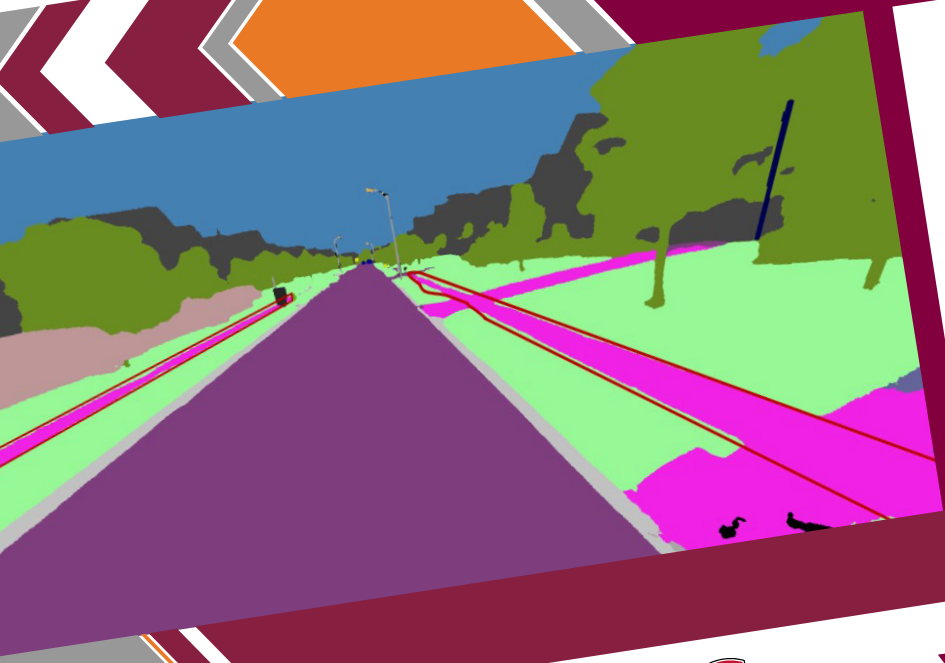
Building Equitable Safe Streets for All: Data-Driven Approach and Computational Tools

August 2023 | Final Report



SAFE-D
SAFETY THROUGH DISRUPTION

Texas A&M Transportation Institute

San Diego State University
Leadership Starts Here

VIRGINIA TECH TRANSPORTATION INSTITUTE
VIRGINIA TECH

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. 06-001	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Building Equitable Safe Streets for All: Data-Driven Approach and Computational Tools		5. Report Date August 2023 6. Performing Organization Code:
7. Author(s) Bahar Dadashova Chunwu Zhu Xinyue Ye Soheil Sohrabi Charles Brown Ingrid Potts		8. Performing Organization Report No.
9. Performing Organization Name and Address: Texas A&M Transportation Institute 3135 TAMU College Station, Texas 77843-3135		10. Work Unit No. 11. Contract or Grant No. 69A3551747115/06-001
12. Sponsoring Agency Name and Address Office of the Secretary of Transportation (OST) U.S. Department of Transportation (US DOT)		13. Type of Report and Period Final Research Report Start 9/2021 End 8/2023 14. Sponsoring Agency Code

15. Supplementary Notes
This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program.

16. Abstract:
Roadway safety in low-income and ethnically diverse U.S. communities has long been a major concern. This research was designed to address this issue by developing a data-driven approach and computational tools to quantify equity issues in roadway safety. This report employed data from Houston, Texas, to explore (1) the relationship between road infrastructure and communities' socioeconomic and demographic characteristics and its association with traffic safety in low-income, ethnically diverse communities and (2) the type of driver behaviors and characteristics that affect crash risks in underserved communities. The team first built an inclusive road infrastructure inventory database by employing remote sensing and image processing techniques. Then, the relationship between communities' socioeconomic and demographic characteristics and traffic safety was investigated through the lens of road infrastructure characteristics using data mining, deep learning tools, and statistical and econometric models. Clustering analysis was used to uncover the role in underserved communities of socioeconomic and demographic characteristics of drivers and victims involved in crashes. Structural equation models were then used to explore the association between neighborhood disadvantage, transportation infrastructure, and roadway crashes. Findings shed light on road safety inequity and sources of these disparities among communities using data-driven methods.

17. Key Words Roadway safety, equity, environmental justice, pedestrian and bicyclist crash, crowdsourced data, street view image, interpretable machine learning, latent class clustering, random forest, structural equation model	18. Distribution Statement No restrictions. This document is available to the public through the Safe-D National UTC website, as well as the following repositories: VTechWorks, The National Transportation Library, The Transportation Library, Volpe National Transportation Systems Center, Federal Highway Administration Research Library, and the National Technical Reports Library.

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 35	22. Price $0

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Abstract

Roadway safety in low-income and ethnically diverse U.S. communities has long been a major concern. This research was designed to address this issue by developing a data-driven approach and computational tools to quantify equity issues in roadway safety. This report employed data from Houston, Texas, to explore (1) the relationship between road infrastructure and communities' socioeconomic and demographic characteristics and its association with traffic safety in low-income, ethnically diverse communities and (2) the type of driver behaviors and characteristics that affect crash risks in underserved communities. The team first built an inclusive road infrastructure inventory database by employing remote sensing and image processing techniques. Then, the relationship between communities' socioeconomic and demographic characteristics and traffic safety was investigated through the lens of road infrastructure characteristics using data mining, deep learning tools, and statistical and econometric models. Clustering analysis was used to uncover the role in underserved communities of socioeconomic and demographic characteristics of drivers and victims involved in crashes. Structural equation models were then used to explore the association between neighborhood disadvantage, transportation infrastructure, and roadway crashes. Findings shed light on road safety inequity and sources of these disparities among communities using data-driven methods.

Acknowledgements

This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program. The authors would like to thank Sue Chrysler for supporting this project, the subject expert matter Tara Goddard for her comments and feedback that helped to improve the quality of the report and co-author Chanam Lee for her inputs and recommendations on one of the studies.

Table of Contents

List of Figures

List of Tables

Introduction

Improving roadway safety in low-income and ethnically diverse communities in the United States has long been a major concern. There are several factors that may contribute to unequal distribution of traffic risks in these communities, but one likely culprit is the transportation infrastructure and urban planning policies. In the 1960s, transportation infrastructure policies had significant impacts on low-income and Black communities. As the interest in building high-capacity roadways took off, these new policies were used to further segregate low-income and Black communities by building high-capacity highways nearby and dividing the neighborhoods [1, 2]. High-capacity roadways, such as interstates and freeways, experience higher traffic volumes, hence increasing the probability of traffic-related risks [3]. Moreover, poor roadway infrastructure and limited access to frequently used modes of transportation, such as vehicles and/or transit systems, may also increase the traffic-related safety risks in these communities [4, 5]. Due to limited access to transit and personally owned vehicles, travelers with lower incomes and members of immigrant communities are more likely to walk and ride bicycles than their higher-income counterparts [5]. However, limitations of the infrastructure (i.e., no sidewalks or bicycle infrastructure) can force users to share the road with motorized vehicles, increasing their exposure to traffic crashes [3, 6]. In fact, riskier driving behavior has been observed in low-income communities (e.g., not wearing seat belt), but the exact reasons for these behaviors have not been researched in detail [7].

The roadway safety concerns in low-income communities are a multi-dimensional, complex problem that requires a solution at various levels. All three major crash-contributing factors—roadway infrastructure, driver characteristics, and vehicle characteristics—are affected, thereby increasing the complexity of addressing roadway safety concerns in these communities. Assessing the presence and quality of roadway infrastructure in low-income communities is not a trivial task, as the existing roadway inventory databases do not include information regarding the quality of the roadway or presence of certain roadway designs such as crosswalks or bike lanes. Moreover, due to the dynamic nature of the problem (e.g., drivers are moving), pinpointing the exact driver-related safety factors may be challenging. Due to these complexities, quantifying equity concerns in roadway safety is not a trivial process and requires an interdisciplinary approach.

This study's objective was to develop a data-driven approach and computational tools to quantify the equity issues in roadway safety. This project assessed two important factors affecting crash frequency and severity in low-income communities: 1) roadway infrastructure and 2) road user characteristics and behaviors. To that end, we addressed the following two research questions:

1. What is the relationship between road infrastructure and communities' socioeconomic and demographic characteristics, and how it can be associated with traffic safety in low-income, ethnically diverse communities?
2. What type of driver characteristics affect the crash risks in underserved communities?

Harris County, Texas, served as the study site to accomplish the goals of this project. This county includes the Houston metro area, which is one of the most diverse cities in the United States.

Background

To establish a knowledge base for this inquiry, we performed a systematic literature review search in three academic databases (i.e., Web of Science, Scopus, and PubMed) commonly used in traffic safety research. The search query was a combination of three groups of terms in the title and abstract of the three databases: "pedestrian OR bicyclist* OR cyclist*" for crash types; "crash* OR accident* OR fatal* OR injur* OR death* OR collision* OR casualt*" for outcome variables, and "macro* OR area OR meso* OR zone OR zonal OR census OR community OR neighborhood" for analysis unit. The literature review search was conducted on Dec 26, 2022, and we limited the publication date of the studies from 2000 to 2022. To ensure consistency in selecting relevant studies, we defined the following set of inclusion and exclusion criteria to select the eligible studies that contained zonal crash prediction models: 1) aggregated pedestrian or bicyclist crashes and crash-contributing factors at the community level; 2) the outcome variable was the frequency of crashes, not crash rate or other metrics; 3) included actual roadway crashes rather than traffic risk perception or risk indicator; and 4) applied quantitative research methodology to model crashes at the zonal level. After retrieving the references, we uploaded them in the Covidence (https://www.covidence.org/)—a web-based tool the team used to conduct and manage the literature review task. We identified the relevant publications by their titles and abstracts in Covidence. The literature review process is shown in Figure 1.

A total of 2,809 unique articles from three databases were identified, and 89 were selected based on their title and abstract. After full text assessment, we identified 63 articles that met our criteria, including six articles according to a bibliographic check. As shown in Figure 2, there was less zonal crash prediction research before 2015, and recent years have witnessed a burst in zonal crash prediction research. Among all included articles, there were 37 (58.7%) publications from the United States, followed by nine (14.3%) publications from Canada and five (8%) publications from the United Kingdom. The remaining publications were from China, Australia, Serbia, South Korea, India, the Kingdom of Saudi Arabia, and Belgium.

We have divided the literature review synthesis into five sections: spatial unit of analysis, safety performance measure, modeling approach, crash-contributing factors, and disparities in roadway safety.

Studies included in the literature review involved conducting analysis based on the following spatial units: census zoning system, traffic analysis zoning system (TAZ), administrative zoning system, and researcher-crafted zoning system. The former three kinds of zoning system vary by country or region where the research took place. The number of articles and reasons for choosing each analysis unit are summarized in the Appendix (Table A-1).

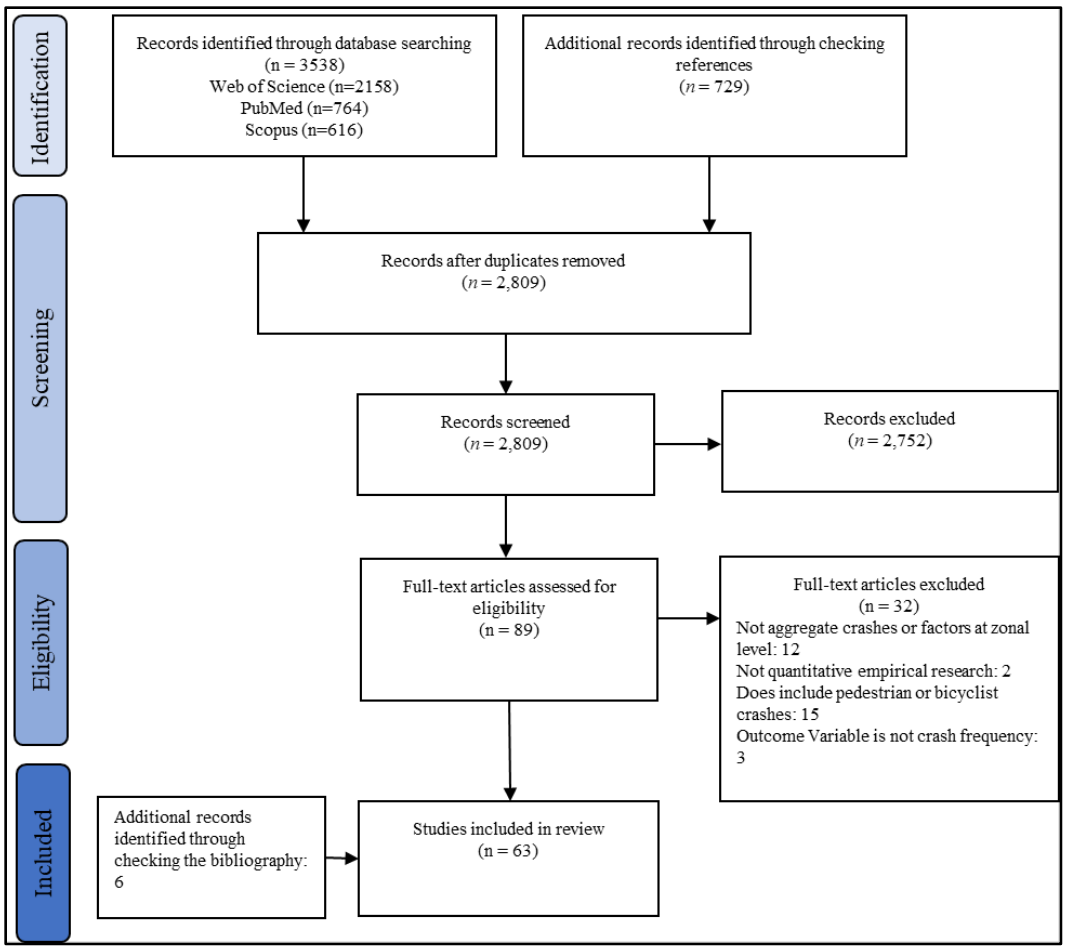


Figure 1. Flowchart. Systematic literature review process.

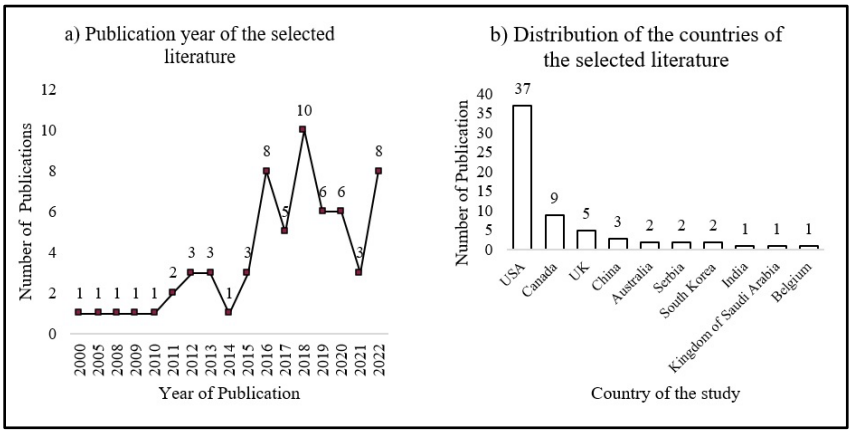


Figure 2. Graphs. Selected literature dates and countries of publication.

Selected studies mostly evaluated pedestrian and bicyclist crashes or both (often referred to as non-motorized crashes). Non-motorists like pedestrians and bicyclists are usually referred to as vulnerable road users (VRUs) in roadway safety research, as they are less protected and prone to

greater severity of injury. Most articles examined all pedestrians/bicyclists as a whole, while some researchers categorized crashes by severity level [8-10] or demographic characteristics like age [11, 12], gender [13], and ethnicity [14], depending on the research interest and study design.

Modeling approaches included multiple linear regression and generalized linear regression, spatial econometrics model, geographically weighted regression, random effect model, bivariate/multivariate model, Bayesian statistical model, and machine learning model. It is worth noting that the Bayesian statistical model and machine learning model have become more popular in recent years among safety researchers.

Crash-contributing factors are categorized into social environment factors, built environment factors, and exposure. The social environment includes demographic and economic characteristics of the communities. Factors from the built environment such as roadway infrastructure, land use, and facilities are commonly investigated. Exposure measures include the volume of different transportation modes such as walking, bicycling, and driving, and surrogate measures such as sociodemographic factors. We have summarized the measurement and influential direction of these factors for pedestrian and bicyclist crashes in the Appendix (Table A-2).

Disparity in roadway crashes is represented by a negative association in the proportion of minority population or income level variables, with other socioeconomic, built environment, and exposure variables controlled. Most researchers answered the question of whether there is a disparity in roadway safety among different neighborhoods by investigating the direction and magnitude of coefficients for equity-related variables after controlling other variables, including exposure. However, only a few studies scrutinized why there may be a disparity in roadway safety among different neighborhoods. We summarized how these studies investigated disparity in roadway safety by the types of disparity and methods to investigate the disparity, major factors influencing disparity, and major results related to disparity in the Appendix (Table A-3).

Methods

To accomplish the objectives of this project, several modeling and machine learning techniques were implemented.

Structural Equation Model

Our team applied a structural equation modeling (SEM) approach to investigate the relationship between neighborhood disadvantage, transportation infrastructure, traffic exposure, and non-motorist crashes. We also applied confirmatory factor analysis (CFA) under the SEM framework to construct latent variables such as neighborhood disadvantage, roadway environment, and lack of active transportation infrastructure. SEM and CFA have both been used in roadway safety research and transportation equity in active transportation research [14-17]. SEM has demonstrated several significant advantages over traditional multiple regression analyses, such as including flexible assumption, reducing measurement error, and testing complex causal paths [18]. The team

chose to use SEM for this research because of the following advantages:

- SEM considers the covariance among latent variables and indicators to construct latent variables. Crash-contributing factors are usually correlated, such as pedestrian exposure, bicyclist exposure, and vehicle exposure, which might cause estimation error when using multiple regression based on an assumption of not having severe multicollinearity.
- SEM can reduce the measurement error by applying CFA; latent variables, such as neighborhood disadvantage, are typically multidimensional and cannot be accurately captured using a single indicator. By combining information on neighborhood disadvantages from multiple indicators, CFA can create a composite factor that reduces measurement errors.
- SEM can model the complex causal paths taken by mediating variables. SEM can test the mediating effect of roadway environment, active transportation infrastructure, and traffic exposure from neighborhood disadvantage to non-motorist crashes, while multiple regression methods can only identify the bivariate association with other variables controlled.

We performed the SEM analysis on STATA 17 using a maximum likelihood estimation approach.

Latent Class Clustering and Random Forest

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 a random forest algorithm to investigate the relative contribution of factors to the crash patterns using 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. The framework for this approach is depicted in Figure 3.

Clustering analysis is an unsupervised machine learning method that can separate the crashes into homogenous subgroups that have the largest similarities within and largest dissimilarity between each subgroup [19]. We used a probability-based clustering approach (i.e., LCA), which has recently been applied in several roadway safety studies [20, 21]. The LCA 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; and 4) can generate statistical criteria afterward to select the best model with a certain number of clusters [20, 22]. The mathematical formula of the LCA approach is as follows [21]:

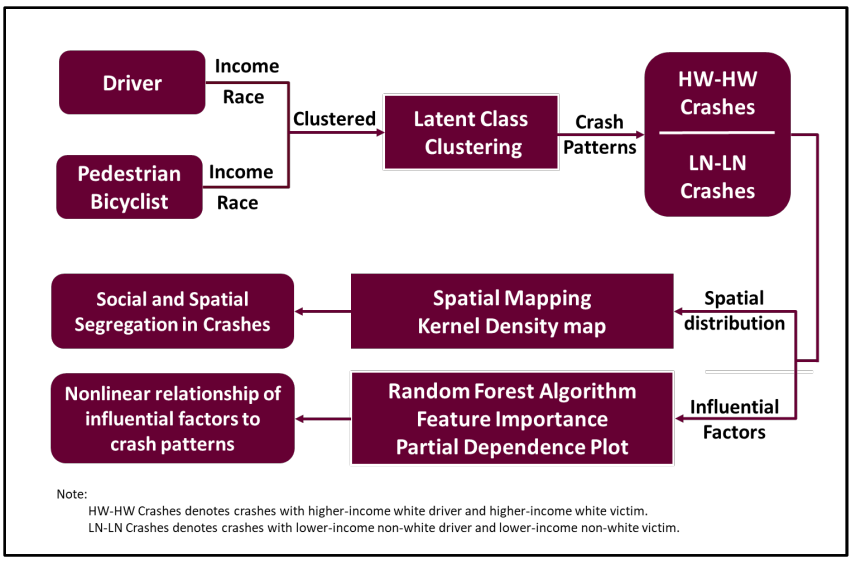


Figure 3. Diagram. Framework for assessing driver and victim pair characteristics.

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

Where $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}, \ldots \boldsymbol{Y}_{iM})$ is the observation (crash) i's responses in M category and the possible values of $\boldsymbol{Y}_{iM}$ are $1, \ldots, r_m$; 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, 1 if y equals n and 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 employed Bayesian information criteria (BIC) to select the appropriate number of clusters. LCA modeling and BIC calculation were conducted using the polPCA package in R.

A random forest algorithm is 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. This algorithm can be applied in both classification and regression; in this task, we used the random forest algorithm for classification. 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 $\{x_1, x_2 \ldots, x_n\}$, target set Y be $\{y_1, y_2 \ldots, y_n\}$, and $i = 1, 2, \ldots I$; the process of random forest can be represented as follows:

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 $\{f_1, f_2 \ldots, f_I\}$;
4) Aggregate 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 $\hat{f}(\hat{x}) = \max_{i=1,2,\ldots,I} f_i(x_i)$

Several parameters can affect the performance of the model, such as 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. 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 [23]. 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 [23]. A random forest algorithm was implemented by Scikit-learn, and PDPs were generated by pdpbox in Python.

Results

Role of Road Infrastructure in Traffic Safety Inequities

Data Collection

To assess the role of transportation infrastructure in safety disparities, we divided Harris County into 2,221 hexagons with a side length of 1 mile and an area of 0.76 square miles to develop SEMs for neighborhood disadvantage, transportation infrastructure, traffic exposure, and non-motorist crashes. We aggregated all the variables including pedestrian and bicyclist crashes, socioeconomic variables, roadway environment, active transportation infrastructure, and traffic exposure on each hexagon by taking the weighted averages. The non-motorist crashes (e.g., pedestrian crashes and bicyclist crashes) were taken from police-recorded crash data in The Texas Department of Transportation's (TxDOT's) Crash Records Information System (CRIS) database from 2018-2020. Socioeconomic variables to measure the neighborhood disadvantage including poverty rate, Hispanic and Black ratio, no high school diploma ratio, public assistance ratio, and no health insurance ratio were collected from the American Community Survey (ACS) 5-year Data in 2019. We obtained roadway environment variables from the roadway inventory from TxDOT (a GIS-based road network database), including roadway length, roadway without median, intersection number, and intersection with four legs or above. We obtained traffic signal data from Houston TranStar (http://www.houstontranstar.org/) to identify signalized intersections, as intersection data is not readily available from the roadway inventory database. We downloaded bike lane geospatial data from Houston Map Viewer（https://mycity.maps.arcgis.com/apps/webappviewer/index.html) and updated it to the most recent status by checking Google Maps. We also identified the presence of sidewalks in Houston's street view images using image segmentation analysis that applied a deep learning model called Seamless Scene Segmentation (Seamseg) [24] and took the proportion of street views without sidewalks as the surrogate measurement of no sidewalk proportion. Finally, we calculated the vehicle miles traveled (VMT), bicycle miles traveled (BMT), and pedestrian miles traveled (PMT) in 2019 to measure the exposure for vehicles, bicyclists, and pedestrians in the hexagon using following equation:

$$VMT/BMT/PMT = \frac{AADT/AADB/AADP \ * road\ segment\ length \ * 365}{1{,}000{,}000}$$

The average annual daily traffic (AADT) data was obtained from TxDOT's roadway inventory, and average annual daily bicycles/pedestrians (AADB/AADP) data was taken from Strava Metro, a crowdsourced database. Because of the bias in the measurement of AADB/AADP, we applied the exposure models developed by researchers in previous studies [25]. Table A-4 shows the descriptive statistics of the variables, including number of non-motorist crashes, socioeconomic variables (poverty rate, Hispanic and Black ratio, no high school diploma ratio, public assistance ratio, no health insurance ratio), transportation infrastructure (roadway length, roadway without median, intersection number, complex intersections, no sidewalk ratio, no bike lane ratio, no signal ratio), and traffic exposure (BMT, PMT, and VMT).

Measurement Model

In our measurement models, we constructed six latent variables to capture the level of neighborhood disadvantage, roadway environment, lack of active transportation infrastructure, active transportation exposure, and pedestrian and bicyclist crashes using single and multiple indicator measurement models (Table 1). We also assessed the contribution of variance explained for each indicator, internal consistency, and convergent validity using factor loading, composite reliability, and average variance extracted. A good measurement model generally has a factor loading for each indicator over 0.7, a composite reliability (CR) over 0.7, and an average variance extracted (AVE) over 0.5 [16]. Our measurement models meet all the criteria for a valid measurement model, except for three indicators' factor loading ranging from 0.637 (no bike lane ratio) to 0.681 (no signal ratio). This is also acceptable because their CR scores were greater than 0.7, indicating good internal consistency among indicators, and AVE was greater than 0.5, indicating the measurement models have good convergent validity.

Table 1. Indicators and Latent Variables in Measurement Model

Latent Variables and Indicators	Factor Loading
Neighborhood disadvantage (CR:0.945; AVE:0.773)	
Poverty rate	0.936
Hispanic and Black ratio	0.934
No high school diploma ratio	0.849
Public assistance ratio	0.854
No health insurance ratio	0.792
Roadway environment (CR: 0.951; AVE:0.829)	
Roadway length	0.982
Roadway without median	0.954
Intersection number	0.910
Complex intersections	0.673
Lack of active transportation infrastructure (CR:0.724; AVE:0.547)	
No sidewalk ratio	0.878

Latent Variables and Indicators	Factor Loading
No bike lane ratio	0.637
No signal ratio	0.681
Active transportation exposure (CR:0.725; AVE:0.568)	
Bicycle miles traveled	0.703
Pedestrian miles traveled	0.809
Vehicle exposure (CR:1.000; AVE:1.000)	
Vehicle miles traveled	1.000
Non-motorist crashes (CR:1.000; AVE:1000)	
Number of pedestrian and bicyclist crashes	1.000

Note. All the indicators are significant at 0.01 level ($p < 0.01$); CR: Composite reliability; AVE: Average Variance Extracted

Structural Equation Models

Our SEM shows a good model fit. Referring to prior aggregated crash SEM research, a good and acceptable model fit is indicated by a comparative fit index (CFI) of over 0.9, a Tucker-Lewis index (TLI) of over 0.8, and a Root Mean Squared Error of Approximation (RMSEA) of less than 0.1 [16]. The unstandardized coefficients of the structural model are shown in Figure 4. As reported in the note of Figure 4, our SEM achieves an RMSEA of 0.098, a TFI of 0.918, and a CFI of 0.994, all of which are within the accepted range of a good and acceptable model fit.

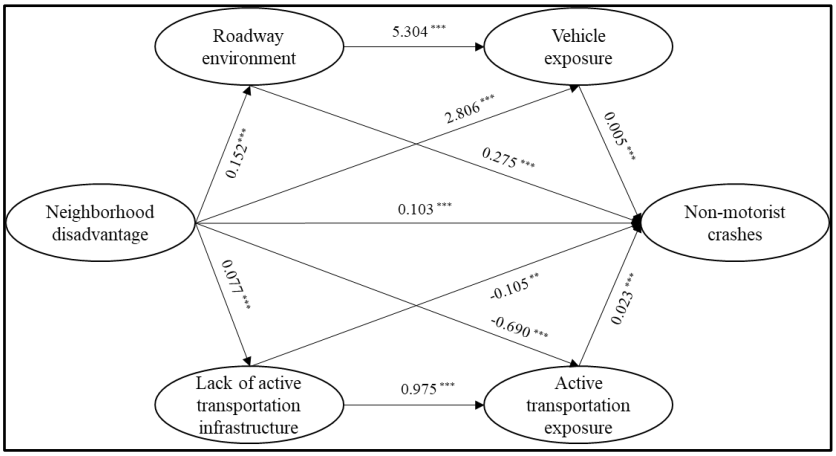


Figure 4. Diagram. Unstandardized direct path coefficients for the structural model.

Note. RMSEA = 0.098; CFI = 0.994; TFI = 0.918; Measurement models are omitted for better visualization; ***$p <$ 0.01; **$p < 0.05$; *$p < 0.1$. Neighborhood disadvantage is measured by poverty rate, Hispanic and Black ratio, no high school diploma ratio, public assistance ratio, and no health insurance ratio.

As illustrated, neighborhood disadvantage is significantly associated with roadway environment (i.e., roadway segment, intersection), vehicle exposure, lack of active transportation infrastructure, active transportation exposure, and non-motorist crashes. The direct path coefficient of neighborhood disadvantage to non-motorist crashes is 0.103 ($p < 0.01$), indicating one factor score increase in neighborhood disadvantage will result in 0.103 (10.3%) increase in the number of non-

motorist crashes. The positive direct relationship of neighborhood disadvantage to non-motorist crashes suggests crashes are not equally distributed across neighborhoods and disadvantaged neighborhoods are more prone to non-motorist crashes, after controlling roadway environment and exposure factors. From the perspective of motor vehicle transportation, disadvantaged neighborhoods are associated with a denser roadway environment and more direct vehicle exposure. Every factor score increase in the neighborhood disadvantage is associated with a 0.152 factor score increase in roadway environment and a 2.806 factor score increase in vehicle exposure. As expected, the denser roadway environment (0.275) and higher vehicle exposure (0.005) both contribute to the increase in non-motorist crashes. For active transportation modes, disadvantaged neighborhoods tend to have less active transportation infrastructure and less direct active transportation exposure. Every factor score increase in the neighborhood disadvantage is associated with a 0.077 factor score increase in lack of active transportation infrastructure and a 0.69 factor score decrease in active transportation exposure. Meanwhile, lack of active transportation infrastructure tends to reduce the number of non-motorist crashes (-0.105), but it does not mitigate the active transportation exposure (0.975). This indicates that in neighborhoods lacking active transportation infrastructure, there is a higher likelihood of pedestrians and bicyclists sharing the road with motor vehicles, which increases their risk of being involved in crashes, as evidenced by the positive direct effect of active transportation exposure to non-motorist crashes (0.023).

Overall, these findings suggest that disadvantaged neighborhoods experienced more non-motorist crashes with other variables controlled, suggesting a disparity in roadway safety between disadvantaged and advantaged neighborhoods. The concentration of motor vehicles in disadvantaged neighborhoods, where there are more motor roads, is one of the reasons for such inequality. On the other hand, there is less active transportation infrastructure and active transportation exposure in disadvantaged neighborhoods, which may have the opposing effects of increasing and decreasing the number of non-motorist crashes.

Direct, Indirect, and Total Effects of Neighborhood Disadvantage

To better interpret the influence of neighborhood disadvantage on non-motorist crashes through the mediating effect of transportation infrastructure and traffic exposure, we have summarized all the regression paths and coefficients, ratio for indirect effect to direct effect (RID), and ratio for indirect effect to total effect (RIT; Table 2). Our model suggests that there are 17 regression paths of neighborhood disadvantages to roadway environment, lack of active transportation infrastructure, vehicle exposure, active transportation exposure, and non-motorist crashes, including direct effects, indirect effects, and total effects.

Table 2. Direct Effect (DE), Indirect Effect (IE), and Total Effect (TE) of Neighborhood Disadvantage

Regression Path	Effect Type	Coefficient	*p*	RIT	RID
Path1: ND→RE	TE/DE	0.152	***	/	/
Path2: ND→LA	TE/DE	0.077	***	/	/

Regression Path	Effect Type	Coefficient	*p*	RIT	RID
Path3: ND→VE	TE	3.613	***	/	/
Path4: ND→VE	DE	2.806	***	/	/
Path5: ND→RE→VE	IE	0.806	***	22.3%	28.7%
Path6: ND→AE	TE	-0.616	***	/	/
Path7: ND→AE	DE	-0.691	***	/	/
Path8: ND→LA→AE	IE	0.075	***	12.2%	10.9%
Path9: ND→CRA	TE	0.141	***	/	/
Path10: ND→CRA	DE	0.103	***	/	/
Path11: ND→CRA	Total IE	0.037	***	26.6%	36.2%
Path12: ND→RE→CRA	IE	0.042	***	29.7%	40.4%
Path13: ND→VE→CRA	IE	0.014	***	10.0%	13.7%
Path14: ND→RE→VE→CRA	IE	0.004	***	2.9%	3.9%
Path15: ND→LA→CRA	IE	-0.008		5.7%	7.8%
Path16: ND→AE→CRA	IE	-0.016	**	11.5%	15.7%
Path17: ND→LA→AE→CRA	IE	0.002	**	1.3%	1.7%

Note. ND = neighborhood disadvantage; VE = vehicle exposure; RD = roadway environment; AE = active transportation infrastructure; LA = lack of active transportation infrastructure; CRA = non-motorist crashes; TE = total effect; DE = direct effect; IE = indirect effect; RIT = ratio of indirect effect to total effect (absolute value); RID = ratio of indirect effect to direct effect (absolute value); *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

Paths 1 and 2 show the positive direct effect of neighborhood disadvantage on the roadway environment and the lack of active transportation infrastructure, as discussed earlier. Paths 3 to 5 represent the direct, indirect, and total impact of neighborhood disadvantage on the level of vehicular exposure. The total effect of neighborhood disadvantage on vehicle exposure (3.613) is larger than the direct effect (2.806) due to the significant positive indirect effect of roadway environment (0.806). The mediating effect of roadway environment is 22.3% of the total effect and 28.7% of the direct effect. This suggests that higher vehicle exposure in disadvantaged neighborhoods is partially due to the denser roadway distribution (and potentially higher vehicular exposure) in these areas.

Paths 6 to 8 show the regression paths for neighborhood disadvantage to active transportation exposure. The total negative effect of neighborhood disadvantage to active transportation exposure (-0.616) is smaller than the direct negative effect (-0.691) because it is reduced by the positive mediating effect of lack of active transportation infrastructure (0.075). The indirect effect of lack of active transportation infrastructure is 12.2% of the total effect and 10.9% of the direct effect for neighborhood disadvantage to active transportation infrastructure. This indicates that the lack of active transportation infrastructure cannot suppress the demand for walking and biking in these areas. There are people who will still walk or bike in neighborhoods with inadequate active transportation infrastructure.

Paths 9 to 17 demonstrate all the regression paths from neighborhood disadvantage to non-motorist crashes. Neighborhood disadvantage has a positive direct effect (0.141), a positive total indirect

effect from all mediating variables (0.037), and a positive total effect (0.103) on non-motorist crashes. The total indirect effect from mediation is 26.6% of the total effect and 36.2% of the direct effect, which suggests that transportation infrastructure and traffic exposure partially improve the crash risk in disadvantaged neighborhoods. A close inspection of different transportation modes shows the discrepancy between motor vehicles and active transportation in their influence on non-motorist crashes. For the motor vehicle mode, the mediating effect of roadway environment (0.042), vehicle exposure (0.014), and both roadway environment and vehicle exposure (0.004) are all significantly positive. Together, they are 42.6% (29.7%+10.0%+2.9% = 42.6%) of the total effect and 58% (40.4%+13.7%+3.9% = 58%) of the direct effect, which means the mediating effect of motor vehicle mode has a very strong influence on non-motorist crashes. However, the mediating effect of active transportation mode shows a counter effect to non-motorist crashes. The total indirect effect of active transportation mode is -0.022 (-0.008-0.046+0.002 = -0.022), with a non-significant mediating effect of lack of roadway infrastructure (-0.008), a negative mediating effect of active transportation exposure (-0.016-), and a small positive mediating effect of lack of active transportation infrastructure and active transportation exposure (0.002). The total indirect effect of active transportation mode is 15.9% (|-5.7%-11.5%+1.3%| = 15.9%) of the total effect and 21.8% (|-7.8%-15.7%+1.7%| = 21.8%) of the direct effect. Compared to the motor vehicle mode, the indirect effect of active transportation mode is both negative and significantly smaller. This means active transportation infrastructure and transportation exposure tend to mitigate the crash risks in disadvantaged neighborhoods (hence the negative impact).

Investigating the Sociodemographic Characteristics of Drivers Involved in Traffic Crashes in Disadvantaged Communities

Data Collection

To assess the sociodemographic and economic characteristics of drivers and victims involved in VRU crashes in disadvantaged communities, we used pedestrian and bicyclist crashes, crash-contributing factors, socioeconomic characteristics of drivers and victims, roadway infrastructure characteristics, and traffic exposure in Harris County. Descriptive information of the variables is shown in Table A-4. We obtained records of pedestrian and bicyclist crashes in a 4-year period (2017-2020) from the TxDOT CRIS database. We identified pedestrian/bicyclist crashes based on the type of primary victim (pedestrian or bicyclist) involved in the crash. Eight factors in crash-specific information were retrieved from the CRIS database, including time of day, whether the crash happened on a weekday, season, weather condition, surface condition, whether the crash happened in a construction zone, whether the crash occurred at an intersection, and years the vehicle had been in use. We also retrieved driver ethnicity, age, and gender and victim ethnicity, age, and gender data from the CRIS database (direct measurements). Since drivers' and victims' income levels were usually not publicly available in police-reported crash data, this research represents the economic status of drivers using aggregated census data from drivers' residential ZIP codes [26,27]. Finally, we recoded the driver's income and victim's income to ordinal variables in five levels: low income (0 to 20th percentile), lower to medium income (20th to 40th

percentile), medium income (40th to 60th percentile), medium to high income (60th to 80th percentile), and high income (80th to 100th percentile), according to the five quintiles of their residential census tracts in the research area.

Roadway infrastructure data was collected from the TxDOT roadway inventory. Data for the roadway inventory is updated annually, and we used the 2020 version, which aligned with the time span of our crash events. We selected 11 characteristics of the roadway infrastructure where the crash happened, including road functional classification, speed limit, whether the crash occurred in an urban area, roadbed width (which comprises shoulder width and surface width), the number of lanes, lane width, median width, inside shoulder width, outside shoulder width, existence of left curb, and existence of right curb.

The team also considered vehicular, pedestrian, and bicyclist exposure variables in this study. Vehicular exposure of the road segments where the crash happened was measured as the AADT, which is available from the TxDOT roadway inventory database for each year of the crash events. In this study, we used a scaling approach to estimate the bicyclist and pedestrian counts. We obtained the observed data from the Texas Bicycle and Pedestrian Data Exchange (BP|CX) (https://mobility.tamu.edu/bikepeddata/). BP|CX is a data repository of pedestrian and bicyclist counts across Texas collected using permanent and short-term counters. We then used the crowdsourced Strava data to estimate the counts for sites with no pedestrian and bicyclist volumes via the scaling approach. In this approach, we first estimated the AADB/AADP using both observed and crowdsourced data obtained for the same sites, and then derived the adjustment factor by dividing the Strava AADB/AADP by observed AADB/AADP [28]. The scaling approach can be adjusted using spatial (e.g., segments vs. intersections) or temporal factors (e.g., weekend vs. weekday). We then multiplied the scaling factor with the Strava AADB/AADP data to estimate the bicyclist and pedestrian counts for sites with no volume data.

Clustering Crashes by LCA

We used the LCA algorithm to automatically group crashes with driver-victim pairs into clusters. According to this algorithm, the optimal number of clusters was two. After exploring the crash victims' and drivers' socioeconomic and demographic backgrounds, the clusters were interpreted as 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; Figure 5; see [29] for more details).

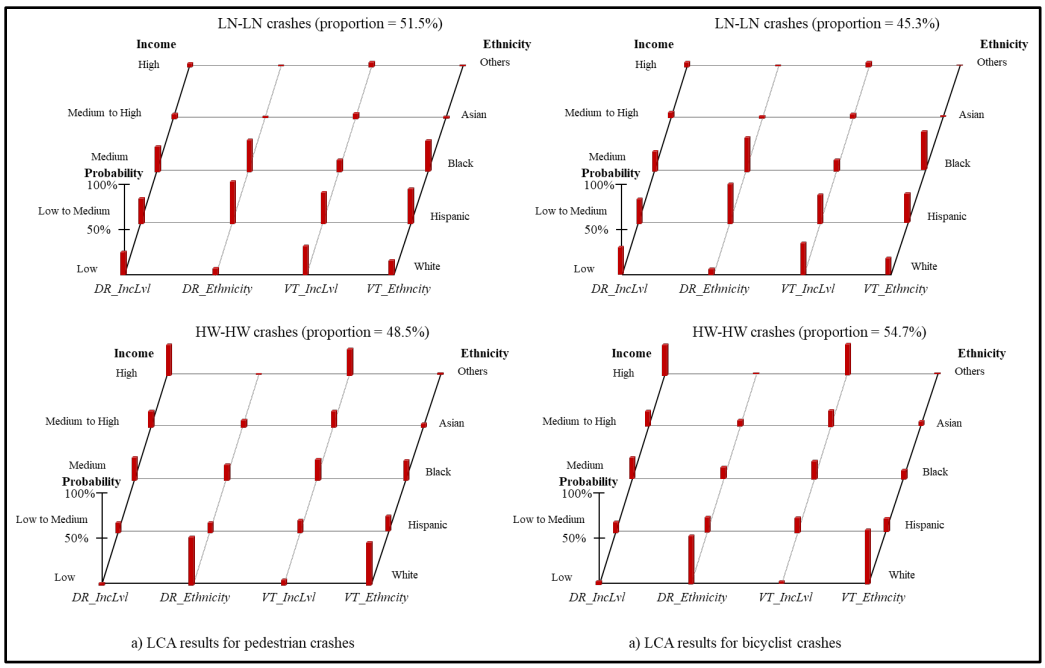


Figure 5. Graphs. Clustering results for pedestrian crashes and bicyclist crashes.

The descriptive statistics of variables included in each cluster are presented in Table A-5. 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 5a shows the two clusters and the corresponding distribution of driver and victim income levels and ethnicity in the pedestrian crash model. For driver characteristics, white drivers comprised 9.2% of LN-LN crashes, while their probability was 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-white (81.9%), while victims in HW-HW crashes have the highest probability of being white (49.1%). Victim income level is also distributed from 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 to 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. Results revealed notable socioeconomic patterns of driver-victim pairs, showing the socioeconomic 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.

Relative Importance of Selected Variables

We used the feature importance to quantify the influence of the variables in a random forest

algorithm to determine whether a crash belonged in LN-LN crashes or HW-HW crashes for pedestrian and bicyclist crashes (Table A-6). Given that we had two clusters in each model, the LN-LN crashes served as the baseline or comparison group. Thus, the higher value of feature importance a variable has, the larger contribution the variables will make in determining whether a crash belongs in the LN-LN crash cluster. 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 and bicyclist volumes each rank first in their respective models. 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 to the classification of LN-LN or HW-HW crashes. The driver and victim ages are also among the most influential variables, while their gender is less influential. Driver age and victim age rank second and third in pedestrian crashes, and driver age and victim age rank third and second in bicyclist crashes. For crash-specific information, the number of years the vehicle was 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 ages. Time of day and season rank eighth and ninth in bicyclist crashes and ninth and eighth in pedestrian crashes, 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 relatively high influence of roadway infrastructure characteristics in determining the crash clusters. However, their feature importance is relatively low compared to previous factors.

PDPs

PDPs are one of the model-agnostic interpretable machine learning approaches to reveal the marginal effect of a feature in machine learning models. They illustrate the change in the probability of a crash being clustered with LN-LN crashes along with the increase of each variable in both the pedestrian and bicyclist models (Figure A-1). For exposure variables, when AADP volumes are less than 2.6 pedestrian trips per day, pedestrian exposure is not influential. When AADP volumes are larger than 2.6 pedestrian trips per day, it becomes positively associated with the probability of a crash being an LN-LN crash. This indicates that LN-LN crashes will likely happen on the road with larger pedestrian exposure, and HW-HW crashes will be less likely. In bicyclist crashes, the positive marginal effect of bicyclist exposure on the probability of a crash being an LN-LN crash will increase when the bicyclist exposure becomes larger, which indicates an increasing non-linear association. 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 the lack of active transportation-friendly infrastructure in low income and minority communities, which may force bicyclists to share the road with oncoming traffic, increasing their crash probability. However, this speculation requires further investigation and validation, after considering bicyclist infrastructure during data analysis. For vehicle volumes, the pedestrian and bicyclist crashes have similar patterns, which shows lower AADT does not have significant influence on the probability

of a crash being an LN-LN crash. Within the highest quantile of the AADT, vehicle volumes 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.

Driver and victim ages 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 an LN-LN crash will decrease. When the driver's age is greater than 64, the probability of a crash being an 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 an 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 an LN-LN crash in its lower quintiles. Driver age only has a positive marginal effect when it exceeds 66 years, indicating that older drivers are more likely to be involved in a bicyclist LN-LN crash. For the victim's age, a victim being 32 or younger will increase the probability of a crash being an LN-LN crash, while a victim being 33 or older will decrease the probability of a crash being an LN-LN crash. This means bicyclist LN-LN crashes are more likely to involve older drivers and younger bicyclists.

For crash-specific information, years the vehicle was in use, time of day, and season rank among the most influential variables. When the age of the vehicle is less than 6 years, it has negligible influence on the probability of a pedestrian LN-LN crash, which is also the case for bicyclist crashes. As the age of the vehicle increases in pedestrian crashes, its marginal effect will become larger in a negative direction (as, again, for bicyclist crashes). This indicates older vehicles are less likely to be involved in an LN-LN crash and more likely to be involved in an HW-HW crash for both pedestrian and bicyclist crashes. In summer and autumn, the probability of being an LN-LN crash is higher than in winter, though the effect of the influence is minimal. For bicyclist crashes, 6:00 a.m. to 12:00 p.m. and 12:00 to 6:00 p.m. have a higher chance of bicyclist crashes. Lower income and non-white groups might choose biking as their mode of transportation to commute during the daytime more frequently than their higher income counterparts due to economic affordability or behavioral difference, which results in 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 an 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 it is 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.

Discussion

Environmental injustice may happen in disadvantaged neighborhoods through two paths related to transportation: motor vehicles and active transportation. The regression path of the motor vehicle mode suggests that disadvantaged neighborhoods are often characterized by high-density roadways with a high volume of motor vehicle traffic. Consistent with most prior research, results show that these both directly and indirectly increase crash risks. Reasons for the denser roadway and higher vehicular exposure in disadvantaged neighborhoods are multiple. Historically, the transportation planning system has tended to place high-capacity roadways near or within disadvantaged communities since the release of Federal Aid Highway Act of 1956, which has shaped the structural racism and discrimination in transportation planning [30]. Spatially, disadvantaged communities tend to be situated in inner-city areas with higher levels of arterial traffic, in contrast to more affluent suburban communities [5]. Socioeconomically, disadvantaged communities may lack the political and economic power to resist construction of new high-capacity roadways or to demand existing roadways be rerouted away from their neighborhoods.

The active transportation mode is less understood and has received less attention in research compared to motor vehicles. In the current study, disadvantaged communities are characterized by less active transportation infrastructure and less active transportation exposure. Both have a negative mediating effect for neighborhood disadvantage on non-motorist crashes. This indicates that, while neighborhood disadvantage is associated with a higher risk of non-motorist crashes, inadequate active transportation infrastructure and fewer active transportation travelers may actually mitigate this risk to some extent. However, this may conflict with the goals of transportation planners and urban planners who prioritize the promotion of active transportation and value its benefits. While significantly reducing the number of pedestrians and bicyclists traveling on the streets could potentially result in fewer non-motorist crashes, this approach would forego the multiple benefits associated with active transportation, including improved health outcomes for travelers and reduced environmental impact from motor vehicles. Some scholars in active transportation have argued that there is a "safety in numbers" effect, which has been investigated in some meta-analyses [31, 32] and multiple empirical studies [33-35]. This refers to the phenomenon that there will be less risk of injury for each pedestrian or bicyclist when there are more people walking or biking [32]. However, the safety in numbers effect may exhibit a non-linear relationship, possibly even a reversed U-shaped curve, which requires further investigation [36]. This effect might appear in Northern European countries, where walking and bicycling are dominant modes of transportation [36], but it clearly does not exist in the Houston case, where motor vehicles remain the primary mode of transportation.

Using crash data from Harris County, we applied a probability-based LCA to classify pedestrian and bicyclist crashes. The clustering results show 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 social segregation in pedestrian and bicyclist crashes, indicating that drivers and victims with 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 longer distance and getting involved in crashes in farther geographic areas than their counterparts.

To explore how the LN-LN and HW-HW crash patterns are shaped, we applied a random forest algorithm and PDPs to model and interpret the clustering outcomes from LCA models. Contributing factors for the crash patterns were selected from crash-specific information, driver and victim age and gender, roadway infrastructure, and traffic exposure. Pedestrian/bicyclist exposure, driver age, victim age, years the vehicle had been in use, AADT, speed limit, roadbed width, time of day, and season are the most influential variables in pedestrian and bicyclist models. We drew PDPs for the most influential variables to interpret how the variables are associated with crash patterns. The results show that LN-LN crashes tend to happen on the road with larger traffic exposure of pedestrians/bicyclists and vehicles, which could be due to the lack of adequate active transportation infrastructure [32]. 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. Older vehicles 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.

Conclusions and Recommendations

Research and data analysis conducted in this project advanced the understanding of environmental justice in roadway safety by exploring the relationship between neighborhood disadvantage, transportation infrastructure, traffic exposure, and non-motorist crashes theoretically and methodologically. It distinguishes the two theoretical pathways of how environmental injustice happened in disadvantaged communities related to the motor vehicle transportation mode and active transportation mode. Methodologically, this study innovatively applied an SEM approach that aligns with the theoretical framework and investigated the direct, indirect, and total effect of neighborhood disadvantage on non-motorist crashes, which is superior to a traditional statistical approach that does not consider the intercorrelation between crash-related factors. Furthermore, to address the limitations of currently available active transportation infrastructure and exposure data sources, the study utilized computer vision models to automatically collect sidewalk information and adjusted pedestrian/bicyclist exposure using recent crowdsourced data. This study also contributes to the foundation of transportation policy aimed at promoting environmental justice by

designing and investing in transportation infrastructure that allocates traffic exposure fairly for both motor vehicle and active transportation modes. Transportation planners should be mindful of the detrimental effects of constructing high-capacity roadways near disadvantaged communities and take measures in transportation investment to mitigate the disproportionate impact of roadway infrastructure in disadvantaged communities. To harness multiple benefits of the safety in numbers effect, transportation planning and policy should prioritize measures to encourage active transportation and shift from car-centric urban lifestyles to more diverse transportation modes.

Additionally, we examined the characteristics of both driver and victim simultaneously by pairing the driver and pedestrian/bicyclist involved in the same crash. This served to broaden understanding of the socioeconomic and demographic make-up of the two parties involved in crashes and the geographic distribution of these crashes and crash-contributing factors. For this purpose, we applied the 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 PDPs to model and interpret the contributing factors of the clusters in both pedestrian and bicyclist models. This research contributes to understanding roadway crashes in several ways. First, this study confirms the long-believed hypothesis that there is a clear sociodemographic and economic segregation of crashes. We also found that crash-contributing factors often vary across different communities. These results can help safety practitioners in both engineering and planning fields to develop and implement practices targeting the main concerns of each community instead of developing one-size-fits-all strategies. The 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 used machine learning techniques to address a largely unexplored research question that involved analyzing driver and victim characteristics simultaneously.

Additional Products

Project Page: https://safed.vtti.vt.edu/projects/building-equitable-safe-streets-for-all-data-driven-approach-and-computational-tools/

Education and Workforce Development Products

Education and workforce development products include:

- Texas A&M Transportation Institute Graduate Assistant Researcher (GAR) Chunwu Zhu developed the methods and algorithms for the work conducted in this project.
- The work initiated in this project will be used in the Ph.D. dissertation of the GAR student Chunwu Zhu.
- A pilot project titled *Smart Information Services for Building Equitable Active Transportation Culture* based on the image segmentation analysis initiated in this project has been funded by the Urban AI Lab of Texas A&M University (TAMU) Data Science

Institute, led by TAMU Landscape Architecture and Urban Planning Professor and research team member Xinyue Ye: https://urbanai.tamids.tamu.edu/2023/02/16/pilot-project-1-smart-information-services/

Technology Transfer Products

The following products were or will be generated:

- Zhu, C., Brown, C. T., Dadashova, B., Ye, X., Sohrabi, S., & Potts, I. (2023). Investigation on the driver-victim pairs in pedestrian and bicyclist crashes by latent class clustering and random forest algorithm. *Accident Analysis & Prevention, 182*, 106964.
- Zhu, C., & Dadashova, B. (2023). *Investigation on the driver-victim pairs in pedestrian and bicyclist crashes by latent class clustering and random forest algorithm*. Presented at the TRB Annual Meeting, Washington DC, 2023.
- Zhu, C., Dadashova, Lee, C., B., Brown, C. T., & Ye, X. Structural equation models for assessing the neighborhood disadvantage and roadway safety [Unpublished manuscript]. In Preparation to be submitted to the Journal of Planning Education and Research.
- Zhu, C., Sohrabi, S., Dadashova, B., Brown, C. T., Ye, X., & Potts, I. A review of zonal crash frequency research for pedestrians and bicyclists and their equity concerns: Systematic review of literature [Unpublished manuscript]. In Preparation to be submitted to the Health & Place.

- Smart Information System Dashboard: https://tamu.maps.arcgis.com/apps/dashboards/49f9e21e68654cda836b41f4bacf0e2e

Data Products

Links to data products from this research are available on project website: https://safed.vtti.vt.edu/projects/building-equitable-safe-streets-for-all-data-driven-approach-and-computational-tools/.

References

1. Golub, A., R. A. Marcantonio, and T. W. Sanchez. Race, Space, and Struggles for Mobility: Transportation Impacts on African Americans in Oakland and the East Bay. *Urban Geography*, Vol. 34, No. 5, 2013, pp. 699–728.

2. Salomons, E. M., and M. B. Pont. Urban Traffic Noise and the Relation to Urban Density, Form, and Traffic Elasticity. *Landscape and Urban Planning*, Vol. 108, No. 1, 2012, pp. 2–16.

3. McAndrews, C., and W. Marshall. Livable Streets, Livable Arterials? Characteristics of Commercial Arterial Roads Associated with Neighborhood Livability. *Journal of the American Planning Association*, Vol. 84, No. 1, 2018, pp. 33–44.

4. Noland, R. B., and M. L. Laham. Are low income and minority households more likely to die from traffic-related crashes? *Accident Analysis & Prevention*, Vol. 120, 2018, pp. 233-238.

5. Barajas, J. M. Not All Crashes Are Created Equal: Associations between the Built Environment and Disparities in Bicycle Collisions. Journal of Transport and Land Use, Vol. 11, No. 1, 2018. https://doi.org/10.5198/jtlu.2018.1145.

6. Forkenbrock, D. J., and L. A. Schweitzer. Environmental justice in transportation planning. *Journal of the American Planning Association*, Vol. 65, No. 1, 1999, pp. 96-112.

7. Elias, W., A. Blank-Gomel, C. Habib-Matar, and Y. Shiftan. Who are the traffic offenders among ethnic groups and why? *Accident Analysis & Prevention*, Vol. 91, 2016, pp. 64-71.

8. Amoh-Gyimah, R., M. Saberi, and M. Sarvi. Macroscopic modeling of pedestrian and bicycle crashes: A cross-comparison of estimation methods. *Accident Analysis & Prevention*, Vol. 93, 2016, pp. 147-159.

9. Alkahtani, K. F., M. Abdel-Aty, and J. Lee. A zonal level safety investigation of pedestrian crashes in Riyadh, Saudi Arabia. *International Journal of Sustainable Transportation*, Vol. 13, No. 4, 2019, pp. 255-267.

10. Dadashova, B., E. S. Park, S. M. Mousavi, B. Dai, and R. Sanders. Assessment of inequity in bicyclist crashes using bivariate Bayesian copulas. *Journal of Safety Research*, Vol. 82, 2022, pp. 221–232.

11. Ha, H.-H., and J.-C. Thill. Analysis of traffic hazard intensity: A spatial epidemiology case study of urban pedestrians. *Computers, Environment and Urban Systems*, Vol. 35, No. 3, 2011, pp. 230-240. https://doi.org/10.1016/j.compenvurbsys.2010.12.004

12. Kim, H. S., S. H. Oh, and Y. Choi. Pedestrian-Vehicle Collision Vulnerability in Senior Citizens' Walking Environment: An Area-Level Investigation of Seoul, South Korea. *KSCE Journal of Civil Engineering*, Vol. 24, No. 11, 2020, pp. 3461-3473.

13. Pirdavani, A., S. Daniels, K. van Vlierden, K. Brijs, and B. Kochan. Socioeconomic and sociodemographic inequalities and their association with road traffic injuries. *Journal of Transport & Health*, Vol. 4, 2017, pp. 152-161.

14. Barajas, J. M. Perceptions, People, and Places: Influences on Cycling for Latino Immigrants and Implications for Equity. *Journal of Planning Education and Research*, Vol. 43, No. 1, 2023, pp. 196–211.

15. Ewing, R., S. Hamidi, and J. B. Grace. Urban sprawl as a risk factor in motor vehicle crashes. *Urban Studies,* Vol. 53, No. 2, 2016, pp. 247-266.

16. Najaf, P., J.-C. Thill, W. Zhang, and M. G. Fields. City-level urban form and traffic safety: A structural equation modeling analysis of direct and indirect effects. *Journal of Transport Geography,* Vol. 69, 2018, pp. 257-270.

17. Kamel, M. B., T. Sayed, and A. Osama. Accounting for mediation in cyclist-vehicle crash models: A Bayesian mediation analysis approach. *Accident Analysis & Prevention,* Vol. 131, 2019, pp. 122-130.

18. Garson, G. D. (2015). *Structural Equation Modeling*. Statistical Publishing Associates.

19. Sivasankaran, S. K., and V. Balasubramanian. Exploring the severity of bicycle–vehicle crashes using latent class clustering approach in India. *Journal of Safety Research*, Vol. 72, 2020, pp. 127–138.

20. Sun, M., X. Sun, and D. Shan. Pedestrian crash analysis with latent class clustering method. *Accident Analysis & Prevention*, Vol. 124, 2019, pp. 50–57.

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

22. Sasidharan, L., K.-F. Wu, and M. Menendez. Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland. *Accident Analysis & Prevention*, Vol. 85, 2015, pp. 219–228.

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

24. Porzi, L., S. R. Bulò, and P. Kontschieder. *Improving Panoptic Segmentation at All Scales* (arXiv:2012.07717). arXiv, 2021.

25. Dadashova, B., K. Dixon, J. Hudson, R. Benz, B. Dai, X. Li, I. Sener, S. Turner, and S. Sarda. Addressing Bicyclist Safety through the Development of Crash Modification Factors for Bikeways. Publication FHWA/TX-22/0-7043-R1. U.S. Department of Transportation, 2022.

26. 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.

27. 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

28. Dadashova, B., G. P. Griffin, S. Das, S. Turner, and B. Sherman. Estimation of Average Annual Daily Bicycle Counts using Crowdsourced Strava Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2674, No. 11, 2020, pp. 390–402.

29. Zhu, C., C. T. Brown, B. Dadashova, X. Ye, S. Sohrabi, and I. Potts. Investigation on the driver-victim pairs in pedestrian and bicyclist crashes by latent class clustering and random forest algorithm. *Accident Analysis & Prevention*, Vol. 182, 2023, 106964. https://doi.org/10.1016/j.aap.2023.106964

30. Stacy, C., K. Ramosm, D. Harvey, S. Rodríguez, J. Morales-Burnett, and S. Morris. Disrupting Structural Racism: Increasing Transportation Equity in South Dallas. Urban Institute, 2022. https://www.urban.org/sites/default/files/2022-12/Disrupting%20Structural%20Racism.pdf. Accessed May 15, 2023.

31. Elvik, R., and T. Bjørnskau. Safety-in-numbers: A systematic review and meta-analysis of evidence. *Safety Science*, Vol. 92, 2017, pp. 274–282.

32. Elvik, R., and R. Goel. Safety-in-numbers: An updated meta-analysis of estimates. *Accident Analysis & Prevention*, Vol. 129, 2019, pp. 136–147. https://doi.org/10.1016/j.aap.2019.05.019.

33. Jacobsen, P. L. Safety in numbers: More walkers and bicyclists, safer walking and bicycling. *Injury Prevention*, Vol. 21, No. 4, 2015, pp. 271–275.

34. Murphy, B., D. M. Levinson, and A. Owen, A. Evaluating the Safety in Numbers effect for pedestrians at urban intersections. *Accident Analysis & Prevention*, Vol. 106, 2017, pp. 181–190.

35. Tasic, I., R. Elvik, and S. Brewer, S. Exploring the safety in numbers effect for vulnerable road users on a macroscopic scale. *Accident Analysis & Prevention*, Vol. 109, 2017, pp.36–46.

36. Bhatia, R., and M. Wier, M. "Safety in Numbers" re-examined: Can we make valid or practical inferences from available evidence? *Accident Analysis & Prevention*, Vol. 43, No. 1, 2011, pp. 235–240.

37. Cottrill, C. D., and P. (Vonu) Thakuriah. Evaluating Pedestrian Crashes in Areas with High Low-Income or Minority Populations. Accident Analysis & Prevention, Vol. 42, No. 6, 2010, pp. 1718–1728. https://doi.org/10.1016/j.aap.2010.04.012.

38. Chen, C., H. Lin, and B. P. Y. Loo. Exploring the Impacts of Safety Culture on Immigrants' Vulnerability in Non-Motorized Crashes: A Cross-Sectional Study. Journal of Urban Health, Vol. 89, No. 1, 2012, pp. 138–152. https://doi.org/10.1007/s11524-011-9629-7.

39. Kravetz, D., and R. B. Noland. 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, Vol. 2320, No. 1, 2012, pp. 10–17. https://doi.org/10.3141/2320-02.

40. Yu, C.-Y., X. Zhu, and C. Lee. Income and Racial Disparity and the Role of the Built Environment in Pedestrian Injuries. Journal of Planning Education and Research, 2018, p. 0739456X18807759. https://doi.org/10.1177/0739456X18807759.

41. Dumbaugh, E., Y. Li, D. Saha, and W. Marshall. Why Do Lower-Income Areas Experience Worse Road Safety Outcomes? Examining the Role of the Built Environment in Orange County, Florida. Transportation Research Interdisciplinary Perspectives, Vol. 16, 2022, p. 100696. https://doi.org/10.1016/j.trip.2022.100696.

Appendix

Table A-1. Summary of Zoning System

Zoning System	Number of Articles using the zonal unit	Major Reasons
Census zoning system	• CBG (20) • CT (13) • SA2 (2) • LSOA (2) • MSOA (1) • Daeguyeog (1)	• Rich available sociodemographic information from census data • Relatively consistent and homogeneous across units • Can be a proxy for neighborhood
Traffic analysis zoning system	• TAZ (20) • TAD (1) • HAY (1)	• Compatible with census unit, thus having rich sociodemographic information • Spatially delineated for traffic analysis
Researcher-defined zoning system	• Police Patrol (1) • Ward (3) • ZIP code (2) • TPU (2)	• Limitation on data availability
Researcher-defined zoning system	• Researcher-defined (6)	• Created for special research purpose • Overcomes the shortage of other zoning systems

Table A-2. Measurement and Direction of Crash-related Factors

Category	Variable	Positive (Pedestrian)	Negative (Pedestrian)	Insignificant (Pedestrian)	Positive (Bicyclist)	Negative (Bicyclist)	Insignificant (Bicyclist)
Demographic Characteristics							
Age	Children (age<18, proportion)			4			
Age	Elderly (age>64, proportion)	1	5	2	2		2
Gender	Male (proportion)			1	1		
Race and ethnicity	White (proportion)				2		
Race and ethnicity	Black (proportion)	4	1	2	1		2
Race and ethnicity	Hispanic (proportion)	2			1		1
Race and ethnicity	Asian (proportion)	1	1				2
Education	College degree or higher (proportion)		1	2		1	1
Language	Speaking limited-English (proportion)			1	1		
Economic Characteristics							
Income	Median household income	1	3	3	1		1
Poverty	Below poverty line (proportion)	3			1		
Vehicle ownership	Households without vehicle (proportion)	7		1	4		1
House ownership	Own house (proportion)	1	1			1	
Roadway infrastructure							
Road	Road (density)	2		1	1		1
Road	Highway (density)						1
Road	Highway (proportion)	1					
Road	Highway (length)		1				1
Road	Arterial road (density)				1		
Road	Arterial road (length)	3					
Road	Arterial road (proportion)	3			2		
Road	Local road (density)					1	
Road	Local road (proportion)	2			2		

Road	Higher speed road (usually >55 mph, proportion)	1	3				2
Road	Lower speed road (usually <30 mph, proportion)	3		1	1	1	
Active transportation infrastructure	Bike lane (density)				2		
Active transportation infrastructure	Sidewalk (length)	3			2		1
Intersection	Intersection (density)			1	1		
Intersection	Intersection (number)	1		1	1		
Traffic signal	Traffic signal (number)	3					
Traffic signal	Traffic signal (density)	2			4		
Land use							
Residential area	Residential area (proportion)	2			1		2
Industrial area	Industrial area (proportion)	2					3
Commercial	Commercial (proportion)	3		1	1		2
Urban area	Urban area (proportion)	1			2		
Land-use mixture	Land-use mixed index	1		1	1		
Facilities							
Bus stop	Bus stop/transit (number)	2					
Bus stop	Bus stop/transit (density)	1			1		
School	School (number)	1					
School	School (density)	3					
Hotel	Hotel (density)	5			2		
Exposure							
Vehicular exposure	VKT/VMT	7		1	8		
Vehicular exposure	ADT/AADT	7		1	1		1
Vehicular exposure	Heavy vehicles millage in VMT (proportion)	1	1			2	
Vehicular exposure	Heavy vehicles (proportion)		2			3	
Pedestrian exposure	Walking trips	2					
Bicyclist exposure	Bicycle Miles Traveled				3		

Surrogate measurement	Population (density)	6	2	1	4	1	2
Surrogate measurement	Population (number)	7		1	5		1
Surrogate measurement	Employment (density)	4			2		1
Surrogate measurement	Employment (number)	4			3		
Surrogate measurement	School enrollment (density)	2	2		1	1	1
Surrogate measurement	Walk commuters (number)	3			3		
Surrogate measurement	Walk commuters (proportion)	3		1	1		
Surrogate measurement	Public transit commuters (number)	2			2		
Surrogate measurement	Public transit commuters (proportion)	1		1			1
Surrogate measurement	Bicycle commuters (number)	1		1	3		1
Surrogate measurement	Bicycle commuters (proportion)	1		1	2		

Table A-3. Summary of Selected Publications Investigating the Disparity in Roadway Safety

Research	Types of disparity	Methods to investigate the disparity	Major factors influencing disparity	Major results related to disparity
[37]	Racial and income disparity	Divided the census tracts into EJ (environmental justice) and non-EJ tracts by income and ethnicity, compared the environmental and behavioral factors among EJ and non-EJ tracts by *t* test and put a dummy variable for whether the tract is EJ area in pedestrian crash frequency model	Environmental factors including number of schools, crime rate, and behavioral factors like income, vehicle ownership, commercial area, proportion of children, etc.	• EJ areas with higher income and majority population are significantly less in pedestrian crashes. • Most of the environmental factors and behavioral factors are significantly different between EJ and non-EJ areas.
[38]	Racial disparity	Compared non-motorized crash model with and without variables related to immigrants (years since entering the U.S., country of origin, etc.)	Immigrants from different country of origin, Immigrants of different entering time	• Areas with more Latin American, Eastern European, or Asian immigrants tend to have more pedestrian and bicyclist crashes after controlling for members of minority groups born in the U.S.

Research	Types of disparity	Methods to investigate the disparity	Major factors influencing disparity	Major results related to disparity
[39]	Income disparity	Manually collected roadway infrastructure data from Google Street View and regressed roadway infrastructure data on the median income of the block groups	Speed limit, Sidewalk presence, Sidewalk buffer, Traffic signal, Intersection	• Areas with lower income and higher minority population have more pedestrian crashes. • Lower speed limit, more sidewalk, less sidewalk buffer, and more pedestrian signals are associated with poorer neighborhoods.
[13]	Gender disparity	Categorized traffic injuries by gender and transportation mode (vehicle driver, vehicle passenger, active mode user), and compared zonal crash prediction models (ZCPMs) for each category	Income level, Vehicle ownership	• Income level and vehicle ownership rates are positively associated with both male and female traffic injuries, but these effects are minor in female traffic injuries compared with male traffic injuries.
[5]	Racial disparity	Developed four bicyclist crash frequency models for White, Black, Hispanic, and Asian and compared the socioeconomic, land use, and transportation characteristics among four models	Arterial road	• Arterial roadways have larger positive effect in Black and Hispanic bicyclist crash models than White bicyclist crash models. • Bicycle infrastructure and low traffic street are not significantly associated with Black and Hispanic bicyclists.
[40]	Racial and income disparity	Used *t* test to compare the difference in crash-related factors between Census Block Groups (CBGs) with percentage of nonwhite higher/lower than median value and with percentage of population below poverty line higher/lower than median value; Compared two sets of ZCPMs for pedestrians with percentage of nonwhite higher/lower than median value and with percentage of population below poverty line higher/lower than median value for total crash, fatal crash, injurious crash, and no-injury crash, respectively.	Schools	• Percentage of school area is positively associated with both injurious and no-injury crashes but is only significant in areas with higher percentage of nonwhites and higher percentage of poverty.

Research	Types of disparity	Methods to investigate the disparity	Major factors influencing disparity	Major results related to disparity
[41]	Income disparity	Divided high-income and low-income block groups and compared influencing factors in pedestrian crash models for low-income, high-income, and all block groups	Arterial road, Buffered sidewalk, Black proportion	• Arterials are a risk factor for all block groups; their negative effects are greater in low-income neighborhoods, and they are insignificant in high-income neighborhoods. • Buffered sidewalks, usually presumed as safety enhancement measures, have no significant effect with pedestrian crashes in low-income areas and positive effect in high-income areas. • Percentage of Black people is not associated with pedestrian crashes in high-income areas but positively associated with pedestrian crashes in low-income areas.

Table A-4. Descriptive Information of the Variables

Variable	Description	Unit	Mean	SD	Min.	Max.
Non-motorized crashes	Number of pedestrian and bicyclist crashes	#	2.28	5.32	0	142.00
Poverty rate	Proportion of population with income below poverty line in the past 12 months	%	13.80	9.17	1.04	46.87
Hispanic and Black ratio	Proportion of population who is Hispanic American or African American	%	56.48	23.93	7.65	99.25
No high school diploma ratio	Proportion of population without high school diploma	%	17.72	12.26	.58	58.53
Public assistance ratio	Proportion of households with cash public assistance or Food Stamps in the past 12 months	%	12.17	8.18	.00	47.13
No health insurance ratio	Proportion of population with no health insurance coverage	%	17.74	9.75	1.15	58.02
Roadway length	Length of arterials, collectors, and local roads	Miles	12.16	8.52	0	59.22
Roadway without median	Length of arterials, collectors, and local roads without median	Miles	11.69	8.40	0	59.13
Intersection number	Number of intersections	#	29.81	30.91	0	231.00
Complex intersections	Intersections with four legs or above	#	9.29	15.81	0	170.00
No sidewalk ratio	Proportion of street view image that is identified without presence of sidewalk	%	43.21	31.64	0	100.00
No bike lane ratio	Proportion of the roadway without bike lane	%	92.33	22.40	0	100.00
No signal ratio	Proportion of intersection without traffic signal	%	77.58	36.39	0	100.00
Bicycle miles traveled	Sum of number of miles traveled by all bicycles	Miles	11.51	37.01	0	618.12
Pedestrian miles traveled	Sum of number of miles traveled by all pedestrians	Miles	23.51	121.53	0	2897.33
Vehicle miles traveled	Sum of number of miles traveled by all vehicles	Miles	96.16	163.76	0	1774.18

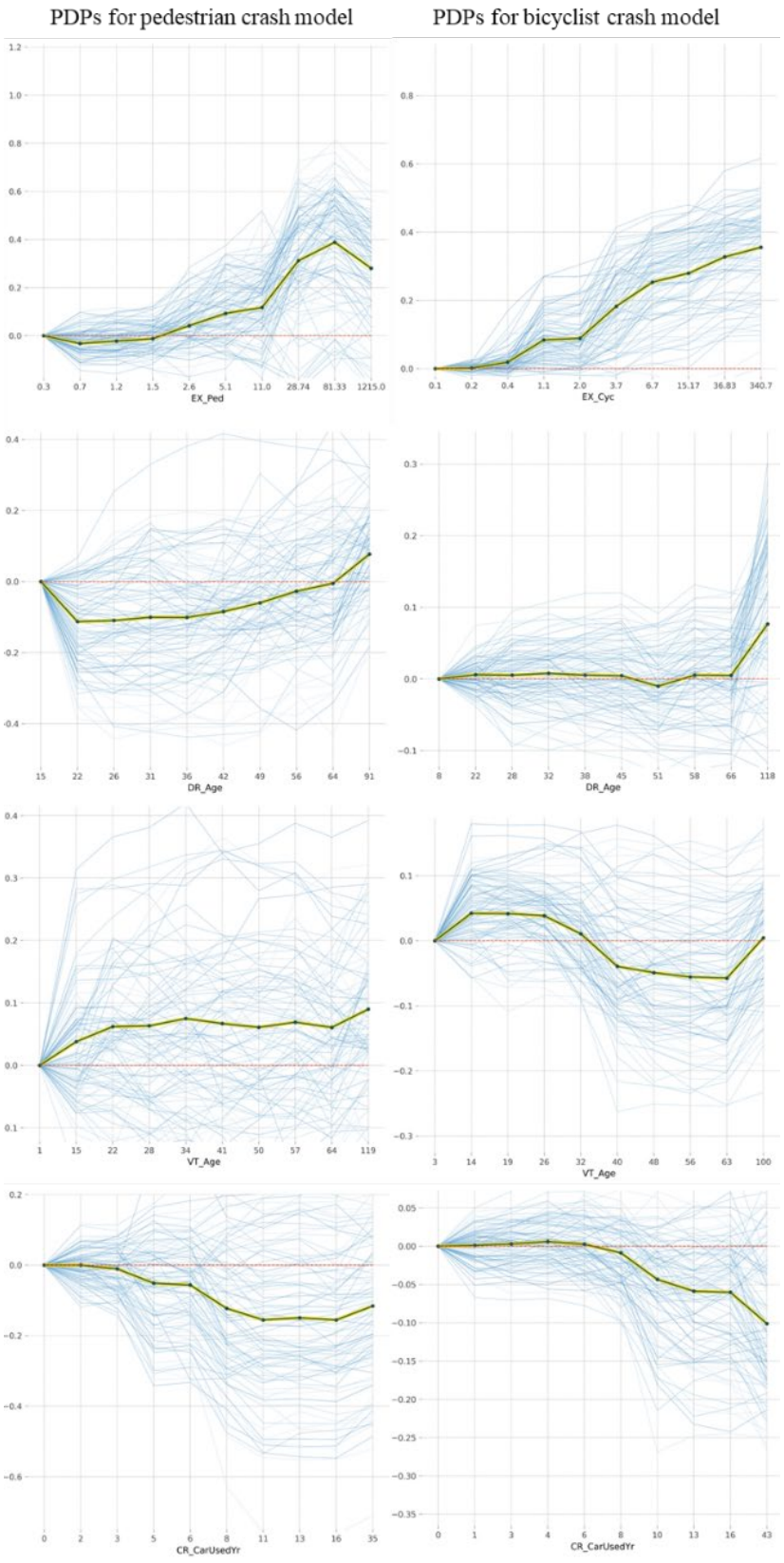
Table A-5. Descriptive Statistics of Driver and Victim Sociodemographic Factors

Variable	Pedestrian Crashes	Bicyclist Crashes
Categorical Variables	**Number (proportion)**	**Number (proportion)**
CR_TimeDay 1 = 0:00-6:00	254 (9.0%)	51 (4.5%)
CR_TimeDay 2 = 6:00-12:00	749 (26.5%)	303 (27.0%)
CR_TimeDay 3 = 12:00-18:00	839 (29.7%)	449 (40.0%)
CR_TimeDay 4 = 18:00-24:00	980 (34.7%)	320 (28.5%)
CR_Workday 1 = Monday to Friday	2199 (77.9%)	856 (76.2%)
CR_Workday 2 = Saturday to Sunday	623 (22.1%)	267 (23.8%)
CR_Season 1 = Spring	746 (26.4%)	243 (21.6%)
CR_Season 2 = Summer	679 (24.1%)	300 (26.7%)
CR_Season 3 = Autumn	606 (21.5%)	293 (26.1%)
CR_Season 4 = Winter	791 (28.0%)	287 (25.6%)
CR_Weather 1 = Clear	2093 (74.2%)	873 (77.7%)
CR_Weather 2 = Others	729 (25.8%)	250 (22.3%)
CR_Surface 1 = Dry	2495 (88.4%)	1039 (92.5%)
CR_Surface 2 = Others	327 (11.6%)	84 (7.5%)
CR_Construct 1 = At construction zone	51 (1.8%)	5 (0.4%)
CR_Construct 2 = Not at construction zone	2771 (98.2%)	1118 (99.6%)
CR_Intersec 1 = At intersection	1034 (36.6%)	661 (58.9%)
CR_Intersec 2 = Not at intersection	1788 (63.4%)	462 (41.1%)
DR_Income 1 = low income	454 (16.1%)	193 (17.2%)
DR_Income 2 = low to medium income	614 (21.8%)	229 (20.4%)
DR_Income 3= medium income	814 (28.8%)	280 (24.9%)
DR_Income 4 =medium to high income	368 (13.0%)	161 (14.3%)
DR_Income 5 = high income	572 (20.3%)	260 (23.2%)
DR_Ethnicity 1 = White	888 (31.5%)	384 (34.2%)
DR_Ethnicity 2 = Hispanic	896 (31.8%)	350 (31.2%)
DR_Ethnicity 3 = Black	800 (28.3%)	297 (26.4%)
DR_Ethnicity 4 = Asian	185 (6.6%)	70 (6.2%)
DR_Ethnicity 5 = Others	47 (1.7%)	20 (1.8%)
DR_Gender 1 = Male	1632 (57.8%)	626 (55.7%)
DR_Gender 2 = Female	1190 (42.2%)	497 (44.3%)
VT_Income 1 = low income	600 (21.3%)	211 (18.8%)
VT_Income 2 = low to medium income	762 (27.0%)	286 (25.5%)
VT_Income 3= medium income	559 (19.8%)	210 (18.7%)
VT_Income 4 =medium to high income	389 (13.8%)	156 (13.9%)
VT_Income 5 = high income	512 (18.1%)	260 (23.2%)
VT_Ethinicity 1 = White	935 (33.1%)	485 (43.2%)
VT_Ethinicity 2 = Hispanic	852 (30.2%)	274 (24.4%)
VT_Ethinicity 3 = Black	843 (29.9%)	299 (26.6%)

Variable	Pedestrian Crashes				Bicyclist Crashes			
VT_Ethinicity 4 = Asian	131 (4.6%)				52 (4.6%)			
VT_Ethinicity 5 = Others	54 (1.9%)				11 (1.0%)			
VT_Gender 1 = Male	1659 (58.8%)				924 (82.3%)			
VT_Gender 2 = Female	1163 (41.2%)				199 (17.7%)			
RD_FuncCls 1 = Collectors	782 (27.7%)				268 (23.9%)			
RD_FuncCls 2 = Local roads	2044 (72.4%)				855 (76.1%)			
RD_Urban 1 = Urban area	2818 (99.9%)				1117 (99.5%)			
RD_Urban 2 = Rural area	4 (0.1%)				6 (0.5%)			
RD_CurbL 1 = Left curb exists	2689 (95.3%)				1092 (97.2%)			
RD_CurbL 2 = No left curb	133 (4.7%)				31 (2.8%)			
RD_CurbR 1 = Right curb exists	2690 (95.3%)				1093 (97.3%)			
RD_CurbR 2 = No right curb	132 (4.7%)				30 (2.7%)			
RD_LnWth (feet) 1 = Less than 10	1527 (54.1%)				187 (16.7%)			
RD_LnWth (feet) 2 = 10 to 12	884 (31.3%)				534 (47.6%)			
RD_LnWth (feet) 3 = 12 to 14	139 (4.9%)				313 (27.0%)			
*RD_LnWth (feet)*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	/	/	/	/
EX_Cyc	/	/	/	/	14.1	0.1	340.7	30.8
AADT	9864.7	50.0	49968.0	9954.5	8601	69	49968	9565.4

Table A-6. Feature Importance of Random Forest Model for Pedestrian and Bicyclist Crashes

Variables for Pedestrian Crash Model	Feature Importance	Rank	Variables for Bicyclist Crash Model	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


PDPs for pedestrian crash model
PDPs for bicyclist crash model
EX_Ped
EX_Cyc
DR_Age
DR_Age
VT_Age
VT_Age
CR_CarUsedYr
CR_CarUsedYr

(continued)

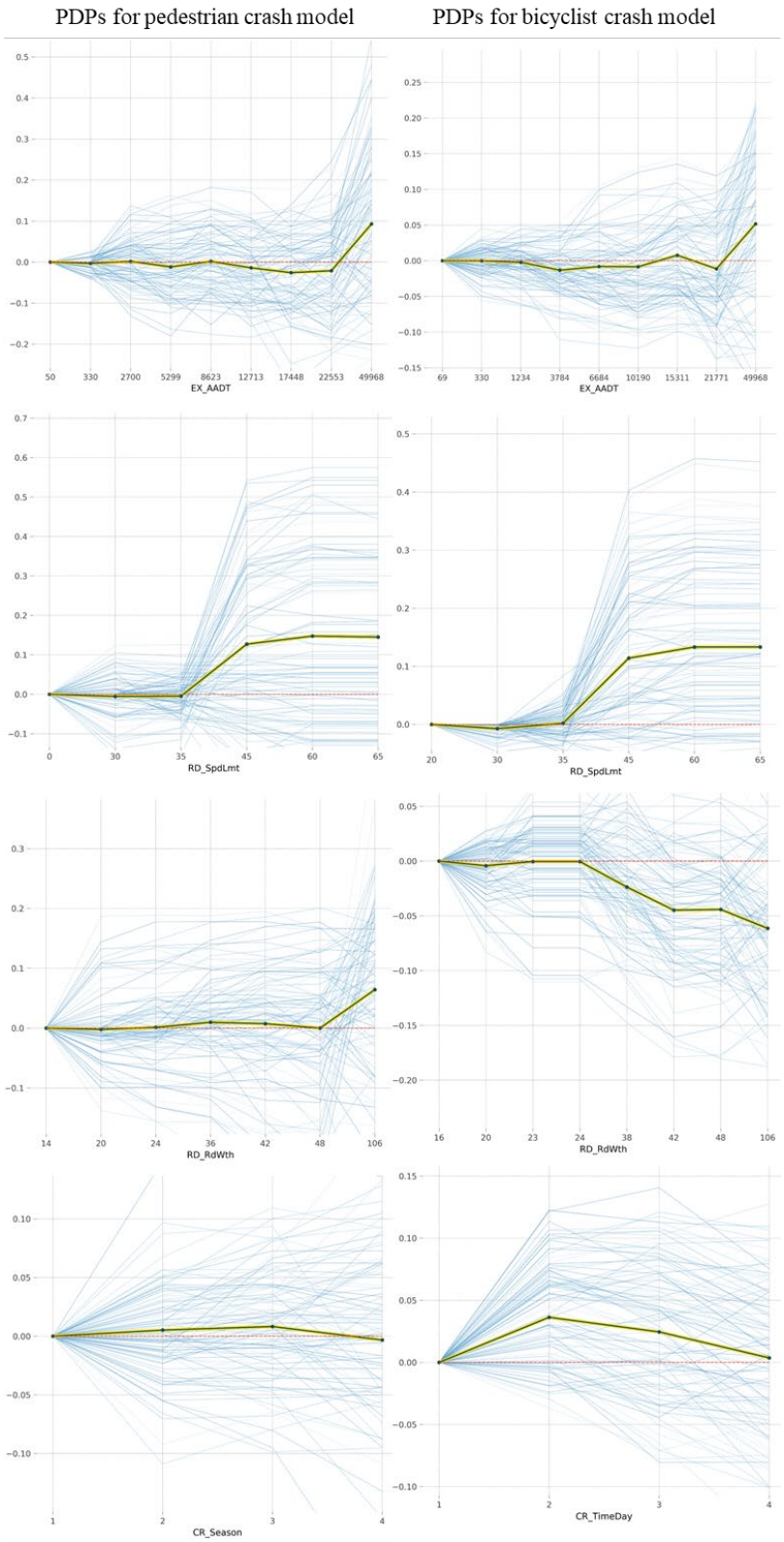


Figure A-1. Graphs. PDPs for variables in pedestrian and bicyclist crash model.