. The first component is a binary model dealing with whether the response crosses the "hurdle", while the second component is a truncated-at-hurdle regression model. In the context of this study the first component models if the driver was speeding (i.e. if the speeding proportion of the driver is greater than 0.05, then speeding = 1, if not speeding = 0). The second component will model the driver speeding proportion greater than 0.05 above the speed limit, given they were speeding (i.e., if speeding = 1 in the first component).

Assume that the first truncated part is governed by function and the second model process follows a truncated-at-hurdle function. Then the Hurdle model can be specified as follows if we set the hurdle as 0 (Cai et al., 2016):

$$f(p_{ij}) = \begin{cases} f_1(\leqslant 0) = p_{ij}, & p \leqslant 0 \\ (1 - f_1(\leqslant 0)\dfrac{f_2}{(1 - f_2(\leqslant 0)}, & p > 0 \end{cases} \quad (2)$$

The logistic regression model is utilized to estimate

$$p_{ij} = \frac{\exp\left(\beta' x + \varepsilon'_{ij}\right)}{1 + \exp\left(\beta' x + \varepsilon'_{ij}\right)} \quad (3)$$

$$\varepsilon'_{ij}\ N(0, \sigma'^2) \quad (4)$$

where, β' is the parameter explanatory variables. As for function f_2, several alternate approaches have been taken to model proportions. These alternatives include Tobit regression, fractional logit, and beta regression models. The previous studies have suggested that the beta regression model is the most preferred approach for modeling proportion data (Meaney and Moineddin, 2014; Moeller, 2013; Ospina and Ferrari, 2012). Hence, the beta regression model was adopted to analyze the positive proportion of speeding 5 % or more above the speed limit.

Within a logit link function, the model is given by:

$$Proportion_{ij}\ B(u_{ij}, \phi_{ij}) \quad (5)$$

$$logit(u_{ij}) = \beta'' x + \varepsilon''_{ij} \quad (6)$$

where, $Proportion_{ij}$ is observed proportion of during timestamp j by driver i, u_{ij} is the distribution mean and ϕ_{ij} plays the role of a precision parameter. x is the set of explanatory variables,

β' is the corresponding parameters. The density function can be specified as

$$f(p, u, \phi) = \frac{\Gamma(\phi)}{\Gamma(u\phi)\Gamma((1-u)\phi))} p^{u\phi - 1}(1 - y)^{(1-u)\phi - 1}, p \in (0, 1) \quad (7)$$

Thus, the hurdle beta model can be specified.

While hurdle beta models could appropriately account for the data structure, a concern that should be considered is that some effects of certain parameters may vary across drivers due to the unobserved heterogeneity. If the unobserved heterogeneity is ignored, the model would be misspecified and the estimated parameters could be biased and inefficient (Mannering et al., 2016). To account for this issue, random parameters can be estimated, allowing for the effect of explanatory variables to vary across drivers (Anastasopoulos and Mannering, 2011; Barua et al., 2016; Cai et al., 2018). The random parameters can be specified as follows:

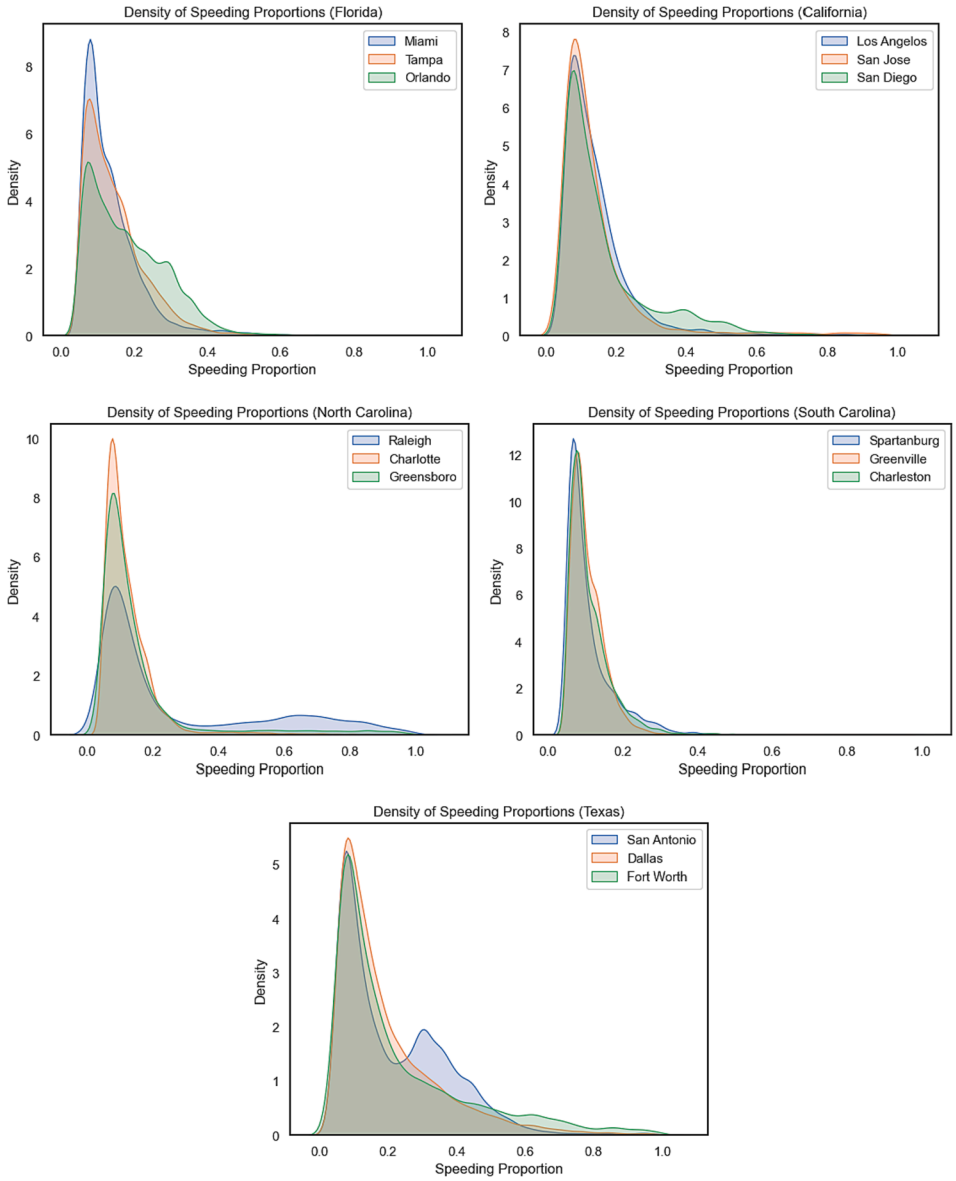

Fig. 9. Density of Speeding Proportions at the City and State Level.

$$\beta_i = \beta + \varepsilon_i \tag{8}$$

where β is the vector of participant-specific parameters and ε_i is the random distributed terms at the participant level which are normally distributed with mean of 0 and variance of σ^2.

As suggested in the previous studies (Cai et al., 2017; Huang et al., 2010; Washington et al., n.d.), Bayesian inference outperforms the traditional maximum likelihood estimation method by incorporating parameter prior information. The freeware WinBUGS has been widely used to estimate models in a fully Bayesian inference using Markov Chain Monte Carlo (MCMC) simulation. Hence, all the candidate models are programmed, estimated, and evaluated in WinBUGS. In the absence of sufficient prior information, non-informative prior are specified for the parameters (Lee et al., 2017; Zeng et al., 2017).

The models' convergence was evaluated by the Gelman-Rubin statistics, visual examination of the MCMC chains, and the ratios of Monte Carlo errors relative to the respective standard deviations of the estimate. As a rule of thumb, the ratios should be less than 0.05 (Xu et al., 2017). The 90 % Bayesian credible interval (BCI) is provided to indicate the significance of the examined variables. The Deviance Information Criteria (DIC) was used for the model performance comparisons.

$$DIC = D(\overline{\theta}) + 2p\mathrm{D} \tag{9}$$

where D($(\overline{\theta})$) is the deviance at the posterior mean of the parameters and pD is the effective number of parameters in the model. Roughly, differences of more than 10 in the value of the DIC would rule out the model with the higher DIC (Zeng et al., 2017).

2.2. Grouped random effect multinomial logit model

Multinomial Logit models have been successfully employed in previous studies related to traffic safety research. Ambo et al. (Ambo et al., 2020) identified and evaluate the major traffic violation with related risk factors using multinomial logit model. Lee et al. (Lee et al., 2018) developed a flexible mixed multinomial logit fractional split model, used to analyze the proportions of accidents by vehicle type at the macro-level. Tay et al. (Tay et al., 2011) estimated a multinomial logit model to identify the factors determining the severity of pedestrian–vehicle accidents.

To understand the factors associated with a specific time-headway, a multinomial logit model was developed considering three classifications

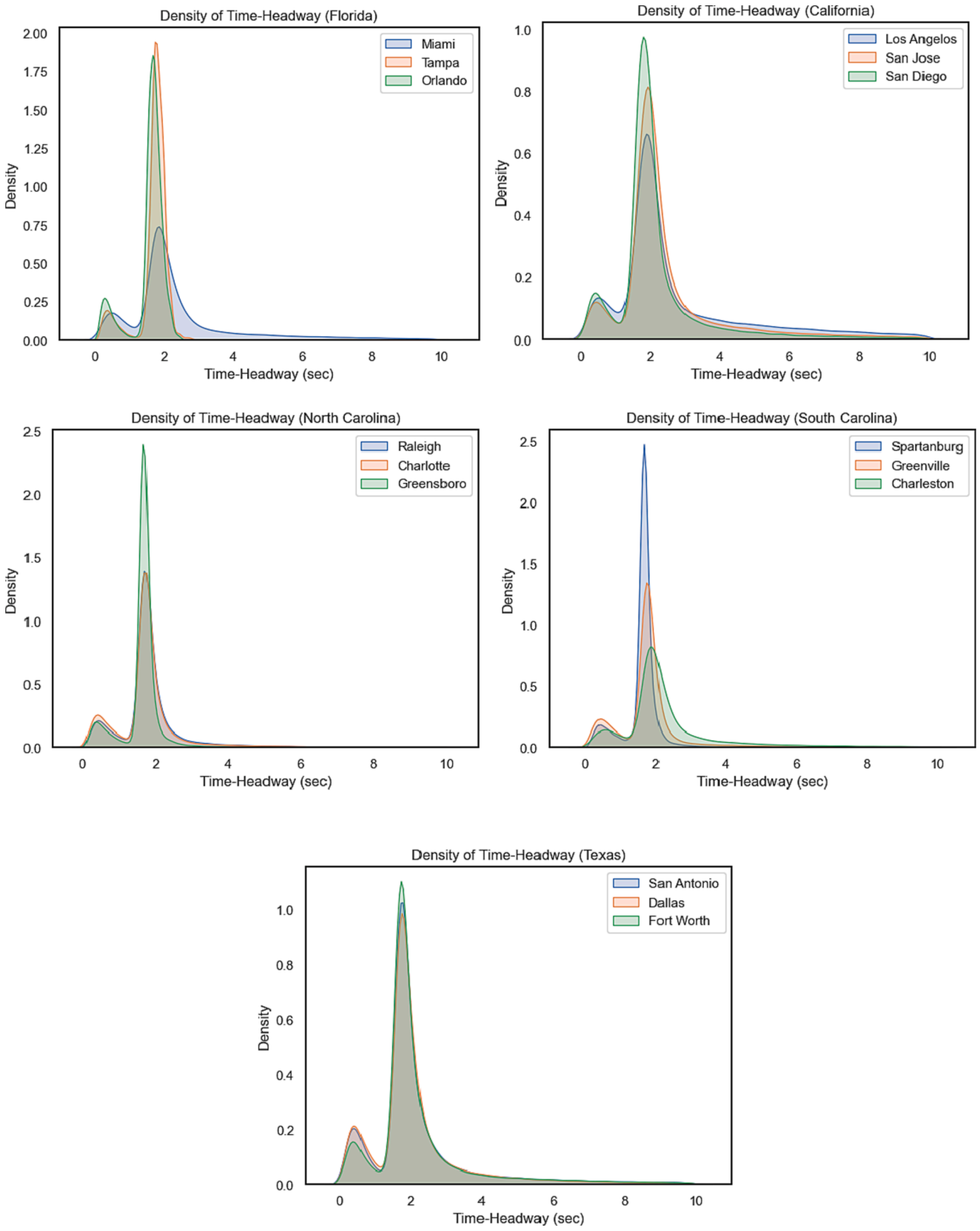

Fig. 10. Density of Time-Headway at the City and State Level.

Table 6
Descriptive Statistics of Aggregated Data.

Variable	Description	mean	std	min	max
ID	Unique driver identifier	–	–	1.00	51620.00
hour	The hour of the timestamp	12.70	3.30	0.00	23.00
day	The day of the timestamp	9.40	6.02	1.00	17.00
speed_limit	The speed limit on the road the vehicle was traveling on	62.54	6.07	50.00	75.00
heading	The direction of the vehicle	184.56	105.22	0.00	360.00
precipprob	Probability of precipitation given the longitude, latitude, and timestamp	38.13	48.57	0.00	100.00
cloudcover	Cloud cover as a percent value given the longitude, latitude, and timestamp	57.09	31.81	0.00	100.00
windgust	A sudden burst in wind speed (mph)	18.00	7.59	4.70	61.90
humidity	Humidity as a percent value given the longitude, latitude, and timestamp	73.62	14.37	26.50	99.70
visibility	The distance at which objects or landmarks can be clearly seen (miles)	8.80	1.76	1.70	9.90
temp	The temperature derived from the average of all available hourly data (F)	55.94	11.16	35.60	80.10
windspeed	The wind speed at 10 m above the surface (mph)	10.92	3.96	2.30	40.10
winddir	The wind direction at 10 m above the surface given the longitude, latitude, and timestamp	162.88	116.40	0.00	360.00
car	The number of surrounding passenger cars in the driver's visual environment	4.29	3.26	0.00	26.00
truck	The number of surrounding trucks in the driver's visual environment	0.51	0.77	0.00	11.00
car_prop	The proportion of cars within the driver's field of vision.	0.02	0.03	0.00	0.20
truck_prop	The proportion of trucks within the driver's field of vision.	0.01	0.02	0.00	0.20
road_prop	The proportion of road within the driver's field of vision.	0.36	0.13	0.00	0.98
tree_prop	The proportion of tree-covered areas within the driver's field of vision	0.10	0.10	0.00	0.50
grass_prop	The proportion of grass-covered areas within the driver's field of vision	0.02	0.04	0.00	0.55
mountain_prop	The proportion of mountain-covered areas within the driver's field of vision	0.00	0.02	0.00	0.20
barrier_prop	The proportion of median barriers areas within the driver's field of vision	0.00	0.01	0.00	0.10
sky_prop	The proportion of sky visible within the driver's field of vision	0.39	0.15	0.00	0.75
building_prop	The proportion of buildings within the driver's field of vision	0.02	0.04	0.00	0.25
prop_speed	The proportion of speed above the speed limit (a vehicle not speeding = 0)	0.03	0.06	0.00	0.91
close_veh_angle	The angle of the closest vehicle to the driver	35.26	25.00	0.00	173.57
close_veh_prop	The proportion of the closest vehicle within the driver's field of vision	0.01	0.02	0.00	0.13
close_veh_depth	The distance of the closest vehicle to the driver (ft)	53.39	57.85	0.00	170.00
speeding	If vehicle is speeding = 1, If not = 0	0.30	0.46	0.00	1.00
headway	The time headway (secs)	2.21	1.63	0.00	23.60
peak	If driver was traveling during the peak hour = 1, If not = 0 Peak hours = 7AM – 9AM, 4PM – 6PM	0.42	0.49	0.00	1.00
dow	The day of the week which the vehicle was traveling. Monday = 0, Sunday = 6	2.79	1.67	0.00	6.00
weekday	If driver was traveling on a weekday = 1, If not = 0	0.24	0.43	0.00	1.00
speed	The speed of the vehicle (mph)	53.69	16.05	10.00	99.57

for time-headway (Short, Typical, and Long) as defined in Table 5. The grouped random effect multinomial logit model, also referred to as the mixed multinomial logit regression, is an extended version of multinomial logistic regression. It enables the coefficients of variables to vary among individuals rather than being fixed, which accounts for the heterogeneity within the population. Unlike the standard logistic regression, where each individual has a fixed probability, in the grouped random effect multinomial logit regression, the probability of driver with a leading vehicle i with a certain time-headway classification l from a set of time-headway classifications J can be calculated as follows (Möhring, 2013):

$$P_{il} = \frac{e^{\beta' x_{il}}}{\sum_{j=1}^{J} e^{\beta' x_{ij}}}$$

In this context, x represents the factor and β' represents the fixed coef-

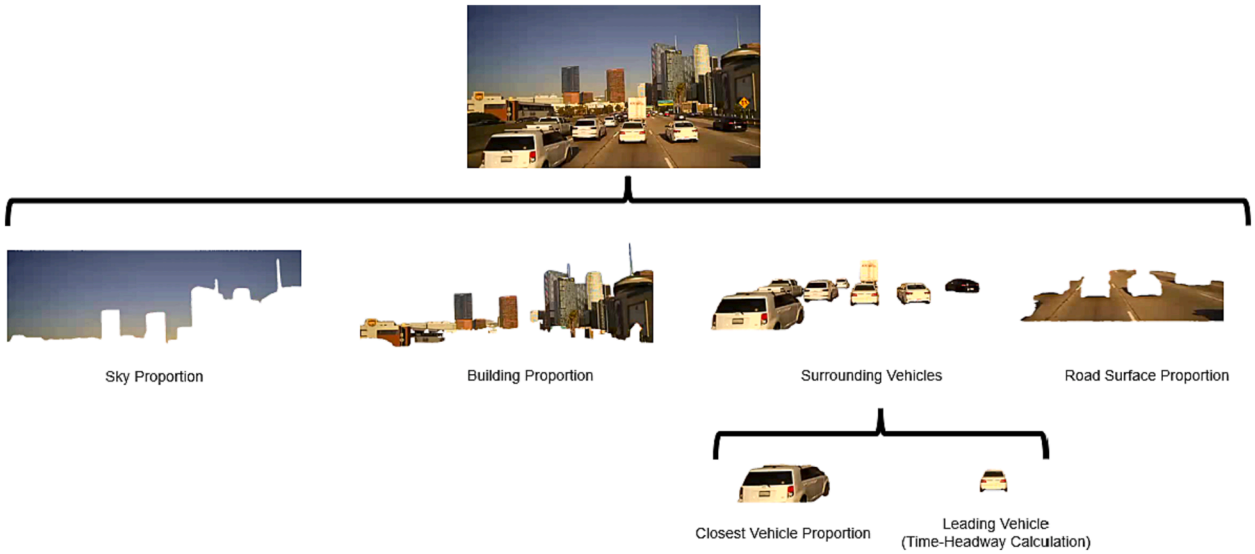

Fig. 11. Elements Within Drivers' Visual Environment.

ficient that applies to all drivers. However, in the mixed multinomial logit model, each individual has their own coefficient denoted as $\beta_i^{'}$. The probabilities are described as the likelihood of a driver with a leading vehicle i maintain a time-headway of level l given their specific vector of individual-specific coefficients, denoted as β_i. These probabilities can be calculated using the formula:

$$P_{il}|\beta_i = \frac{e^{\beta_i^{'} x_{il}}}{\sum_{j=1}^{J} e^{\beta_i^{'} x_{ij}}}$$

2.3. Grouped random effect bivariate probit model

A bivariate probit model is used to investigate the interplay between time-headway and speeding, aiming to uncover their relationship more seamlessly. Our analysis focuses on two binary outcomes: first, whether the driver was speeding, and second, whether there existed a time-headway, indicating the presence of a leading vehicle. The bivariate probit model operates under the assumption that the unobserved factors impacting both outcomes exhibit correlation. Put differently, there exists a connection between the error terms in the two equations. This correlation effectively encapsulates the association or dependence between the two binary outcomes. Bivariate models have been used in traffic safety research to simultaneously analyze interconnected variables (Aidoo, 2019; Fountas et al., 2018; Guo et al., 2017; Tarko and Azam, 2011; Yuan et al., 2020; Lee et al., 2021). Tarko and Azam (Tarko and Azam, 2011) successfully linked accident data from both police and hospitals and investigated contributing factors using a bivariate probit model. Guo et al. (Guo et al., 2017) applied a bivariate probit model to assess factors associated with e-bike involved accident and e-bike license plate and to account for the correlations between them. The license plate can regulate e-bikes aberrant riding behavior such as red-light running and violating traffic laws. Fountas et al. (Fountas et al., 2018) utilized segment-based and accident-based latent class ordered probit models to estimate injury severity. Aidoo et al. (Aidoo, 2019) conducted a bivariate probit analysis on child passengers' seating behavior and restraint use. Lee et al. (Lee et al., 2021) applied a bivariate probit model to analyze the injury severity of highway traffic accidents involving school-age children. Yuan et al. (Yuan et al., 2020) adopted a Bayesian bivariate probit approach to explore the severity of injuries in side accidents involving striking and struck vehicles.

The bivariate probit model is specified as follows (Plum, 2016):

$$y_{i1}^{*} = \beta_1^{'} x_{i1} + \varepsilon_{i1}, y_{i1} = 1 if y_{i1}^{*} > 0, y_{i1} = 0 otherwise \quad (10)$$

$$y_{i2}^{*} = \beta_2^{'} x_{i2} + \varepsilon_{i2}, y_{i2} = 1 if y_{i2}^{*} > 0, y_{i2} = 0 otherwise \quad (11)$$

where y_{1j}^{*} are latent response variables, $j = 1$ for when the driver is speeding (1 = speeding, 0 = not speeding, $j = 2$ for the when the driver has a time-headway (i.e., leading vehicle is present), and ε_{i1} and ε_{i2} are error terms that capture the effect of unobserved variables. The error terms are assumed to follow a bivariate normal distribution:

$$[\varepsilon_{i1}, \varepsilon_{i1}]\ N_2[0, 0, 1, 1, \rho], -1 < \rho < 1 \quad (11)$$

where ρ is the cross-equation error term indicating the unobserved shared factors. To capture the heterogeneity across the observations, a random parameter approach is applied. The random parameter, β_{ijk}, for ith observation, jth dependent variable, and jth coefficient is defined as follows:

$$\beta_{ijk} = \beta_{jk} + \varepsilon_{ijk} \quad (12)$$

where β_{ijk} is the mean of the coefficient and ε_{ijk} is a normally distributed error term with a mean zero and a variance σ^2. The model is developed using STATA (Plum, 2016).

2.4. Empirical analysis

2.4.1. Parameter variation across various samples

During the initial phase of our model development exercise, we focused on examining the variability of parameters across different samples. It has been observed in studies that large sample sizes have a tendency to amplify minor differences, making them appear statistically significant, even if they are truly insignificant. This can lead both researchers and transportation agencies astray (Faber and Fonseca, 2014). To address this issue, we obtained 10 samples, each consisting of 20,000 observations, from the dataset. We ensured that each sample maintained the same proportions as the original dataset. For each of these samples, we estimated the grouped random effect hurdle beta model, which accounted for both random and fixed effects. Various variable specifications based on the described variables in the data preparation section were tested.

In order to compare the parameters obtained from the models generated by each sample, we selected one of the ten samples as the benchmark. We then evaluated whether the parameters of the other models were statistically different from those of the benchmark sample. To facilitate this comparison, we generated a revised Wald test statistic

Fig. 12. Top: Sample Image Collected by Lytx, Middle: Sample Image[1] Segmented by Detectron2, Bottom: Sample Image Depth Visualization by Monodepth2 [1]The percentage on each object (e.g., car, truck) represents the confidence or probability associated with the detection of each object from the identification and segmentation of the Detectron2 output.

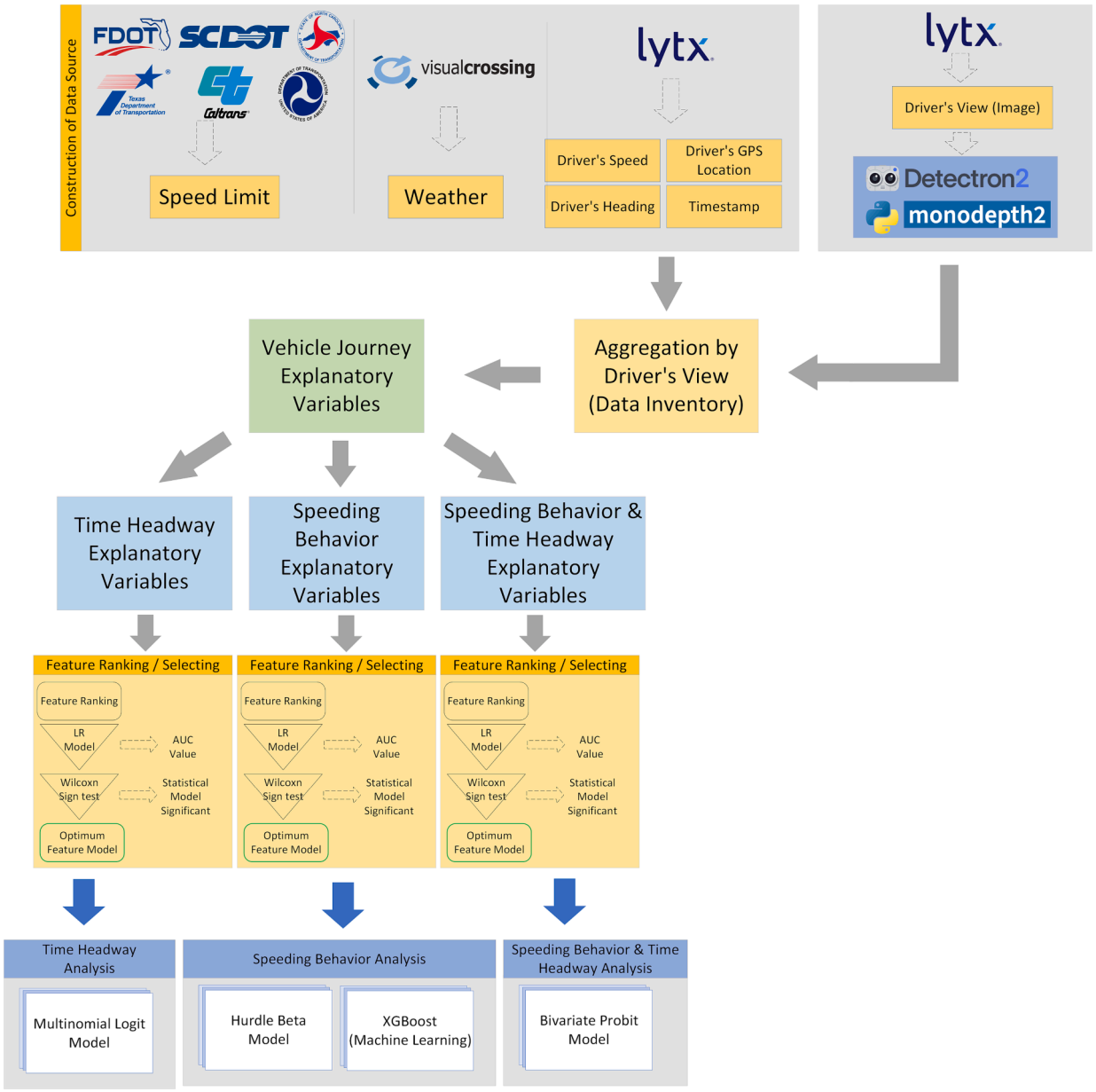

Fig. 13. Flowchart illustrating the utilization of data for each model in the analysis of the study.

specific to the benchmark model. The procedure for generating this statistic is as follows:

$$\text{Parameter test statistic} = abs\left[\frac{(sampleparameter - benchmarkparameter)}{\sqrt{SE_{sample}{}^2 + SE_{benchmark}{}^2}}\right]$$

If the computed test statistic for a parameter is higher than the 90 % t-statistic, it indicates a significant difference in that parameter. Using the test statistic calculation described above, revised t-statistics were computed for all parameters across all samples. Fig. 14 presents a box plot summary illustrating the variations in these test statistics across samples for all parameters. The figure clearly shows that the range of the test statistic is relatively narrow for most parameters, exceeding the 90 % significance threshold only for three parameters. Two of the parameters, headway and cloud cover, which exceeded the 90 % significance threshold was excluded from the grouped random effect hurdle beta model, headway was further analyzed in a grouped random effect bivariate probit regression model to ensure the relationship was well understood. The other variable was retained due to its intuitive nature. Given the overall stability observed across all samples, the benchmark model was selected for further analysis and discussion.

2.5. Speeding behavior

Table 7 and Table 8 present the outcomes of the grouped random effect hurdle beta model, which examines the modeling results related to if the driver was speeding (Table 7) and the proportion of speeding that exceeds the speed limit (Table 8). The modeling results are divided into two sections: the logistic regression component and the beta regression component. The outcomes of the logistic regression section determine whether drivers were speeding (i.e., if the proportion of speeding was above 0.05), while the results from the beta regression section disclose the actual proportion of speeding when drivers engage in speeding behavior. Table 9 presents the outcomes of the grouped random effect multinomial logit model, where the base outcome was the long time-headway category. Table 10 presents the outcomes of the grouped random effect bivariate probit model, where the relationship between

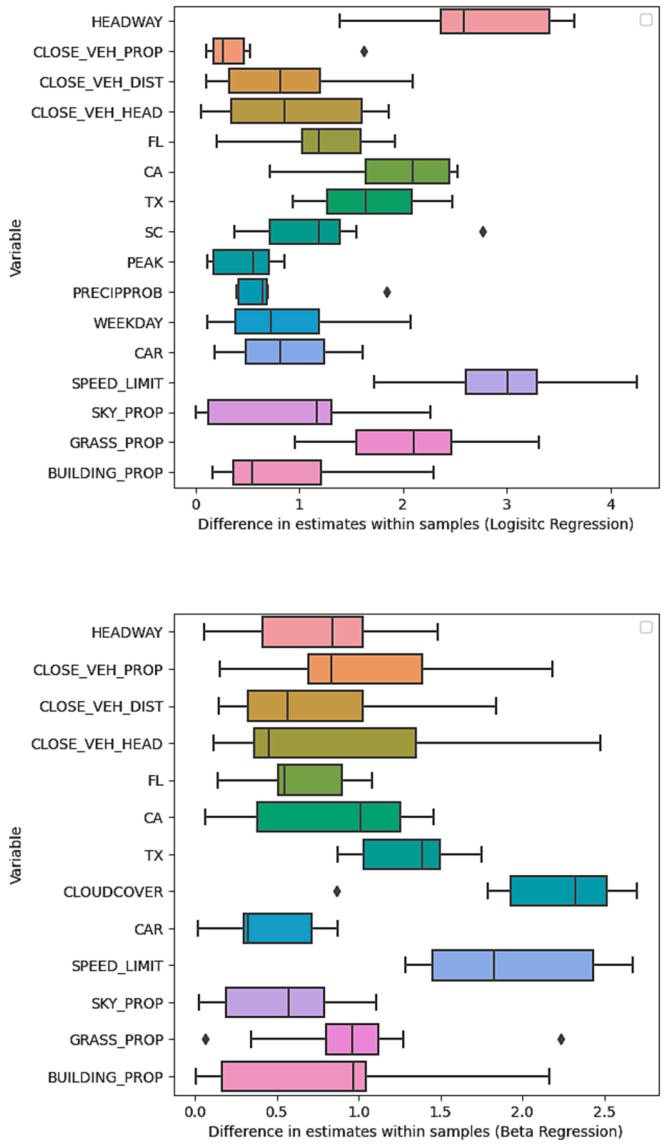

Fig. 14. Test Statistics (t-statistics) for Parameter Estimates Across Samples for Each Variable.

Table 7
Modeling the proportion of speeding instances exceeding the speed limit (Logistic Regression).

	Grouped Random Effect Hurdle Beta Model		Fixed Effect Hurdle Beta Model	
Logistic Regression	Mean	S.E.	Mean	S.E.
Constant	8.10**	0.37	5.80**	0.33
Peak Hour (1 = yes; 0 = no)	−0.43**	0.06	−0.28**	0.03
Weekday (1 = yes; 0 = no)	−0.23**	0.05	−0.16**	0.03
Probability of Precipitation	−2.00E-03**	5.38E-04	−1.42E-03**	3.36E-04
Speed Limit	−0.13**	0.01	−8.97E-02**	4.62E-03
FL (1 = yes; 0 = no)	−1.00**	0.08	−0.69**	0.05
CA (1 = yes; 0 = no)	−2.08**	0.13	−1.39**	0.06
TX (1 = yes; 0 = no)	−1.10**	0.09	−0.71**	0.05
SC (1 = yes; 0 = no)	−0.44**	0.09	−0.31**	0.06
Number of Surrounding Vehicles	−0.11**	0.01	−0.08**	0.01
Proportion of Open Sky	0.94**	0.20	0.55**	0.13
Proportion of Grass	−4.82**	0.73	−3.26**	0.47
Proportion of Buildings	−6.76**	0.77	−4.93**	0.50
DIC	399,241,000		399,246,000	

Table 8
Modeling the proportion of speeding instances exceeding the speed limit (Beta Regression).

	Grouped Random Effect Hurdle Beta Model		Fixed Effect Hurdle Beta Model	
Beta Regression	Mean	S.E.	Mean	S.E.
Constant	3.84**	0.14	3.02	0.13
Speed Limit	−1.08E-01**	2.47E-03	−8.81E-02	2.12E-03
FL (1 = yes; 0 = no)	−0.14**	0.03	−0.16	0.03
TX (1 = yes; 0 = no)	0.12**	0.03	0.11	0.03
Proportion of Open Sky	0.28**	0.06	0.21	0.07
Proportion of Grass	−1.29**	0.34	−1.20	0.31
Proportion of Buildings	−0.96**	0.32	−1.30	0.30
DIC	399,241,000		399,246,000	

speeding and headway is shown.

The analysis conducted in both parts demonstrates that factors related to a driver's visual environment have a significant impact on their speeding behavior. The proportion of sky visible to the driver serves as an indicator of the amount of open space (Table 6, *Proportion of Open Sky*; Table 7, *Proportion of Open Sky*). The findings suggest that drivers with more open space are more likely to speed and exhibit a higher proportion of speeding compared to the speed limit. Several significant factors contribute to reducing speed. The proportion of grass, referring to the median strip that separates opposing lanes, plays a role in decreasing speeding behavior and the likelihood of speeding (Table 6, *Proportion of Grass*; Table 7, *Proportion of Grass*). This can be attributed to drivers needing to exercise more caution when driving near median strips. Similarly, the proportion of buildings in the driver's view also has a similar effect. An increase in the proportion of buildings reduces the amount of speeding and the likelihood of speeding (Table 6, *Proportion of Buildings*; Table 7, *Proportion of Buildings*). The presence of building along roadways can have a calming effect on drivers, leading to reduced speeds. These features may serve as a visual cue for drivers to perceive a more urban or populated area, which tends to prompt them to drive at lower speeds. On the other hand, monotonous or featureless environments, such as long stretches of highway without much visual variety, can contribute to speeding. Lack of visual interest or stimulation may lead to decreased attention and increased monotony, which could potentially result in higher speeds.

The analysis of the logistic regressions reveals several noteworthy findings. Firstly, during peak hours, specifically from 7 AM to 9 AM and 4 PM to 6 PM, there is a lower likelihood of speeding (Table 6, *Peak Hour*). This result is consistent with expectations, as peak hours are associated with heavy traffic, which naturally limits the speed of drivers. Similarly, weekdays exhibit a lower probability of speeding compared to weekends, as more drivers are typically on the road for work during weekdays (Table 6, *Weekday*). This implies that drivers are more likely to adhere to speed limits when commuting for professional reasons. This finding is also confirmed with the fact that the presence of a higher number of vehicles surrounding the driver reduces the likelihood of speeding, and if the driver does choose to speed, the proportion of speeding is lower (Table 6, *Number of Surrounding Vehicles*; Table 7, *Number of Surrounding Vehicles*). The number of surrounding vehicles serves as a reliable surrogate measure for traffic density. As the number of vehicles around a driver increases, it indicates a higher level of traffic congestion and a greater density of vehicles on the road. Another significant finding is that higher probabilities of precipitation in the area lead to a decreased likelihood of speeding (Table 6, *Probability of Precipitation)*. This aligns with previous research, which indicates that driving at high speeds in rainy conditions can result in dangerous situations such as skidding and hydroplaning. Consequently, drivers tend to be more cautious and avoid excessive speeding during such weather conditions. Furthermore, the analysis demonstrates that roadways with higher speed limits are associated with drivers who are less likely to speed (Table 6, *Speed Limit)*. This finding is consistent with a previous

Table 9
Modeling results of time-headway (Multinomial Logit Model).

Base Outcome (Long)	Grouped Random Effect Multinomial Logit Model				Fixed Effect Multinomial Logit Model			
	Typical		Short		Typical		Short	
	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Constant	−1.56**	0.18	−0.37**	0.18	−1.47**	0.18	−0.32*	0.17
Speed Limit	4.36E-02**	2.79E-03	4.05E-02**	2.81E-03	4.26E-02**	2.69E-03	3.87E-02**	2.59E-03
TX (1 = yes; 0 = no)	−2.97E-03	5.02E-02	0.39**	0.05	4.68E-03	4.86E-02	0.36**	0.05
SC (1 = yes; 0 = no)	0.17**	0.07	0.04	0.07	0.16**	0.07	0.05	0.06
Number of Surrounding Vehicles	−0.04**	0.01	0.05**	0.01	−0.04**	0.01	0.04**	0.01
Proportion of Open Sky	−0.96**	0.14	−4.68**	0.16	−1.00**	0.14	−4.28**	0.14
Proportion of Buildings	−4.19**	0.50	−6.65**	0.52	−4.16**	0.49	−6.32**	0.48
Log-likelihood at convergence	−20,014				−20,071			
BIC	40,187				40,280			

* Significant at the 90% confidence level.
** Significant at the 95% confidence level.

Table 10
Modeling results of relationship between speeding and time-headway (Bivariate Probit Model).

	Grouped Random Effect Bivariate Probit Model				Fixed Effect Bivariate Probit Model			
	Speeding		Headway		Speeding		Headway	
	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Constant	5.19**	0.15	2.41**	0.15	4.28**	0.10	2.36**	0.13
Peak Hour (1 = yes; 0 = no)	−0.23**	0.03	−0.07**	0.02	−0.19**	0.02	−0.07**	0.02
Weekday (1 = yes; 0 = no)	−0.33**	0.04	−0.08**	0.03	−0.27**	0.03	−0.08**	0.03
Probability of Precipitation	−1.33E-03**	2.39E-04	–	–	−1.12E-03**	1.96E-04	–	–
Speeding (1 = yes; 0 = no)	–	–	−1.06**	0.07	–	–	−1.06**	0.06
Speed Limit	−7.45E-02**	2.15E-03	−1.54E-02**	1.97E-03	−6.15E-02**	1.52E-03	−1.57E-02**	1.73E-03
FL (1 = yes; 0 = no)	−0.62**	0.04	–	–	−0.52**	0.03	–	–
CA (1 = yes; 0 = no)	−1.15**	0.05	−0.50**	0.03	−0.94**	0.03	−0.46**	0.03
TX (1 = yes; 0 = no)	−0.64**	0.04	–	–	−0.52**	0.03	–	–
SC (1 = yes; 0 = no)	−0.24**	0.04	–	–	−0.20**	0.03	–	–
Number of Surrounding Vehicles	−5.50E-02**	3.92E-03	1.12E-01**	4.11E-03	−4.78E-02**	3.22E-03	1.01E-01**	3.71E-03
Proportion of Open Sky	–	–	−3.48**	0.08	–	–	−3.22**	0.07
Proportion of Grass	−1.99**	0.34	−0.59*	0.28	−1.59**	0.28	−0.55*	0.26
Proportion of Buildings	−2.85**	0.35	−2.51**	0.27	−2.49**	0.29	−2.48**	0.24
Error-term correlation***	0.60**	0.04	–	–	0.66**	0.03	–	–
Log-likelihood at convergence	−21,133				−21,268			
BIC	42,524				42,763			

* Significant at the 90% confidence level.
** Significant at the 95% confidence level.
*** The statistically significant error-term correlation indicates that two outcomes are inter-related.

study, as it is generally easier for drivers to exceed lower speed limits (Ugan et al., 2303). When the speed limit is already set relatively high, drivers are less inclined to engage in further speeding. Examining the various state binary variables, it can be observed that drivers from California are less likely to speed compared to drivers from other states (Table 6, *CA)*. On the other hand, drivers in Florida and Texas exhibit similar probabilities of speeding (Table 6, *FL, TX)*. South Carolina stands out with the highest coefficient, indicating that drivers from this state are more likely to engage in speeding (Table 6, *SC)*. This finding is also supported by NHTSA 2020 Speeding report, which has highlighted the highest incidence of speeding were among drivers in South Carolina (NHTSA, 2020).

In the beta regression analysis, the effect of speed limits on speeding behavior is consistent with the logistic regression findings. Drivers tend to not engage in speeding when the speed limit is already set high, as they are already driving at a relatively high speed (Table 7, *Speed Limit)*. The state binary variables yield an interesting result in the beta regression analysis. Although the logistic regression indicated that drivers from Florida and Texas have a similar likelihood of speeding compared to drivers from other states, the beta regression reveals that if a Florida driver does choose to speed, they are less likely to speed at a high proportion (Table 7, *FL)*. On the other hand, drivers in Texas tend to speed at a higher proportion if they choose to speed (Table 7, *TX)*.

The grouped random effect multinomial logit analysis aimed to investigate the factors influencing the classification of time-headways into three categories: Short, Typical, and Long. The results indicate that as the number of vehicles in a driver's field of view increases, there is a higher likelihood of maintaining a short headway (Table 8, *Number of Surrounding Vehicles)*. This finding supports the idea that drivers may opt for a short time-headway due to the insufficient number of safe gaps between vehicles. Notably, the analysis of state binary variables revealed intriguing results. Drivers from Texas displayed a higher likelihood of maintaining a short time-headway, while drivers from South Carolina were more likely to have a typical time-headway (Table 8, *TX, SC)*. These findings suggest the presence of region-specific factors that influence drivers' choices regarding headway lengths.

The grouped random effect bivariate probit model analysis aimed to understand the relationship between speeding and time-headway. The first outcome pertained to whether the driver was speeding, while the second outcome focused on the presence or absence of a time-headway, which indicates the existence of a leading vehicle. It is worth noting that the second outcome differs somewhat from the classification utilized in grouped random effect multinomial logit model. In the grouped random effect multinomial logit model, the classification revolves around time-

headway (Short, Typical, or Long), implying that each observation corresponds to a scenario with a leading vehicle. However, in the grouped random effect bivariate probit model, the second outcome is binary and models the presence or absence of a time-headway, allowing for scenarios without a leading vehicle as well. The results of the grouped random effect bivariate probit model revealed a strong association between the presence of speeding and an increased likelihood of drivers having a time-headway (i.e., presence of a leading vehicle). It appears that drivers who engage in speeding intentionally create a longer time headway as a safety measure. By keeping a greater distance from the vehicle ahead, they provide themselves with additional time and space to react and maneuver effectively in unexpected events or emergencies. Ayres, T. J., et al. also discovered that when traveling at speeds greater than approximately 50 mph, a wide range of time-headways can be observed (Ayres et al., 2001). The results revealed a strong association between speeding and time-headway with the peak hour, weekday, and the surrounding vehicles. During the peak hour, weekday, and many surrounding vehicles, i.e. congestion, which means less likelihood of speeding and a longer time-headway (Table 9, *Peak Hour, Weekday, Number of Surrounding Vehicles)*. As confirmed by the previous models, drivers are less likely to speed at higher speed limits, which also means they are more likely to have a long time-headway (Table 9, *Speed Limit)*. California drivers are more likely to leave a long time-headway when compared to the drivers of other states (Table 9, *CA)*. Generally, we can conclude if a driver engages in speeding behavior, the driver is more likely to have a longer headway.

Two models proposed for speeding are the grouped random effect hurdle beta model (Model 1) and the fixed effect hurdle beta model (Model 2). The mean, standard error, and the DIC values for Models 1 and 2 are provided in Tables Table 7 and Table 8, respectively. According to Spiegelhalter et al. (Spiegelhalter et al., 2005) it is difficult to determine what would constitute an important difference in DIC. Very roughly, differences of more than 10 might definitely rule out the model with the higher DIC. Differences between 5 and 10 are substantial. If the difference in DIC is less than 5, and the models make very different inferences, then it could be misleading just to report the model with the lowest DIC. The DIC value for Model 1 (399,241,000) is smaller than Model 2 (399,246,000); subsequently, this indicates that Model 1 is superior to Model 2. Two models proposed for time-headway are the grouped random effect multinomial logit model (Model 3) and the fixed effect multinomial logit model (Model 4). The mean and standard error, and the BIC values for Models 3 and 4 are provided in Table 9. The BIC value for Model 3 (40,187) is smaller than Model 4 (40,280); subsequently, this indicates that that Model 3 is superior to Model 4. Therefore, the grouped random effect multinomial logit model is superior to the fixed effect multinomial logit model. Two models proposed for the speeding and time-headway relationship are grouped random effects bivariate probit model (Model 5) and the fixed effect bivariate probit model (Model 6). The mean and standard error, and the BIC values for Models 5 and 6 are provided in Table 10. The BIC value for Model 5 (42,524) is smaller than Model 6 (42,763); subsequently, this indicates that Model 5 is superior to Model 6. Therefore, the grouped random effect bivariate probit model is superior to the fixed effect bivariate probit model.

3. Results and discussion

3.1. Machine learning

The significance of XGBoost in this study lies in its ability to effectively model and estimate different levels of speeding (Low, Medium, and High Speeding) based on a dataset comprising 1,340,035 driver view images. XGBoost, short for eXtreme Gradient Boosting, was chosen as the modeling technique for its robustness and efficiency in handling complex relationships within the data. The input to the XGBoost model consists of various features extracted from the driver view images, allowing the model to capture patterns and associations that contribute to different levels of speeding behavior. These features may include visual cues from the driver's environment, road conditions, and other relevant factors. The output of the XGBoost model is a prediction of the likelihood of different speeding levels for each instance in the dataset, ranging from Low to Medium, High, and also including a non-speeding level for valid comparison.

In previous studies, XGBoost has consistently proven to be a superior model for predicting speeding behavior. Its effectiveness in handling complex relationships within the data, capturing nonlinear patterns, and providing accurate predictions has positioned it as a robust choice for modeling driving behaviors. The selection of XGBoost in this study aligns with its track record of outperforming other models in similar contexts. However, one challenge with machine learning models, including XGBoost, is their reputation as "black boxes," meaning that they lack interpretability. To address this challenge and enhance the interpretability of the model, the authors employed SHAP (SHapley Additive exPlanations) values.

SHAP values provide insights into the importance of each feature for a given prediction, allowing the researchers to understand the impact of individual parameters on the model's output. This interpretability is crucial for translating the model's predictions into actionable insights and understanding the factors that contribute to different levels of speeding behavior. In Fig. 15, a summary plot showcases the SHAP values for the XGBoost model. High and low feature values are represented by red and blue dots, respectively, with a vertical line at 0.0 separating positive and negative predictions. This visual representation allows researchers and stakeholders to grasp the impact of specific features on the model's predictions. For instance, the finding that higher speed limits are associated with a greater likelihood of non-speeding or low speeding aligns with outcomes observed in statistical models, providing a consistent and comprehensible narrative. Moreover, the convergence of interpretations derived from SHAP values with coefficients obtained from statistical methods reinforces the reliability and representativeness of the insights derived from the model. This convergence not only enhances confidence in the model's interpretability but also makes the interpretations applicable to the entire dataset, ensuring a broader and more generalizable understanding of speeding behaviour.

4. Conclusions

In this study, we aim to uncover both the factors influencing whether a driver engages in speeding and the factors affecting the extent or degree of speeding for regions known for high speeding. To conduct the analysis, a vast dataset of approximately 3,400,000 images was collected from 15 cities located in five states known for having high rates of speeding in the United States. The uniqueness of the image dataset lies in its ability to provide not only the driver's viewpoint captured in an image but also the precise GPS location, speed, and timestamp associated with it. These images were carefully verified using specific criteria to ensure they accurately represented the driver's view. Each image was then linked to its corresponding speed limit on the road using GPS coordinates. To extract relevant information from the images, a detection and segmentation algorithm was applied. This algorithm allowed for the determination of proportions associated with various key elements within the images. Additionally, a depth estimation model was utilized to estimate the distance between the driver and the elements within each image. The collected data was aggregated, resulting in a dataset of approximately 1,340,000 images. This aggregated database encompassed information pertaining to various elements within the driver's view, the driver's location, timestamp, weather conditions, and speed limit data. Overall, the study utilized a large-scale dataset and sophisticated algorithms to investigate the impact of the driver's visual environment on speeding behavior. The collected information was then analyzed using a grouped random effect hurdle beta model to gain

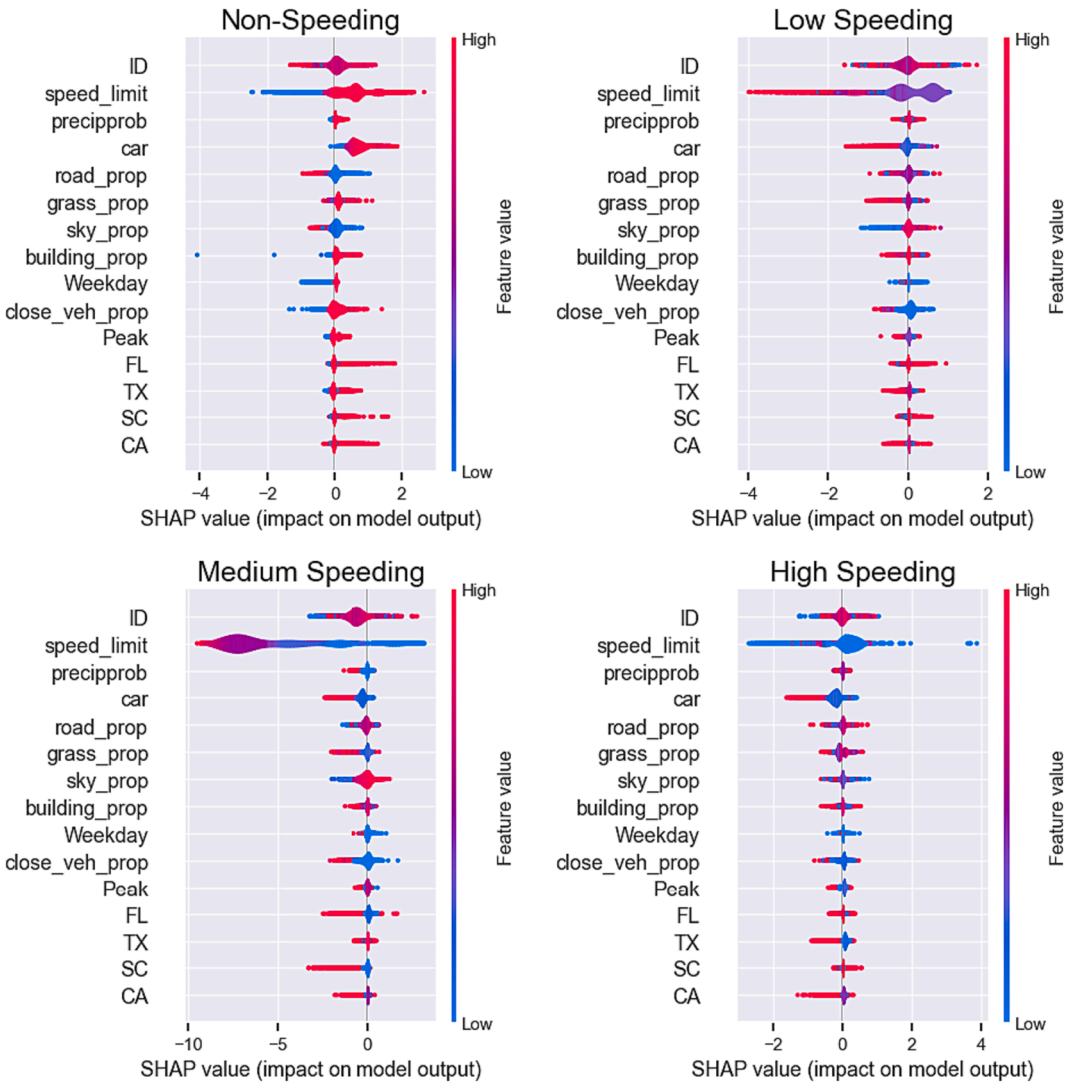

Fig. 15. Summary plot of XGBoost model with SHAP values.

insights into both the factors influencing whether a driver speeds and the factors affecting the degree of speeding.

In each of the statistical models (grouped random effect hurdle beta model, grouped random effect multinomial logit model, and grouped random effect bivariate probit model), considering the fixed effects model for comparison, were utilized to estimate the speeding proportion and identify significant contributing factors. Factors associated with the driver's visual environment and location were taken into account during the model estimation and found to be statistically significant. By incorporating random effects into the models, it became possible to estimate both fixed effects (representing the average effect across all drivers) and random effects (capturing specific effects within each driver). These random effects components accounted for driver heterogeneity and improved the ability to explain the observed variation in the data. Various factors were examined, including the driver's visual environment, weather conditions, and location. The results indicated that drivers in areas with more open space and fewer surrounding elements such as trees and buildings were more likely to engage in speeding and exhibit higher proportions of speeding relative to the speed limits. The probability of precipitation increased the likelihood of drivers reducing their speeds due to safety concerns and the desire to avoid hazardous situations like hydroplaning. The state in which the driver was located also had a significant effect. Drivers in South Carolina were more likely to speed compared to drivers in other states. Drivers in California were less likely to speed, however Texas and Florida drivers were similar in the likelihood of speeding, but speed at different proportions if they choose to speed. Furthermore, the study delved into the correlation between speed and headway, emphasizing the significance of comprehending this connection. The inherent dangers of speeding are compounded when coupled with a reduced time-headway. Findings from the research revealed that drivers who were more inclined to engage in speeding also tended to maintain longer time-headways. This suggests a positive correlation between the proportion of speeding instances and the length of time-headway chosen by drivers. In other words, as drivers exhibited a higher propensity for speeding, they were more likely to adopt a longer time-headway when driving. Additionally, the length of the time-headway was found to be closely associated with geographical regions, with South Carolina drivers exhibiting a propensity for shorter time-headways, while Texas drivers tended to maintain longer time-headways. In conclusion, this study effectively examined the various factors that influence speeding behavior and established a clear understanding of the interplay between speeding and time-headway. The insights gained from this research have significant implications for the development and implementation of advanced traffic management systems focused on mitigating instances of speeding. These systems can leverage the findings to effectively address speeding incidents on highways, such as by implementing intelligent speed monitoring technologies or dynamic speed limit signs. Open spaces on

highways may lead to a perception of wider and more open roads, which can inadvertently encourage drivers to speed. By strategically adding visual elements such as vegetation, landscaping, or architectural features alongside the road, drivers' perception of the road width can be altered, creating an impression of narrower lanes and promoting slower and safer driving speeds. Installing weather-activated speed warning signs that detect rain or wet conditions can be effective in reminding drivers to slow down. These signs can display messages like "Reduce Speed in Rain" or "Slippery When Wet," providing visual cues to encourage drivers to drive at appropriate speeds during inclement weather. By leveraging these findings, strategies can be devised to effectively reduce speeding incidents and enhance overall road safety.

CRediT authorship contribution statement

Mohamed Abdel-Aty: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing. **Jorge Ugan:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Zubayer Islam:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

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

Data availability

The data that has been used is confidential.

Acknowledgments

The authors also acknowledge Lytx for providing vehicle data. This paper and its contents, including conclusions and results, are solely those of the authors; they do not represent opinions or policies of Lytx.

References

Abdelraouf, A., Abdel-Aty, M., Mahmoud, N., 2022. Sequence-to-Sequence Recurrent Graph Convolutional Networks for Traffic Estimation and Prediction Using Connected Probe Vehicle Data. IEEE Transactions on Intelligent Transportation Systems.

Adu-Gyamfi, Y.O., Sharma, A., Knickerbocker, S., Hawkins, N., Jackson, M., 2017. Framework for evaluating the reliability of wide-area probe data. Transportation Research Record 2643, 93–104.

Ahsani, V., Amin-Naseri, M., Knickerbocker, S., Sharma, A., 2019. Quantitative analysis of probe data characteristics: Coverage, speed bias and congestion detection precision. Journal of Intelligent Transportation Systems 23, 103–119.

Aidoo, E.N., et al., 2019. A bivariate probit analysis of child passenger's sitting behaviour and restraint use in motor vehicle. Accident Analysis & Prevention 129, 225–229.

Ambo, T.B., Ma, J., Fu, C., 2020. Investigating influence factors of traffic violation using multinomial logit method. International Journal of Injury Control and Safety Promotion 28, 78–85.

Anastasopoulos, P.C., Mannering, F.L., 2011. An empirical assessment of fixed and random parameter logit models using crash-and non-crash-specific injury data. Accident Analysis & Prevention 43, 1140–1147.

Atombo, C., Wu, C., Zhong, M., Zhang, H., 2016. Investigating the motivational factors influencing drivers intentions to unsafe driving behaviours: Speeding and overtaking violations. Transportation Research Part f: Traffic Psychology and Behaviour 43, 104–121.

Ayres, T., Li, L., Schleuning, D. & Young, D. in *ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No. 01TH8585)*. 826-829 (IEEE).

Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Analytic Methods in Accident Research 9, 1–15.

Bassani, M., Dalmazzo, D., Marinelli, G. & Cirillo, C. The effects of road geometrics and traffic regulations on driver-preferred speeds in northern Italy. An exploratory analysis. *Transportation research part F: traffic psychology and behaviour* **25**, 10-26 (2014).

Bhowmik, T., Yasmin, S., Eluru, N., 2019. A multilevel generalized ordered probit fractional split model for analyzing vehicle speed. Analytic Methods in Accident Research 21, 13–31.

Boucher, J.-P., Santolino, M., 2010. Discrete distributions when modeling the disability severity score of motor victims. Accident Analysis & Prevention 42, 2041–2049.

Cai, Q., Lee, J., Eluru, N., Abdel-Aty, M., 2016. Macro-level pedestrian and bicycle crash analysis: Incorporating spatial spillover effects in dual state count models. Accident Analysis & Prevention 93, 14–22.

Cai, Q., Abdel-Aty, M., Lee, J., Eluru, N., 2017. Comparative analysis of zonal systems for macro-level crash modeling. Journal of Safety Research 61, 157–166.

Cai, Q., Abdel-Aty, M., Lee, J., Wang, L., Wang, X., 2018. Developing a grouped random parameters multivariate spatial model to explore zonal effects for segment and intersection crash modeling. Analytic Methods in Accident Research 19, 1–15.

Cai, Q., Abdel-Aty, M., Mahmoud, N., Ugan, J., Ma'en, M., 2021. Developing a grouped random parameter beta model to analyze drivers' speeding behavior on urban and suburban arterials with probe speed data. Accident Analysis & Prevention 161, 106386.

Cai, Q., Abdel-Aty, M., Zheng, O., Wu, Y., 2022. Applying machine learning and google street view to explore effects of drivers' visual environment on traffic safety. Transportation Research Part c: Emerging Technologies 135, 103541.

Crossing, V. (2023).

Edquist, J., Rudin-Brown, C.M., Lenné, M.G., 2012. The effects of on-street parking and road environment visual complexity on travel speed and reaction time. Accident Analysis & Prevention 45, 759–765.

Eluru, N., Chakour, V., Chamberlain, M., Miranda-Moreno, L.F., 2013. Modeling vehicle operating speed on urban roads in Montreal: A panel mixed ordered probit fractional split model. Accident Analysis & Prevention 59, 125–134.

Faber, J., Fonseca, L.M., 2014. How sample size influences research outcomes. Dental Press Journal of Orthodontics 19, 27–29.

Fountas, G., Anastasopoulos, P.C., Mannering, F.L., 2018. Analysis of vehicle accident-injury severities: A comparison of segment-versus accident-based latent class ordered probit models with class-probability functions. Analytic Methods in Accident Research 18, 15–32.

Ghasemzadeh, A., Ahmed, M.M., 2019. Quantifying regional heterogeneity effect on drivers' speeding behavior using SHRP2 naturalistic driving data: A multilevel modeling approach. Transportation Research Part c: Emerging Technologies 106, 29–40.

Guo, Y., Zhou, J., Wu, Y., Chen, J., 2017. Evaluation of factors affecting e-bike involved crash and e-bike license plate use in China using a bivariate probit model. Journal of Advanced Transportation 2017.

Hu, J., Fontaine, M.D., Park, B.B., Ma, J., 2016. Field evaluations of an adaptive traffic signal—using private-sector probe data. Journal of Transportation Engineering 142, 04015033.

Huang, H., Abdel-Aty, M.A., Darwiche, A.L., 2010. County-level crash risk analysis in Florida: Bayesian spatial modeling. Transportation Research Record 2148, 27–37.

Isola, P.D., et al., 2019. Google Street View assessment of environmental safety features at the scene of pedestrian automobile injury. Journal of Trauma and Acute Care Surgery 87, 82–86.

Lee, J., Abdel-Aty, M., Cai, Q., 2017. Intersection crash prediction modeling with macro-level data from various geographic units. Accident Analysis & Prevention 102, 213–226.

Lee, J., Yasmin, S., Eluru, N., Abdel-Aty, M., Cai, Q., 2018. Analysis of crash proportion by vehicle type at traffic analysis zone level: A mixed fractional split multinomial logit modeling approach with spatial effects. Accident Analysis & Prevention 111, 12–22.

Lee, J., Mao, S., Abdel-Aty, M., Fu, W., 2021. Use of bivariate random-parameter probit model to analyze the injury severity of highway traffic crashes involving school-age children. Transportation Research Record 2675, 530–537.

Ma, L., Yan, X., Weng, J., 2015. Modeling traffic crash rates of road segments through a lognormal hurdle framework with flexible scale parameter. Journal of Advanced Transportation 49, 928–940.

Ma, L., Yan, X., Wei, C., Wang, J., 2016. Modeling the equivalent property damage only crash rate for road segments using the hurdle regression framework. Analytic Methods in Accident Research 11, 48–61.

Mahmoud, N., Abdel-Aty, M., Cai, Q., 2021. Factors contributing to operating speeds on arterial roads by context classifications. Journal of Transportation Engineering, Part a: Systems 147, 04021040.

Mahmoud, N., Abdel-Aty, M., Cai, Q., Abuzwidah, M., 2022. Analyzing the difference between operating speed and target speed using mixed-effect ordered logit model. Transportation Research Record 2676, 596–607.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1–16.

Meaney, C., Moineddin, R., 2014. A Monte Carlo simulation study comparing linear regression, beta regression, variable-dispersion beta regression and fractional logit regression at recovering average difference measures in a two sample design. BMC Medical Research Methodology 14, 1–22.

Moeller, M. M. Methods for analyzing proportions. (2013).

Möhring, K., 2013. & Schmidt-Catran, A. Stata module to provide multilevel tools, MLT.

Mooney, S.J., et al., 2020. Development and validation of a Google Street View pedestrian safety audit tool. Epidemiology (cambridge, Mass.) 31, 301.

Mullahy, J., 1986. Specification and testing of some modified count data models. Journal of Econometrics 33, 341–365.

NHTSA. (National Center for Statistics and Analysis, 2020).

Ospina, R., Ferrari, S.L., 2012. A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis 56, 1609–1623.

PE, W. E. M., Coppola, N. & Golombek, Y. Urban clear zones, street trees, and road safety. *Research in Transportation Business & Management* **29**, 136-143 (2018).

Plum, A., 2016. bireprob: An estimator for bivariate random-effects probit models. The Stata Journal 16, 96–111.

Rundle, A.G., Bader, M.D., Richards, C.A., Neckerman, K.M., Teitler, J.O., 2011. Using Google Street View to audit neighborhood environments. American Journal of Preventive Medicine 40, 94–100.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. Winbugs User Manual. (2005).

Tarko, A., Azam, M.S., 2011. Pedestrian injury analysis with consideration of the selectivity bias in linked police-hospital data. Accident Analysis & Prevention 43, 1689–1695.

Tay, R., Choi, J., Kattan, L., Khan, A., 2011. A multinomial logit model of pedestrian–vehicle crash severity. International Journal of Sustainable Transportation 5, 233–249.

Ugan, J., Abdel-Aty, M., Cai, Q., Mahmoud, N., Al-Omari, M., e., 2022. Effect of various speed management strategies on bicycle crashes for urban roads in central Florida. Transportation Research Record 2676, 544–555.

Ugan, J., Abdel-Aty, M. & Islam, Z. Using Connected Vehicle Trajectory Data to Evaluate the Effects of Speeding. *arXiv preprint arXiv:2303.16396* (2023).

Washington, S., Congdon, P., Karlaftis, M. & Mannering, F. in *Transportation Research Board Annual Conference, TRB, Washington, DC.*

Wu, Y., Abdel-Aty, M., Park, J., Selby, R.M., 2018. Effects of real-time warning systems on driving under fog conditions using an empirically supported speed choice modeling framework. Transportation Research Part c: Emerging Technologies 86, 97–110.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y. & Girshick, R. Detectron2. (2019).

Xu, P., Huang, H., Dong, N., Wong, S., 2017. Revisiting crash spatial heterogeneity: a Bayesian spatially varying coefficients approach. Accident Analysis & Prevention 98, 330–337.

Young, S. in *Proceedings of the 2007 Mid-Continent Transportation Research Symposium.* 16-17 (Citeseer).

Yu, S.-Y., Malawade, A.V., Muthirayan, D., Khargonekar, P.P., Al Faruque, M.A., 2021. Scene-graph augmented data-driven risk assessment of autonomous vehicle decisions. IEEE Transactions on Intelligent Transportation Systems 23, 7941–7951.

Yuan, Q., Xu, X., Xu, M., Zhao, J., Li, Y., 2020. The role of striking and struck vehicles in side crashes between vehicles: Bayesian bivariate probit analysis in China. Accident Analysis & Prevention 134, 105324.

Zeng, Q., Wen, H., Huang, H., Abdel-Aty, M., 2017. A Bayesian spatial random parameters Tobit model for analyzing crash rates on roadway segments. Accident Analysis & Prevention 100, 37–43.