Investigation on the Driver-Victim Pairs in Pedestrian and Bicyclist Crashes by Latent Class Clustering and Random Forest Algorithm

Chunwu Zhu*
Texas A&M Transportation Institute (TTI), Texas A&M University, Texas, USA
Department of Landscape Architecture and Urban Planning, School of Architecture, Texas A&M University, Texas, USA
Email: chunwu.zhu@tamu.edu

Charles T. Brown
Equitable Cities LLC, New Jersey, USA
Email: charlesbrown@equitablecities.com

Bahar Dadashova*
Texas A&M Transportation Institute (TTI), Texas A&M University, Texas, USA
Email: b-dadashova@tti.tamu.edu

Xinyue Ye
Department of Landscape Architecture and Urban Planning, School of Architecture, Texas A&M University, Texas, USA
Email: xinyue.ye@tamu.edu

Soheil Sohrabi
Safe Transportation Research and Education Center (SafeTREC), University of California, Berkeley
Email: sohrabi@berkeley.edu

Ingrid Potts
Texas A&M Transportation Institute (TTI), Texas A&M University, Texas, USA
Email: i-potts@tti.tamu.edu

Acknowledgment

The work conducted in this paper was funded by the University Transportation Center Safety through Disruption (Safe-D) Project ID 06-001. The authors would like to thank the three anonymous reviewers whose feedback and comments helped to improve the quality of the manuscript significantly.

***Corresponding Author:** *Bahar Dadashova, Ph.D. b-dadashova@tti.tamu.edu*

ABSTRACT

Pedestrians and bicyclists from marginalized and underserved populations experienced disproportionate fatalities and injury rates due to traffic crashes in the US. This disparity among road users of different races and the increasing trend of traffic risk for underserved racial groups called for an urgent agenda for transportation policy making and research to ensure equity in roadway safety. Pedestrian and bicyclist crashes involved drivers and pedestrians/bicyclists; the latter were usually victims. Traditional safety studies did not account for the interaction between the two parties and assumed that they were independent from each other. In this study we paired the driver and pedestrian/bicyclist involved in the same crash to understand the socioeconomic and demographic make-up of the two parties involved in crashes and assessed the geographic distribution of these crashes and crash-contributing factors. For this purpose, we applied the latent class clustering analysis (LCA) to classify different crash types and analyze the patterns of the crashes based on the income and ethnicity of both drivers and victims involved in pedestrian and bicyclist crashes. We then used random forest algorithms and partial dependence plots (PDPs) to model and interpreted the contributing factors of the clusters in both pedestrian and bicyclist models. The clustering results showed a pattern of social segregation in pedestrian and bicyclist crashes that drivers and victims with similar socioeconomic characteristics tend to be involved in one crash. Pedestrian/bicyclist exposure, driver's age, victim's age, year of the car in use, annual average daily traffic (AADT), speed limit, roadbed width, and lane width were the most influential factors contributing to this pattern. Crashes that involved drivers and victims with lower income and non-white ethnicity tended to happen in the location with higher pedestrian/bicyclist exposure, higher speed limit, and wider road. The findings of this research can help to inform the decision-making process for improving safety to ensure equitable and sustainable safety for all road users and communities.

Keywords: driver-victim pairs, pedestrian crashes, bicyclist crashes, latent class clustering, random forest

1. INTRODUCTION

Crash statistics in the US showed that vulnerable road users (VRUs) from marginalized and underserved populations experienced disproportionate fatalities and injury rates due to traffic crashes. According to a report from Governors Highway Safety Association (GHSA), Black, Indigenous, and People of Color (BIPOC) experienced disproportionate traffic crash fatalities in the US from 2015-2019. The nationwide total traffic deaths were 145.6 and 68.5 per 100,000 population for American Indian/Alaska Native and Black, respectively, higher than 58.1 per 100,000 for total population (GHSA, 2021). For pedestrian crashes, a report from National Highway Traffic Safety Administration (NHTSA) showed that the pedestrian fatality rate for the white population is 1.5/100,000 in 2018, while the pedestrian fatality rate for the Black population is 2.94/100,000, which is twice than the white pedestrian fatality rate (Glassbrenner et al., 2022). Meanwhile, the motor vehicle traffic fatality for the Black population has increased by 23 percent from 2019 to 2020, while for total population, it only increased by 7 percent (NHTSA, 2021). These disparities in the distribution of traffic crashes among VRUs of different races and the increasing trend of traffic risk for underserved racial groups suggested an urgent agenda for transportation policy and research to ensure equity in roadway safety.

Traditional approaches to roadway safety, such as predictive and systemic tools safety analysis, usually studied various road users and roadway infrastructure characteristics to predict the crash frequency and severity and develop implementable solutions for preventing crashes. At the individual level, roadway safety research investigated the influential factors such as the demographic and economic and behavioral features of both parties involved in the crash (Hasheminejad et al., 2018; Balakrishnan et al., 2019; Mokhtarimousavi et al., 2020), roadway environment such as speed limit, number of lanes, and traffic control of the road segment where the crash happened (Sivasankaran & Balasubramanian, 2020; Xiao et al., 2022), and other circumstances of the crash like weather condition and surface condition (Weiss et al., 2014; Li et al., 2018). Human factors play an important role in a traffic crash for both parties. Typical human factors like belligerent driving behavior and violations of traffic rules are deeply rooted in the road users' socioeconomic backgrounds, which shape the different levels of vulnerability for road users with different socioeconomic characteristics. Age, gender, income, and ethnicity were found to be major demographic and socioeconomic features in the disparity of crash vulnerabilities (Boufous et al., 2011; Zhao et al., 2013; Lombardi et al., 2017; Barajas, 2018; Billah et al., 2022). The difference in income and ethnicity for both parties not only have a potential influence on the road user's driving, walking and cycling behavior, but also result in an environmental difference in roadway infrastructure of a traffic crash due to residential segregation of road users. For example, disadvantaged communities with more minority populations and populations of lower socioeconomic status were found to have less access to bike lanes across 22 large US cities (Braun et al., 2019). This disparity in crash risks among income and ethnic groups was one of the major concerns for scholars and practitioners who want to ensure the principle of environmental justice by mitigating the crash risk for low-income and minority groups through improving the roadway infrastructure for them (Kravetz and Noland, 2012; Rebentisch et al., 2019).

VRU crashes usually involve two parties: drivers and VRUs like pedestrians or bicyclists. Drivers are typically reported as the party at-fault in pedestrian/bicyclist-involved crashes, and pedestrians/bicyclists are the victims. Previous research has investigated both parties' demographic and behavioral factors in disaggregated analysis (Hasheminejad et al., 2018; Salon and Mclntyre, 2018; Balakrishnan et al., 2019;). Although these studies do include the characteristics of both drivers and VRUs in the analysis, they usually treated the characteristics of drivers and victims as unrelated independent variables in their theoretical assumptions and modeling process, which might overlook the potential interaction between two parties. The close-to-home effect in roadway crashes suggested that the drivers and VRUs involved in the same crash might live near each other and might share similar socioeconomic and demographic characteristics (Burdett et al., 2017; Ulak et al., 2019). Under this assumption, the characteristics of drivers and victims might be correlated, and understanding the occurrence of a crash should consider the similarity of drivers and victims. This raised research questions about the socioeconomic patterns of drivers and victims involved in one crash: To what extent the drivers and victims involved in one crash share similar demographic and economic features? Are there potential crash patterns that can be found based on their demographic and economic features? How are the different crash patterns distributed geographically? And what factors shape the distribution of these patterns of crashes?

The remainder of this paper is organized as follows. In the next section we provide a literature review. In section 3 we describe the data which is then followed by the methodological approach and modeling techniques. In section 5 we describe the study findings and provide discussions. The paper ends with the conclusions, references and Appendix.

2. LITERATURE REVIEW

2.1 Vulnerable Road Users in Traffic Crashes

Crashes and their consequences are not created equally for all road users. VRUs, such as pedestrians and bicyclists, are more likely to be injured than drivers since they are less protected. There are also “implicit” VRUs of certain demographic and economic groups who are usually found to have a higher chance of getting involved in a crash or receiving more severe consequences. For example, children and the elderly were considered more vulnerable than adult pedestrians and bicyclists (Braver, 2004; Ivan et al., 2019; Ding et al., 2020). Behavioral and environmental differences were two major reasons contributing to the vulnerability of implicit VRUs. Behavioral difference refers to the particular groups of VRUs who showed riskier behavior when driving, walking, or biking. For example, younger drivers were more likely to intentionally engage in risky driving behaviors such as mobile phone use (Scott-Parker and Oviedo-Trespalacios, 2017; Oviedo-Trespalacios and Scott-Parker, 2018; Eren and Gauld. 2022). Environmental difference refers to specific groups of VRUs who might live and travel in places with higher traffic exposure and more unsafe roadway infrastructure. For example, Rothman et al. (2020) compared the road infrastructure for low-income and high-income communities and found that fewer speed humps and lower road classification might result in higher rates of child pedestrian crashes in low-income communities in Toronto, Canada.

2.2 Vulnerability of Pedestrians and Bicyclists

Pedestrians/bicyclists are usually considered as VRUs in road safety literature, but certain groups of pedestrians/bicyclists are more vulnerable according to their age (Boufous et al., 2011; Koopmans et al.; 2015, Boele-Vos et al.; 2017, Das et al., 2019), gender (Zhao et al., 2013; Toran Pour et al., 2018; Algurén and Rizzi, 2022)., income (Siddiqui et al. 2012; Barajas, 2018), ethnicity (Kravetz and Noland; 2012, Steinbach et al. 2016, Barajas, 2018), among others. Nearly one-third of pedestrian crashes and two-thirds of bicyclist crashes involved school-aged children, according to police-reported crash data in 26 states in the US (Wheeler-Martin et al., 2020). Significant higher crash risks have also been found in bicyclists younger than 30 years and older than 65 years of age when controlling for exposure in Spain from 1993 to 2019 (Martínez-Ruiz et al., 2014). Though there was no solid evidence showing that male pedestrians or bicyclists have higher crash risks than their counterparts, a few studies found male pedestrians and bicyclists have less rule compliance and lower risk perception than females (Tom and Granié, 2011; Prati et al., 2019). The behavioral differences among age and gender groups play a major role in the disparity of roadway crashes, while the environmental differences better explained the disparity among income and ethnic groups. Research has found that low-income and minority groups were exposed to higher crash risk in pedestrian and bicyclist crashes in regions and cities of the United States (Kravetz and Noland, 2012; Barajas, 2018). Scholars have also linked the disparity between low-income and minority communities and high-income and majority communities with traffic exposure and quality of roadway infrastructure and provided a potential explanation for this disparity from the environmental difference (Fuller and Winters, 2017, Wang and Lindsey, 2017, Braun et al., 2019). For example, Ferenchak and Marshall (2021) investigated the installation of bicycling facilities across 29 US cities and found a lower rate of bicycling facility installation in the block groups with more people of color. Recent research in Oregon has found that lower median income and a higher proportion of the BIPOC population are associated with more pedestrian crashes at the census tract level considering factors from roadway infrastructure, land use, and socioeconomic background (Roll and McNeil, 2022). The disproportionate share of low-income and minority groups in traffic crashes has called for equity and environmental justice considerations in transportation planning and policy (Kravetz and Noland, 2012; Rebentisch et al., 2019). Besides, the population with lower educational attainment and limited English speaking has also been found to have higher crash risks at the aggregated level. (Barajas, 2018; Saha et al., 2018).

2.3 Vulnerability of Drivers

Drivers' socioeconomic features, attitudes toward driving, and driving behavior are primary contributing factors to the occurrence of roadway crashes (Adanu et al., 2017; Kemnitzer et al., 2019). Like pedestrians and bicyclists, certain groups of drivers are more vulnerable to roadway crashes, primarily due to differences in behavior and environmental factors. In the safety literature, these groups of drivers were divided mainly by their socioeconomic and demographic characteristics in the literature, like age (Lombardi et al., 2017; Gong and Fan, 2017; Liang and Yang, 2022) and gender (Russo et al., 2014; Pulido et al., 2016; Billah et al., 2022). Regev et al. (2018) found that crash risk is highest for drivers aged 21 to 29 in single-vehicle and multi-vehicle crashes from 2002 to 2012 in Great Britain when controlling an exposure measurement considering the driver's trip number and population size. Billah et al. (2022) found a more significant association between male drivers and the likelihood of crashes mainly due to riskier driving behaviors of male drivers compared to their counterparts, such as speeding, driving under the influence, and lane departure. Since drivers' income level and ethnicity were usually not publicly available in police-reported crash data, research represents the economic status of drivers using aggregated census data of drivers' residential ZIP code (Lee et al., 2021; Sagar et al., 2021). Though it was not without bias, this surrogate measurement provides a feasible way to investigate the driver's economic status in police-reported crashes. In the region where drivers' ethnic information was unavailable, some researchers have also developed alternative approaches to estimate the drivers' race and ethnicity. For example, Sartin et al. (2021) employed a Bayesian Improved Surname Geocoding (BISG) method to estimate the population-level ethnic information for drivers in New Jersey.

2.4 Linking Drivers and Victims in Crash Analysis

Demographic and economic characteristics of drivers and victims should be considered in the crash analysis since specific demographic and economic groups of drivers and victims are more vulnerable than others. Existing literature considered demographic and economic features from both parties (Salon and Mclntyre, 2018; Balakrishnan et al., 2019). For example, Behnood and Mannering (2017) incorporated both bicyclists' characteristics (gender, age, ethnicity, etc.) and drivers' characteristics (gender, age, ethnicity, etc.) in their crash severity model of bicyclist crashes and found bicyclists' and drivers' race and gender are the most important determinants of injury severity. However, these studies treated the characteristics of drivers and victims as unrelated variables independent in their quantitative analysis. This assumption might be problematic since potential spatial association might exist between drivers and victims, which might lead to the similarity of social characteristics between drivers and victims. A series of research investigating the proximity of crashes to the residential location of drivers/victims found a close-to-home effect in crashes in which most of the crashes happened near the residence of both drivers and victims (Burdett et al. 2017; Ulak et al., 2019). This close-to-home effect indicated that drivers and victims involved in a crash might share the same neighborhood and similar socioeconomic and demographic characteristics. Treating the socioeconomic characteristics of drivers and victims as uncorrelated variables might ignore the spatial similarity of both parties and may lead to potential bias in estimation. Thus, it is vital to consider the similarity of their characteristics in crash analysis.

Linking the characteristic of drivers and victims as driver-victim pairs and finding the hidden crash patterns within driver-victim pairs can reveal the similarity between the drivers and victims in the same crash. Clustering approaches have been usually employed to classify crashes by maximizing similarity and minimizing dissimilarity among clusters to find the potential patterns in roadway crashes. Latent class clustering analysis is one of the most popular approaches for revealing different crash patterns recently. Sun et al. (2019) employed a latent class clustering method to classify pedestrian crashes in Louisiana and found five clusters based on the factors from pedestrians' demographic features, crash-related factors, and environmental factors. Samerei et al. (2021) also used latent class clustering analysis to classify bicyclist crashes in Australia and found two clusters of crashes with different characteristics of bicyclists, road environment, traffic control, and crash circumstance.

3. DATA PREPARATION

Data used in this study includes counts of pedestrian and bicyclist crashes, crash specific information, socioeconomic characteristics of drivers and victims, roadway infrastructure characteristics, and traffic exposure. Descriptive information of the variables is shown in Table 1.

3.1 Crashes and Crash Specific Information

This research aimed to investigate the spatial distribution and contributing factors for driver-victim pairs in pedestrian/bicyclist crashes in Harries County, Texas, whose county seat is Houston. To collect the crashes and related information, we obtained four-year (2017-2020) records of pedestrian and bicyclist crashes from the Crash Records Information System (CRIS) of the Texas Department of Transportation (TxDOT). We identified pedestrian/bicyclist crashes based on the type of primary victim (pedestrian or bicyclist) involved in the crash. After removing redundant information and crash cases with missing critical information, we kept only one driver and one primary victim for each crash event. As a result, 2,822 pedestrian crashes and 1,123 bicyclist crashes were identified with both the driver's and victim's economic and demographic information available. There were 1,659 (58.8%) male and 1,163 (41.2%) female victims in pedestrian crashes, with an average age of 39.3. For bicyclist crashes, there were 924 (82.3%) male and 199 (17.7%) female victims with an average age of 37.5. Eight factors in crash specific information were retrieved from the CRIS database, including time of the day (*CR_TimeDay*), whether the crash happened on a workday (*CR_Workday*), season (*CR_Season*), weather condition (*CR_Weather*), surface condition (*CR_Surface*), whether crash happened on construction zone (*CR_Construct*), whether the crash occurred at the intersection (*CR_Intersec*), and years of the car was in use (*CR_CarUsedYr*).

3.2 Economic and Demographic Characteristics

The drivers' and victims' economic and demographic characteristics were usually missing in a publicly accessible crash database for privacy and liability concerns. From the CRIS database, we retrieved the driver's ethnicity (*DR_Ethincity*), age (*DR_Age*), and gender (*DR_Gender*), and victim's ethnicity (*VT_Ethincity*), age (*VT_Age*), and gender (*VT_Gender*). However, the CRIS database did not include the income information of drivers and victims. Thus, we estimated the driver's and victim's income information by the income level of their residential census tract based on median household income in 2019 American Community Survey (ACS) 5-year estimates. To obtain the driver's census tract, we matched the ZIP code of drivers with their census tract in ArcGIS Pro. The corresponding census tract of the driver was where the centroid of the ZIP code is situated. The victim's residential census tract was hypothesized to be the same as the census tract where the crash happened. There were 786 census tracts in Harris County with an average area of 5.9 km^2, which is within the range of acceptable walking and biking distance (1,750-2,122 meters) (Rahul and Verma, 2014). Besides, most pedestrian and bicyclist crashes happened near the victim's home (Steinbach et al., 2013; Ulak et al., 2019). Therefore, we assumed the crash location was the same as the victim's residential census tract. Finally, we recoded the driver's income (*DR_Income*) and victim's income (*VT_Income*) to ordinal variables in five levels: low income, lower to medium income, medium income, medium to high income, and high income, according to the five quintiles of their residential census tracts in the research area. In our dataset, there are 73 driver-victim pairs in which the driver and victim live in the same census tract, which accounts for the 3.14% of the total number of pedestrian crashes; for bicyclist crashes, there are 50 driver-victim pairs in which the driver and victim live in the same census tract, which accounts for 5.18% of the total bicyclist crashes. This small proportion of driver-victim pairs shows that assigning the driver-victim pair to the same income level (despite the potential differences) does not introduce significant error to overall model performance.

3.3 Roadway Infrastructure Characteristics

Characteristics of roadway infrastructure were collected from the Roadway Inventory of TxDOT, which was a GIS-based road network database storing roadway information in Texas. Data for the roadway inventory was updated annually, and we used the 2020 version, which conformed with the time span of our crash events. We selected 11 characteristics of the roadway infrastructure where the crash happened, including road functional classification (*RD_FuncCls*), speed limit (*RD_SpdLmt*), whether the crash occurred in an urban area (*RD_Urban*), roadbed width which comprises shoulder width and surface width) (*RD_RdWth*), the number of lanes (*RD_LnNum*), lane with (*RD_LnWth*), median width (*RD_MedWth*), inside shoulder width (*RD_SWthIn*), outside shoulder width (*RD_SWthOut*), existence of left curb (*RD_CurbL*), and existence of right curb (*RD_CurbR*).

3.4 Exposure

Vehicular, pedestrian, and bicyclist exposure variables were also taken into consideration in this study. Vehicular exposure of the road segments where the crash happened was measured as the Annual Average Daily Traffic (AADT) from the TxDOT Roadway Inventory database for each year of the crash events. However, the scarcity of pedestrian and bicyclist exposure data was one of the primary limitations in crash modeling. Scholars have used emerging crowdsource data to estimate pedestrian and bicyclist exposure, such as bicycle count data from Strava (Dadashova and Griffin, 2020; Dadashova et al., 2020). In this study, we used a scaling approach to estimate the bicyclist and pedestrian exposure leveraging observed pedestrian and bicyclist count data available from Texas Bicycle and Pedestrian Data Exchange (BP|CX) (https://mobility.tamu.edu/bikepeddata/) and crowed-sourced pedestrian and bicyclist count data from Strava. The estimated pedestrian and bicyclist counts were averaged daily and calculated annually using the scaling approach for Strava data by Dadashova et al. (2020). Using this approach, we calculated the scaling factors (See Table S1) for each year by dividing the observed pedestrian/bicyclist counts with the crowed-sourced pedestrian/bicyclist counts and calculated the estimated pedestrian/bicyclist counts by multiplying the scaling factors with the crowed-sourced pedestrian/bicyclist counts on other road segments.

Table 1. Descriptive Statistics of Variables

Categorical Variables		Pedestrian crashes	Bicyclist crashes
CR_TimeDay	1 = 0:00-6:00	254(9.0%)	51(4.5%)
	2 = 6:00-12:00	749(26.5%)	303(27.0%)
	3 = 12:00-18:00	839(29.7%)	449(40.0%)
	4 = 18:00-24:00	980(34.7%)	320(28.5%)
CR_Workday	1 = Monday to Friday	2199(77.9%)	856(76.2%)
	2 = Saturday to Sunday	623(22.1%)	267(23.8%)
CR_Season	1 = Spring	746(26.4%)	243(21.6%)
	2 = Summer	679(24.1%)	300(26.7%)
	3 = Autumn	606(21.5%)	293(26.1%)
	4 = Winter	791(28.0%)	287(25.6%)
CR_Weather	1 = Clear	2093(74.2%)	873(77.7%)
	2 = Others	729(25.8%)	250(22.3%)
CR_Surface	1 = Dry	2495(88.4%)	1039(92.5%)
	2 = Others	327(11.6%)	84(7.5%)
CR_Construct	1 = At construction zone	51(1.8%)	5(0.4%)
	2 = Not at construction zone	2771(98.2%)	1118(99.6%)
CR_Intersec	1 = At intersection	1034(36.6%)	661(58.9%)
	2 = Not at intersection	1788(63.4%)	462(41.1%)
DR_Income	1 = low income	454(16.1%)	193(17.2%)
	2 = low to medium income	614(21.8%)	229(20.4%)
	3= medium income	814(28.8%)	280(24.9%)
	4 =medium to high income	368(13.0%)	161(14.3%)
	5 = high income	572(20.3%)	260(23.2%)
DR_Ethnicity	1 = White	888(31.5%)	384(34.2%)
	2 = Hispanic	896(31.8%)	350(31.2%)
	3 = Black	800(28.3%)	297(26.4%)
	4 = Asian	185(6.6%)	70(6.2%)
	5 = Others	47(1.7%)	20(1.8%)
DR_Gender	1 = Male	1632(57.8%)	626(55.7%)
	2 = Female	1190(42.2%)	497(44.3%)
VT_Income	1 = low income	600(21.3%)	211(18.8%)
	2 = low to medium income	762(27.0%)	286(25.5%)
	3= medium income	559(19.8%)	210(18.7%)
	4 =medium to high income	389(13.8%)	156(13.9%)
	5 = high income	512(18.1%)	260(23.2%)
VT_Ethinicity	1 = White	935(33.1%)	485(43.2%)
	2 = Hispanic	852(30.2%)	274(24.4%)

Categorical Variables		Pedestrian crashes	Bicyclist crashes
	3 = Black	843(29.9%)	299(26.6%)
	4 = Asian	131(4.6%)	52(4.6%)
	5 = Others	54(1.9%)	11(1.0%)
VT_Gender	1 = Male	1659(58.8%)	924(82.3%)
	2 = Female	1163(41.2%)	199(17.7%)
RD_FuncCls	1 = Collectors	782(27.7%)	268(23.9%)
	2 = Local roads	2044(72.4%)	855(76.1%)
RD_Urban	1 = Urban area	2818(99.9%)	1117(99.5%)
	2 = Rural area	4(0.1%)	6(0.5%)
RD_CurbL	1 = Left curb exists	2689(95.3%)	1092(97.2%)
	2 = No left curb	133(4.7%)	31(2.8%)
RD_CurbR	1 = Right curb exists	2690(95.3%)	1093(97.3%)
	2 = No right curb	132(4.7%)	30(2.7%)
RD_LnWth (feet)	1 = Less than 10	1527(54.1%)	187(16.7%)
	2 = 10 to 12	884(31.3%)	534(47.6%)
	3 = 12 to 14	139(4.9%)	313(27.0%)
	4 = Greater than 14	272(9.6%)	89(7.9%)

Continuous Variables	Mean	Min	Max	SD	Mean	Min	Max	SD
CR_CarUsedYr	8.3	0.0	43.0	5.8	8.2	0	43	6.1
DR_Age	41.6	15.0	118.0	16.6	43.4	8	118	17.3
VT_Age	39.3	1.0	100.0	19.5	37.5	3	100	19
RD_SpdLmt (miles per hour)	36.7	20.0	65.0	11.2	37.4	20	65	12
RD_RdWth (feet)	33.3	14.0	106.0	14.1	30.7	16	106	12.8
RD_LnNum	2.9	1.0	6.0	1.1	2.7	2	6	1
RD_LnWth	11.4	5.0	27.0	2.8	11	5	27	2.3
RD_MedWth (feet)	0.3	0.0	138.0	3.6	1.1	0	138	9.4
RD_SWthIn (feet)	0.1	0.0	10.0	0.6	0.1	0	10	0.6
RD_SWthOut (feet)	0.1	0.0	10.0	0.8	0.1	0	10	1
EX_Ped	36.3	0.3	340.7	104.0	Not Applicable			
EX_Cyc	Not Applicable				14.1	0.1	340.7	30.8
AADT	9864.7	50.0	49968.0	9954.5	8601	69	49968	9565.4

4. METHODOLOGY

In this study, we applied a latent class clustering analysis (LCA) to identify the patterns in driver-victim pairs according to the driver's and victim's income and ethnicity in pedestrian and bicyclist crashes. We also mapped the crash patterns in the study area to reveal their spatial distribution. Then, we used random forest algorithm to investigate the relative contribution of factors to the crash patterns from crash specific information, economic and demographic characteristics of drivers and victims, roadway infrastructure, and exposure. Finally, we drew partial dependence plots (PDPs) for the most important factors to interpret their influences on certain crash patterns.

4.1 Latent Class Clustering Analysis

To investigate the possible patterns in driver-victim pairs, we applied LCA to divide the pedestrian and bicyclist crashes according to the victim's and driver's socioeconomic characteristics. Clustering analysis is an unsupervised machine learning method that can separate the crashes into homogenous subgroups, which have the largest similarities within each subgroup and the largest dissimilarity between each subgroup (Sivasankaran and Balasubramanian, 2020). We used a probability-based clustering approach (i.e., Latent Class Clustering), which has recently been applied in several roadway safety research studies (Sun et al., 2019, Samerei et al., 2021). The Latent Class Clustering approach has several advantages over other clustering approaches (e.g., K-means) in that it 1) can calculate the probability of a crash of being in a certain cluster by maximum likelihood method; 2) does not necessarily need to standardize the variables beforehand; 3) does not need to specify the number of clusters before performing the clustering; 4) can generate statistical criteria afterward to select the best model with a certain number of clusters (Sasidharan et al., 2015, Sun et al., 2019). The mathematical formula of the LCA approach is shown below (Samerei et al., 2021):

$$P(\boldsymbol{Y}_i = y) = \sum_{k=1}^{K_c} \rho \prod_{m=1}^{M} \prod_{n=1}^{r_m} \theta_{mn|l}^{l(y_m=n)}$$

Where $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}, \dots \boldsymbol{Y}_{iM})$ is the observation (crash) i's responses in M category, where the possible values of $\boldsymbol{Y}_{iM}$ are $1, \dots, r_m$; r_m represents the crash i's rth attribute in m category; K_c represents the number of latent classes to be estimated; $l(y_m = n)$ is the indicator function to be 1 if y equals n and to be 0 when y is not 1; ρ is the probability of latent class membership probability and θ is the conditional probabilities of responses on latent class membership. The number of clusters can influence the goodness-of-fit of the latent class clustering model. We employ Bayesian Information Criteria (BIC) to select the appropriate number of clusters. LCA modeling and BIC calculation were conducted by package *polPCA* in R.

4.2 Random Forest Algorithm and Partial Dependence Plot

Random forest algorithm is known as a tree-based ensemble machine learning technique. It is built upon a multitude of weak decision tree models to form a strong "forest" by averaging the predictions from all the individual regression trees or by taking the majority vote from the classification tree. It can be applied in both classification and regression, and we use the random forest algorithm for classification in this task. The random forest algorithm employs a bagging technique to repeatedly select a random sample from the training dataset and use the sample to fit a decision tree. Let feature set X be $\{x_1, x_2 \dots, x_n\}$, target set Y be $\{y_1, y_2 \dots, y_n\}$, and $i = 1,2, \dots I$, the process of random forest can be represented as:

1) Select a random sample set from $\{X,Y\}$, which is denoted as $\{x_i,y_i\}$;
2) Train a decision tree f_i on the sample set $\{x_i,y_i\}$;
3) Repeat procedures 1 and 2 for I times to get I decision trees $\{f_1, f_2 \dots, f_I\}$;
4) Aggregating the prediction results for any random sample $\hat{x}$ to get function $\hat{f}$ for the random forest. For classification, it takes the majority vote of the target from all individual decision trees, denoted as $\hat{f}(\hat{x}) = \max_{i=1,2,\dots,I} f_i(x_i)$.

Several parameters can affect the performance of the model: for example, the number of decision trees (I). To optimize performance, we employed a random search method for optimal parameters with successive halving to automatically find the best combination of parameters (Scikit-Learn, 2022). To investigate the impact of variables in clusters of driver-victim pairs, we calculated the feature importance for each variable to assess the relative contribution of all the variables (Masís, 2021). Furthermore, we used the PDPs, which is one of the model-agnostic interpretable machine learning approaches to reveal the marginal effect of a feature in machine learning models (Masís, 2021). Random forest algorithm was implemented by *Scikit-learn*, and PDPs are generated by *pdpbox* in Python.

5. RESULTS

5.1 Results of Latent Class Clustering Analysis

Identifying the Number of Clusters

The LCA models were performed on the economic and demographic characteristics of the driver-victim pairs to find the patterns within the driver-victim pairs regarding their income level and ethnicity. Figure 1 shows the graphs of BIC value with a different number of clusters for pedestrian crashes and bicyclist crashes. As indicated by the graph, the BIC values in the LCA of pedestrian crashes increase along with the increasing number of clusters, and the minimum BIC value is generated when the number of clusters is set as two. The trend of BIC values in the LCA of bicyclist crashes is similar to pedestrian crashes, achieving its lowest value when there are two clusters. Thus, we report the results of the LCA for pedestrian crashes and bicyclist crashes when there are two clusters in each model.

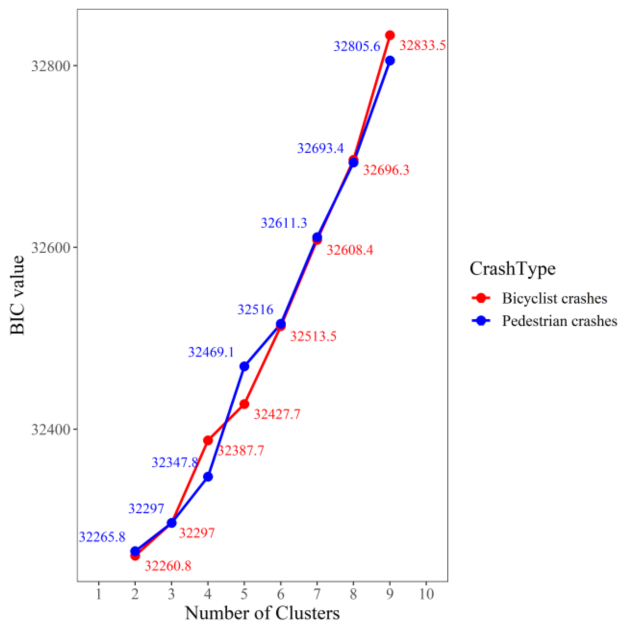
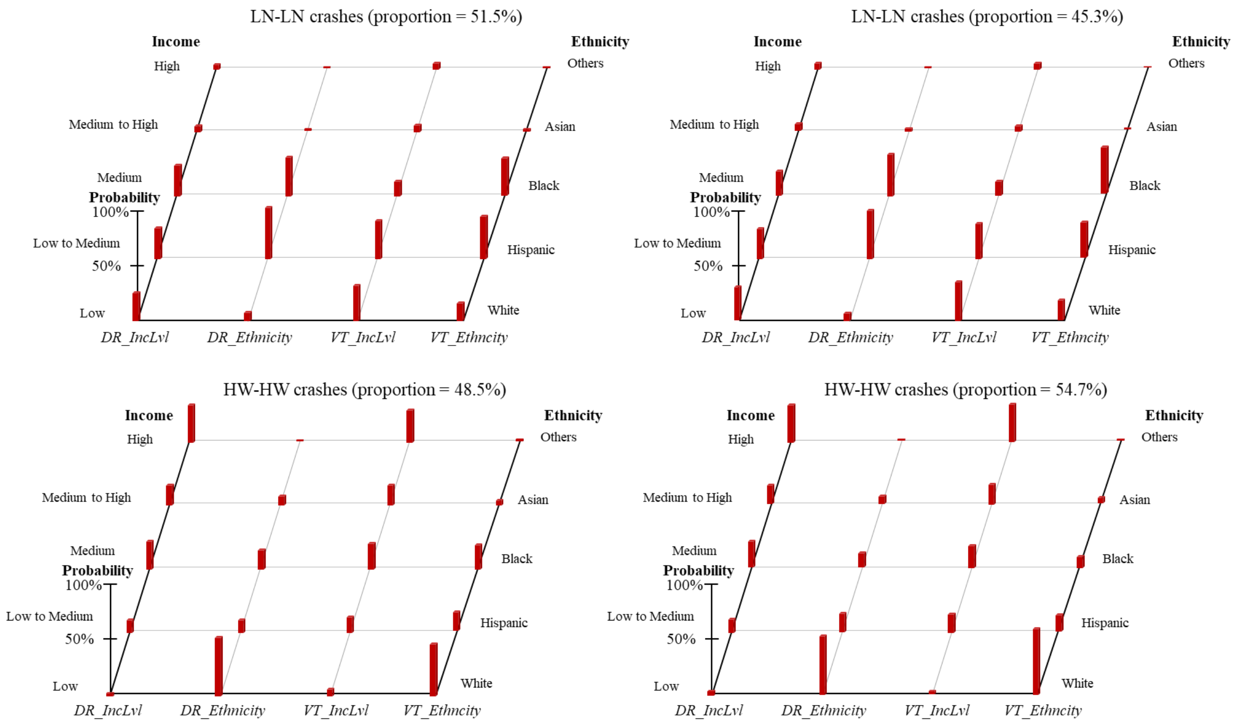


Figure 1. Number of Clusters and Their Respective BIC Value in Pedestrian and Bicyclist Crashes

Clustering Crashes by LCA

Figure 2 shows the distribution of the driver's and victim's income level and ethnicity in each cluster in both models. Detailed information about the clustering results can be found in Table S2. For pedestrian/bicyclist crashes, the driver-victim pairs are clustered into crashes involving "lower income non-white driver and lower income non-white victim" (LN-LN crashes) and crashes involving "higher income white driver and higher income white victim" (HW-HW crashes), respectively. In the pedestrian crash model, two clusters are almost evenly divided (51.5% for LN-LN crashes and 48.5% for HW-HW crashes). Figure 2.a shows the two clusters and the corresponding distribution of drivers' and victims' income levels and ethnicity in the pedestrian crash model. For driver's characteristics, white drivers make 9.2% of LN-LN crashes, while its probability is 55% in HW-HW crashes. The income level of drivers in LN-LN crashes concentrates in the low income to medium income categories. In contrast, the income level of drivers in HW-HW crashes is distributed in medium income to high income categories. Victims in LN-LN crashes have a higher probability of being non-whites (81.9%), while victims in HW-HW crashes have the highest probability of being white (49.1%). Victims' income level is also distributed on low income to medium income in LN-LN crashes and medium income to high income in HW-HW crashes. Clustering results in bicyclist crashes appear to have similar patterns of economic and demographic characteristics for drivers and victims with pedestrian crashes. The bicyclist LN-LN crashes have a higher probability of involving non-white drivers (90.6%), drivers from lower income levels and non-white victims (79.7%), and victims from lower income levels. In comparison, bicyclist HW-HW crashes have a higher chance of involving white drivers (54.7%), drivers from higher income levels, white victims (62.1%), and victims from higher income levels. Notable socioeconomic patterns of driver-victim pairs have been revealed in these results, which show the social segregation of pedestrian and bicyclist crashes. This social segregation of crashes demonstrates that the driver and victim involved in a crash are likely to be similar regarding their income and ethnicity. Non-white and low-income drivers and non-white and low-income victims are more likely to be involved in one crash, while white and high-income victims are more likely to get into crashes by white and high-income drivers.

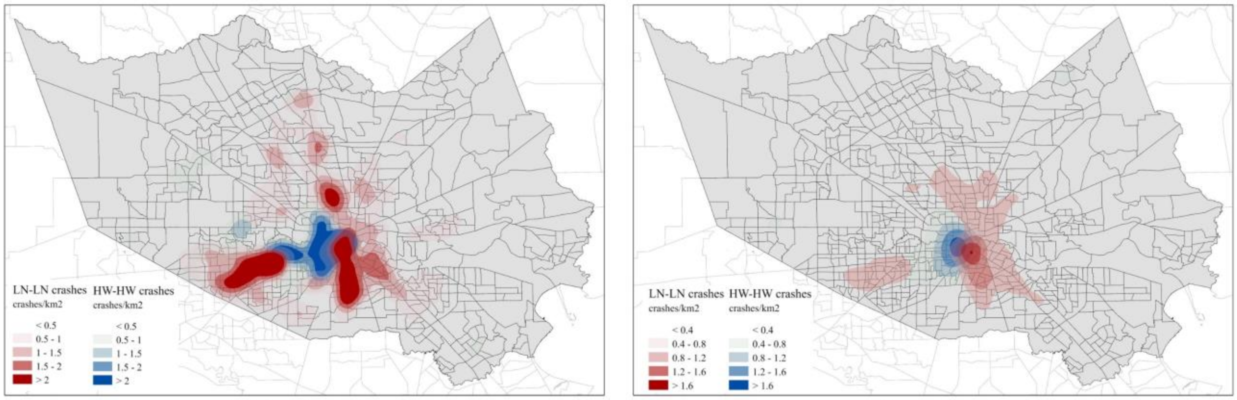


a) LCA results for pedestrian crashes

a) LCA results for bicyclist crashes

Figure 2. Clustering Results for Pedestrian Crashes and Bicyclist Crashes

As discussed before, there are potential spatial patterns due to the spatial proximity of crashes, so we plot the density maps of pedestrian and bicyclist crashes based on crash location to observe the spatial distribution of LN-LN and HW-HW crashes (See Figure 3). Figure 3.a and Figure 3.b show the density of pedestrian LN-LN crashes and HW-HW crashes of pedestrian crashes and bicyclist crashes, respectively. The concentrated area of both LN-LN and HW-HW crashes overlay in the downtown area, which suggests that downtown is the nucleus for crashes of all kinds. Except for downtown Houston, two types of crashes show spatial segregation in which the LN-LN crashes concentrate on three major areas, including southern, northern, and further southwest areas near downtown. In contrast, HW-HW crashes happened more in a closer west region near downtown. Bicyclist crashes show a more segregated pattern for which LN-LN crashes are denser in the western area near downtown, and HW-HW crashes happened more in the eastern, southern, northern, and further southwest areas near downtown. Compared with bicyclist crashes, pedestrian crashes have a denser distribution within the research area.



a) Pedestrian crashes b) Bicyclist crashes

Figure 3. Density Map of LN-LN and HW-HW Crashes for Pedestrian and Bicyclist Crashes

Distance between the crash and residence of both parties is important in understanding LN-LN crashes and HW-HW crashes due to the difference in travel behaviors and activity space between drivers and victims from different demographic and socioeconomic backgrounds. We map the trajectory of driver-victim pairs based on the crash location and the centroid of the driver's ZIP code to investigate the spatial characteristics of driver-victim pairs for each cluster (See Figure S1). Downtown Houston is the nucleus of all four types of crashes according to the distribution of driver-victim pairs. Compared with LN-LN crashes, HW-HW crashes trajectory are sparser for both pedestrian and bicyclist crashes. Figure 4 plots the probability density for the distance of driver-victim pairs, showing that LN-LN crashes have a more positively skewed distribution than their counterparts. The geographical and probability distribution of driver-victims pairs indicate that LN-LN crashes are more likely to involve a crash location when a driver lives nearby, while drivers in HW-HW crashes might live farther away from the crash location. This might be because drivers from higher-income and majority-white communities have more resources and capability to travel further away, while drivers in lower-income and minority communities are limited in their activity space hence getting into a crash within their community.

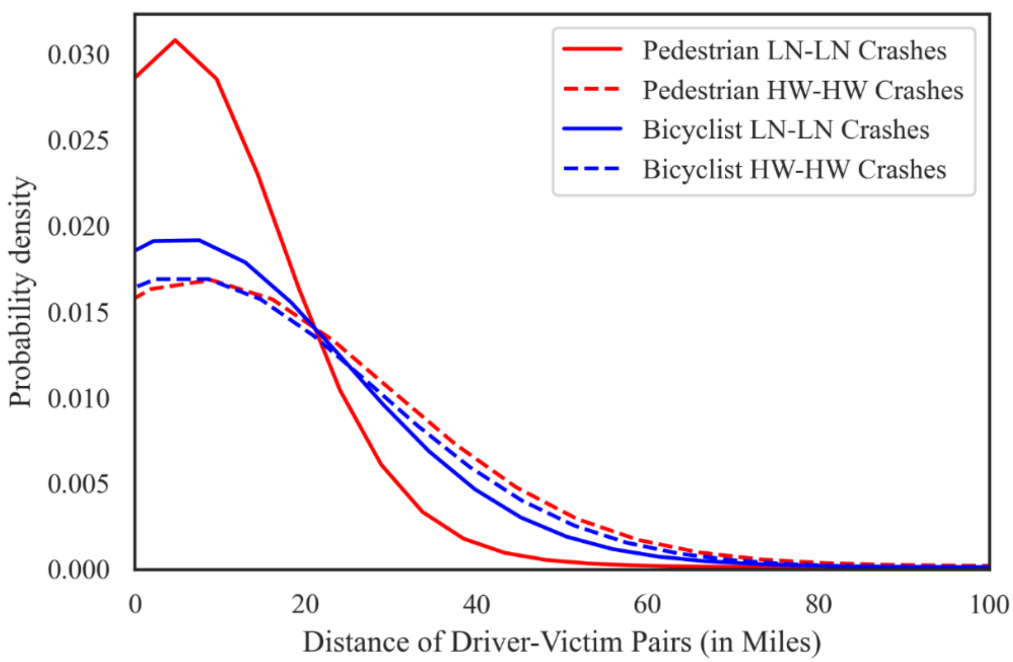


Figure 4. Probability Density Plot for Distance of Driver-Victim Pairs

5.2 Random Forest Results

Relative importance of selected variables

Table 3 shows the variables' feature importance of the random forest algorithm to classify whether a crash belongs to LN-LN crashes or HW-HW crashes for pedestrian and bicyclist crashes, respectively.

Table 3. Feature Importance of Random Forest Model for Pedestrian and Bicyclist Crashes

Pedestrian Crash model			Bicyclist Crash model		
Variables	**Feature Importance**	**Rank**	**Variables**	**Feature Importance**	**Rank**
EX_Ped	0.260	1	*EX_Cyc*	0.227	1
DR_Age	0.132	2	*VT_Age*	0.114	2
VT_Age	0.117	3	*DR_Age*	0.103	3
CR_CarUsedYr	0.100	4	*EX_AADT*	0.091	4
EX_AADT	0.098	5	*CR_CarUsedYr*	0.089	5
RD_SpdLmt	0.049	6	*RD_SpdLmt*	0.066	6
RD_RdWth	0.045	7	*RD_RdWth*	0.056	7
CR_Season	0.033	8	*CR_TimeDay*	0.037	8
CR_TimeDay	0.031	9	*CR_Season*	0.034	9
RD_LnWth	0.024	10	*RD_LnWth*	0.031	10
VT_Gender	0.014	11	*VT_Gender*	0.025	11
CR_Intersec	0.013	12	*RD_LnNum*	0.019	12
RD_LnNum	0.013	13	*CR_Workday*	0.015	13
DR_Gender	0.012	14	*CR_Weather*	0.015	14
CR_Surface	0.012	15	*CR_Intersec*	0.015	15
CR_Workday	0.011	16	*DR_Gender*	0.014	16
CR_Weather	0.010	17	*RD_FuncCls*	0.014	17
RD_FuncCls	0.010	18	*CR_Surface*	0.011	18
CR_Construt	0.005	19	*RD_MedWth*	0.006	19
RD_CurbR	0.004	20	*RD_SWthIn*	0.005	20
RD_CurbL	0.004	21	*RD_SWthOut*	0.005	21
RD_SWthIn	0.002	22	*RD_CurbL*	0.004	22
RD_SWthOut	0.001	23	*RD_CurbR*	0.003	23
RD_MedWth	0.001	24	*RD_Urban*	<0.001	24
RD_Urban	<0.001	25	*CR_Construt*	<0.001	25

Since we only got two clusters in each model, the LN-LN crashes are taken as the reference group. Thus, the higher value of feature importance a variable has, the larger contribution the variables will make in determining whether a crash belongs to the LN-LN crash. The ranks of feature importance imply the relative contribution of a feature in the random forest model. Exposures are the most relevant factors in determining crash clusters. The estimated pedestrian exposure and estimated pedestrian exposure rank first in their respective model. AADT ranks fifth in pedestrian crashes and ranks fourth in bicyclist crashes. The high rank of exposure variables indicates a strong association between the traffic volume of both vehicles and pedestrians/bicyclists with patterns of driver-victim pair. The driver's age and victim's age are also among the most important variables, while their gender is less influential. Driver's age and victim's age rank second and third in pedestrian crashes, and driver's age and victim's age rank third and second in bicyclist crashes. For crash specific information, the year of the car in use ranks fourth in pedestrian crashes, and fifth in bicyclist crashes, indicating the vehicles involved in LN-LN and HW-HW crashes might have different used years. Time of the day and season ranks eighth in both models, indicating a relatively sizeable temporal variation of the crash pattern. For road infrastructure characteristics, speed limit and roadbed width rank sixth and seventh in both pedestrian and bicyclist crashes, showing the considerable influence of roadway infrastructure characteristics in determining the crash clusters. However, their feature importance is relatively low compared to previous factors.

5.3 Partial Dependence Plots

To investigate variables' impact on driver-victim pairs' patterns, we draw the PDPs for the top eight variables in feature importance for both pedestrian and bicyclist models (Figure 5). For exposure variables, when annual daily pedestrian volumes are less than 2.6, pedestrian exposure is not influential. When it is larger than 2.6, it becomes positively associated with the probability of a crash being a LN-LN crash. This indicates that LN-LN crashes will likely happen on the road with larger pedestrian exposure. HW-HW crashes will be less likely to occur on the road with larger pedestrian exposure. In bicyclist crashes, the positive marginal effect of bicyclist exposure on the probability of a crash being a LN-LN crash will increase when the bicyclist exposure becomes larger. This indicates that LN-LN crashes will be more likely to happen on the road with larger bicyclist exposure, and the larger the bicyclist exposure, the higher the probability of LN-LN crashes. One of the potential explanations for this could be missing active transportation-friendly infrastructure in low income and minority communities, which may force the bicyclists to share the road with oncoming traffic, increasing their crash probability. However, this speculation needs further explored and proven by accounting for the bicyclist infrastructure in data analysis. For vehicular exposure, the pedestrian and bicyclist crashes have similar patterns, which shows lower AADT does not have significant influence on the probability of a crash being LN-LN crash. While within the highest quantile of the AADT, it will have a larger positive association for both pedestrian and bicyclist crashes. This means both pedestrian and bicyclist LN-LN crashes tend to occur on the road with a larger vehicular volume.

The driver's and victim's age are among the most influential factors in socioeconomic characteristics for both crash types. In the pedestrian crash model, when the driver's age is less than 64, the probability of a crash being a LN-LN crash will decrease. When the driver's age is larger than 64, the probability of a crash being a LN-LN crash will increase. This means younger drivers are less likely to be involved in a pedestrian LN-LN crash, while older drivers are more likely to be involved in a pedestrian LN-LN crash. The PDP shows that as the victim's age increases, the marginal effect of the probability of being in a LN-LN crash will rise, indicating that older victims are more likely to be involved in a pedestrian LN-LN crash. In the bicyclist crash model, the driver's age does not have much influence on the probability of a LN-LN crash in its lower quintiles. It only has a positive marginal effect when the driver's age exceeds 66 years old, indicating that older drivers will be more likely to be involved in a bicyclist LN-LN crash. For the victim's age, a victim aged 32 or younger will increase the probability of a crash being a LN-LN crash, while a victim aged 33 or higher will decrease the probability of a crash being a LN-LN crash. This means bicyclist LN-LN crashes are more likely to involve older drivers and younger bicyclists.

For crash specific information, the year of the car in use, time of the day, and season rank among the most influential variables. When the year of the car in use is less than six, it has negligible influence on the probability of a pedestrian LN-LN crash for both crash types. As the year of the car in use increases in pedestrian crashes, its marginal effect will become larger in a negative direction for both crash types. This indicates older cars are less likely to be involved in a LN-LN crash and more likely to be involved in an

HW-HW crash for both pedestrian and bicyclist crashes. The influence of time of the day on pedestrian LN-LN crashes is positive in summer and autumn and negative in winter, but the effect of the influence is minimal. For bicyclist crashes, from 6:00 am to 12:00 pm and from 12:00 pm to 6:00 pm, there will be a higher chance of bicyclist crashes. Lower-income and non-white groups might choose biking as their mean of transportation to commute during the daytime more frequently than their counterparts due to economic affordability or behavioral difference, which forms a higher bicyclist crash probability.

For road infrastructure characteristics, the road speed limit has the same patterns in its influence on pedestrian and bicyclist crashes. When the road speed limit is less than 35 miles per hour, its impact on the crash clusters is negligible. When the road speed limit exceeds 45 miles per hour, the probability of a crash being a LN-LN crash will increase in both pedestrian and bicyclist crashes. This indicates that LN-LN crashes for pedestrians and bicyclists are more likely to happen on the road with a higher speed limit. Roadbed width has little effect when less than 40 feet and only has a positive marginal effect on the highest quantile, indicating that LN-LN pedestrian crashes are more likely to happen on wider roads. In the bicyclist crash model, the effect of roadbed width is not influential when it is less than 24 feet but becomes negative when it is larger than 24 feet, suggesting that LN-LN bicyclist crashes are less likely to happen on wider roads.

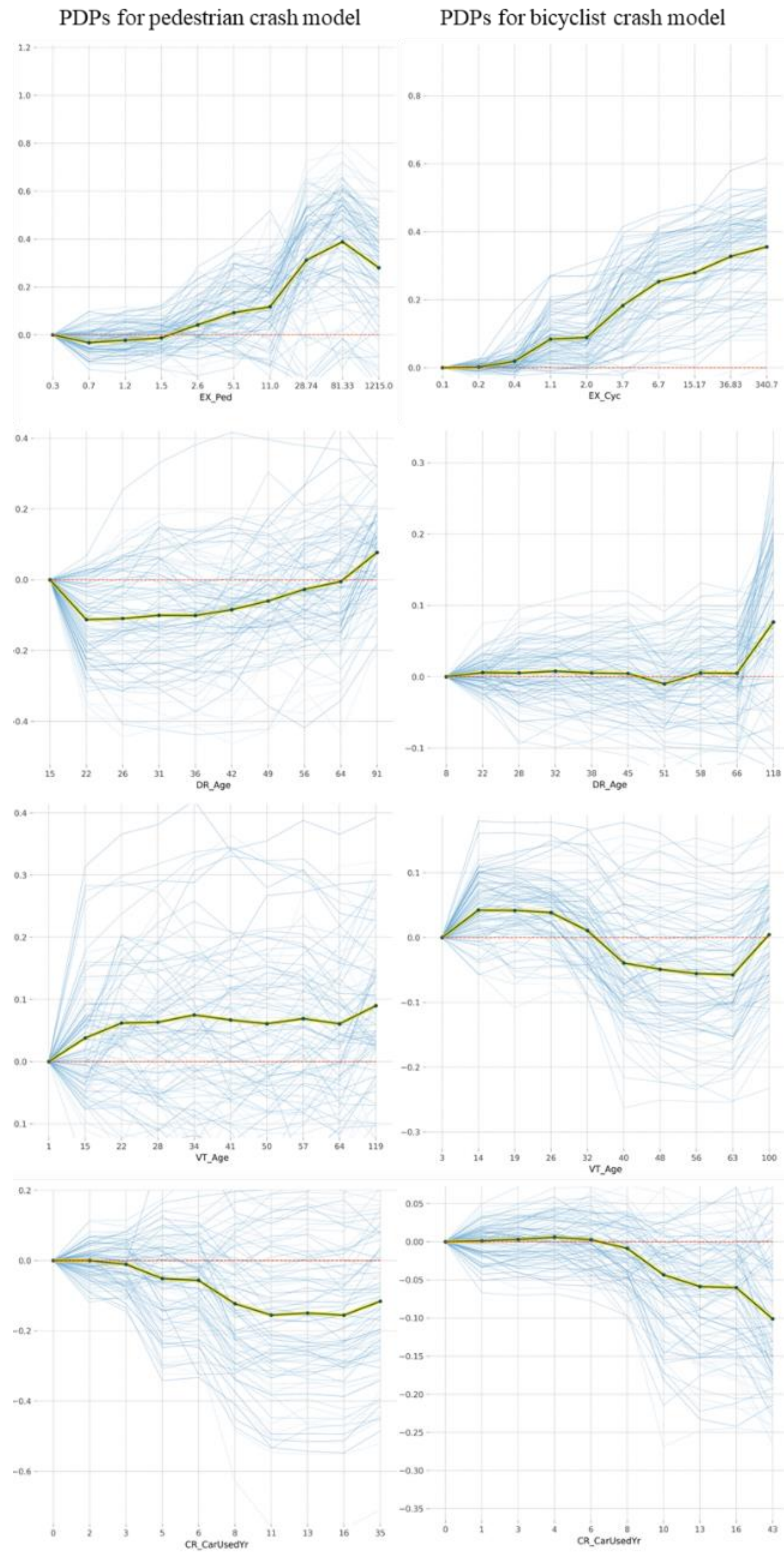
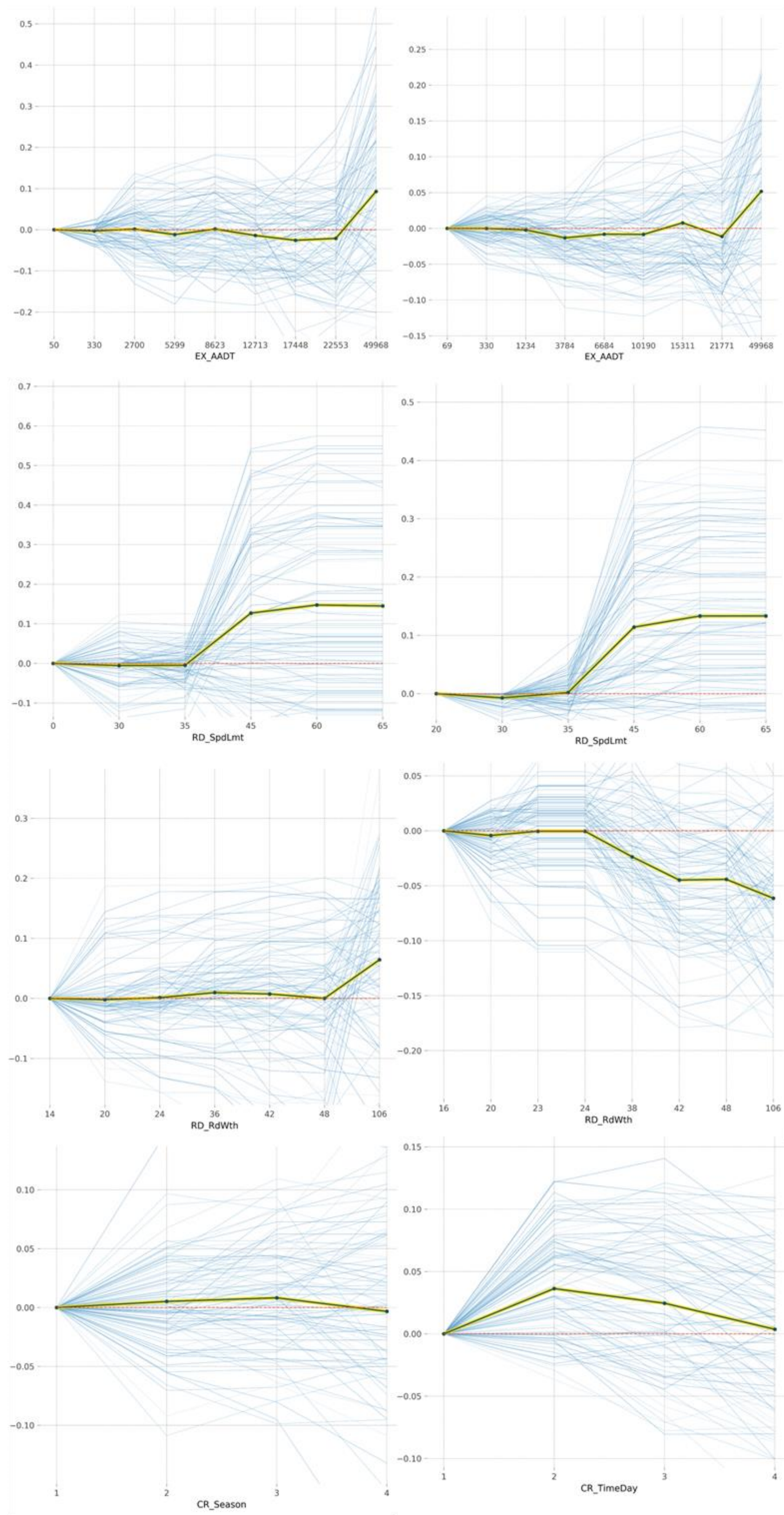


Figure 5. PDPs for Variables in Pedestrian and Bicyclist Crash Model

(Figure 5 continued)

PDPs for pedestrian crash model PDPs for bicyclist crash model



6. CONCLUSION

In this study, we used driver-victim pairs to reveal the crash patterns based on clustering drivers' and victims' ethnicity and income level. Using crash data from Harris County, we applied a probability-based latent class clustering analysis to classify pedestrian and bicyclist crashes. The clustering results showed that lower income and non-white drivers tend to be involved in crashes with lower income and non-white victims (LN-LN crashes). While higher income and white drivers tend to be involved in crashes with higher income and white victims (HW-HW crashes). This result showed a certain degree of social segregation in pedestrian and bicyclist crashes, indicating that drivers and victims of similar socioeconomic characteristics are more likely to be involved in the same crash, while those from different socioeconomic backgrounds are not. We further analyzed the trajectories of driver-victim pairs and found all crash types tend to concentrate in downtown Houston. The trajectories of HW-HW crashes are sparser in their geographic distribution, which suggests higher income and white drivers are driving a long distance and getting involved in a crash in farther geographic areas than their counterparts.

To explore how the LN-LN and HW-HW crash patterns were shaped, we employed a random forest algorithm and partial dependence plots to model and interpret the clustering outcomes from LCA models. Contributing factors for the crash patterns were selected from crash specific information, drivers' and victims' age and gender, roadway infrastructure, and traffic exposure. Pedestrian/bicyclist exposure, driver's age, victim's age, year of the car in use, AADT, speed limit, roadbed width, time of the day, and season are the most influential variables in pedestrian and bicyclist models. We drew partial dependence plots for the most influential variables to interpret how the variables are associated with crash patterns. The results showed that LN-LN crashes tend to happen on the road with larger traffic exposure of pedestrians/bicyclists and vehicle, which is contradictory to safety in number theory, indicating that the European model of bicycling/walking is not always implementable for underserved communities in the US (Elvik and Bjørnskau, 2017). Older drivers and older pedestrians are more likely to be in the same LN-LN crash, while older drivers and younger bicyclists are more likely to be in the same LN-LN crash. Longer years of the car in use will increase the probability of HW-HW crashes. Higher speed limits and wider roads are associated with a higher probability of LN-LN crashes for both pedestrian and bicyclist crashes. The results indicated the coexistence of LN-LN crashes and road conditions of higher traffic exposure, higher speed limit, and wider roads. The communities where low-income and ethnic minorities are concentrated might have higher traffic exposure and less safe road environments, which shapes the distribution of LN-LN crashes.

This study contributes to the existing body of literature in several ways. First, from a planning and engineering perspective, this study confirms long-believed hypotheses that there is a clear sociodemographic and economic segregation of crashes. We also find that the crash-contributing factors are not usually the same across different communities. These results can help safety practitioners in both engineering and planning fields to develop and implement practices that will target the main concerns of each community instead of developing one size fits all strategies. Safe systems approach can be one of the potential strategies to accomplish this goal. Another significant contribution of this study concerns the methodological approach. We innovatively use machine learning techniques to address a largely unexplored research question where the driver's and victim's characteristics are analyzed simultaneously.

Despite these contributions, the study does have limitations. In this study, we used the police-reported crash data, which have been considered to underestimate the actual number of crashes. Besides, the police-reported crash data also lacks other economic and demographic information for drivers and victims, such as educational level and occupation. The detailed income level is also not reported by the police agents. On the other hand, collecting individual level income data is not be feasible in an observational study and may require additional data collection efforts by implementing experimental design studies. The success of such experimental design study however is not guaranteed given that many drivers may be reluctant to share their crash history due to potential liabilities. We therefore use the surrogate measurement of income level based on drivers' and victims' residential census tract. This approach may be biased, but it is an acceptable alternative in the absence of readily available data on income measurement. Another limitation of the study is related to the exposure data. Although we account for bicycle and pedestrian exposure by developing scaling factors, the measurement of exposure can be improved by implementing more rigorous models.

In this study we used Harris County as the pilot site, which might not be robust, but the analytical methods can be generalized to other cities and regions with the availability of data. We also do not account for the bicycle and pedestrian infrastructure such as the quality of sidewalk or bike lane, which can help to explain some of the findings of this research. Future studies will try to address these limitations by implementing rigorous statistical models and image analysis tools to obtain the infrastructure information.

REFERENCES

Adanu, E. K., Smith, R., Powell, L. & Jones, S. 2017. Multilevel analysis of the role of human factors in regional disparities in crash outcomes. Accident Analysis & Prevention, 109, 10-17.

Algurén, B., & Rizzi, M. 2022. In-depth understanding of single bicycle crashes in Sweden—Crash characteristics, injury types and health outcomes differentiated by gender and age-groups. Journal of Transport & Health, 24,

Balakrishnan, S., Moridpour, S., & Tay, R. 2019. Sociodemographic influences on injury severity in truck-vulnerable road user crashes. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 5(4), 04019015.6.

Barajas, J. M. 2018. Not all crashes are created equal: Associations between the built environment and disparities in bicycle collisions. Journal of Transport and Land Use, 11(1), 865-882.

Behnood, A., & Mannering, F. 2017. Determinants of bicyclist injury severities in bicycle-vehicle crashes: A random parameters approach with heterogeneity in means and variances. Analytic Methods in Accident Research, 16, 35–47.

Billah, K., Sharif, H. O. & Dessouky, S. 2022. How gender affects motor vehicle crashes: A case study from San Antonio, Texas. Sustainability, 14(12), 7023.

Boele-Vos, M. J., Van Duijvenvoorde, K., Doumen, M. J. A., Duivenvoorden, C. W. A. E., Louwerse, W. J. R., & Davidse, R. J. 2017. Crashes involving cyclists aged 50 and over in the Netherlands: An in-depth study. Accident Analysis & Prevention, 105, 4–10.

Boufous, S., Rome, L. D., Senserrick, T., & Ivers, R. 2011. Cycling Crashes in Children, Adolescents, and Adults—A Comparative Analysis. Traffic Injury Prevention, 12(3), 244–250.

Braun, L. M., Rodriguez, D. A., & Gordon-Larsen, P. (2019). Social (in)equity in access to cycling infrastructure: Cross-sectional associations between bike lanes and area-level sociodemographic characteristics in 22 large U.S. cities. Journal of Transport Geography, 80, 102544.

Braver, E. R. 2004. Are older drivers actually at higher risk of involvement in collisions resulting in deaths or non-fatal injuries among their passengers and other road users? Injury Prevention, 10(1), 27–32.

Burdett, B. R. D., Starkey, N. J., & Charlton, S. G. 2017. The close to home effect in road crashes. Safety Science, 98, 1–8.

Dadashova, B., & Griffin, G. P. 2020. Random parameter models for estimating statewide daily bicycle counts using crowdsourced data. Transportation Research Part D: Transport and Environment, 84, 102368.

Dadashova, B., Griffin, G. P., Das, S., Turner, S., & Sherman, B. 2020. Estimation of Average Annual Daily Bicycle Counts using Crowdsourced Strava Data. Transportation Research Record: Journal of the Transportation Research Board, 2674(11), 390–402.

Das, S., Bibeka, A., Sun, X., Zhou, H. "Tracy," & Jalayer, M. 2019. Elderly pedestrian fatal crash-related contributing factors: applying empirical Bayes geometric mean method. Transportation Research Record: Journal of the Transportation Research Board, 2673(8), 254–263.

Ding, H., Sze, N. N., Li, H., & Guo, Y. 2020. Roles of infrastructure and land use in bicycle crash exposure and frequency: A case study using Greater London bike sharing data. Accident Analysis & Prevention, 144, 105652.

Elvik, R. and Bjørnskau, T., 2017. Safety-in-numbers: a systematic review and meta-analysis of evidence. Safety science, 92, 274-282.

Eren, H., & Gauld, C. 2022. Smartphone use among young drivers: Applying an extended Theory of Planned Behaviour to predict young drivers' intention and engagement in concealed responding. Accident Analysis & Prevention, 164, 106474.

Ferenchak, N. N., & Marshall, W. E. 2021. Bicycling facility inequalities and the causality dilemma with socioeconomic/sociodemographic change. Transportation Research Part D: Transport and Environment, 97, 102920.

Fuller, D., & Winters, M. 2017. Income inequalities in Bike Score and bicycling to work in Canada. Journal of Transport & Health, 7, 264–268.

Glassbrenner, D., Herbert, G., Reish, L., Webb, C., & Lindsey, T. 2022. Evaluating disparities in traffic fatalities by race, ethnicity, and income (Report No. DOT HS 813 188). https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813188. Accessed October 5th, 2022.

Gong, L. & Fan, W. D. 2017. Modeling single-vehicle run-off-road crash severity in rural areas: Accounting for unobserved heterogeneity and age difference. Accident Analysis & Prevention, 101, 124-134.

Governors Highway Safety Association (GHSA). 2021. An Analysis of Traffic Fatalities by Race and Ethnicity. https://www.ghsa.org/resources/Analysis-of-Traffic-Fatalities-by-Race-and-Ethnicity21. Accessed October 5th, 2022.

Hasheminejad, S. H.-A., Zahedi, M., & Hasheminejad, S. M. H. (2018). A hybrid clustering and classification approach for predicting crash injury severity on rural roads. International Journal of Injury Control and Safety Promotion, 25(1), 85–101.

Ivan, K., Benedek, J., & Ciobanu, S. 2019. School-aged pedestrian–vehicle crash vulnerability. Sustainability, 11(4), 1214.

Kemnitzer, C. R., Pope, C. N., Nwosu, A., Zhao, S., Wei, L. & Zhu, M. 2019. An investigation of driver, pedestrian, and environmental characteristics and resulting pedestrian injury. Traffic Injury Prevention, 20, 510-514.

Koopmans, J. M., Friedman, L., Kwon, S., & Sheehan, K. 2015. Urban crash-related child pedestrian injury incidence and characteristics associated with injury severity. Accident Analysis & Prevention, 77, 127–136.

Kravetz, D. & Noland, R. B. 2012. Spatial analysis of income disparities in pedestrian safety in northern New Jersey: is there an environmental justice issue? Transportation Research Record: Journal of the Transportation Research Board, 2320, 10-17.

Lee, J., Li, X., Mao, S., Fu, W. & Moridpour, S. 2021. Investigation of contributing factors to traffic crashes and violations: A random parameter multinomial logit approach. Journal of Advanced Transportation, 2021, 1-11.

Li, Z., Chen, C., Ci, Y., Zhang, G., Wu, Q., Liu, C., & Qian, Z. (Sean). (2018). Examining driver injury severity in intersection-related crashes using cluster analysis and hierarchical Bayesian models. Accident Analysis & Prevention, 120, 139–151.

Liang, O. S. & Yang, C. C. 2022. How are different sources of distraction associated with at-fault crashes among drivers of different age gender groups? Accident Analysis & Prevention, 165, 106505.

Lombardi, D. A., Horrey, W. J., & Courtney, T. K. 2017. Age-related differences in fatal intersection crashes in the United States. Accident Analysis & Prevention, 99, 20–29.

Martínez-Ruiz, V., Jiménez-Mejías, E., Luna-del-Castillo, J. de D., García-Martín, M., Jiménez-Moleón, J. J., & Lardelli-Claret, P. 2014. Association of cyclists' age and sex with risk of involvement in a crash before and after adjustment for cycling exposure. Accident Analysis & Prevention, 62, 259–267.

Masís, S. 2021. Interpretable machine learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples, Packt Publishing Ltd.

Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., & Hadi, M. (2020). Factors affecting injury severity in vehicle-pedestrian crashes: A day-of-week analysis using random parameter ordered response models and Artificial Neural Networks. International Journal of Transportation Science and Technology, 9(2), 100–115.

National Highway Traffic Safety Administration (NHTSA). 2021. Early Estimates of Motor Vehicle Traffic Fatalities and Fatality Rate by Sub-Categories in 2020. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813118. Accessed October 5th, 2022.

Oviedo-Trespalacios, O., & Scott-Parker, B. (2018). Young drivers and their cars: Safe and sound or the perfect storm? Accident Analysis & Prevention, 110, 18–28.

Prati, G., Fraboni, F., De Angelis, M., Pietrantoni, L., Johnson, D., & Shires, J. 2019. Gender differences in cycling patterns and attitudes towards cycling in a sample of European regular cyclists. Journal of Transport Geography, 78, 1–7.

Pulido, J., Barrio, G., Hoyos, J., Jimenez-Mejias, E., Martin-Rodriguez Mdel, M., Houwing, S. & Lardelli-Claret, P. 2016. The role of exposure on differences in driver death rates by gender and age: Results of a quasi-induced method on crash data in Spain. Accident Analysis Prevention, 94, 162-7.

Rahul, T. M., & Verma, A. 2014. A study of acceptable trip distances using walking and cycling in Bangalore. Journal of Transport Geography, 38, 106–113.

Rebentisch, H., Wasfi, R., Piatkowski, D. P. & Manaugh, K. 2019. Safe streets for all? Analyzing infrastructural response to pedestrian and cyclist crashes in New York City, 2009–2018. Transportation Research Record: Journal of the Transportation Research Board, 2673, 672-685.

Regev, S., Rolison, J. J. & Moutari, S. 2018. Crash risk by driver age, gender, and time of day using a new exposure methodology. Journal of Safety Research, 66, 131-140.

Roll, J., & McNeil, N. 2022. Race and income disparities in pedestrian injuries: factors influencing pedestrian safety inequity. Transportation Research Part D: Transport and Environment, 107, 103294.

Rothman, L., Cloutier, M.-S., Manaugh, K., Howard, A. W., Macpherson, A. K., & Macarthur, C. 2020. Spatial distribution of roadway environment features related to child pedestrian safety by census tract income in Toronto, Canada. Injury Prevention, 26(3), 229–233.

Russo, F., Biancardo, S. A. & Dell'acqua, G. 2014. Road safety from the perspective of driver gender and age as related to the injury crash frequency and road scenario. Traffic Injury Prevention, 15, 25-33.

Sagar, S., Stamatiadis, N. & Stromberg, A. 2021. Effect of socioeconomic and demographic factors on crash occurrence. Transportation Research Record: Journal of the Transportation Research Board, 2675, 80-91.

Saha, D., Alluri, P., Gan, A. & Wu, W. 2018. Spatial analysis of macro-level bicycle crashes using the class of conditional autoregressive models. Accident Analysis & Prevention, 118, 166-177.

Salon, D., & McIntyre, A. 2018. Determinants of pedestrian and bicyclist crash severity by party at fault in San Francisco, CA. Accident Analysis & Prevention, 110, 149–160.

Samerei, S. A., Aghabayk, K., Shiwakoti, N. & Mohammadi, A. 2021. Using latent class clustering and binary logistic regression to model Australian cyclist injury severity in motor vehicle-bicycle crashes. Journal of Safety Research, 79, 246-256.

Sartin, E. B., Metzger, K. B., Pfeiffer, M. R., Myers, R. K. & Curry, A. E. 2021. Facilitating research on racial and ethnic disparities and inequities in transportation: Application and evaluation of the

Bayesian Improved Surname Geocoding (BISG) algorithm. Traffic Injury Prevention, 22, S32-S37.

Scikit-Learn. 2022. Tuning the hyper-parameters of an estimator [Online]. Available: https://scikit-learn.org/stable/modules/grid_search.html [Accessed 07/30/2022].

Scott-Parker, B., & Oviedo-Trespalacios, O. 2017. Young driver risky behaviour and predictors of crash risk in Australia, New Zealand and Colombia: Same but different? Accident Analysis & Prevention, 99, 30–38.

Siddiqui, C., Abdel-Aty, M., & Choi, K. 2012. Macroscopic spatial analysis of pedestrian and bicycle crashes. Accident Analysis & Prevention, 45, 382–391.

Sivasankaran, S. K., & Balasubramanian, V. (2020). Exploring the severity of bicycle–vehicle crashes using latent class clustering approach in India. Journal of Safety Research, 72, 127–138.

Steinbach, R., Edwards, P. & Grundy, C. 2013. The road most travelled: The geographic distribution of road traffic injuries in England. International Journal of Health Geographics, 12(1), 1-7.

Steinbach, R., Green, J., Kenward, M. G., & Edwards, P. 2016. Is ethnic density associated with risk of child pedestrian injury? A comparison of inter-census changes in ethnic populations and injury rates. Ethnicity & Health, 21(1), 1–19.

Sun, M., Sun, X. & Shan, D. 2019. Pedestrian crash analysis with latent class clustering method. Accident Analysis Prevention, 124, 50-57.

Tom, A., & Granié, M.-A. 2011. Gender differences in pedestrian rule compliance and visual search at signalized and unsignalized crossroads. Accident Analysis & Prevention, 43(5), 1794–1801.

Toran Pour, A., Moridpour, S., Tay, R., & Rajabifard, A. 2018. Influence of pedestrian age and gender on spatial and temporal distribution of pedestrian crashes. Traffic Injury Prevention, 19(1), 81–87.

Ulak, M. B., Kocatepe, A., Ozguven, E. E., & Horner, M. W. 2019. How far from home do crashes occur? A network based analysis. Safety Science, 118, 298–308.

Wang, J., & Lindsey, G. 2017. Equity of bikeway distribution in Minneapolis, Minnesota. Transportation Research Record: Journal of the Transportation Research Board, 2605(1), 18–31.

Weiss, H. B., Kaplan, S., & Prato, C. G. (2014). Analysis of factors associated with injury severity in crashes involving young New Zealand drivers. Accident Analysis & Prevention, 65, 142–155.

Wheeler-Martin, K. C., Curry, A. E., Metzger, K. B. & Dimaggio, C. J. 2020. Trends in school-age pedestrian and pedalcyclist crashes in the USA: 26 states, 2000-2014. Injury Prevention, 26, 448-455.

Xiao, D., Šarić, Ž., Xu, X., & Yuan, Q. (2022). Investigating injury severity of pedestrian–vehicle crashes by integrating latent class cluster analysis and unbalanced panel mixed ordered probit model. Journal of Transportation Safety & Security, 1–20. https://doi.org/10.1080/19439962.2022.2033900

Zhao, H., Yang, G., Zhu, F., Jin, X., Begeman, P., Yin, Z., Yang, K. H., & Wang, Z. 2013. An Investigation on the Head Injuries of Adult Pedestrians by Passenger Cars in China. Traffic Injury Prevention, 14(7), 712–717.

APPENDIX

Table S1. Scaling Factors for Pedestrian and Bicyclist Crash from 2017 to 2019

Crash type	Year	Scaling factor
Pedestrian crashes	17	9.99
	18	9.97
	19	10.29
	20	6.06
Bicyclist crashes	17	48.31
	18	47.7
	19	46.53
	20	21.32

Table S2. Latent Class Cluster Results for Pedestrian and Bicyclist Crashes

Variables		Pedestrian Crashes			Bicyclist Crashes		
		Total	**LN-LN**	**HW-HW**	**Total**	**LN-LN**	**HW-HW**
DR_IncLvl	Low	456(16.2%)	396(27.1%)	60(4.4%)	192(17.1%)	160(31.7%)	32(5.2%)
	Low to medium	615(21.8%)	439(30.0%)	176(13.0%)	229(20.4%)	144(28.5%)	85(13.7%)
	Medium	815(28.9%)	445(30.4%)	370(27.2%)	280(24.9%)	123(24.3%)	157(25.4%)
	Medium to high	366(13.0%)	98(6.7%)	268(19.8%)	161(14.3%)	42(8.3%)	119(19.3%)
	High	569(20.2%)	84(5.8%)	485(35.7%)	260(23.2%)	36(7.2%)	224(36.4%)
DR_Ethnicity	White	883(31.3%)	135(9.2%)	748(55.0%)	385(34.3%)	47(9.4%)	338(54.7%)
	Hispanic	900(31.9%)	722(49.3%)	178(13.1%)	349(31.1%)	230(45.5%)	119(19.3%)
	Black	802(28.4%)	539(36.8%)	263(19.3%)	296(26.4%)	202(39.9%)	94(15.3%)
	Asian	184(6.5%)	45(3.1%)	139(10.2%)	70(6.2%)	18(3.5%)	52(8.5%)
	Other	53(1.9%)	21(1.5%)	32(2.3%)	22(2.0%)	9(1.7%)	13(2.2%)
VT_IncLvl	Low	602(21.3%)	493(33.7%)	109(8.0%)	210(18.7%)	188(37.2%)	22(3.6%)
	Low to medium	764(27.1%)	545(37.3%)	219(16.1%)	285(25.4%)	171(33.9%)	114(18.5%)
	Medium	558(19.8%)	217(14.8%)	341(25.1%)	210(18.7%)	75(14.9%)	135(21.9%)
	Medium to high	388(13.7%)	112(7.7%)	276(20.3%)	156(13.9%)	34(6.8%)	122(19.8%)
	High	510(18.1%)	96(6.5%)	414(30.5%)	261(23.2%)	37(7.3%)	224(36.3%)
VT_Ethnicity	White	932(33.0%)	265(18.1%)	667(49.1%)	486(43.3%)	103(20.3%)	383(62.1%)
	Hispanic	854(30.3%)	600(41.0%)	254(18.7%)	273(24.3%)	171(33.8%)	102(16.6%)
	Black	844(29.9%)	525(35.9%)	319(23.5%)	298(26.5%)	221(43.8%)	77(12.4%)
	Asian	130(4.6%)	50(3.4%)	80(5.9%)	52(4.6%)	9(1.8%)	43(7.0%)
	Other	61(2.2%)	23(1.5%)	38(2.8%)	13(1.2%)	2(0.3%)	11(1.8%)
Total		2822(100.0%)	1463(51.5%)	1359(48.5%)	1123(100.0%)	506(45.3%)	617(54.7%)

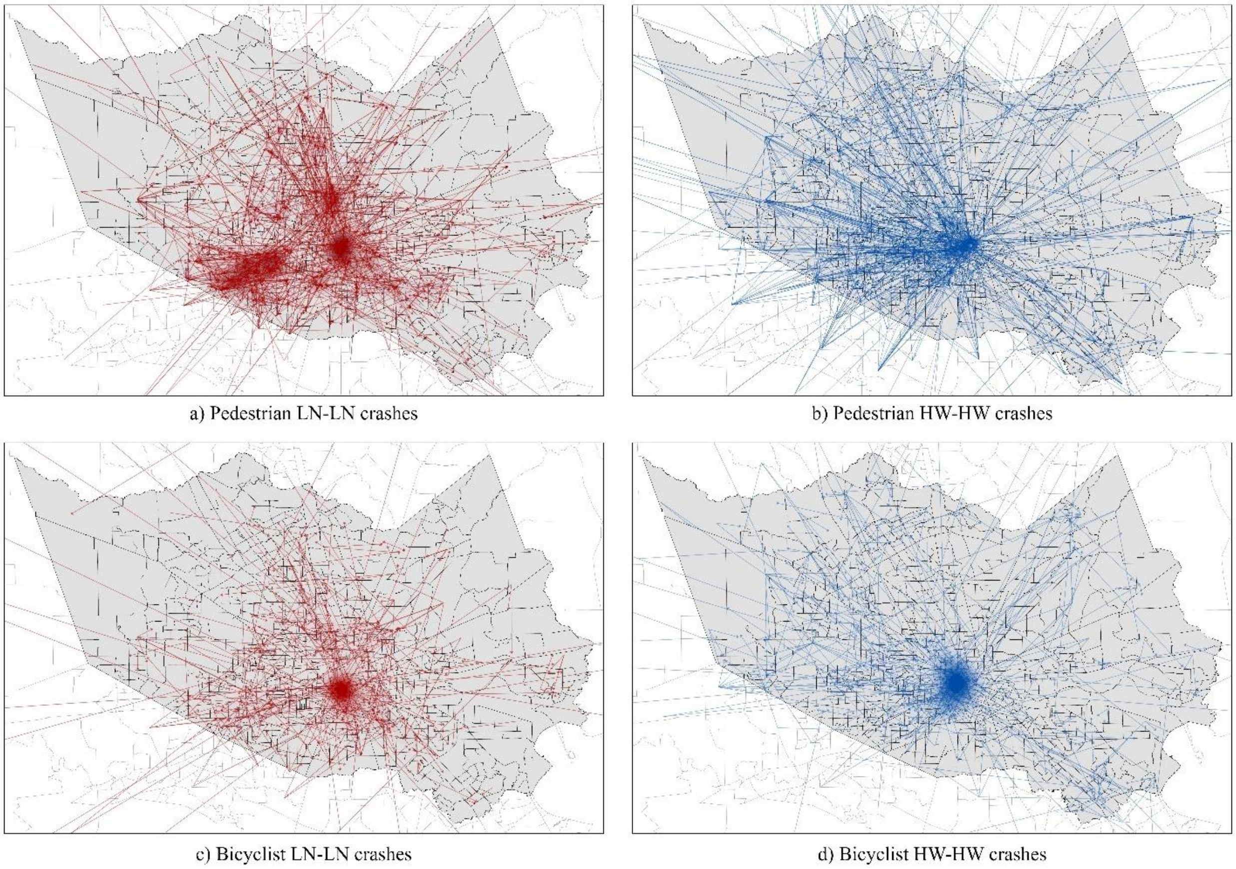


Figure S1. The Trajectory of Driver-Victim Pairs for LN-LN and HW-HW Crashes