ANALYSIS OF ASSOCIATION BETWEEN DEMOGRAPHIC, SOCIOECONOMIC, AND BUILT ENVIRONMENT FACTORS AND PEDESTRIAN SAFETY USING TRADITIONAL AND AI APPROACHES

by

Jinli Liu, B.A., M.S.

A dissertation submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a Major in Geographic Information Science
December 2024

Committee Members:

F. Benjamin Zhan, Co-Chair

Subasish Das, Co-Chair

Tzee-Kiu Edwin Chow

Yihong Yuan

ACKNOWLEDGMENTS

As I write this acknowledgment, many faces come to mind, and my heart fills with warmth and gratitude. First and foremost, I would like to express my deepest gratitude to my advisors, Dr. F. Benjamin Zhan and Dr. Subasish Das. Dr. Zhan has been incredibly supportive, offering patient guidance and invaluable wisdom, his encouragement gives me great strength. Dr. Das has been an inspiring mentor, whose knowledge and passion have profoundly influenced me. I have learned so much from him, not only academically but also through his unwavering support. Dr. Das went above and beyond to provide full support when I was facing a crisis. His kindness helped me navigate through it. He is the kind of advisor who truly cares about your growth and offers unwavering support to ensure it.

I would also like to extend my heartfelt thanks to my committee members, Dr. Tzee-Kiu Edwin Chow and Dr. Yihong Yuan. Their continual support, encouragement, and insightful questions have been instrumental in shaping my research. I have admired their knowledge and am grateful for their guidance. Additionally, I want to acknowledge Dr. Antariksa Gian for his crucial help in providing me with valuable ideas to improve my dissertation.

A special thanks to my master's advisor, Dr. Yi Qi, whose continuous support has had a lasting impact on my life. I am also thankful to Dr. Feng Wang, Dr. Jennifer Jensen, Dr. Yongmei Lu for their valuable help throughout this process. I would especially like to thank Allison Glass-Smith for her warmth and support. Her comforting presence and love have made this journey much smoother, and her hugs were always a source of reassurance and strength.

To my dear friends, thank you for your laughter, and companionship, and for always being there to listen. I am truly blessed to have shared so many great moments with you. Your support has been invaluable in helping me get through this process.

Finally, I want to thank my family - Jueqiang Tao, Saiping Liu, Xianyu Li, Xiali Liu, Weiqiang Pan, and Yuran Pan. Their love and belief in me have been my anchor throughout this journey, and I miss them dearly.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

Abbreviation	Description
AI	Artificial Intelligence
CBGs	Census Block Groups
SHAP	SHapley Additive exPlanations
MGWR	Multiscale Geographically Weighted Regression
EPDO	Equivalent Property Damage Only
SLD	Smart Location Database
MAUP	Modifiable Areal Unit Problem
ETC	Equitable Transportation Community
OLS	Ordinary Least Squares

ABSTRACT

Pedestrian safety is a critical concern, particularly in urbanized areas where increasing population densities and heavy reliance on motorized transportation elevate risks for vulnerable road users who travel on foot. Creating safe, walkable environments is not only a public health priority but also vital for sustainable urban development. To better understand pedestrian crash risks, this dissertation explores the relationship between built-environment and socio-economic factors and their influence on crash risks at the Census Block Groups (CBGs) level in Austin, San Antonio, and Dallas. Given the spatial nature of CBGs, the spatial distribution of pedestrian crash risks within urban areas, such as Austin, is also examined to understand how built-environment and socio-economic factors contribute to this variability. Additionally, the study identifies distinct individual-scale pedestrian crash clustering patterns by severity using data from California.

The dissertation is organized around three major studies, each addressing a specific research question: What demographic, socioeconomic, and built environment factors are associated with pedestrian safety? What are the spatial variations in pedestrian crash risks at the census block group level in urban areas? Are individual-level pedestrian crashes spatially clustered? Specifically, the first study investigates the impact of socio-economic, built-environment, transit, and trip characteristics on pedestrian crashes in the Texas cities of Austin, San Antonio, and Dallas. It seeks to identify key factors influencing pedestrian crash risks across CBGs and evaluates the effectiveness of machine learning models, specifically SHapley Additive exPlanations (SHAP), in explaining how these factors affect Equivalent Property Damage Only (EPDO) rates. The findings reveal that auto-oriented network density is consistently associated with higher pedestrian crash risks, while pedestrian-oriented network density and sidewalk

coverage generally have negative associations. Transit frequency and socio-economic factors, such as the percentage of zero-car households, also significantly impact pedestrian safety. The study underscores the need for targeted interventions in disadvantaged CBGs with higher levels of zero-car households, more mixed land use, and denser transit networks, but lower percentages of high-wage workers and certain community services.

The second study examines the spatial variability of pedestrian crash risks within urban areas, focusing on Austin using Multiscale Geographically Weighted Regression (MGWR). The analysis reveals that higher percentages of two-plus-car households are associated with lower crash risks, possibly due to reduced pedestrian exposure. Conversely, areas with higher employment rate and household entropy, auto network density, and higher transit frequency exhibit greater crash risks, likely driven by increased interactions between pedestrians and vehicles in more densely populated areas, transit-oriented environments.

The third study identifies patterns in pedestrian crash severity using a clustering framework enhanced by explainable artificial intelligence (AI) techniques. This analysis, based on data from California, uncovers distinct patterns in crash severity by examining factors associated with fatal, injury, and non-injury crashes. It also explores how societal and demographic factors differ in their association with varying levels of crash severity, highlighting the disparities between underserved and more resilient communities. The most impactful factors in fatal crashes include pedestrian sobriety impairment, lighting conditions, and macro-scale traffic fatalities. In less severe crashes, broader societal and demographic influences are more prominent. The dissertation may provide insight for policymakers seeking to improve pedestrian traffic safety.

1. INTRODUCTION

1.1. Motivation

Pedestrian safety is a pressing public health concern. According to the World Health Organization (WHO), 92% of global road fatalities occur in low- and middle-income countries, which only possess 60% of the world's vehicles (World Health Organization 2024). Traffic fatalities predominantly affect vulnerable road users such as pedestrians, cyclists, and motorcyclists, who account for over half of all road traffic deaths. Most pedestrian fatalities occur on high-capacity urban roads, and in 60% of these cases in 2020 in the US, alcohol consumption by either the driver or the pedestrian was involved. Additionally, non-Hispanic American Indian and Alaska Native, as well as Black individuals, experienced the highest pedestrian fatality rates among all racial and ethnic groups in 2021. Addressing this issue, the Department of Transportation (DOT) is committed to maintaining a safe, efficient, and accessible transportation system nationwide (US Department of Transportation 2024). This commitment includes bridging gaps in transportation infrastructure and public services to ensure equitable access for all communities.

Many studies have explored various aspects of pedestrian safety, including pedestrian crash severity analysis (Haleem, Alluri, and Gan 2015; Pour-Rouholamin and Zhou 2016), pedestrian crash frequency analysis (Chen and Zhou 2016; Ding, Chen, and Jiao 2018), and pedestrian pattern analysis (Das et al. 2019; Kim and Yamashita 2007; Sasidharan, Wu, and Menendez 2015; Sun, Sun, and Shan 2019). There is also a growing body of research exploring the relationship between roadway safety and socioeconomic and demographic factors at various scales from a social justice perspective. Studies have focused on pedestrian and vehicle crashes and have employed a variety of analytical approaches, including non-spatial statistical models,

spatial regression analysis, and machine learning models. Several studies have suggested that sociodemographic factors such as population density, poverty, street connectivity, and land use patterns influence roadway safety. There is limited exploration of the broader regional or macro-scale factors that may influence roadway safety from a social justice perspective. Therefore, this study emphasizes the importance of considering both individual and macro factors in understanding the risks of pedestrian crashes and developing effective safety policies.

Meanwhile, in recent years, machine learning has shown promise in predicting crash injury severity, with an increasing focus on explainable AI to address the opacity of such models (Atakishiyev et al. 2021; Karim, Li, and Qin 2022). The incorporation of explainable AI ensures that these models are not just black boxes; instead, it makes their decision-making processes accessible and interpretable. For example, Kong et al. (2023) adopted an interpretable machine learning framework using SHAP to understand the factors associated with critical pedestrian-involved near-crash events. Chang et al. (2022) used XGBoost and SHAP to explore the effects of the built environment on fatal pedestrian accidents at a location-specific scale. Explainable AI has shown great potential in explaining the influence of variables on pedestrian safety.

Therefore, this dissertation presents a comprehensive examination of pedestrian crash risks at the CBGs and individual scale, focusing on understanding the impact of various risk factors, exploring spatial variability, and identifying distinct patterns in crash severity. By employing advanced analytical techniques, including machine learning, spatial modeling, and clustering enhanced by explainable AI, this research provided insights into the complex dynamics of pedestrian safety in urban settings.

1.2. Problem Statement, Purpose of the Study, and Research Questions

This dissertation is structured around three major studies, each focusing on a different aspect of pedestrian crash risks. The first study investigates the impact of various built-environment and socio-economic factors on pedestrian crashes in Austin, San Antonio, and Dallas. It seeks to understand which key factors significantly influence pedestrian crash risks across different Census Block Groups (CBGs). Furthermore, it explores the effectiveness of machine learning models, particularly through SHapley Additive exPlanations (SHAP), in capturing and explaining the influence of these risk factors on EPDO rates. The study also attempts to compare the influence of variables across different cities. This study attempts to answer the following question:

- **RQ 1**: What demographic, socioeconomic, and built environmental factors are associated with pedestrian safety?

The second study aims to deepen the understanding of spatial variability in pedestrian crash risks within urban areas, using Austin as a case study. It examines how built-environment and socioeconomic factors contribute to this spatial variability and identifies which areas within Austin are at the highest risk for pedestrian crashes. Additionally, the study compares the effectiveness of the Multiscale Geographically Weighted Regression (MGWR) and Ordinary Least Squares (OLS) model in capturing local and global variations in crash risks, providing insights into the complex interactions influencing pedestrian crashes. This study attempts to answer the following question:

- **RQ 2**: What are the spatial variations of pedestrian crash risks at the census block groups level in urban areas?

The third study focuses on identifying patterns in pedestrian crash severity using a clustering framework enhanced by explainable AI techniques. This study aims to uncover distinct patterns in pedestrian crash severity in California by analyzing factors associated with fatal, injury, and non-injury crashes. It also explores how broader societal and demographic factors differ in their association with varying scales of crash severity, highlighting the differences between crashes that occur in vulnerable communities versus those in more resilient areas. By identifying these patterns, the study seeks to provide insights that can inform the development of comprehensive safety measures, addressing both immediate situational risks and the underlying socio-economic landscape. This study attempts to answer the following question:

- **RQ 3**: Are individual-level pedestrian crashes clustered?

1.3. Dissertation Organization

Chapter 1 outlines the motivation and contributions of the dissertation, while Chapter 2 provides an up-to-date literature review. Chapter 3 investigates the impact of built-environment and socio-economic factors on pedestrian crashes in urbanized areas of Texas, aiming to identify key influences on pedestrian crashes and variations in pedestrian crash risk across CBGs. It also evaluates the effectiveness of machine learning models, particularly SHAP, in explaining the influence of these factors on pedestrian crash risk. Building on the data from Chapter 3, Chapter 4 focuses on spatial variability in pedestrian crash risks in Austin, exploring how built-environment and socio-economic factors contribute to this variability. It also assesses the effectiveness of MGWR in capturing local and global variations in crash risks. Chapter 5 shifts to individual-scale analysis, using crash data from California to identify patterns in pedestrian crash severity through a clustering framework enhanced by explainable AI techniques. This study uncovers patterns in fatal, injury, and non-injury crashes while examining how societal and

demographic factors influence crash severity in vulnerable versus resilient communities. These insights aim to inform the development of comprehensive safety measures that address both situational risks and broader socio-economic factors. Chapter 6 provides a conclusion of the dissertation.

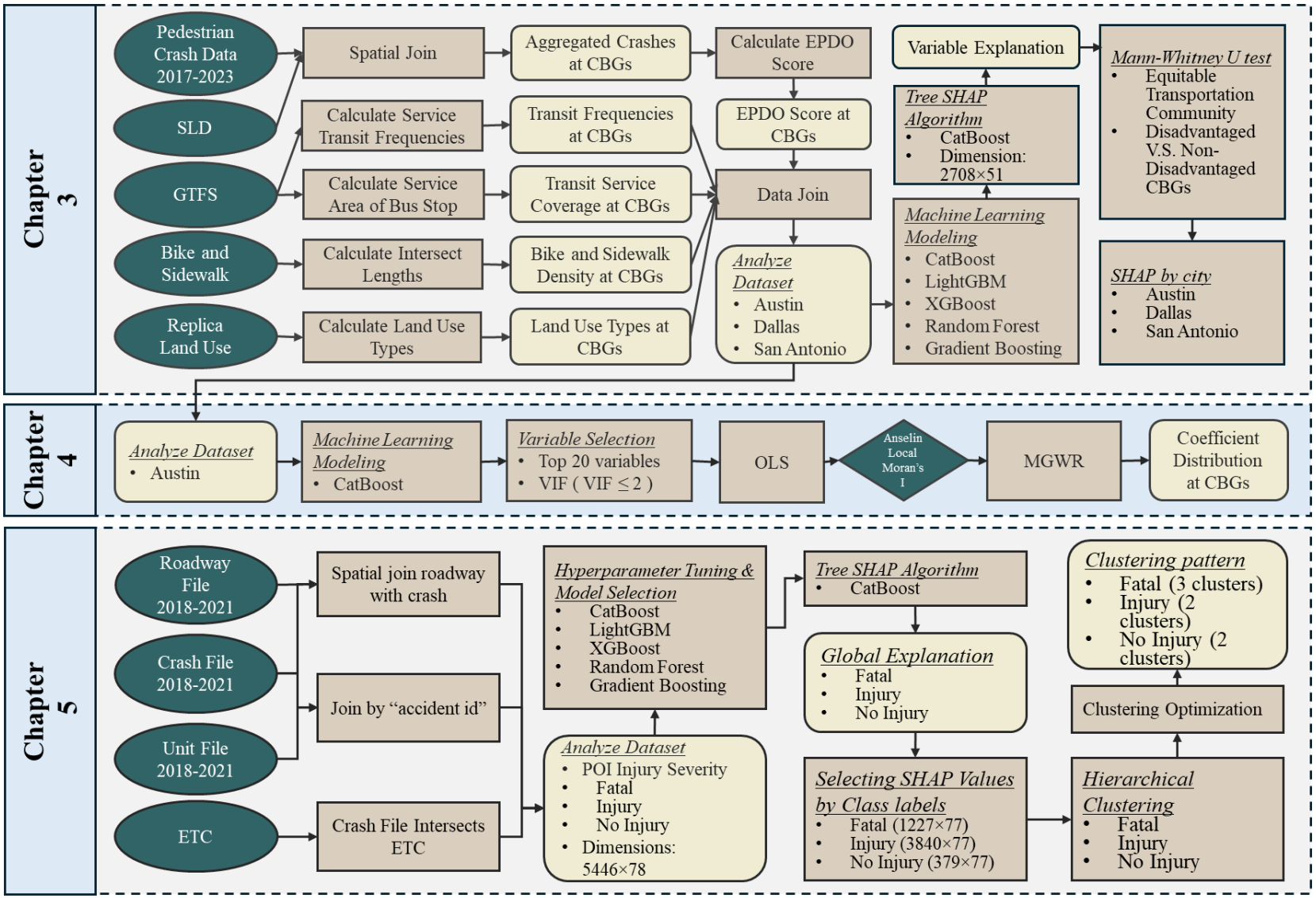


Figure 1. Dissertation organization

1.4. Dissertation Contributions

While a significant body of research has examined pedestrian crash risks, key gaps remain. One prominent gap involves the lack of detailed transit-related information in many safety studies. Transit services, including bus stops, transit service frequencies, and transit service areas, can significantly influence pedestrian safety. For instance, areas with heavy transit use often experience higher pedestrian volumes, which can increase exposure to crash risks if adequate safety measures are not in place. To our knowledge, no studies have utilized walking distance from bus stops to precisely define transit service areas, nor have they calculated transit frequencies based on different days of the week and specific time slots. Instead, previous studies typically rely on buffered distances around bus stops or the number of stops to represent transit accessibility. As a result, the interaction between these more refined transit characteristics and pedestrian crash risks remains underexplored.

Another critical gap pertains to pedestrian facilities, such as crosswalks, and sidewalks, which are often treated as secondary variables in crash risk studies. The presence or absence of such facilities directly impacts pedestrian safety, influencing both pedestrian behaviors and vehicle-pedestrian interactions. However, existing studies frequently overlook the quality and condition of pedestrian facilities, focusing instead on more general road network characteristics. This gap highlights the need for more comprehensive analyses that incorporate detailed pedestrian facility data into pedestrian crash models.

Moreover, studies have not adequately addressed the influence of land use patterns on pedestrian crash risks. The mix of residential, commercial, and recreational land uses affects pedestrian exposure and behavior, as well as the nature of vehicular traffic. Highly urbanized areas with mixed land uses may experience different pedestrian crash dynamics compared to

more suburban environments. Yet, few studies have systematically analyzed how land use interacts with transit infrastructure and pedestrian facilities to affect crash risks.

Furthermore, while many studies use traditional crash metrics, the use of the EPDO score as a more nuanced measure of crash severity remains underutilized. The EPDO score assigns different weights to crashes based on their severity, offering a clearer understanding of crash impacts beyond simple frequency counts. By incorporating EPDO into safety analyses, this study will provide a more accurate assessment of the severity of pedestrian crashes, highlighting areas that may not only have high crash rates but also more severe outcomes.

In addition, spatial modeling in crash studies has advanced significantly, but there is still room for improvement. While various spatial models have been employed MGWR approach remains underexplored. MGWR offers the flexibility to allow variables to vary at different spatial scales, providing a more nuanced understanding of the factors influencing crash occurrences. Through these refinements - focusing on detailed transit data, pedestrian facilities, EPDO scoring, and advanced spatial modeling - this study aims to fill critical gaps in the analysis of pedestrian crash risks, providing a more comprehensive and precise framework for improving pedestrian safety.

This study contributes by providing both individual-scale and CBG-scale analyses to enhance understanding of the differences between these scales. For the individual-scale analysis, the dissertation introduces a clustering method based on SHAP values, shifting the focus from grouping crashes by raw feature values to clustering based on the contribution of each feature to model predictions. This approach enables a deeper exploration of factors driving pedestrian crashes, revealing patterns that may not be evident from raw data. Since SHAP values are derived from tree-based models, they effectively capture non-linear interactions between

features, resulting in clusters that better reflect the complexity of real-world data. Additionally, this method reduces the influence of outliers, leading to more robust and stable clustering outcomes across diverse datasets.

Furthermore, this study integrates equitable transportation data to examine the relationship between transportation equity and pedestrian safety. By including this data, the analysis goes beyond traditional crash risk factors to investigate how transportation inequities disproportionately affect vulnerable populations, providing a more comprehensive view of pedestrian safety and highlighting disparities in crash risks that impact underserved communities.

This dissertation contributes to the field of traffic safety by presenting a data-driven, evidence-based approach to understanding and addressing pedestrian crash risks. By leveraging machine learning, spatial analysis, and clustering, it offers actionable insights for urban planners, policymakers, and public safety officials. These insights not only advance academic knowledge but also provide practical guidance for enhancing pedestrian safety and reducing crash-related harm in urban areas. Future research should continue to refine these models and extend their applicability to diverse urban contexts, incorporating real-time data and longitudinal analyses to further validate and enhance their predictive capabilities.

2. LITERATURE REVIEW

2.1. Key Variables in Pedestrian Safety Studies

Pedestrian safety studies have consistently identified a wide array of variables that influence crash risks in a place, which can be broadly classified into socio-economic, demographic, and built environment factors. These variables provide insights into the environmental and human characteristics that contribute to pedestrian safety issues in urbanized areas. Socio-economic conditions, such as income levels (Noland, Klein, and Tulach 2013; Dumbaugh et al. 2022; Roll and McNeil 2022a), employment status (Bernhardt and Kockelman 2021), and access to resources (Su, Sze, and Bai 2021; Sung et al. 2022), are correlated with pedestrian crash risks. Demographic variables, such as age (M. Rahman, Kockelman, and Perrine 2022; Park and Bae 2020), race (Sanders and Schneider 2022; Mwende et al. 2024), and population density (Almasi, Behnood, and Arvin 2021; Su, Sze, and Bai 2021), are also pivotal in determining pedestrian crash risks.

For build-environment factors, the density and configuration of road networks are essential factors in pedestrian safety studies (X. Wang et al. 2016). Su, Sze, and Bai (2021) found that areas with higher road density often correlate with more pedestrian crashes, as they create more points of interaction between vehicles and pedestrians. Public transportation plays a key role in influencing pedestrian crash risks. Areas with extensive public transit networks often experience higher levels of pedestrian activity around transit stops (Su, Sze, and Bai 2021), increasing the likelihood of crashes in these high-exposure zones. Turner et al. (2018) outlined methodologies for evaluating pedestrian and bicyclist safety risks using scalable data-driven approaches that integrate exposure metrics, crash data, and environmental factors.

The availability and condition of pedestrian facilities, such as sidewalks, crosswalks, and traffic-calming measures, are also crucial determinants of pedestrian safety. Dumbaugh et al. (2020) revealed a complex relationship between sidewalks and crash risks for all types of road users. Land use characteristics are also an important part of the analysis. Zafri and Khan (2022) found that pedestrian crash density had a positive relationship with employed person density, mixed land use density, and recreational land use density.

In summary, a variety of socio-economic, demographic, and built environment factors contribute to pedestrian crash risks in urbanized areas. These variables must be considered in combination to fully understand the spatial distribution and pedestrian crash risk. The aggregation of these variables, as discussed in the following sections, is crucial to modeling and predicting crash risks across different geographic scales.

2.2. Independent Variables in Pedestrian Safety Research

Aggregated pedestrian safety studies rely on various independent variables to model and predict crash risks effectively. These variables are typically derived from aggregated data, which involves summarizing detailed information into broader spatial units. Understanding the procedures for data aggregation, the choice of spatial units, and the metrics used is essential for accurate modeling and interpretation of pedestrian crash risks. Data aggregation involves compiling detailed pedestrian crash data and predictor variables into larger spatial units, such as block groups (X. Li et al. 2022; Liu, Das, and Khan 2024), census tracts (Tasic, Elvik, and Brewer 2017; Sanders and Schneider 2022; Roll and McNeil 2022b; Patwary et al. 2024), or traffic analysis zones (Osama and Sayed 2017; M. S. Rahman et al. 2019; Merlin et al. 2020; Zafri and Khan 2022). This process facilitates the analysis of spatial patterns and relationships that may not be apparent at more granular scales. However, aggregation can introduce

challenges, including the potential for ecological fallacies-wherein relationships observed at the aggregated scale do not necessarily hold at the individual scale the masking of localized variations in crash risks (Anselin 1995).

Metrics used to quantify crash risk in aggregated studies are fundamental for modeling and comparison purposes. Two primary metrics are commonly employed: crash frequencies (Haddad et al. 2023; Almasi and Behnood 2022) and EPDO scores (Oh, Washington, and Lee 2010; Ma et al. 2016). Crash frequency refers to the total number of pedestrian crashes within a specified time frame and spatial unit. This metric provides a straightforward measure of crash occurrence but does not account for the severity of each crash. EPDO is a composite metric that assigns different weights to crashes based on their severity, such as fatalities, injuries, and property damage. By incorporating severity, EPDO offers a more nuanced representation of crash impacts. EPDO allows researchers to prioritize areas with not only higher crash frequencies but also more severe outcomes, enabling more targeted interventions. Studies have demonstrated that using EPDO can enhance the predictive power of safety models by capturing the varying impacts of different crash types (Miao, Chen, and Zhang 2024).

2.3. Modeling Methods in Pedestrian Safety Studies

The study of pedestrian crash risks has traditionally employed models, such as ordinary least squares regression (Bernhardt and Kockelman 2021; Russo et al. 2018), logistic regression (Ammar et al. 2022; Nasri et al. 2022), and random parameter models (Mokhtarimousavi et al. 2020; Y. Li, Song, and Fan 2021). These models assume that relationships between dependent variables (e.g., pedestrian crash frequency) and independent variables (e.g., socio-economic or built environment factors) are constant across space. These models may oversimplify the relationship between predictors and outcomes, particularly in urban environments where

pedestrian crash risks vary widely across different neighborhoods and road conditions. To account for spatial heterogeneity Many studies have employed spatial models (Pljakić, Jovanović, and Matović 2022; Liu et al. 2024). Lizarazo and Valencia (2018) used a spatial lag model to analysis pedestrian crashes in Medellin. Zafri and Khan (2022) explored the association between the built environment and pedestrian crash occurrences using multiscale geographically weighted regression.

Previous studies adopted a wide range of modeling techniques, each tailored to specific aspects of traffic safety, with a common emphasis on spatial factors. Many studies utilize Bayesian spatial models to account for spatial correlation in crash data. For example, Siddiqui, Abdel-Aty, and Choi (2012) applied a Bayesian Poisson-lognormal model to assess pedestrian and bicycle crashes, demonstrating that models incorporating spatial effects perform better than those without spatial consideration. Similarly, Wang and Kockelman (2013) used a Poisson-lognormal conditional autoregressive (CAR) model to examine pedestrian crashes across neighborhoods, capturing regional heterogeneity. Cai et al. (2016) Cai et al. (2016) also explored dual-state count models, such as zero-inflated and hurdle models, integrating spatial spillover effects to improve the accuracy of crash frequency predictions. The importance of spatial heterogeneity was further emphasized in studies like Xu and Huang (2015), who compared geographically weighted Poisson regression (S-GWPR) and random parameter models, finding that the S-GWPR model provided superior predictive accuracy. Meanwhile, Guo et al. (2017) explored the role of road network patterns using Bayesian spatial models, while Stipancic et al. (2020) used a Full Bayes spatial Poisson-lognormal model to analyze pedestrian safety at intersections. Some recent studies, such as Rahman et al. (2019), have introduced machine learning techniques like decision tree regression (DTR) and ensemble methods (gradient

boosting and random forest) to enhance crash prediction models, showcasing the expanding toolkit of spatial traffic safety analysis. Several studies have examined Geographically Weighted models. For example, Xu and Huang (2015) used geographically weighted Poisson regression (S-GWPR) to explore spatial heterogeneity in crash data. Zafri and Khan (2022) applied multiscale geographically weighted regression (MGWR) to model spatial heterogeneity in pedestrian crash occurrences in Dhaka. Mathew, Pulugurtha, and Duvvuri (2022) used geographically weighted negative binomial regression (GWNBR) to model teen crashes. Li et al. (2022) examined the relationship between social vulnerability and crash risks in Texas using a spatial analysis approach using MGWR model. Ling et al. (2023) analyzed right-turn lane crashes using a geographically and temporally weighted negative binomial model.

Recently, machine learning models have gained prominence in pedestrian safety studies due to their ability to handle large datasets and capture complex, non-linear relationships between variables. Unlike traditional regression-based methods, machine learning models do not rely on pre-specified functional forms, making them particularly useful for uncovering hidden patterns in pedestrian crash data (Ali, Hussain, and Haque 2024). Common machine learning techniques applied to pedestrian safety include decision-tree models (Wu, Misra, and Bao 2023; Papathanasopoulou et al. 2021), neural network models (Rahim and Hassan 2021; Khan, Das, and Liu 2024; Al-Ani et al. 2024), other models (Das et al. 2021; Katanalp and Eren 2022). These models offer several advantages over traditional methods, including the ability to handle high-dimensional data and the flexibility to model non-linear interactions between predictors.

Despite their predictive power, machine learning models are often criticized for their lack of interpretability-a challenge in public policy and urban planning contexts, where decision-makers need to understand the relationships between variables. To address this issue, recent advances in

explainable machine learning techniques, such as SHapley Additive exPlanations (SHAP), have been incorporated into pedestrian safety studies. SHAP values provide a clear, interpretable framework for understanding how individual predictors contribute to model outputs (Chang et al. 2022). Given the flexibility of machine learning models in processing diverse data inputs-such as socio-economic, demographic, road infrastructure, and traffic volume data-these models are highly beneficial for pedestrian safety research. To further enhance the interpretability of these complex models, SHAP values are utilized, offering insights into the contribution of each variable to the model's predictions.

3. MODELING PEDESTRIAN CRASH RISKS IN URBAN AREAS AT THE BLOCK GROUP SCALE: A SHAP-BASED ANALYSIS OF BUILT ENVIRONMENT AND SOCIOECONOMIC FACTORS IN TEXAS

Pedestrian crashes account for a significant portion of annual traffic incidents in urban areas. Understanding the impact of various risk factors associated with pedestrian crashes is essential for developing effective and cost-efficient safety countermeasures. This study examines urban cities in Texas, utilizing a comprehensive data collection approach that integrates built-environment and socio-economic factors. Focusing on the EPDO rate at CBGs, this research addresses a critical knowledge gap by analyzing its distributional characteristics and constructing a machine-learning model based on Texas crash data from 2017 to 2023. SHAP is employed to evaluate and compare the influence of different risk factors on EPDO. Additionally, the study compares disadvantaged and non-disadvantaged CBGs using the Mann-Whitney U test. The SHAP results indicate that income is significantly associated with pedestrian crash risks, and transit frequency is a critical factor influencing crash severity. Significant differences are observed between disadvantaged and non-disadvantaged CBGs in terms of transportation, employment, and land use variables. Disadvantaged CBGs are characterized by higher scales of pedestrian crash risk, zero-car households, more diverse land use, and denser transit networks, but lower percentages of high-wage workers and reduced access to community services. These results provide insights into pedestrian crash risks and offer a robust framework for developing targeted safety intervention

3.1. Motivation

The analysis of pedestrian crashes in Texas from 2017 to 2023 reveals significant differences in crash distribution between urban and rural areas. Out of approximately 53,727 pedestrian-

involved crashes during this period, 41,988 crashes have geospatial data, and upon mapping, it is evident that a majority - around 39,408 crashes - occurred within urbanized areas, compared to just 2,279 crashes in rural regions. Figure 2 further highlights the distribution of pedestrian crashes by city. Houston, the largest city in Texas, unsurprisingly leads with the highest number of crashes, followed by other major metropolitan areas like Dallas-Fort Worth-Arlington, San Antonio, and Austin. The figure indicates a steep drop-off in crash numbers as we move down the list of cities, with most crashes concentrated in the largest urban centers. This spatial distribution of crashes emphasizes the importance of urban design, traffic management, and pedestrian infrastructure improvements in densely populated areas to mitigate the risks and improve overall pedestrian safety in Texas.

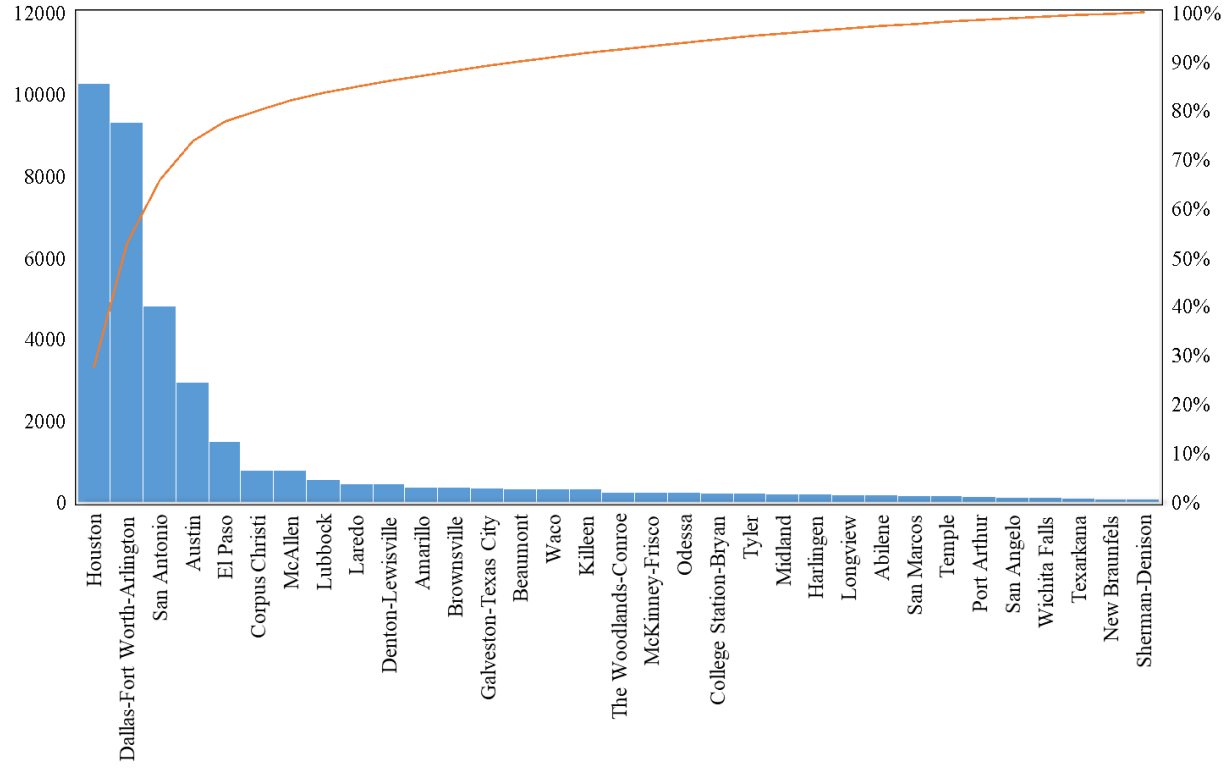


Figure 2. Number of pedestrian crashes by city from 2017 to 2023

While a significant body of research has examined pedestrian crash risks, key gaps remain. One prominent gap involves the lack of detailed transit-related information in many safety

studies. One prominent gap involves the lack of detailed transit-related information in many safety studies. Transit services, including bus stops, transit service frequencies, and transit service areas, can significantly influence pedestrian safety. For instance, areas with heavy transit use often experience higher pedestrian volumes, which can increase exposure to crash risks if adequate safety measures are not in place. To our knowledge, no studies have utilized walking distance from bus stops to precisely define transit service areas, nor have they calculated transit frequencies based on different days of the week and specific time slots. Instead, previous studies typically rely on buffered distances around bus stops or the number of stops to represent transit accessibility. As a result, the interaction between these more refined transit characteristics and pedestrian crash risks remains underexplored.

Another critical gap pertains to pedestrian facilities, such as crosswalks, and sidewalks, which are often treated as secondary variables in crash risk studies. The presence or absence of such facilities directly impacts pedestrian safety, influencing both pedestrian behaviors and vehicle-pedestrian interactions. However, existing studies frequently overlook the quality and condition of pedestrian facilities, focusing instead on more general road network characteristics. This gap highlights the need for more comprehensive analyses that incorporate detailed pedestrian facility data into pedestrian crash models.

Moreover, studies have not adequately addressed the influence of land use patterns on pedestrian crash risks. The mix of residential, commercial, and recreational land uses affects pedestrian exposure and behavior, as well as the nature of vehicular traffic. Highly urbanized areas with mixed land uses may experience different pedestrian crash dynamics compared to more suburban environments. Yet, few studies have systematically analyzed how land use interacts with transit infrastructure and pedestrian facilities to affect crash risks.

Furthermore, while many studies use traditional crash metrics, the use of the EPDO score as a more nuanced measure of crash severity remains underutilized. The EPDO score assigns different weights to crashes based on their severity, offering a clearer understanding of crash impacts beyond simple frequency counts (Oh, Washington, and Lee 2010). By incorporating EPDO into safety analyses, this study will provide a more accurate assessment of the severity of pedestrian crashes, highlighting areas that may not only have high crash rates but also more severe outcomes.

In addressing these gaps, this study makes several key contributions. First, it integrates transit-related information, pedestrian facility details, and land use patterns to create a more holistic model of pedestrian crash risks in urbanized cities. By leveraging SHAP values, a novel approach in this context, the study provides an interpretable analysis of how different factors contribute to crash risks. Second, this research expands on existing safety models by using the EPDO score, enabling a more nuanced understanding of crash severity patterns. Lastly, the study compares disadvantaged and non-disadvantaged CBGs to identify disparities in pedestrian crash risks, offering further insights for more equitable safety interventions. By doing so, this study aims to identify the demographic, socioeconomic, and built environment factors associated with pedestrian safety and provide insights to inform the development of more effective policies for creating safer, more walkable urban spaces.

3.2. Methodology

3.2.1. Study Design

This study aims to investigate pedestrian crash risks in urbanized areas at the CBG scale. The study flowchart (as shown in Figure 3) offers an overview of the evaluation process for pedestrian crash risks at CBGs, organized into two primary sections: data preparation and

modeling. During the data preparation phase, in addition to the available attributes from the Smart Location Database (SLD) (US EPA 2014), key variables such as transit frequency, transit coverage, land use metrics, and pedestrian facility characteristics were calculated. The subsequent modeling phase involved analyzing these variables to assess their impact on pedestrian crash risks. This structured approach ensures a thorough evaluation of the factors contributing to these crashes. The modeling process begins with hyperparameter tuning and model selection, focusing on tree-based machine-learning models, specifically CatBoost, LightGBM, XGBoost, Random Forest, and Gradient Boosting. This step ensures the selection of the best-performing model by optimizing its parameters to meet the study's specific requirements. CatBoost was selected as the optimal model based on its better predictive performance. The SHAP algorithm explains the contributions of different variables to the model's predictions. The final dataset for SHAP analysis comprises 2,708 records and 51 variables. The final stage of the modeling phase involves interpreting the SHAP analysis results to explain the impact of various factors on crash risks. Additionally, the distributional differences between disadvantaged and non-disadvantaged CBGs were analyzed using the Mann-Whitney U test. The classification of disadvantaged and non-disadvantaged CBGs follows the definition provided by the Equitable Transportation Community framework (US Department of Transportation 2024). The analysis focuses on pedestrian EPDO scores, providing insights into how different variables influence crash outcomes. This structured and detailed approach ensures a comprehensive evaluation of the factors contributing to crash risks on urban CBGs.

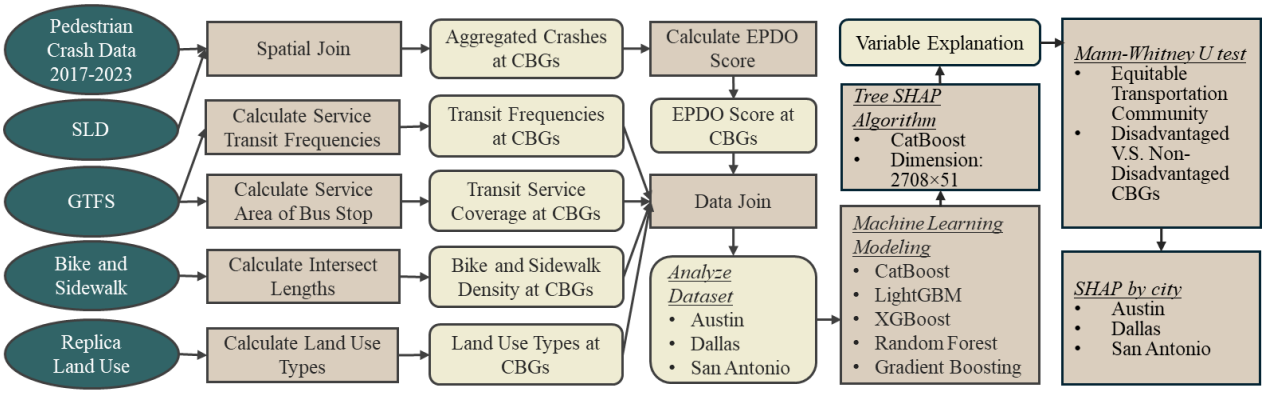


Figure 3. Study flowchart

3.2.2. Data Preparation

Pedestrian crash data from 2017 to 2023 were sourced from the Texas Crash Record Information System (CRIS), maintained by the Department of Transportation. During this period, a total of 53,727 pedestrian crashes were recorded. This study specifically focuses on urban areas, with the boundaries of urbanized areas derived from the local classifications provided by NCES Education Demographic and Geographic Estimates. Due to the coverage of data on transit, sidewalks, and bicycle infrastructure for entire cities, only CBGs with complete data across these variables were included in the final analysis. Additionally, pedestrian facility data for Houston was not accessible, resulting in Houston being excluded from the study. The geographic coverage of the study area is illustrated in Figure 4.

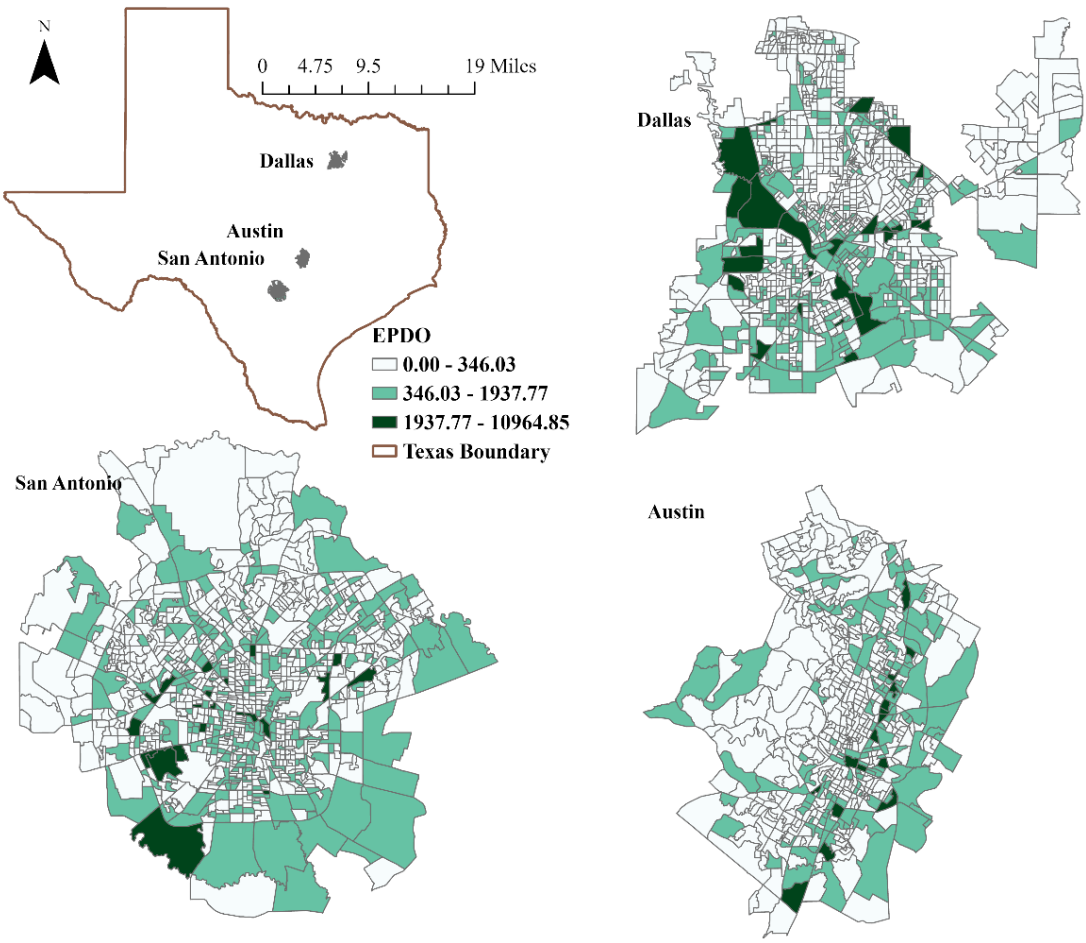


Figure 4. Study area

As presented in Table 1, the final dataset includes 2,708 CBGs, distributed among Austin (604), Dallas (1,047), and San Antonio (1,057), where a total of 11,152 pedestrian crashes occurred. The data are categorized by five scales of crash severity using the KABCO injury scale: K for fatal, A for incapacitating, B for non-incapacitating, C for possible, and O for no injury or property damage-only (PDO) crashes. The severity distribution is as follows: K = 1,212; A = 2,269; B = 719; C = 4,368; O = 2,544; and Unknown = 40. In this study, crashes with unknown severity are treated as O crashes. It can be observed that CBGs in the Dallas region experience more incapacitating crashes compared to the San Antonio region, while the Austin region has more non-incapacitating pedestrian crashes, despite having fewer CBGs than

the Dallas and San Antonio regions. The descriptive statistics of the variables are presented in Table 2.

Table 1. Crash statistics at CBGs by city

Severity	Austin (604 CBGs)	Dallas (1,047 CBGs)	San Antonio (1,057 CBGs)	Total
K	274	489	449	1212
A	484	1082	703	2269
B	1036	1431	1901	4368
C	601	948	995	2544
O	107	214	398	719
Unknown	7	16	17	40

3.2.2.1. Pedestrian Crash EPDO Score at CBGs

Pedestrian crash data were assigned to CBGs using the intersect function of ArcGIS. The study collected seven years (2017-2023) of CRIS data. Given that pedestrian crashes frequently occur on roads, which often delineate the boundary lines of CBGs as depicted in Figure 5, a strategic approach was employed involving the creation of a polygon buffer. Specifically, a 10-meter buffer zone was established surrounding each CBG. The 10-meter buffer was determined based on practical considerations. Since crash points do not always align perfectly with road centerlines, the 10-meter buffer ensures that crashes occurring near the centerline are accurately captured, while preventing crashes on nearby roads from being incorrectly assigned to adjacent CBGs. This approach guarantees that crashes are counted in both adjacent CBGs when the CBG boundary coincides with the road's centerline, providing a more comprehensive and accurate representation of crash data.

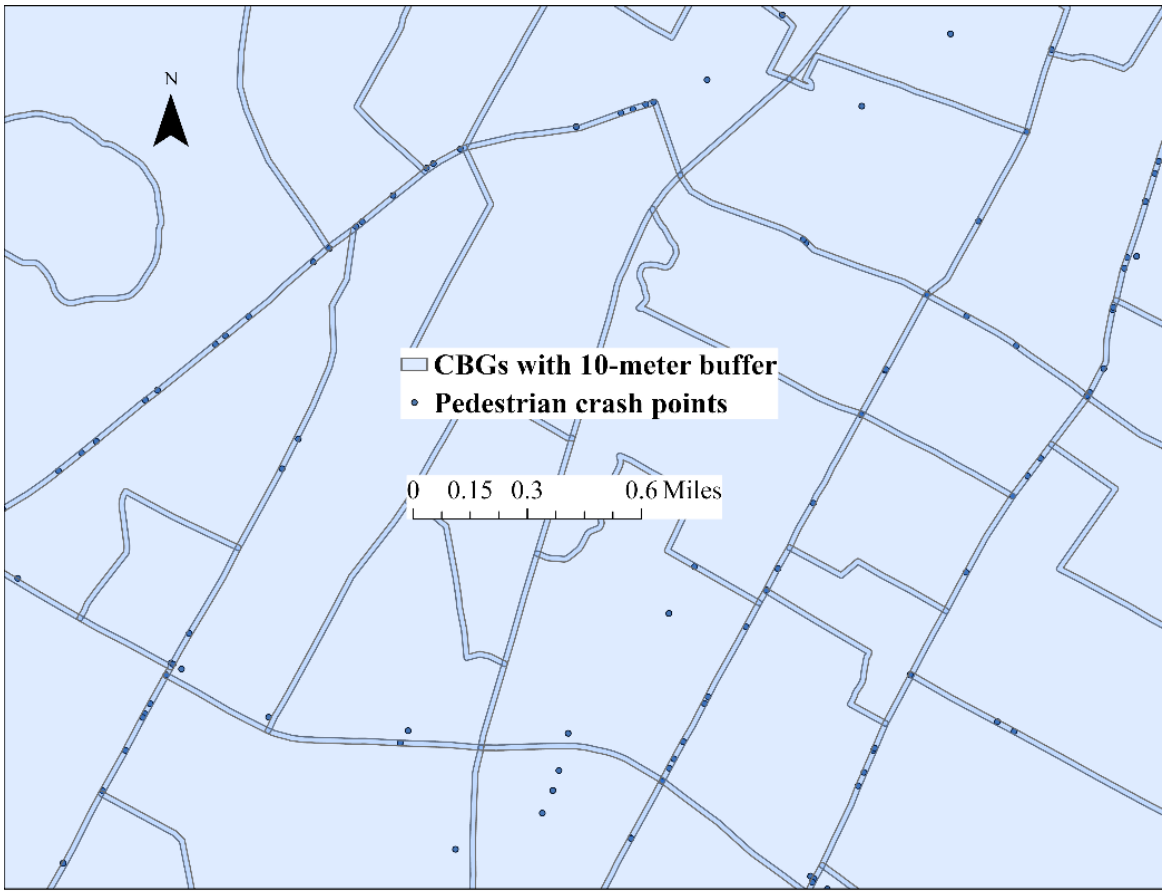


Figure 5. Pedestrian crash data points

The EPDO risk score of CBGs was calculated based on Equation 1 from the 2010 AASHTO Highway Safety Manual, Chapter 7 (2010 AASHTO 2010).

$$EPDO = \frac{4{,}008{,}900}{7{,}400} * \mathrm{C}_K + \frac{82{,}600}{7{,}400} * \mathrm{C}_{ABC} + \mathrm{C}_{PDO} \quad (1)$$

Where, C_K is the number of fatal crashes, C_{ABC} is the number of injury crashes, C_{PDO} is the number of PDO crashes. The distribution of EPDO score of CBGs is presented in Figure 4.

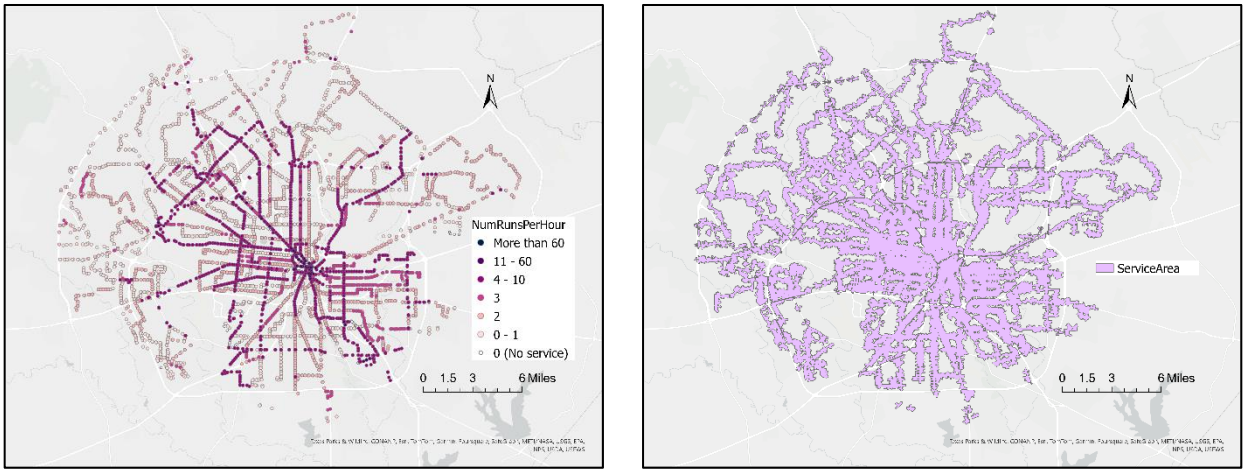
3.2.2.2. Transit Accessibility

Previous studies have demonstrated a strong link between public transit and pedestrian crash risks, but many have relied on simplified metrics such as the number of bus stops or buffer zones around stops to estimate service areas. This approach misses the detailed effects of transit

services. To fill this gap, this study examines the relationship between public transit and pedestrian crash risk in greater depth, using more comprehensive data such as transit frequencies and service areas measured by walking distance.

Transit-related data for major Texas cities (San Antonio, Austin, and Dallas) was obtained from the GTFS (General Transit Feed Specification) files available on the Open Mobility website (MobilityData 2024). GTFS is a standardized format for public transportation schedules and geographic information, enabling developers to build applications like trip planners and maps. For this study, data from May to August 2022 was extracted, and weekday and weekend transit service frequencies were calculated. Service areas were mapped based on a 0.4 km walking distance from transit stops, providing a more precise understanding of transit accessibility. 0.4 km walking distance is commonly used to identify bus stop service area (El-Geneidy et al. 2014; Zuo, Wei, and Chen 2020). Both analyses were conducted using ArcGIS Pro.

To streamline the analysis, specific time intervals were used to represent typical transit frequencies. For weekdays, data from Monday between 5 am and 6 pm was used for daytime service, and from 6 pm to 5 am for nighttime service. For weekends, Saturday's data from the same time periods was used. These intervals capture key patterns of transit availability during different times of day and week, offering deeper insights into how public transit patterns influence pedestrian safety. An example of weekday transit frequency and service area mapping for San Antonio is shown in Figure 6.

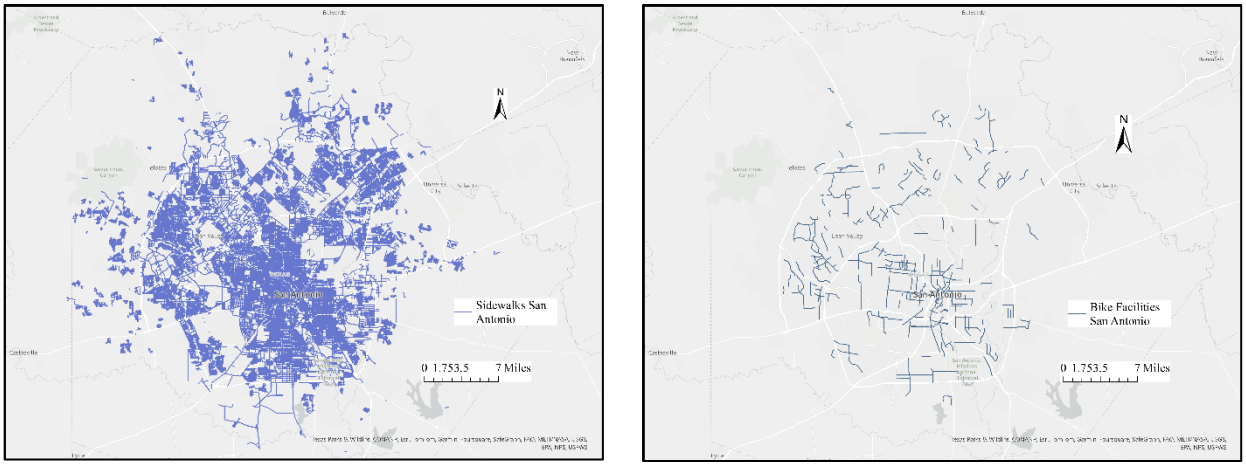


a) Weekday Transit Frequency b) Transit Service Area Coverage

Figure 6. Example of transit accessibility in San Antonio

3.2.2.3. Pedestrian Facilities

Sidewalk and bicycle shapefiles were obtained from the open data portals of Austin, Dallas, and San Antonio. The sidewalk lengths were calculated using the Intersect tool. Four different length indices were derived: planned sidewalks, existing sidewalks, acceptable sidewalks, and deficient sidewalks. Bike lanes are typically assigned only to the centerline of roads, and since CBGs often align with these centerlines, a 10-meter buffer was created around the CBGs to intersect with bike lanes, a 10-meter buffer distance is based on experiments. This buffer accounts for any potential discrepancies in alignment, ensuring an accurate representation of bike lane coverage within the surrounding areas. The length of existing bike lanes within these buffered areas was calculated to assess the current bicycle infrastructure. Figure 7 presents pedestrian facility coverage in San Antonio.



a) Sidewalk Coverage b) Bike Lane Coverage

Figure 7. Example of pedestrian-oriented facility coverage in San Antonio. a) sidewalk coverage, b) bike lane coverage

Table 2. Variable descriptives

Variables	Description	Source	Mean	STD	Min	50%	Max
Dependent Variables							
EPDO Score	Pedestrian crash EPDO Score	CRIS	280.587	646.622	0	22.32	10964.85
Independent Variables							
Demographics and socio-economic							
Worker Pop Density	Density of workers (people/acre)	SLD	4.553	3.967	0.003	3.779	56.968
Resident Pop density	Density of residents (people/acre)	SLD	2.380	1.833	0.111	2.216	88.000
Pct of WrkAge Pop	Percentage of the working-age population (18-64)	SLD	0.628	0.110	0.000	0.619	1.000
Pct of Zero-car	Percentage of households that do not own a car	SLD	0.078	0.097	0.000	0.043	0.702
Pct of Two-plus-car	Percentage of households that own two or more cars	SLD	0.550	0.201	0.000	0.562	1.000
Pct Lowwage (home location)	Percentage of low-wage workers based on their home location	SLD	0.105	0.078	0.000	0.098	2.313
Pct Highwage (home location)	Percentage of high-wage workers based on their home location	SLD	0.212	0.144	0.000	0.189	2.971
Pct Lowwage (workplace)	Percentage of low-wage workers based on their workplace location	SLD	0.282	0.161	0.000	0.258	1.000
Pct Highwage (workplace)	Percentage of high-wage workers based on their workplace location	SLD	0.332	0.191	0.000	0.314	1.000
5-tier Employment Entropy	A measure of employment diversity across five employment categories	SLD	0.634	0.268	0.000	0.707	1.000
Employment and Household Entropy	A combined measure of diversity and balance in employment types and household characteristics	SLD	0.447	0.230	0.000	0.474	0.917
Household Workers per Job Equilibrium	A ratio indicating the balance between the number of working household members and the number of jobs available in a given area	SLD	0.268	0.313	0.000	0.078	0.997
Trip Characteristics							
Trip productions and trip attractions equilibrium	The balance between the number of trip productions and trip attractions	SLD	0.378	0.297	0.000	0.406	1.000
Employment within Transit Stop (0.5 mile)	Number of jobs located within a half-mile radius of a transit stop	SLD	0.054	0.192	0.000	0.000	1.000
Jobs within 45 minutes auto travel time	Number of jobs accessible within a 45-minute drive	SLD	236234.299	124864.564	8931.000	188809.500	551730.000
Working age population within 45 minutes auto travel	Number of working-age individuals accessible within a 45-minute drive	SLD	177341.109	77580.104	10694.000	142092.000	360942.000
Non Commute VMT	Vehicle miles traveled for non-commuting purposes per worker	SLD	5.004	1.554	1.319	4.863	13.721
Commute VMT	Vehicle miles traveled for commuting purposes per worker	SLD	19.041	5.712	4.564	18.070	70.986
Work related VMT	Vehicle miles traveled for work-related purposes per worker	SLD	24.045	6.660	6.982	23.229	82.636
Network Characteristics							
Auto Ntwrk Density	Density of the automobile network per square mile (41 mph or more)	SLD	1.911	3.372	0.000	0.000	31.457
Multi Ntwrk Density	Density of the automobile network per square mile (21 to 54 mph)	SLD	3.056	2.974	0.000	2.348	20.122
Ped Ntwrk Density	Density of the automobile network per square mile (20 mph or less)	SLD	15.694	6.738	0.000	15.828	53.884
Street Intersection Density	The number of street intersections per square mile	SLD	97.030	67.932	0.000	82.025	864.31
Public Transit Characteristics							
Wrk Day Runs	Frequency of transit runs per hour during workdays	GTFS	15.081	26.617	0.000	8.462	597.538
Wrk Night Runs	Frequency of transit runs per hour during work nights	GTFS	6.404	10.845	0.000	3.545	219.000
Wknd Day Runs	Frequency of transit runs per hour during weekend days	GTFS	11.055	17.408	0.000	6.462	290.615
Wknd Night Runs	Frequency of transit runs per hour during weekend nights	GTFS	5.265	8.493	0.000	2.909	152.091
Pct Transit Service Area	Percentage of the area covered by transit service	GTFS	0.507	0.353	0.000	0.527	1.782

Pedestrian Facility Characteristics							
Planned Sidewalk Density	Density of planned sidewalks (after April 2023)	City open data portal	1.283	1.477	0.000	0.752	8.592
Existing Sidewalk Density	Density of existing sidewalks (before April 2023)	City open data portal	3.360	2.478	0.000	3.012	14.488
Bike Lane Density	Density of bike lanes	City open data portal	0.598	0.923	0.000	0.122	10.735
Land Use Characteristics							
Pct Single Residential	Percentage of single-family residential use	Replica	0.488	0.232	0.000	0.497	0.999
Pct Multi Residential	Percentage of multi-family residential use	Replica	0.069	0.071	0.000	0.050	0.714
Pct Commercial Retail	Percentage of commercial retail purposes	Replica	0.134	0.124	0.000	0.093	0.679
Pct Commercial Office	Percentage of commercial office purposes	Replica	0.049	0.053	0.000	0.034	0.483
Pct Commercial Non-Retail Attraction	Percentage of commercial purposes that are not retail	Replica	0.020	0.034	0.000	0.009	0.544
Pct Mixed Use Residential	Percentage of mixed residential and other uses	Replica	0.014	0.027	0.000	0.004	0.381
Pct Mixed Use Commercial	Percentage of mixed commercial uses	Replica	0.045	0.062	0.000	0.026	0.735
Pct Mixed Use Industrial	Percentage of mixed industrial purposes	Replica	0.009	0.021	0.000	0.001	0.277
Pct Mixed Use Other	Percentage of other mixed-use purposes not classified under residential, commercial, or industrial.	Replica	0.029	0.054	0.000	0.014	0.804
Pct Industrial	Percentage of industrial purposes	Replica	0.018	0.039	0.000	0.001	0.480
Pct Civic Healthcare	Percentage of healthcare facilities	Replica	0.001	0.010	0.000	0.000	0.332
Pct Civic Education	Percentage of educational facilities	Replica	0.039	0.067	0.000	0.023	0.792
Pct Civic Other	Percentage of other civic purposes	Replica	0.005	0.010	0.000	0.000	0.169
Pct Transportation Utilities	Percentage of transportation infrastructure and utilities,	Replica	0.009	0.050	0.000	0.000	0.815
Pct Open Space	Percentage of open space	Replica	0.028	0.052	0.000	0.008	0.504
Pct Agriculture	Percentage of agricultural purposes	Replica	0.037	0.089	0.000	0.000	0.644
Land Use Entropy	A measure of the diversity and distribution of different land uses	Replica	1.460	0.435	0.000	1.498	2.456

3.2.3. Machine Learning Models

Machine learning's prediction function has a variety of customized models, including random forest algorithms (Random Forest), extreme random tree algorithms (Extremely Randomized Trees), two boosting tree algorithms (Catboost and Light GBM), and ensemble learning approaches. The random forest and extremely randomized tree algorithms select the best variables using Entropy or Gini coefficients (Breiman 2001). Entropy, a concept adopted from information theory, measures the uncertainty inherent in a collection of random variables. Similarly, the Gini coefficient is a measure of data impurity. Similar to information entropy, it is a useful tool for feature selection, assisting in the identification of the most discriminative variables (Geurts, Ernst, and Wehenkel 2006; Kullback and Leibler 1951).

The boosting tree technique is a machine learning method that uses decision trees and is optimized with the forward stepwise algorithm and linear tree combination (James et al. 2023). The gradient boosting decision tree strategy is often used to optimize the complexity of boosting tree models (T. Chen and Guestrin 2016). Boosting is a mechanism for turning weak learners into strong learners. Tree based boosting methods include Catboost and Light GBM, which are based on gradient boosting decision trees and can handle massive amounts of data (Prokhorenkova et al. 2018). Furthermore, these algorithms are well-known for their resistance to overfitting and ability to handle categorical features directly, making them a popular choice for a variety of machine learning tasks (Prokhorenkova et al. 2018; Ke et al. 2017; Erickson et al. 2020).

3.2.4. Explainable AI

Explainable AI methods analyze feature attributions to determine each feature's contribution to a machine learning model's predictions. SHAP allows users to explain the predictions of sophisticated machine learning models using input variables (Linardatos, Papastefanopoulos, and

Kotsiantis 2020). In this experiment, SHAP values provide a measure of each feature's importance in the model. The goal of this method is to clarify pixel anomalies by evaluating each feature's contribution to severity. To explain a prediction for a single case x using tree-based model $f(x)$. $f(x)$ can be expressed as a sum of the importance of individual pixel features ϕ_i. The model's prediction can be described as follows:

$$f(x) = g(x') = \phi_0 + \phi_i x_i' \tag{2}$$

Here, ϕ_0 is what the prediction would be if this case had no features (a baseline), and x' is a vector indicating which features are included in the severity prediction. The function g represents an additive explanation model, where the overall prediction is broken down by the contribution of each feature.

SHAP is a method for such additive feature attribution proposed by Lundberg (Lundberg and Lee 2017), inspired by the Shapley value from cooperative game theory (Shapley 1997), which measures the contribution each player brings to the game. The Shapley value $\phi_i(v)$ is defined as:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \tag{3}$$

This equation considers all subsets S of features that do not include feature i, where n is the total number of features. It computes the average marginal contribution of feature i, across all possible combinations of features. In this context, assumed of this like a game where each feature value is a player, and the payout is the prediction made by the model f. For an input instance x and a feature set S, the Shapley value $\phi_i(f, x)$ quantifies the contribution of feature i to the prediction $f(x)$ and is calculated as:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(n - |z'| - 1)!}{n!} \left(f(z^{'}) - f(z^{'} \setminus \{i\})\right) \tag{4}$$

where z' is a binary vector representing a subset of features. This formula averages the impact of including feature i in the model across all subsets of features. By utilizing this method, this research seeks to explain severity data by computing the contribution of each severity label prediction. In this study, the SHAP value was calculated using Lundberg's (2019) explainer approach, which provides a speedy and accurate feature attribution method by employing an ensemble-based decision tree structure. The best machine learning model used in this approach is compatible with SHAP Explainer because it makes predictions using an ensemble-based decision tree.

3.3. Results and Analysis

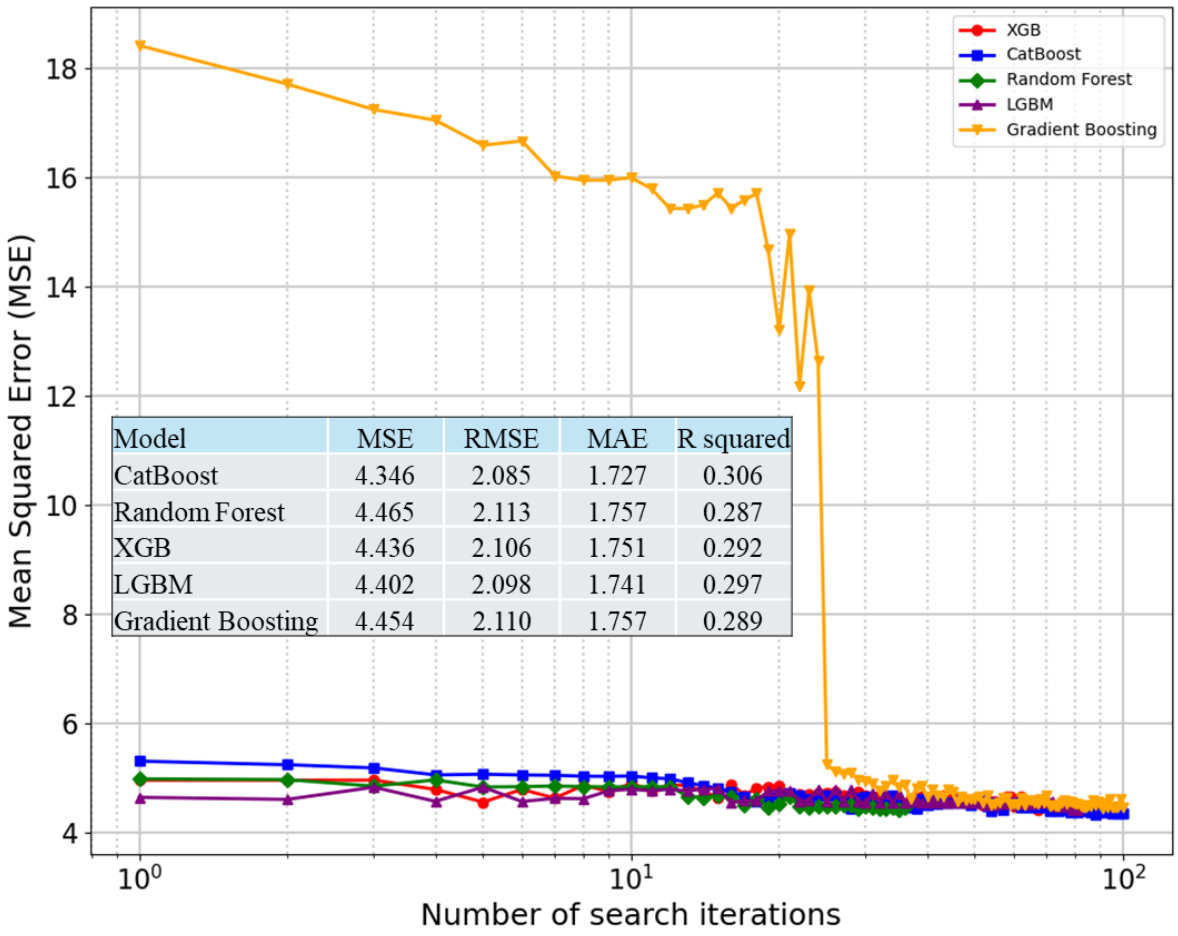
3.3.1. Model Optimization and Selection

The first step of the pedestrian crash severity analysis involved fine-tuning the parameters and selecting the machine learning models. These models include CatBoost, LightGBM, XGBoost, Random Forest, and Gradient Boosting. For this process, the dataset was split into two subsets: 80% for training and the remaining 20% for testing. Furthermore, 20% of the training data was set aside for model validation. During the training phase, the models were validated at every three iterations. The aim of the optimization was minimizing the mean squared error (MSE), which signifies the predictive capability of the model. Each model requires different parameters to be optimized. Grid search is a widely adopted approach for hyperparameter optimization. To enhance computational efficiency, the hyperparameter tuning was confined to parameters that significantly affect model performance. Table 3 presents the hyperparameter tuning space for each model. Parameters such as learning rate, number of estimators, criterion, batch size, max depth, and others were fine-tuned. It is important to note that parameter tuning is not guaranteed for global optimal but local optimal.

Table 3. Hyperparameter tuning space

Models	Hyperparameters
CatBoost	Iterations: 100, 200, 300; Learning rate: 0.01, 0.05, 0.1; Depth: 4, 6, 10; L2_leaf_reg: 1, 3, 5, 7
LightGBM	Number of leaves: 31, 63, 127; Learning rate: 0.01, 0.05, 0.1; Number of estimators: 100, 200, 300; Min split gain: 0.0, 0.1, 0.2
XGBoost	Number of estimators: 100, 200, 300; Learning rate: 0.01, 0.05, 0.1; Max depth: 3, 5, 7; Min child weight: 1, 3, 5
Random Forest	Number of estimators: 100, 300, 500; criterion: Friedman MSE, Absolute error, Poisson, Squared error; Max depth: 2, 5, 7
Gradient Boosting	Number of estimators: 100, 200, 300; Learning rate: 0.01, 0.05, 0.1; Loss: Huber, Quantile, Absolute error, Squared error; Max depth: 2, 3, 5

Figure 8 presents the MSE for the same models over search iterations, a metric where lower scores indicate better model performance. There is a downward trend in log loss values across all models, suggesting a general improvement in model accuracy as the search progresses. Sharp increases in log loss at certain points can be interpreted as instances where the model parameters did not align well with the prediction task, potentially leading to poorer performance. Notably, CatBoost showed stability in log loss in the later iterations. Overall, CatBoost achieved the highest training accuracy and lowest training log loss. Therefore, CastBoost is selected for further sensitivity analysis.



Model	MSE	RMSE	MAE	R squared
CatBoost	4.346	2.085	1.727	0.306
Random Forest	4.465	2.113	1.757	0.287
XGB	4.436	2.106	1.751	0.292
LGBM	4.402	2.098	1.741	0.297
Gradient Boosting	4.454	2.110	1.757	0.289

Figure 8. Model comparison

3.3.2. Sensitivity Analysis of Urbanized CBGs in Texas

Using SHAP values, the global importance of the top 20 variables influencing pedestrian crash risk at the CBG scale is presented in Figure 9, illustrating how each feature contributes to either increasing or decreasing the likelihood of pedestrian crashes. The y-axis lists the top 20 features, including aspects of urban infrastructure, transportation, employment, and demographic factors. The x-axis displays the SHAP values, which indicate the impact of each feature on the model's output, with positive values suggesting an increased risk of pedestrian crashes and negative values suggesting a decreased risk. The color gradient of the dots, ranging from green to pink, represents the actual feature values. Green indicates higher feature values, while pink denotes lower values. For example, green dots with positive SHAP values show that higher

feature values increase pedestrian crash risk, whereas pink dots with negative SHAP values imply the opposite. The spread of dots along the x-axis for each feature indicates how the impact of that feature varies across different CBGs, with a widespread suggesting significant variation.

Several key features stand out in this figure. Auto-oriented network density (Auto Ntwrk Density) has both negative and positive SHAP values. High auto-oriented network density (green dots) tends to shift SHAP values more positively, meaning higher auto network density is associated with higher pedestrian crash risk. Low auto-oriented network density (pink dots) has a smaller, negative effect.

In contrast, the percentage of high-wage income households, represented by "Pct Highwage (home location)," shows both positive and negative SHAP values. A higher percentage of high-wage households is associated with negative SHAP values, suggesting that higher-income areas tend to have a lower risk of pedestrian crashes. On the other hand, lower percentages of high-wage households exhibit positive SHAP values, indicating a positive association with pedestrian crash risk. This pattern is further supported by the "Pct Lowwage (workplace)" variable, which, although less influential, follows a similar trend: a higher percent of workplace low wage correspond to positive SHAP values, indicating an increased crash risk, while lower values are associated with negative SHAP values, reflecting a reduced risk. Roll and McNeil (2022b) found that areas with lower median income and higher proportions of BIPOC residents are linked to an increased frequency of pedestrian injuries. Both weekend daytime transit frequency (Wknd Day Runs) and weekend nighttime transit frequency (Wknd Night Runs) exhibit mixed SHAP values, indicating that their impact on pedestrian crash risk can be either positive or negative depending on contextual factors. However, overall, there is a positive association: as the frequency of transit increases, the likelihood of pedestrian crashes tends to rise. P. Chen and Zhou (2016) revealed

that regions with a higher density of bus stops tend to experience an increased number of pedestrian crashes.

Features related to land use and socioeconomic conditions, such as “Employment and Household Entropy” and “Pct of Zero-car,” also display notable patterns. “Employment and Household Entropy,” associated with diverse land use, shows a broad spread of SHAP values, suggesting that its impact on crash risk can vary widely. The presence of both pink and green dots indicates that this feature can sometimes increase crash risk and sometimes decrease it, highlighting the complexity of its relationship with pedestrian safety. In contrast, “Pct of Zero-car,” which represents the percentage of zero-car households, generally shows positive SHAP values, suggesting that areas with more zero-car households are linked to higher pedestrian crash risk, possibly due to increased pedestrian activity. In conclusion, this SHAP summary plot offers valuable insights into the factors contributing to pedestrian crash risk across different urban areas. By highlighting the varying impact of different features, the figure aids in understanding the complex interplay of factors that influence pedestrian safety. These insights can be crucial for urban planners and policymakers aiming to design safer, more pedestrian-friendly environments.

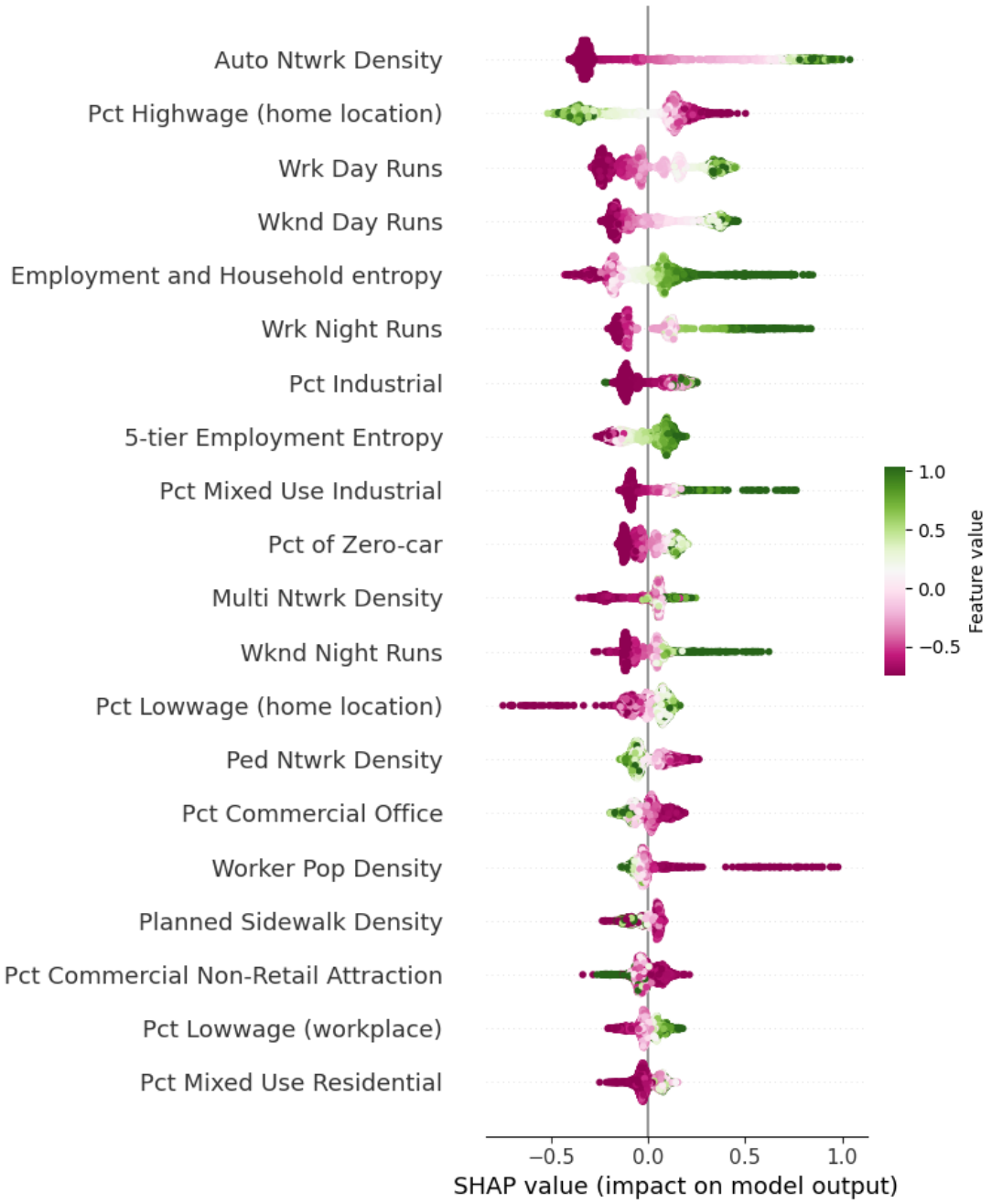


Figure 9. Global importance of the top 20 variables for CBG pedestrian crash risk

A more detailed analysis of feature influence is presented. Figure 10 highlights the relationship between various road network and socio-economic factors and pedestrian crash risk, with SHAP values illustrating each feature's impact on the model's predictions. Each scatter plot represents a specific factor, showing its values against corresponding SHAP values to depict how

changes in that feature influence pedestrian crash risk. The plot for auto network density reveals a positive correlation with crash risk, suggesting that as auto network density increases, so does the likelihood of pedestrian crashes. This relationship stabilizes once the density exceeds 8. The multi-modal network density plot also shows a generally positive correlation, though with more variability, indicating that areas with denser multi-modal networks may experience higher pedestrian crash risk. In contrast, pedestrian-oriented network density exhibits a negative correlation with crash risk. Higher densities are associated with lower crash risks, with SHAP values turning negative once the density surpasses 14, indicating a reduced risk. However, this effect diminishes as the density approaches 30.

For infrastructure and household characteristics, planned sidewalk density exhibits a non-linear relationship with SHAP values. When planned sidewalk density is near zero, SHAP values fluctuate between -0.25 and 0.1, indicating a mixed influence. This could suggest that areas with zero planned sidewalk density either have comprehensive existing sidewalks or are severely lacking in pedestrian infrastructure. As sidewalk density increases, the SHAP value steadily decreases, becoming negative around a density of 1.5, which is associated with reduced crash risk. The percentage of zero-car households shows a clear negative correlation with crash risk, suggesting that areas with more households lacking cars tend to experience higher pedestrian crash risks. However, this trend stabilizes after reaching a certain percentage of zero-car households. Worker population density initially correlates with a sharp decline in crash risk, which then levels off. This version improves clarity, flow, and precision, making the relationships you’re describing easier to understand.

For infrastructure and household characteristics, the planned sidewalk density shows a non-linear relationship with SHAP values, when the planned sidewalk density is around zero, the

SHAP value varies from -0.25 to 0.1, suggesting a mixed influence, which indicates that 0 planned sidewalk might indicate a comprehensive sidewalk or a very lacking sidewalk. after that, with the increase of density, the SHAP value decreased, after around 1.5, the value Shao value is negative, which is associated with reduced crash risk. The percentage of zero-car households shows a negative correlation with crash risk, implying that areas with more households lacking cars tend to have higher pedestrian crash risks. This trend stabilizes after a certain percentage of zero-car households. Worker population density indicates a strong initial decrease in crash risk, then stabilized.

An examination of economic factors reveals that the percentage of high-wage workers (by workplace) shows a slightly negative correlation with pedestrian crash risk. This suggests that areas with a greater proportion of high-wage workers tend to experience fewer pedestrian crashes. However, this effect plateaus at higher percentages of high-wage workers. Notably, some CBGs with the highest proportion of high-wage workers exhibit extreme SHAP values, indicating a stronger positive association with crash risk in those cases. In contrast, the plot for the percentage of low-wage workers (by workplace) presents a more nuanced pattern. Moderate levels of low-wage workers are linked to reduced crash risk, but when the proportion exceeds a certain threshold, the risk increases. Similarly, the percentage of low-wage workers (by home location) shows a slightly positive correlation with pedestrian crash risk, meaning areas with more low-wage residents tend to experience higher crash risks. This effect, however, diminishes at higher percentages. In summary, Figure 6 illustrates the complex relationship between pedestrian crash risks and both road network configurations and socio-economic characteristics. While improvements in pedestrian infrastructure and certain socio-economic factors can reduce crash risks, overly dense networks and specific economic compositions may have the opposite

effect. These findings underscore the importance of balanced urban planning, where both network density and socio-economic context are carefully considered to enhance pedestrian safety effectively.

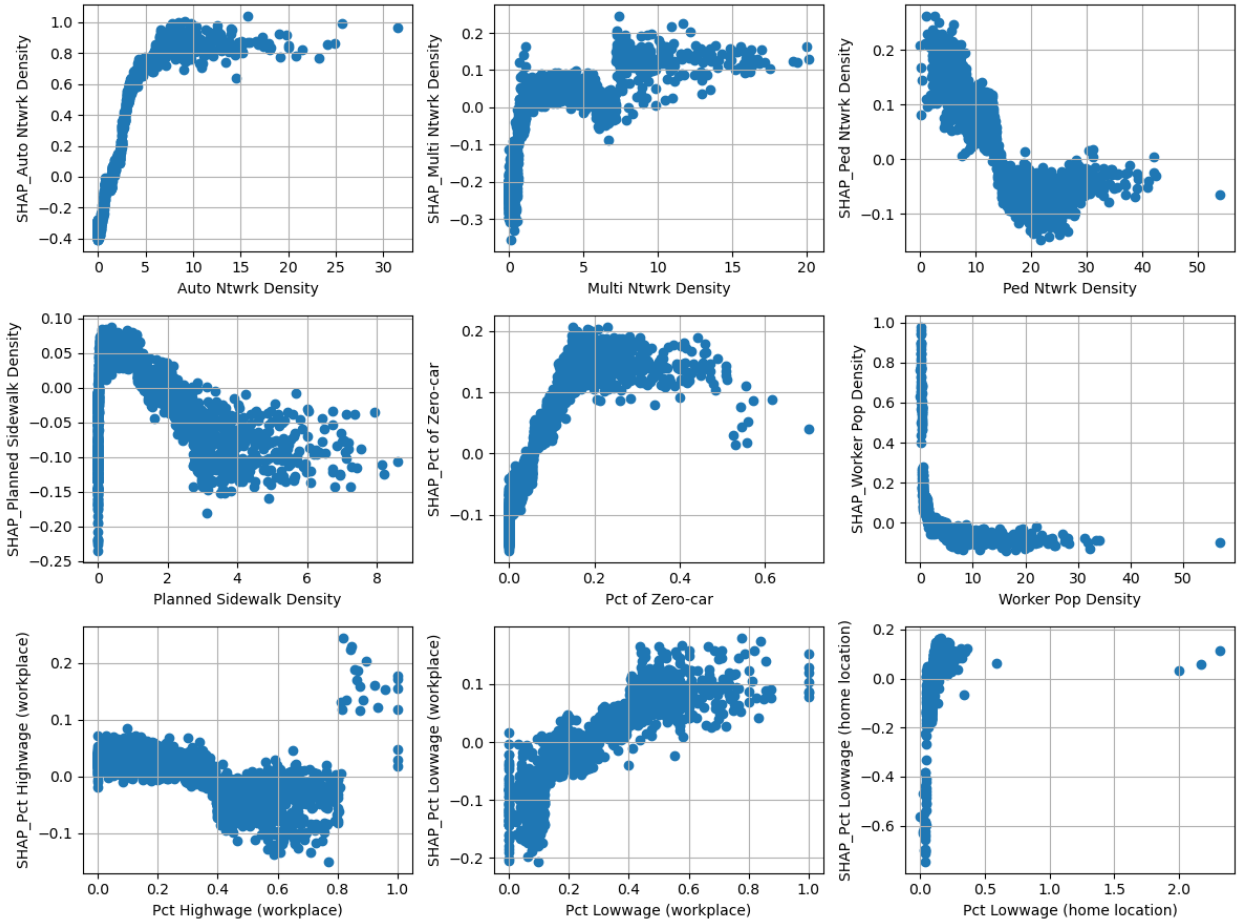


Figure 10. SHAP values of road network and socio-economic factors on pedestrian crash risk

Scatter plots that depict the SHAP values for various transit frequencies and land use characteristics are presented in Figure 11. The first set of plots focuses on transit frequency. Workday runs show a positive correlation with pedestrian crash risk, indicating that as the number of weekday transit frequency increases, so does the risk of pedestrian crashes. This effect is more pronounced at higher numbers of weekday runs. Similarly, weekday runs and work night runs also show positive relationships with crash risk, suggesting that increased transit activity

during weekend days and weekday nights contributes to higher pedestrian crash risks. Weekend night runs show a similar positive trend, indicating that weekend night transit activity is also associated with elevated pedestrian crash risks.

The next set of plots examines land use characteristics. percent of mixed use industrial shows a non-linear relationship, where moderate percentages of mixed-use industrial areas are associated with increased pedestrian crash risks. However, this effect diminishes at higher percentages. pct industrial indicates a positive correlation, suggesting that areas with a higher percentage of industrial land use have an increased risk of pedestrian crashes, especially when industrial land use exceeds a certain threshold.

The final set of plots explores different commercial and residential land use types. Percent of commercial offices demonstrates a negative correlation with pedestrian crash risk, implying that areas with a higher percentage of commercial office space tend to have lower pedestrian crash risks. This trend is most significant at lower percentages of commercial office space. pct commercial non-retail attraction shows a similar negative trend, suggesting that non-retail commercial attractions are also associated with reduced pedestrian crash risks. Lastly, percent of mixed use residential indicates a slight negative correlation, meaning that areas with mixed-use residential development might experience lower pedestrian crash risks, though this effect is less pronounced compared to other land use types.

In summary, the results highlight how both transit frequency and land use characteristics can significantly impact pedestrian crash risk. Increased transit frequency, particularly during peak times and in areas with mixed-use industrial land, is associated with higher pedestrian crash risks. Conversely, certain types of commercial land use, such as offices and non-retail attractions, may contribute to reduced crash risks. These findings suggest that urban planning strategies

should carefully consider transit schedules and land use composition to mitigate pedestrian crash risks effectively.

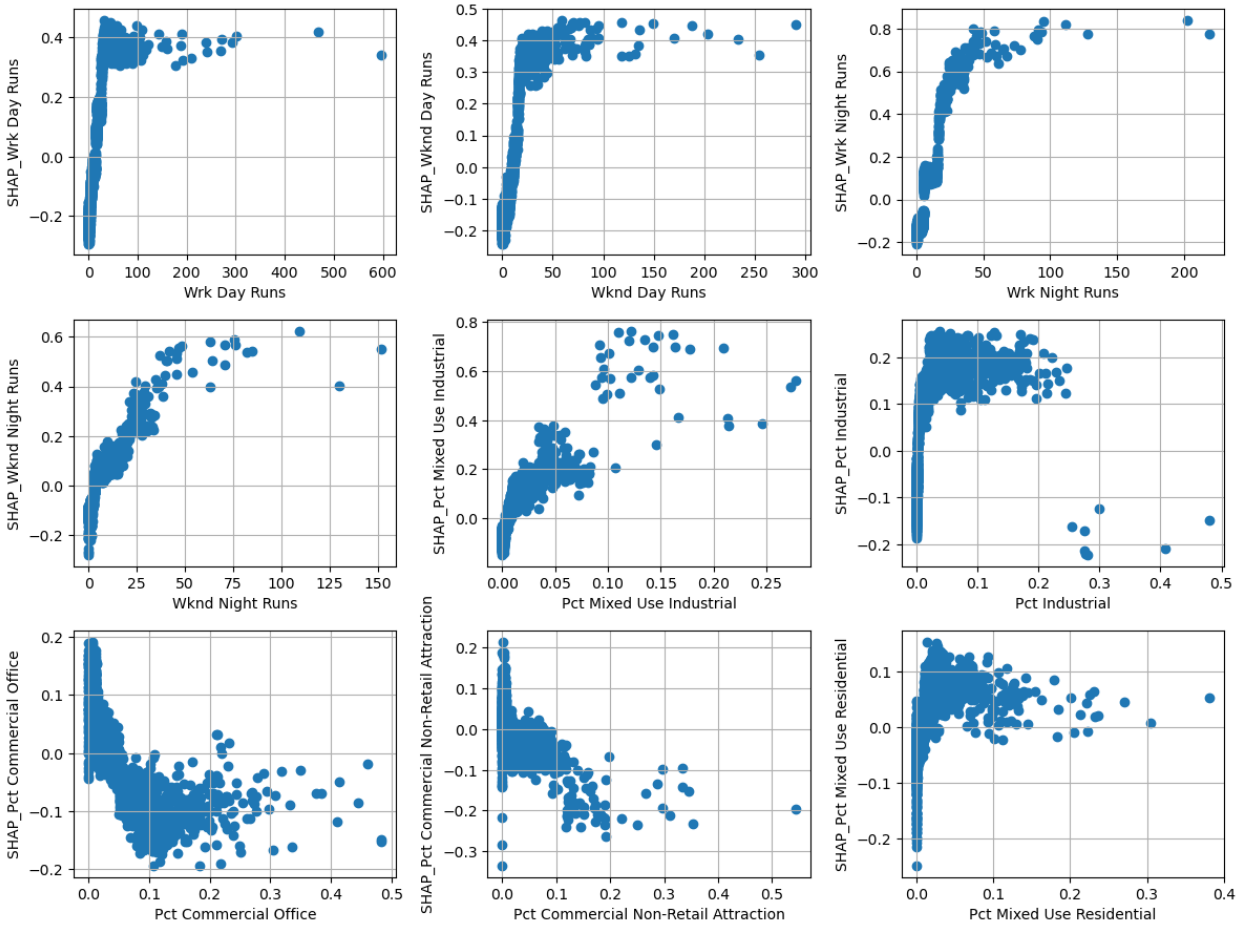


Figure 11. SHAP values of transit frequency and land use characteristics on pedestrian crash risk

The impact of employment characteristics on pedestrian crash risk is presented in Figure 12. The plot for employment and household entropy shows a positive correlation, indicating that higher entropy, which reflects a more diverse mix of employment and household types, is associated with increased pedestrian crash risk. Similarly, the plot for 5-tier employment entropy reveals a positive relationship, suggesting that areas with a more diverse employment mix tend to have higher pedestrian crash risks. Overall, the results highlight the complex interplay between transit frequency, land use, and employment characteristics in influencing pedestrian crash risk.

High transit frequency, especially during night-time, and certain land use types, such as mixed-use industrial areas, are associated with increased crash risks. In contrast, areas with higher proportions of commercial office and non-retail commercial spaces tend to have lower pedestrian crash risks. Understanding these relationships can inform urban planning and policy decisions aimed at enhancing pedestrian safety.

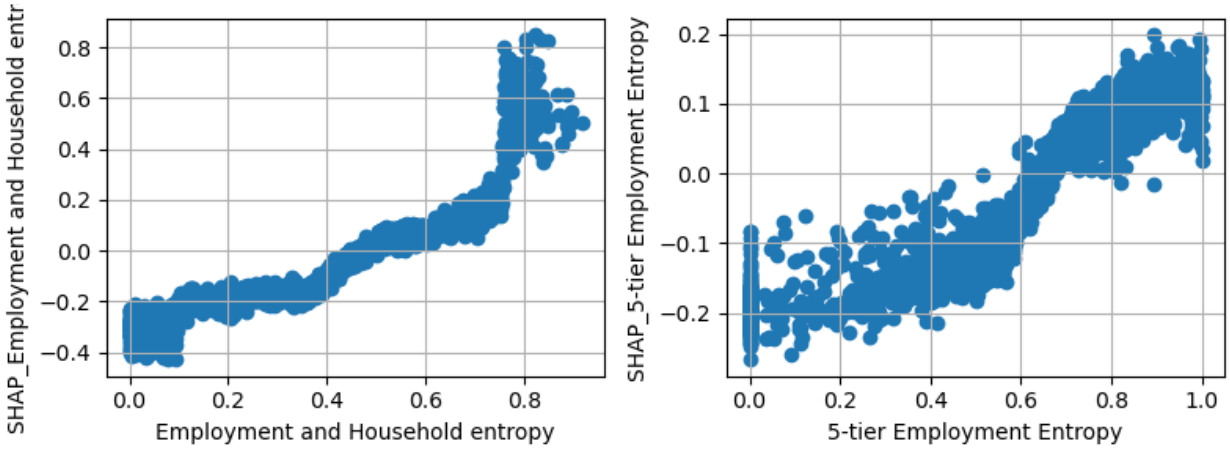


Figure 12. SHAP values of employment characteristics on pedestrian crash risk

3.3.3. Disadvantaged and Non-disadvantaged CBGs

The study also compared disadvantaged and non-disadvantaged CBGs across multiple categories, including transportation, employment, and land use variables. Disadvantaged and non-disadvantaged communities were defined based on the Equitable Transportation Community framework. A Mann-Whitney U test was employed for the analysis. These findings provide insight into how the characteristics of disadvantaged CBGs differ from those of non-disadvantaged areas. There are significant disparities in EPDO scores between disadvantaged and non-disadvantaged CBGs, with respective values of 451.305 and 190.559, highlighting a substantial difference. Additionally, the analysis shows notable variations in income-related variables: disadvantaged CBGs have a higher percentage of low-wage workers at their home location (0.117) compared to non-disadvantaged CBGs (0.098), while the percentage of high-

wage workers is lower (0.129 vs. 0.255). At the workplace, the percentage of low-wage workers is also higher in disadvantaged CBGs (0.295) than in non-disadvantaged ones (0.276), whereas the percentage of high-wage workers is lower (0.296 vs. 0.351). These findings emphasize the economic inequalities between the two groups.

The transportation-related variables indicate disparities in vehicle miles traveled (VMT) and network-oriented density between the two groups. For instance, disadvantaged CBGs have a higher mean work-related VMT (24.916 compared to 23.586), which is significant with a p-value of 0.0442. The difference in commute VMT is also significant ($p = 0.0036$), with disadvantaged areas showing a higher mean (19.909 compared to 18.583). However, the non-commute VMT shows no significant difference ($p = 0.0841$). Network density is also higher in disadvantaged CBGs (21.352 vs. 20.296), and the difference is statistically significant ($p = 0.0002$). Differences in auto-oriented, multi-oriented, and pedestrian-oriented network densities show that disadvantaged areas have a denser auto network (2.334 compared to 1.688) and a slightly higher multi-modal network density (3.250 vs. 2.954), both with statistically significant p-values of 0.0113 and 0.0052, respectively. However, there is no significant difference in pedestrian network density ($p = 0.8786$), suggesting similar levels of pedestrian infrastructure between the two groups.

In terms of employment and population characteristics, several differences stand out. Disadvantaged CBGs have a lower percentage of the working-age population (0.596 vs. 0.646, $p = 0.0000$), higher percentages of zero-car households (0.127 vs. 0.053, $p = 0.0000$), and lower percentages of one-car households (0.392 vs. 0.355, $p = 0.0000$). Employment density, however, does not show a significant difference ($p = 0.0968$), though employment within transit stops is higher in disadvantaged CBGs (0.066 vs. 0.048, $p = 0.0031$). Disadvantaged CBGs also show

lower five-tier employment entropy (0.582 vs. 0.662, p = 0.0000), reflecting less diversity in employment sectors. Additionally, there is a higher proportion of low-wage workers at home locations in disadvantaged CBGs (0.117 vs. 0.098, p = 0.0000), while the proportion of high-wage workers is lower (0.129 vs. 0.255, p = 0.0000).

Differences are also evident in land use and infrastructure variables. Disadvantaged CBGs have lower levels of planned sidewalk density (1.227 vs. 1.313, p = 0.0011) and higher levels of acceptable sidewalk density (0.416 vs. 0.709, p = 0.0000). However, existing sidewalk density is higher in disadvantaged areas (3.861 vs. 3.095, p = 0.0000), suggesting that while planned sidewalks may be lower, current infrastructure is more developed in these areas. Other infrastructure-related differences include higher bike lane density in non-disadvantaged CBGs (0.298 vs. 0.756, p = 0.0000), and a higher percentage of transit service area coverage in disadvantaged areas (0.635 vs. 0.440, p = 0.0000), indicating better access to public transportation. Street intersection density shows no significant difference between the two groups (p = 0.1565).

Lastly, there are significant differences in land use and zoning between disadvantaged and non-disadvantaged CBGs. The land use entropy (which measures the diversity of land use types) is higher in disadvantaged areas (1.522 vs. 1.427, p = 0.0000), indicating more mixed land use. Disadvantaged areas have a lower percentage of single residential land use (0.432 vs. 0.518, p = 0.0000), but similar percentages of multi-residential land use (0.070 vs. 0.068, p = 0.0012). Additionally, disadvantaged CBGs have a higher percentage of commercial retail land (0.175 vs. 0.112, p = 0.0000) but a lower percentage of high-wage workplace land use (0.296 vs. 0.351, p = 0.0000), reflecting the economic and commercial landscape in these areas. Furthermore, disadvantaged areas show lower percentages of civic healthcare, civic education, and civic other

land uses, with significant p-values across these variables, indicating disparities in access to community services and public institutions.

In summary, the results highlight significant differences between disadvantaged and non-disadvantaged CBGs across transportation, employment, and land use variables. Disadvantaged areas show higher levels of zero-car households, more mixed land use, and denser transit networks, but lower percentages of high-wage workers and certain community services. These differences emphasize the socio-economic and infrastructural disparities between the two groups.

Table 4. Mann-Whitney U test results between disadvantaged and non-disadvantaged CBGs

Variables	Mean (Disadvantaged)	Mean (Non-Disadvantaged)	U-Statistic	P-Value	Variables	Mean (Disadvantaged)	Mean (Non-Disadvantaged)	U-Statistic	P-Value
EPDO	451.305	190.559	1047728	0.0000	*Land Use Characteristics*				
Demographics					Pct Single Residential	0.432	0.518	659040	0.0000
Worker Pop Density	4.342	4.664	775845	0.0061	Pct Multi Residential	0.07	0.068	891381	0.0012
Resident Pop density	2.752	2.184	1153201	0.0000	Pct Commercial Retail	0.175	0.112	1111160	0.0000
Pct of WrkAge Pop	0.596	0.646	602420	0.0000	Pct Commercial Office	0.054	0.047	913635	0.0000
Pct of Zero-car	0.127	0.053	1204800	0.0000	Pct Commercial Non-Retail Attraction	0.018	0.021	749883	0.0000
Pct of Two-plus-car	0.472	0.591	545012	0.0000	Pct Mixed Use Residential	0.012	0.015	685355	0.0000
Pct Lowwage (home location)	0.117	0.098	1037802	0.0000	Pct Mixed Use Commercial	0.044	0.046	852465	0.2227
Pct Highwage (home location)	0.129	0.255	212024	0.0000	Pct Mixed Use Industrial	0.013	0.007	931165	0.0000
Pct Lowwage (workplace)	0.295	0.276	855189	0.1738	Pct Mixed Use Other	0.036	0.025	943675	0.0000
Pct Highwage (workplace)	0.296	0.351	689291	0.0000	Pct Industrial	0.029	0.011	1105463	0.0000
5-tier Employment Entropy	0.582	0.662	722397	0.0000	Pct Civic Healthcare	0	0.001	731597	0.0000
Employment and Household entropy	0.422	0.46	756678	0.0002	Pct Civic Education	0.046	0.036	900680	0.0002
Household Workers per Job Equilibrium	0.271	0.266	804046	0.1990	Pct Civic Other	0.004	0.005	796647	0.0776
Pedestrian Facility Characteristics Public Transit Characteristics					Pct Transportation Utilities	0.012	0.007	883543	0.0015
Wrk Day Runs	20.782	12.074	1112070	0.0000	Pct Open Space	0.021	0.032	704246	0.0000
Wrk Night Runs	8.568	5.263	1098284	0.0000	Pct Agriculture	0.027	0.042	688753	0.0000
Wknd Day Runs	15.144	8.899	1111421	0.0000	Land Use Entropy	1.522	1.427	927049	0.0000
Wknd Night Runs	7.114	4.29	1107674	0.0000	*Pedestrian Facility Characteristics*				
Pct Transit Service Area	0.635	0.44	1094291	0.0000	Planned Sidewalk Density	1.227	1.313	891845	0.0011
Trip Characteristics					Existing Sidewalk Density	3.861	3.095	978593	0.0000
Trip productions and trip attractions equilibrium	0.358	0.388	765696	0.0011	Bike Lane Density	0.298	0.756	614498	0.0000
Employment within Transit Stop (0.5 mile)	0.066	0.048	860991	0.0031	*Network Characteristics*				
Jobs within 45 minutes auto travel time	238157.47	235220.11	883263	0.0049	Auto Ntwrk Density	2.334	1.688	874176	0.0113
Working age population within 45 minutes auto travel	183416.19	174137.39	958341	0.0000	Multi Ntwrk Density		3.25	2.954	882892
Non Commute VMT	5.007	5.003	795458	0.0841	Ped Ntwrk Density	15.768	15.654	831832	0.8786
Commute VMT	19.909	18.583	885191	0.0036	Street Intersection Density	100.295	95.308	856292	0.1565
Work related VMT	24.916	23.586	867807	0.0442					

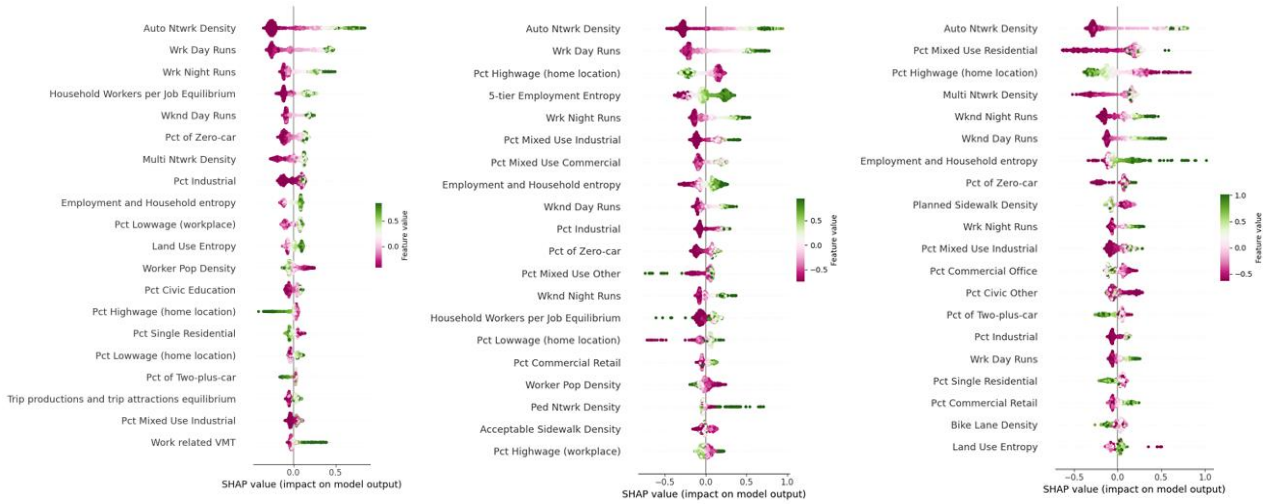
3.3.4. Sensitivity Analysis by City

The SHAP value plots depicting the feature importance across three different urban areas are presented in Figure 13. As shown in Figure 13 a), the Auto Network Density stands out as the most influential feature with a predominantly positive association with pedestrian crash risk. Features like Work Day Runs, Work Night Runs, and Household Workers per Job Equilibrium also contribute positively to the model output. There are negative influences from features like Pct of Two-plus-car Households Pct Single Residential, Worker Pop Density, and Pct Highway (home location), indicating these variables reduce crash risk in these CBGs.

Figure 13 b) highlights Auto Network Density as a consistently important variable, with a strong positive relationship to crash risk, similar to the San Antonio area. Interestingly, Pct Highway (home location) appears to play a much stronger positive role in the Dallas area compared to the San Antonio area. Another difference is the influence of Acceptable Sidewalk Density, which emerges as an important variable with a slight positive association in this context, indicating the role of infrastructure in affecting pedestrian crash risks. 5-tier Employment Entropy also appears as an important variable associated with crash risk, in contrast, Pct Commercial Retail has a more significant negative impact compared to the other city, possibly reflecting different land-use patterns and pedestrian behaviors. More land use features appeared as important variables compared to the San Antonio region.

Figure 13 c) shows Auto Network Density as a primary feature impacting crash risk. Besides, the Pct of Mixed Use Residential feature now shows a much more pronounced positive effect compared to the first two cities, suggesting that mixed land-use planning in this city might increase pedestrian-vehicle interactions. Planned Sidewalk Density appears as an important variable with slightly negative associations. Similarly, Bike Lane Density makes a notable

appearance in the Austin area, indicating that bike lane infrastructure might have complex effects, either increasing or reducing risk based on local conditions. In contrast, Land Use Entropy exhibits a positive relationship with crash risk, reflecting a potential increase in crash occurrences in areas with a more balanced land-use mix.



a) San Antonio b) Dallas c) Austin

Figure 13. Feature importance plot by city

San Antonio

Transportation Infrastructure

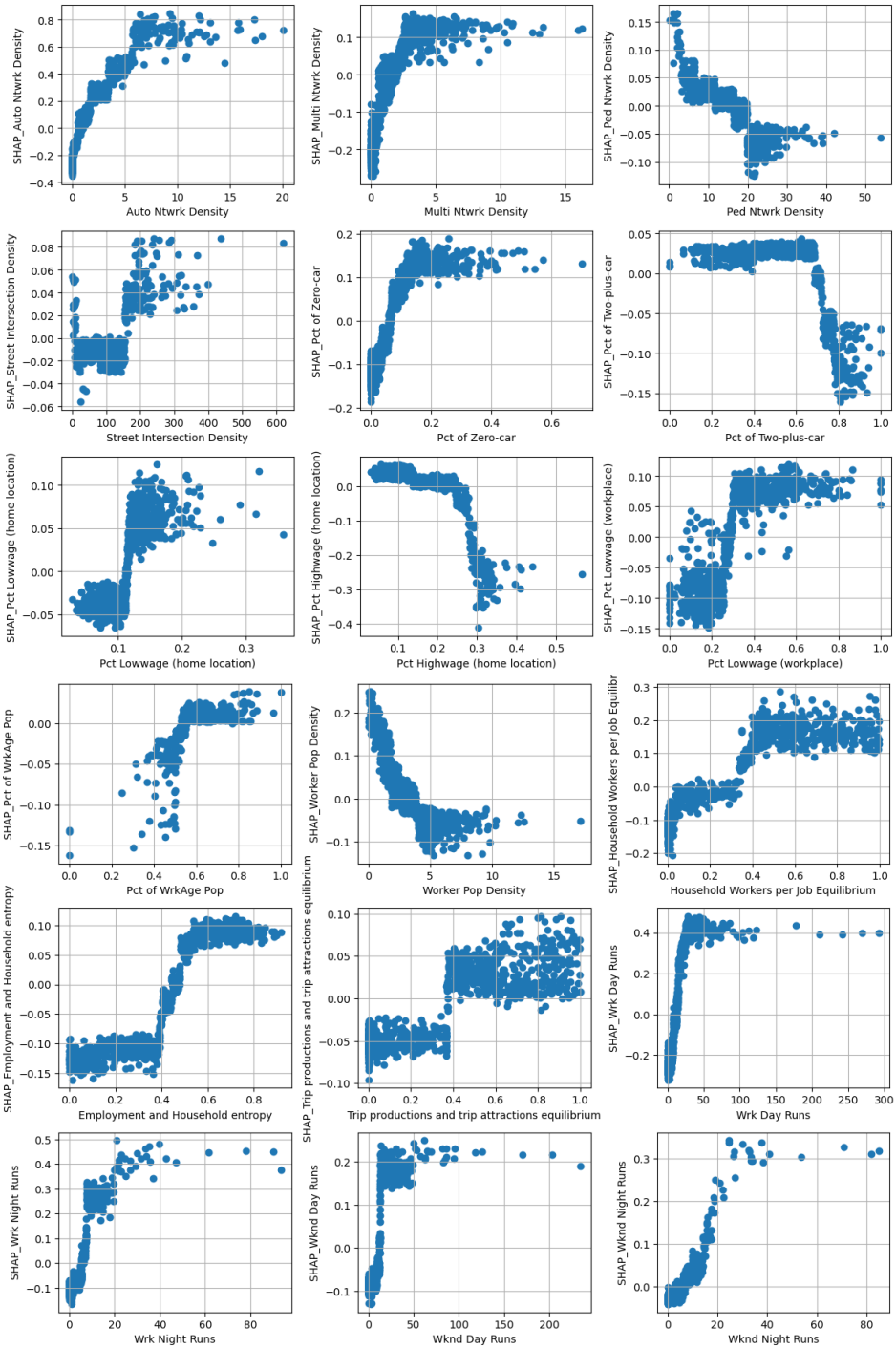
Auto Network Density and Multi-Network Density have strong positive impacts on pedestrian crash risk. The higher the density of these networks, the more likely the risk of pedestrian crashes increases. This is logical, as denser networks often indicate higher vehicular traffic, leading to greater exposure for pedestrians. Conversely, Pedestrian Network Density shows a negative relationship with pedestrian crash risk, implying that better lower speed network density tends to reduce crashes. Street Intersection Density shows a complex relationship, with an initial increase in SHAP values followed by variability. Intersections are typically hotspots for pedestrian-vehicle interactions, which may explain the heightened crash risks in areas with more intersections.

Socio-demographic Variables

Pct Highwage (Home Location) has a negative relationship with crash risk, suggesting that areas with higher income experience fewer crashes. The percentage of Zero-car Households exhibits a strong positive association with pedestrian crashes, suggesting that areas with fewer car-owning households may have more pedestrians on the road, increasing their exposure to crash risk. On the other hand, Percentage of Two-plus-car Households shows a negative relationship, indicating that areas with more car ownership tend to reduce pedestrian exposure, likely because fewer individuals rely on walking or public transit. Percentage of Working Age Population also contributes positively to crash risk, as areas with a higher proportion of working-age individuals may have increased activity levels, leading to greater pedestrian exposure.

Land Use and Employment

Employment and Household Entropy and Land Use Entropy show strong positive relationships with pedestrian crash risk. Mixed land use, which often increases pedestrian-vehicle interactions, leads to higher crash risks in these areas. This reflects the potential dangers of pedestrian activity in densely mixed-use environments. Pct Industrial has a slight negative effect on crash risk, indicating that highly industrialized areas, where pedestrian activity is generally lower, are associated with reduced crash risks. Similarly, Pct Commercial Retail shows varying effects, with SHAP values decreasing for higher percentages. Work-related VMT, Commute VMT, and Non-Commute VMT consistently show positive relationships with crash risk, suggesting that higher vehicle mileage in an area increases the likelihood of pedestrian crashes. This aligns with expectations, as higher traffic volumes typically correlate with greater pedestrian exposure and risk.


SHAP_Auto Ntwrk Density
Auto Ntwrk Density
SHAP_Multi Ntwrk Density
Multi Ntwrk Density
SHAP_Ped Ntwrk Density
Ped Ntwrk Density
SHAP_Street Intersection Density
Street Intersection Density
SHAP_Pct of Zero-car
Pct of Zero-car
SHAP_Pct of Two-plus-car
Pct of Two-plus-car
SHAP_Pct Lowwage (home location)
Pct Lowwage (home location)
SHAP_Pct Highwage (home location)
Pct Highwage (home location)
SHAP_Pct Lowwage (workplace)
Pct Lowwage (workplace)
SHAP_Pct of WrkAge Pop
Pct of WrkAge Pop
SHAP_Worker Pop Density
Worker Pop Density
SHAP_Household Workers per Job Equilibr
Household Workers per Job Equilibrium
SHAP_Employment and Household entropy
Employment and Household entropy
SHAP_Trip productions and trip attractions equilibrium
Trip productions and trip attractions equilibrium
SHAP_Wrk Day Runs
Wrk Day Runs
SHAP_Wrk Night Runs
Wrk Night Runs
SHAP_Wknd Day Runs
Wknd Day Runs
SHAP_Wknd Night Runs
Wknd Night Runs

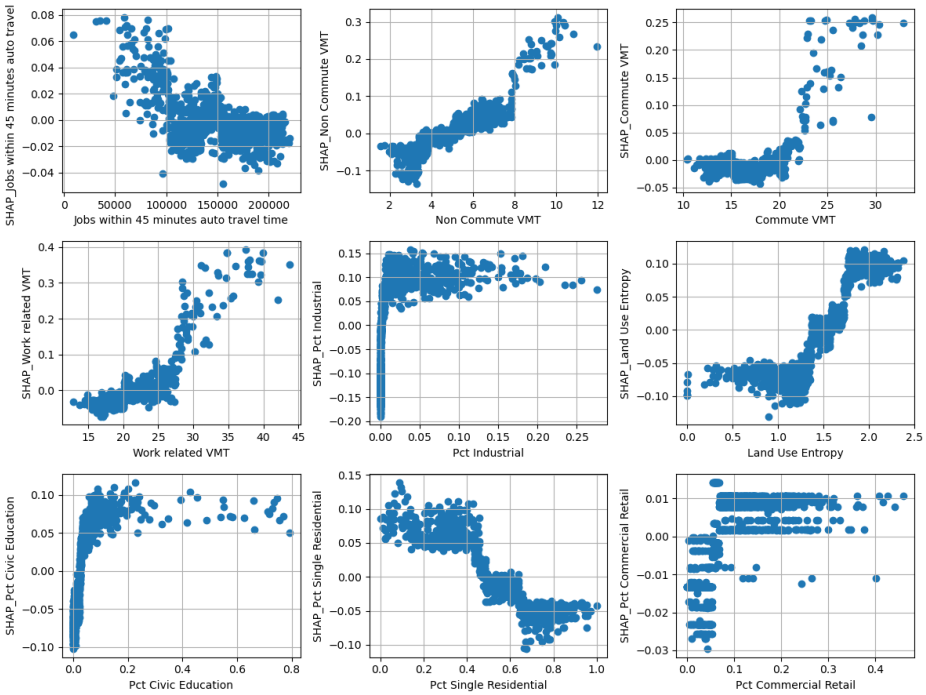


Figure 14. SHAP values of top factors on pedestrian crash risk in the San Antonio region

Dallas

Transportation Infrastructure

Auto Network Density is strongly associated with an increased risk of pedestrian crashes. Multi-Network Density follows a similar pattern, showing a positive relationship with crash risk. In contrast, Pedestrian Network Density has a negative SHAP value, meaning areas with higher low-speed network density tend to have reduced crash risks, as expected with safer walking environments. Street Intersection Density shows a mixed relationship with pedestrian crashes. Higher intersection density initially increases SHAP values, indicating a higher risk of crashes due to the frequent crossing points for pedestrians, but with some variability in impact. Socio-demographic Variables

Socio-demographic Variables

Pct Highwage has a negative relationship with crash risk. Percentage of Zero-car Households shows a positive relationship with crash risk. Percentage of Low-wage Jobs (Workplace) also has

a positive relationship with crash risk. Low-wage job density, which may correlate with specific commuting patterns and pedestrian exposure, affects crash risks differently depending on location. In areas where fewer residents own cars, pedestrian traffic is likely higher, thus increasing the likelihood of crashes. On the other hand, Percentage of Two-plus-car Households has a negative effect on crash risk. Percentage of Working Age Population also contributes positively to crash risk.

Employment and Land Use

Employment and Household Entropy and 5-tier Employment Entropy have strong positive associations with crash risk. These findings highlight that mixed land use and employment diversity tend to increase pedestrian exposure to vehicular traffic, leading to more crashes. Mixed Use Industrial and Mixed Use Other exhibit mixed effects. While some industrial areas may have heightened crash risks, others may have lower pedestrian activity, leading to varied impacts. Percentage of Commercial Retail shows a positive relationship with crash risk. Retail areas often have more foot traffic, increasing pedestrian-vehicle interactions and thereby increasing crash risk Day Runs and Work Night Runs both show strong positive relationships with crash risk. As the number of trips during the day or night increases, pedestrian exposure increases, elevating the likelihood of crashes.

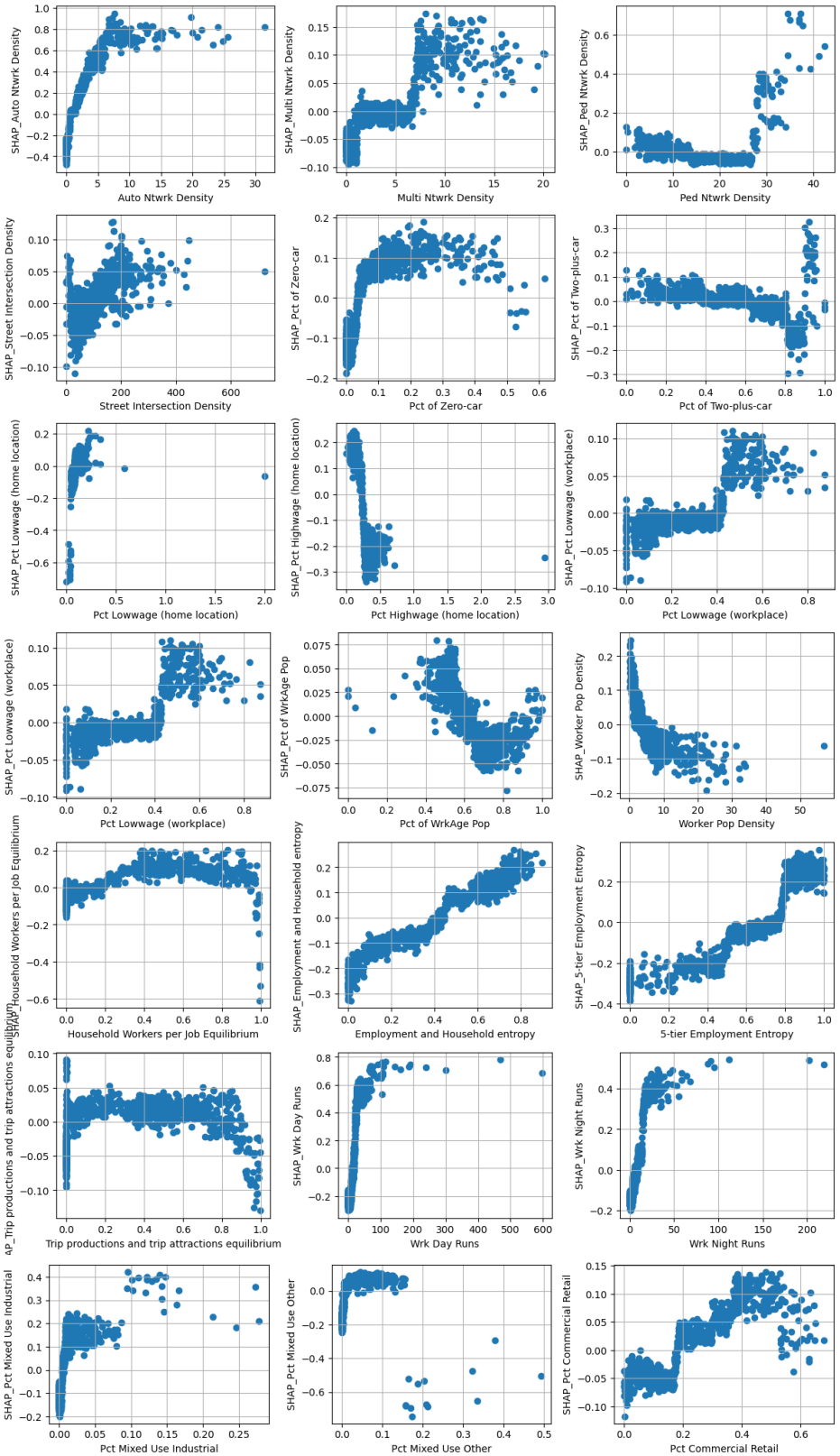


Figure 15. SHAP values of top factors on pedestrian crash risk in the Dallas region

Austin

Transportation Infrastructure

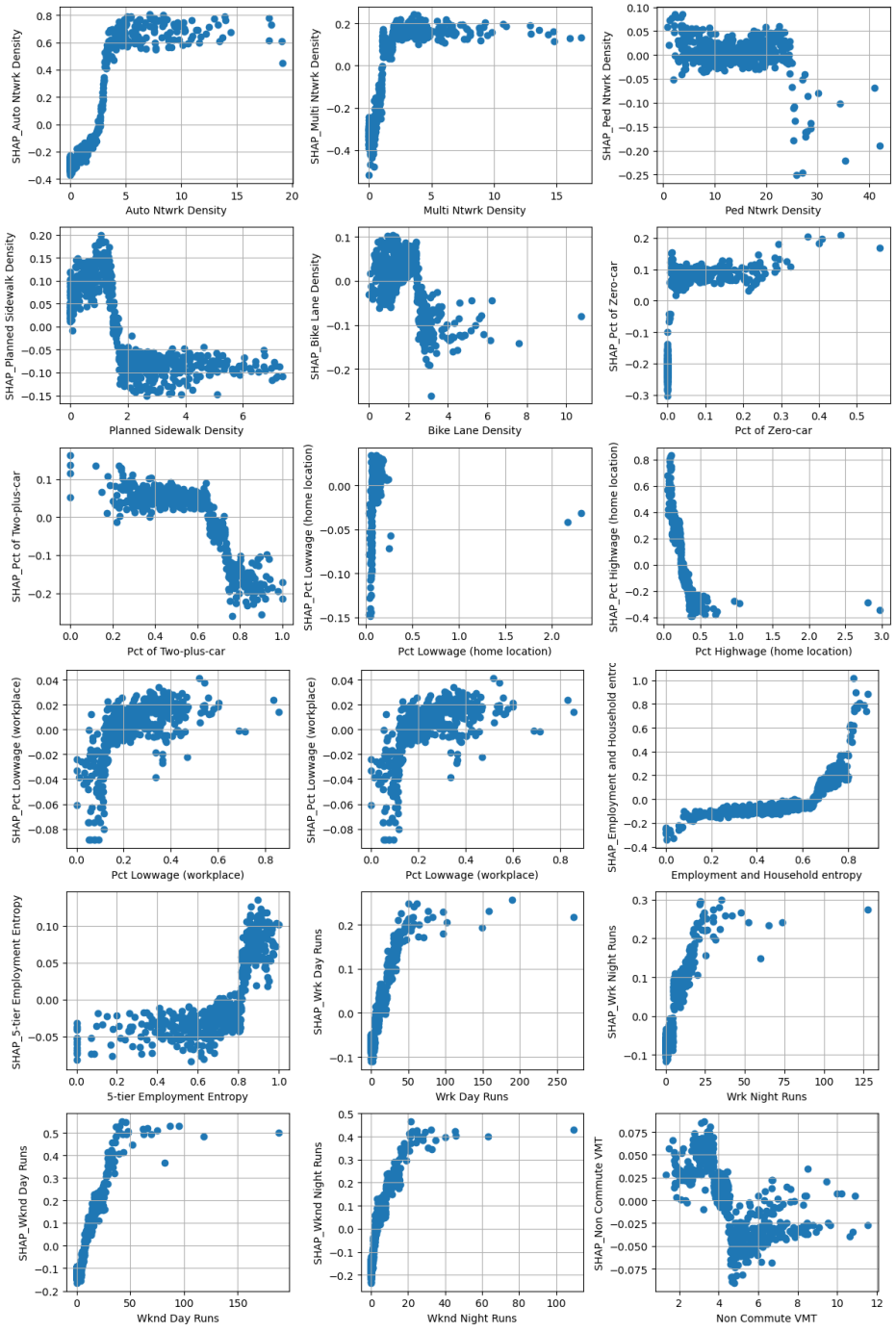
Auto Network Density has a strong positive association with crash risk. As auto network density increases, pedestrian crash risk rises due to more vehicular traffic. Multi Network Density also shows a positive influence on crash risk. This suggests that areas with a combination of different transportation modes experience higher pedestrian crash exposure. Pedestrian Network Density shows a negative SHAP value, meaning that well-developed pedestrian networks tend to lower crash risk by providing safer pedestrian infrastructure. Bike Lane Density has a negative impact at lower densities but flattens at higher values, suggesting that low bike lane coverage reduces crash risk slightly, but after a certain point, this effect stabilizes. Planned Sidewalk Density has a similarly negative relationship, confirming that better sidewalk coverage reduces pedestrian crash risk.

Socio-demographic Variables

Percentage of Zero-car Households shows a significant positive relationship with crash risk, indicating that areas with more carless households tend to have more pedestrian activity and, therefore, higher crash risk. Conversely, Percentage of Two-plus-car Households has a strong negative association with crash risk, suggesting that more car ownership reduces pedestrian exposure. Pct Highwage has a negative relationship with crash risk. Percentage of Low-Wage Workers (both home and workplace) has a positive relationship, with a general trend towards increased risk where low-wage workers are concentrated, likely due to increased pedestrian traffic in areas where residents rely more on walking or public transit.

Land Use and Employment

Employment and Household Entropy and 5-tier Employment Entropy both show strong positive relationships with crash risk. These results highlight that areas with more mixed land use and employment variety often have more pedestrian activity, raising the risk of crashes. Mixed Use Residential exhibits a moderate positive association with crash risk, while Mixed Use Industrial demonstrates a less consistent but generally positive relationship. These findings suggest that more diversified land use increases pedestrian-vehicle interactions. Percentage of commercial retail shows a strong positive relationship with crash risk, likely because retail areas attract more foot traffic, which increases interactions between pedestrians and vehicles. Percentage of civic other has a somewhat negative relationship, which might suggest that civic areas (e.g., parks, community centers) tend to reduce pedestrian exposure to traffic. Commute VMT and Work-Related VMT both have a positive relationship with crash risk, reflecting that areas with higher vehicular travel for work and commuting lead to more pedestrian exposure to crashes. Non-Commute VMT exhibits a complex relationship, where the effect fluctuates, but areas with moderate VMT levels seem to reduce pedestrian crash risk.


SHAP_Auto Ntwrk Density
Auto Ntwrk Density
SHAP_Multi Ntwrk Density
Multi Ntwrk Density
SHAP_Ped Ntwrk Density
Ped Ntwrk Density
SHAP_Planned Sidewalk Density
Planned Sidewalk Density
SHAP_Bike Lane Density
Bike Lane Density
SHAP_Pct of Zero-car
Pct of Zero-car
SHAP_Pct of Two-plus-car
Pct of Two-plus-car
SHAP_Pct Lowwage (home location)
Pct Lowwage (home location)
SHAP_Pct Highwage (home location)
Pct Highwage (home location)
SHAP_Pct Lowwage (workplace)
Pct Lowwage (workplace)
SHAP_Pct Lowwage (workplace)
Pct Lowwage (workplace)
SHAP_Employment and Household entrc
Employment and Household entropy
SHAP_5-tier Employment Entropy
5-tier Employment Entropy
SHAP_Wrk Day Runs
Wrk Day Runs
SHAP_Wrk Night Runs
Wrk Night Runs
SHAP_Wknd Day Runs
Wknd Day Runs
SHAP_Wknd Night Runs
Wknd Night Runs
SHAP_Non Commute VMT
Non Commute VMT

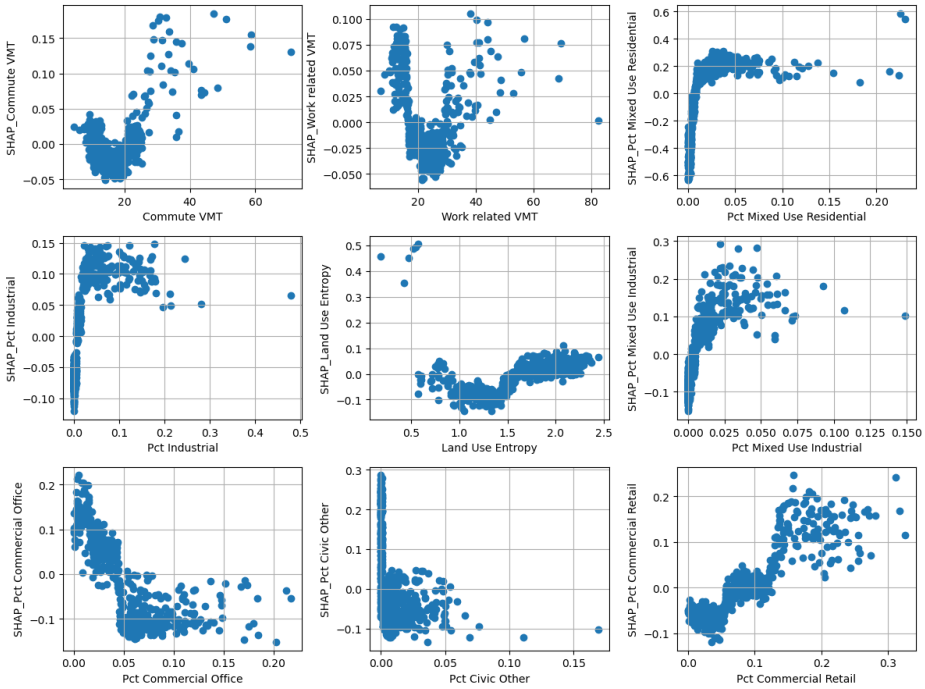


Figure 16. SHAP values of top factors on pedestrian crash risk in the Austin region

3.4. Summary

This chapter investigated the pedestrian crash risks in urbanized CBGs in Texas, incorporating key elements identified in previous research, including transit characteristics, pedestrian facilities and land use characteristics. By analyzing the distributional characteristics of EPDO and developing optimal machine learning models, the study aimed to enhance our understanding of these risk factors. The analysis shows that auto network density consistently contributes to higher pedestrian crash risks, as indicated by predominantly positive SHAP values. This finding underscores the need for targeted interventions in areas with dense auto networks to mitigate risks, such as implementing traffic calming measures or enhancing pedestrian crossings. In contrast, features such as pedestrian network density and planned sidewalk coverage generally show negative SHAP values, suggesting that increasing pedestrian infrastructure can effectively reduce crash risks. This highlights the importance of investing in dedicated pedestrian pathways and ensuring that pedestrian infrastructure keeps pace with urban development.

Transit frequency is another critical factor influencing pedestrian crash risk, with SHAP values indicating that increased transit activity-especially during peak times-correlates with higher crash risks. This suggests that urban planners need to carefully manage transit operations, possibly by optimizing transit stop locations, improving lighting and signage around transit areas, and enhancing pedestrian safety measures near high-frequency transit routes. Socio-economic factors also significantly impact pedestrian safety. The study finds that areas with a higher percentage of zero-car households tend to have increased pedestrian crash risks, as these areas likely see more pedestrian activity. This suggests a need for more robust pedestrian safety measures in neighborhoods with higher numbers of zero-car households, such as improving crosswalks and traffic signals. Moreover, mixed-use environments, reflected in variables like

employment and household entropy, exhibit a broad range of SHAP values, indicating that while diversity in land use can bring vibrancy, it also requires careful planning to manage the associated risks. Land use characteristics play a dual role. While commercial office spaces and non-retail commercial attractions are associated with reduced pedestrian crash risks, likely due to better infrastructure or lower pedestrian-vehicle interaction in these zones, areas with a significant industrial presence or high mixed-use industrial land show increased risks. This suggests that zoning regulations and land use planning should consider the safety implications of industrial and mixed-use developments on pedestrian environments.

The group comparison highlights significant differences between disadvantaged and non-disadvantaged CBGs, particularly in pedestrian crash risks, with disadvantaged CBGs facing higher levels of risk. These disparities extend across income, transportation, employment, and land use variables. Disadvantaged areas have higher proportions of zero-car households, larger low-income populations, more mixed land use, and denser transit networks, but lower percentages of high-wage workers and access to certain community services. These contrasts underscore the socio-economic and infrastructural inequalities between the two groups.

In conclusion, the SHAP value analysis provides a robust framework for understanding the diverse factors that contribute to pedestrian crash risks in urban areas. These findings highlight the need for a comprehensive, multi-faceted approach to urban planning that integrates pedestrian infrastructure development, transit planning, and socio-economic considerations. By utilizing SHAP values, city planners and policymakers can identify high-risk areas and prioritize interventions that effectively reduce pedestrian crash risks. Future research should continue refining these models and investigating additional variables to further illuminate the complex factors influencing pedestrian safety, ensuring that urban environments are both functional and

safe for all users. The choice of spatial units significantly affects the study's ability to identify and interpret relationships between independent variables and pedestrian crash risks. Researchers must account for the modifiable areal unit problem, which demonstrates how different aggregation scales and zoning can produce varying analytical results. In CBG-level analysis, spatial models may help capture spatial influences more effectively.

4. INVESTIGATING THE SPATIAL VARIATION OF PEDESTRIAN CRASH RISKS IN URBAN CENSUS BLOCK GROUPS USING MULTISCALE GEOGRAPHICALLY WEIGHTED REGRESSION

Pedestrian safety is a critical concern in urban environments, where various socio-demographic and infrastructural factors contribute to crash risk. This study explores the spatial variability of these factors influencing pedestrian crash risk by applying a Multiscale Geographically Weighted Regression (MGWR) model. Using the EPDO score as the measure of pedestrian crash risk, the relationships between key explanatory variables are analyzed across an urban study area. The results showed that the percentage of two-plus-car households, employment and household entropy, auto-oriented network density, and transit frequency, exhibit stable and significant effects on pedestrian crash risk across the entire region. In contrast, variables like the percentage of high wages at home location and bike lane density demonstrate significant spatial variation, with localized effects that vary considerably by location. The findings underscore the importance of using spatially adaptive models like MGWR to capture both global and local relationships, offering insights for developing targeted, location-specific policy interventions aimed at enhancing pedestrian safety.

4.1. Motivation

In the literature on spatial analysis of traffic safety, several modeling techniques are employed to analyze crash occurrences, with an increasing focus on spatial correlation and heterogeneity. However, there are three notable gaps in macro-level pedestrian crash analysis. First, while GWR is frequently used to capture spatial heterogeneity, MGWR, a newer and more flexible approach, remains underexplored. MGWR allows for different variables to vary at different spatial scales, which provides a more nuanced understanding of the relationships

between crash occurrences and their influencing factors. Despite its potential to offer more precise modeling, its adoption in pedestrian crash analysis is still limited. Second, the use of EPDO as a response variable is another area with untapped potential. Most studies tend to focus on pedestrian crash counts or severity levels. EPDO, however, offers a more comprehensive measure by accounting for both the frequency and severity of crashes, providing a more balanced and detailed understanding of crash risks. Incorporating EPDO could yield deeper insights into the risk levels of different locations and their associated characteristics. Finally, the detailed inclusion of public transit and pedestrian infrastructure variables is often lacking in previous research. Many studies account for broad categories like land use and road network density, but finer details - such as transit frequencies, service areas, and the availability of sidewalks and bike lanes - are critical in understanding pedestrian crash dynamics. These variables are especially relevant for urban areas where pedestrian and transit interactions are more frequent, and their inclusion could significantly improve model accuracy. Addressing these gaps could lead to a more thorough and precise understanding of pedestrian crash risks at the macro level.

4.2. Data Preparation

Figure 17 presents the study area located within the city of Austin, Texas, as highlighted on the map of the state of Texas. The map is zoomed into Austin and provides a detailed view of the study region, segmented into different zones based on EPDO values. The EPDO is categorized into three classes, each represented by varying shades of green. The spatial distribution of EPDO values across the study area shows that higher EPDO values are concentrated in the central and southern parts of Austin, particularly along major transportation corridors. These areas may experience higher traffic volumes or more intense use of the road network, contributing to

increased property damage risk. This figure provides a geographical context for the study, illustrating the variability in EPDO values across the different regions of Austin, Texas.

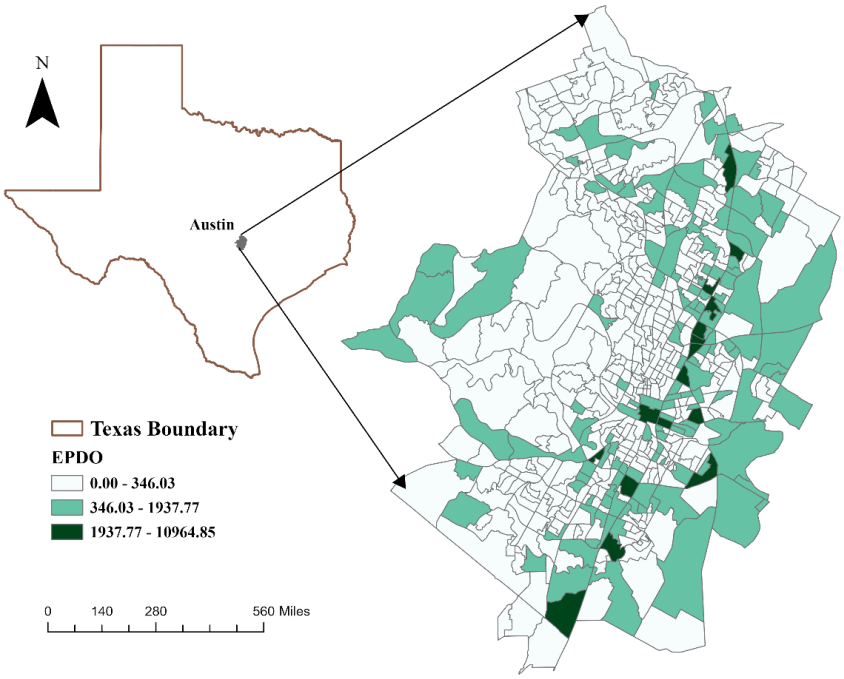


Figure 17. Study area

Table 5. Variable descriptives

Variables	Description	Source	Mean	STD	Min	50%	Max
Dependent Variables							
EPDO Score	Pedestrian crash EPDO Score	CRIS	285.143	596.1307	0	22.32	4534.86
Independent Variables							
Demographics							
Worker Pop Density	Density of workers (people/acre)	SLD	4.388	2.914	0.015	4.012	22.561
Resident Pop density	Density of residents (people/acre)	SLD	2.123	0.785	0.111	1.998	8.010
Pct of WrkAge Pop	Percentage of the working-age population (18-64)	SLD	0.678	0.107	0	0.670	1
Pct of Zero-car	Percentage of households that do not own a car	SLD	0.058	0.074	0	0.031	0.560
Pct of Two-plus-car	Percentage of households that own two or more cars	SLD	0.565	0.190	0	0.571	1
Pct Lowwage (home location)	Percentage of low-wage workers based on their home location	SLD	0.102	0.127	0.030	0.092	2.313
Pct Highwage (home location)	Percentage of high-wage workers based on their home location	SLD	0.282	0.187	0.051	0.277	2.971
Pct Lowwage (workplace)	Percentage of low-wage workers based on their workplace location	SLD	0.240	0.125	0	0.225	0.857
Pct Highwage (workplace)	Percentage of high-wage workers based on their workplace location	SLD	0.417	0.179	0	0.415	1
5-tier Employment Entropy	A measure of employment diversity across five employment categories	SLD	0.690	0.205	0	0.740	1
Employment and Household Entropy	A combined measure of diversity and balance in employment types and household characteristics	SLD	0.511	0.204	0	0.546	0.886
Household Workers per Job Equilibrium	A ratio indicating the balance between the number of working household members and the number of jobs available in a given area	SLD	0.337	0.325	0	0.312	0.995
Trip Characteristics							
Trip productions and trip attractions equilibrium	The balance between the number of trip productions and trip attractions	SLD	0.433	0.279	0	0.448	0.999
Employment within Transit Stop (0.5 mile)	Number of jobs located within a half-mile radius of a transit stop	SLD	0.028	0.136	0	0.000	0.998
Jobs within 45 minutes auto travel time	Number of jobs accessible within a 45-minute drive	SLD	161192.043	43417.161	31596	163734	247583
Working age population within 45 minutes auto travel	Number of working-age individuals accessible within a 45-minute drive	SLD	118703.520	23120.054	24937	122854	159515
Non Commute VMT	Vehicle miles traveled for non-commuting purposes per worker	SLD	4.679	1.717	1.319	4.516	11.556
Commute VMT	Vehicle miles traveled for commuting purposes per worker	SLD	18.085	6.805	4.564	17.969	70.986
Work related VMT	Vehicle miles traveled for work-related purposes per worker	SLD	22.765	7.971	6.982	22.504	82.542
Network Characteristics							
Auto Ntwrk Density	Density of the automobile network per square mile (41 mph or more)	SLD	2.250	3.408	0	0.499	19.129
Multi Ntwrk Density	Density of the automobile network per square mile (21 to 54 mph)	SLD	2.360	2.460	0	1.770	16.941
Ped Ntwrk Density	Density of the automobile network per square mile (20 mph or less)	SLD	14.083	6.276	0.841	14.546	42.029
Street Intersection Density	The number of street intersections per square mile	SLD	87.784	67.752	0	76.097	864.311
Public Transit Characteristics							
Wrk Day Runs	Frequency of transit runs per hour during workdays	GTFS	12.518	21.524	0	5.385	272.077
Wrk Night Runs	Frequency of transit runs per hour during work nights	GTFS	5.801	9.753	0	2.545	127.909
Wknd Day Runs	Frequency of transit runs per hour during weekend days	GTFS	9.465	15.344	0	3.885	188.308

Wknd Night Runs	Frequency of transit runs per hour during weekend nights	GTFS	4.939	8.191	0	2.182	109.455
Pct Transit Service Area	Percentage of the area covered by transit service	GTFS	0.390	0.360	0	0.319	1.621
Pedestrian Facility Characteristics							
Planned Sidewalk Density	Density of planned sidewalks (after April 2023)	City open data portal	2.039	1.702	0	1.700	7.368
Existing Sidewalk Density	Density of existing sidewalks (before April 2023)	City open data portal	2.609	1.907	0	2.365	10.375
Acceptable Sidewalk Density	Density of acceptable sidewalks	City open data portal	0.706	0.638	0	0.560	2.899
Deficient Sidewalk Density	Density of deficient sidewalks	City open data portal	1.604	1.395	0	1.352	8.428
Bike Lane Density	Density of bike lanes	City open data portal	1.850	1.090	0	1.688	10.735
Land Use Characteristics							
Pct Single Residential	Percentage of single-family residential use	Replica	0.488	0.197	0	0.486	0.964
Pct Multi Residential	Percentage of multi-family residential use	Replica	0.081	0.069	0	0.067	0.478
Pct Commercial Retail	Percentage of commercial retail purposes	Replica	0.083	0.065	0	0.063	0.326
Pct Commercial Office	Percentage of commercial office purposes	Replica	0.049	0.037	0	0.041	0.218
Pct Commercial Non-Retail Attraction	Percentage of commercial purposes that are not retail	Replica	0.017	0.018	0	0.011	0.115
Pct Mixed Use Residential	Percentage of mixed residential and other uses	Replica	0.022	0.029	0	0.012	0.231
Pct Mixed Use Commercial	Percentage of mixed commercial uses	Replica	0.052	0.057	0	0.033	0.333
Pct Mixed Use Industrial	Percentage of mixed industrial purposes	Replica	0.009	0.016	0	0.003	0.149
Pct Mixed Use Other	Percentage of other mixed-use purposes not classified under residential, commercial, or industrial.	Replica	0.020	0.027	0	0.012	0.231
Pct Industrial	Percentage of industrial purposes	Replica	0.024	0.048	0	0.002	0.480
Pct Civic Healthcare	Percentage of healthcare facilities	Replica	0.004	0.021	0	0.000	0.332
Pct Civic Education	Percentage of educational facilities	Replica	0.033	0.037	0	0.023	0.319
Pct Civic Other	Percentage of other civic purposes	Replica	0.009	0.015	0	0.004	0.169
Pct Transportation Utilities	Percentage of transportation infrastructure and utilities,	Replica	0.005	0.038	0	0.000	0.774
Pct Open Space	Percentage of open space	Replica	0.049	0.061	0	0.027	0.385
Pct Agriculture	Percentage of agricultural purposes	Replica	0.043	0.084	0	0.003	0.449
Land Use Entropy	A measure of the diversity and distribution of different land uses	Replica	1.618	0.421	0.187	1.660	2.442

4.3. Methodology

4.3.1. Ordinary Least Squares

The Ordinary Least Squares (OLS) model is a linear regression method used to estimate the relationship between a dependent variable and one or more independent variables (Liu et al. 2024). It assumes that the relationship between the variables is linear and aims to minimize the sum of the squared differences between the observed values and the predicted values. In the simplest form, the OLS model is written as:

$$y = X\beta + \varepsilon \tag{5}$$

Where, y is the dependent variable, X is a matrix of independent variables (predictors), β represents the regression coefficients to be estimated, ε is the error term.

4.3.2. Anselin Local Moran's I

Anselin Local Moran's I is a widely used spatial statistical measure that identifies local clusters or spatial outliers in a dataset by evaluating the degree of spatial autocorrelation at each individual location (Anselin 1995). It is an extension of Global Moran's I, which measures spatial autocorrelation across the study area and is especially useful for exploring spatial non-stationarity, where spatial relationships vary across locations. The primary goal of Anselin's Local Moran's I is to identify areas of significant spatial clustering (both high-high and low-low) as well as spatial outliers (high-low or low-high) in the data. This tool helps detect areas where a variable's values are more spatially dependent on neighboring values than would be expected under random spatial distribution. Inverse distance weights is used to define neighbors

The Local Moran's I for location i is calculated as:

$$I_i = (x_i - \bar{x}) \sum_j w_{ij}(x_j - \bar{x}) \tag{6}$$

Where, x_i is the value of the variable of interest at location i, $\bar{x}$ is the mean of the variable over all locations, x_j is the value at neighboring location j, w_{ij} is the spatial weight between locations i and j, typically defined by the spatial proximity or contiguity between these locations.

The Z-score for Local Moran's I is computed as:

$$\boldsymbol{Z}(I_i) = \frac{I_i - \boldsymbol{E}(I_i)}{\sqrt{Var(I_i)}} \tag{7}$$

Where, $\boldsymbol{E}(I_i)$ is the expected value of Local Moran's I under the null hypothesis (no spatial autocorrelation), $Var(I_i)$ is the variance of Local Moran's I.

4.3.3. Multiscale Geographically Weighted Regression (MGWR)

GWR is a spatial regression technique commonly employed in geography and other disciplines. It examines the local model of the variable or process of interest by fitting a regression equation to each feature in the dataset. These individual equations are constructed by considering the dependent and explanatory variables of features within the neighborhood of each target feature. For the GWR model, the specification is as follows (Bo 2018):

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_{ij} + \varepsilon_i, \tag{8}$$

Where y_i is the predicted value at location i. x_{ij} is the value of jth independent variable for location i, β_{ij} is a location-specific coefficient corresponding to x_{ij}, and ε_i is a random error at location i.

The MGWR is an extension of the GWR model. While GWR uses a fixed neighborhood around each feature to create a local linear regression model, MGWR allows each explanatory variable's neighborhood size to vary (ESRI 2023). This variation is essential as different variables may operate on different spatial scales. Some variables may have gradual changes across the study area, while others may exhibit rapid changes. By matching the neighborhood of each explanatory variable to its spatial scale, MGWR can more accurately estimate the

coefficients of the local regression model for interpretation and prediction. This study uses the GWR tool and MGWR tool available in ArcGIS Pro software to perform the modeling.

4.4. Results and Analysis

4.4.1. Variable Selection Using SHAP

Machine learning models have demonstrated significant efficiency in variable selection, particularly when dealing with large numbers of variables. In this study, CatBoost was employed to model the relationship between all the explanatory variables and the EPDO score. Following the modeling process, SHAP values were used to interpret the importance of each feature and assess the contribution of individual predictors to the model's output. Figure 18 presents the top 20 most important variables from the CatBoost model, as determined by SHAP values. The horizontal axis represents the SHAP value, which quantifies the effect of each variable on the EPDO score. Positive SHAP values indicate that the feature increases the prediction, while negative SHAP values suggest a decrease. The magnitude of the SHAP value reflects the strength of the variable's influence. The color gradient from green to pink represents the feature value (high or low) for each variable, with green indicating lower feature values and pink indicating higher feature values.

Auto Network Density and Pct Mixed Use Residential are among the highest-ranked variables, indicating that these factors have the most substantial impact on the model predictions. Pct Highway (home location) and Multi Network Density also have significant contributions, suggesting that transportation network density and land-use factors play critical roles in influencing the model output. Employment and Household Entropy and Pct of Zero-car households also show notable influence. Planned Sidewalk Density, Bike Lane Density, and

Land Use Entropy show varying effects, with mixed contributions (both positive and negative) across the range of SHAP values.

Figure 18. Top 20 important variables based on the CatBoost model

4.4.2 OLS Modeling Results

After variable selection, the chosen predictors were included in the OLS modeling. Variables with a Variance Inflation Factor (VIF) score higher than 3, indicating potential multicollinearity issues (Kock and Lynn 2012), were removed from the model to ensure more reliable results. Since transit frequencies are highly correlated, in further analysis, the sum of transit frequency by time is used as a new measure. The final set of variables, all with VIF values less than 3, as presented in Table 6. The model achieved an adjusted R-squared of 0.27, indicating that

approximately 26.6% of the variance in the EPDO score is explained by the included variables. Additionally, the model's Akaike Information Criterion (AIC) is 2695.70, providing a measure of the model's relative goodness-of-fit while accounting for model complexity.

The proportion of two-plus-car households is negatively associated with the EPDO score, with a coefficient of -1.642 and a p-value of 0.012, indicating that higher percentages of multi-car households are linked to lower values of the EPDO score. Additionally, Employment and Household Entropy show a strong positive relationship (coefficient = 2.348, p<0.001). Employment and Household Entropy is a measure of the diversity or balance in the distribution of employment types and occupied housing units within a given area. A higher entropy value suggests that the area has a more balanced mix of different types of employment (across the five-tier employment categories) and households. Conversely, a lower entropy value indicates that one type of employment or household dominates.

In addition, Auto-oriented Network Density is positively associated with the dependent variable (coefficient = 0.139, p<0.001), meaning that locations with denser auto-oriented networks are likely to have higher dependent variable values. Conversely, Pct Highway (home location) exhibits a negative association (coefficient = -1.751), indicating that higher highway percentages near home locations correspond to lower dependent variable values. Transit Frequency is another strong predictor with a positive and significant effect (coefficient = 0.012, p<0.001), highlighting the association of public transit frequency in this model. However, several variables, such as Planned Sidewalk Density, Bike Lane Density, and Pct Single Residential, are not statistically significant in the OLS model. Despite this, these variables are retained because they could contribute to explaining local variability in the MGWR analysis. Variables that are

insignificant globally might still have spatially varying effects and removing them prematurely could overlook important localized relationships.

Table 6. OLS modeling results

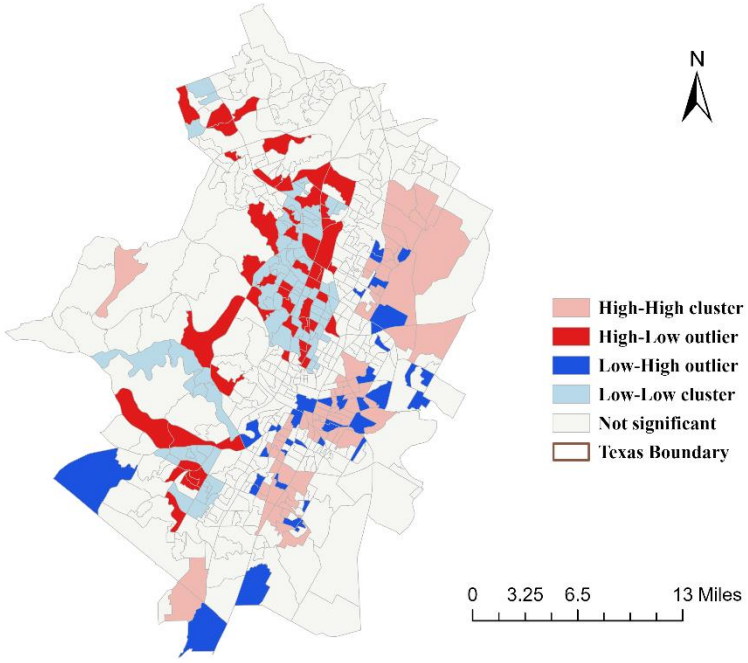
Variable	Coefficient	StdError	t-Statistic	P-value	VIF
Intercept	3.053	0.746	4.095	0.000*	--------
Pct of Zero-car	1.167	1.498	0.779	0.436	1.527
Pct of Two-plus-car	-1.642	0.649	-2.531	0.012*	1.875
Employment and Household Entropy	2.348	0.480	4.890	0.000*	1.177
Auto Ntwrk Density	0.139	0.030	4.609	0.000*	1.297
Multi Ntwrk Density	0.032	0.043	0.752	0.452	1.343
Pct Highwage (home location)	-1.751	0.507	-3.452	0.001*	1.102
Planned Sidewalk Density	-0.049	0.061	-0.813	0.417	1.311
Existing Sidewalk Density	-0.047	0.058	-0.807	0.420	1.520
Bike Lane Density	-0.145	0.111	-1.309	0.191	1.781
Pct Single Residential	-0.229	0.720	-0.318	0.751	2.477
Pct Commercial Office	-2.495	3.101	-0.805	0.421	1.572
Pct Mixed Use Residential	6.166	3.283	1.878	0.061	1.135
Pct Mixed Use Industrial	7.289	7.701	0.946	0.344	1.810
Pct Industrial	2.431	2.705	0.899	0.369	2.069
Pct Civic Other	-5.317	6.312	-0.842	0.400	1.045
TransitFrq	0.012	0.002	6.029	0.000*	1.385
Number of Observations	604				
Multiple R-Squared	0.285				
Akaike's Information Criterion (AICc)	2695.700				
Adjusted R-Squared	0.266				

4.4.3. Anselin Local Moran's I

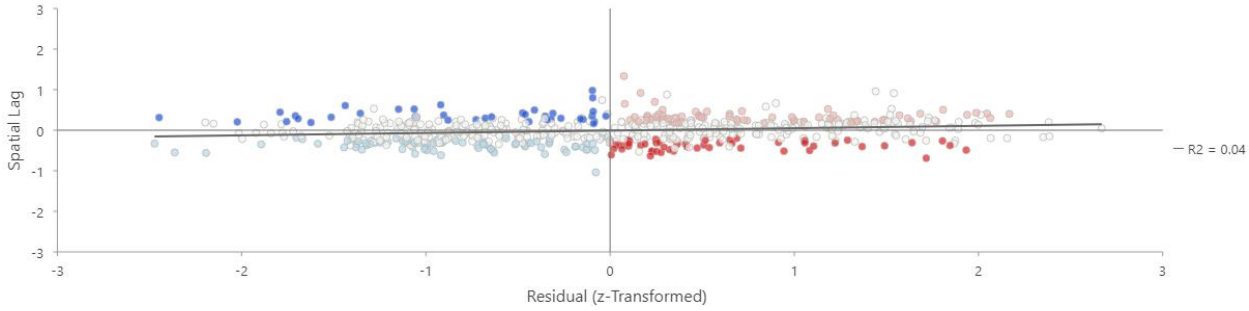
The Anselin Local Moran's I analysis was conducted on the residuals of the OLS model to investigate the presence of spatial non-stationarity. If significant clusters are identified in the analysis, it indicates the existence of spatial non-stationarity. In other words, the relationships between the variables vary across space, and the model does not fully capture these localized variations. Figure 19 a) illustrates the spatial distribution of significant clusters and outliers across the study area. Areas marked in red represent high-high clusters, indicating locations where high residuals are surrounded by other high residuals. These areas suggest regions where the OLS model systematically overpredicts or underfits, leaving a concentration of high residual values. The blue areas show low-high outliers, where low residuals are surrounded by high residuals, indicating spatial outliers with lower than expected outcomes in their neighboring

context. Low-low clusters (light blue) represent areas where low residuals are surrounded by other low residuals, reflecting regions where the OLS model consistently underpredicts. High-low outliers (dark red) are locations where high residuals are surrounded by low residuals, representing localized spatial anomalies.

Figure 19 b) visualizes the relationship between the residuals and their spatial lags, categorized into the four types of clusters and outliers. The scatterplot shows positive spatial autocorrelation, particularly with the red dots clustered in the upper right quadrant (high-high) and the blue dots in the bottom left (low-low). The R-squared value (0.04) indicates a minor spatial autocorrelation in the residuals. The presence of significant clusters and outliers suggests that the OLS model does not fully capture the spatial variation in the data, warranting further spatial modeling such as MGWR to address the spatial non-stationarity.



a)



b)

Figure 19. Anselin Local Moran's I of OLS residuals, a) cluster outliers, b) Moran's scatterplot

4.4.4. MGWR Modeling Results

The MGWR model results, presented in Table 7, demonstrate the spatial variability in the relationships between the explanatory variables and EPDO score. Prior to modeling, the explanatory variables were scaled, and non-significant variables were removed through a backward elimination process. Unlike the OLS model, MGWR allows the coefficients of the explanatory variables to vary across space, capturing localized effects that a global model like OLS cannot detect. The model achieved an R-squared value of 0.395 and an Adjusted R-squared of 0.362, indicating that the MGWR model explains approximately 36.2% of the variance in the dependent variable - a clear improvement over the OLS model. Additionally, the AICc value of 1478.498 is significantly lower than that of the OLS model, further confirming that the MGWR provides a better fit while accounting for model complexity.

Among the key variables, the Pct of Two-plus-car Households has a mean coefficient of -0.120, indicating a negative relationship with pedestrian crash risk across the study area. This relationship is spatially stable, with 99.83% of the features being statistically significant, and the variable maintains influence across the entire region (100% of neighbors). The spatial consistency of this variable reflects its widespread influence on the model. Employment and Household Entropy exhibits a positive mean coefficient of 0.222, indicating that areas with

greater entropy (i.e., a more balanced mix of employment and household types) are associated with higher values of pedestrian crash risk. This variable is significant for 100% of the features, indicating that it plays an important and consistently positive role across the entire study area.

Similarly, Auto-oriented Network Density has a mean coefficient of 0.210, with very little spatial variation (standard deviation = 0.003), and is significant for all features (100%). This suggests that the presence of a denser auto network is consistently associated with higher values of crash risk across space. In contrast, Pct Highway (home location) shows greater spatial variability, with a mean coefficient of -0.203 and a wide range from -0.608 to 0.025. This indicates that the relationship between highway proximity and the dependent variable is not uniform across space, and is statistically significant for only 27.48% of the features, highlighting that highways have localized effects.

Bike Lane Density demonstrates substantial spatial variation, with a mean coefficient of -0.095 and a standard deviation of 0.090. Only 19.87% of the features are statistically significant, and the range of coefficients (-0.303 to 0.038) indicates that this variable has different effects depending on the location. Transit Frequency shows a strong and positive relationship with the dependent variable, with a mean coefficient of 0.204. Like employment entropy and auto network density, it is significant for all features, indicating its widespread influence across the study area.

Overall, the MGWR results show that while some variables, such as Pct of Two-plus-car Households, Employment and Household Entropy, Auto Network Density, and Transit Frequency, exhibit relatively stable relationships across space, others, such as Pct Highway (home location) and Bike Lane Density, show more localized, spatially varying effects. This

underscores the importance of using MGWR to capture both global and local spatial relationships that are not fully captured by the OLS model.

Table 7. MGWR modeling results

Explanatory Variables	Mean	Standard Deviation	Min	Median	Max	Neighbors (% of Features)[a]	Significant (% of Features)[b]
Intercept	-0.048	0.227	-0.456	-0.064	0.390	114 (18.87)	130 (21.52)
Pct of Two-plus-car	-0.120	0.015	-0.142	-0.121	-0.092	604 (100.00)	603 (99.83)
Employment and Household Entropy	0.222	0.015	0.199	0.222	0.248	604 (100.00)	604 (100.00)
Auto Ntwrk Density	0.210	0.003	0.201	0.211	0.215	604 (100.00)	604 (100.00)
Pct Highwage (home location)	-0.203	0.166	-0.608	-0.156	0.025	166 (27.48)	166 (27.48)
Bike Lane Density	-0.095	0.090	-0.303	-0.057	0.038	281 (46.52)	120 (19.87)
Transit Frequency	0.204	0.005	0.198	0.201	0.217	604 (100.00)	604 (100.00)
R-Squared		0.395					
Adjusted R-Squared		0.362					
AICc		1478.498					
Sigma-Squared		0.638					
Sigma-Squared MLE		0.605					

Figure 20 a) displays the results of the Auto Network Density variable from the MGWR analysis, focusing on the spatial variation in the relationship between auto network density and the dependent variable across the study area. The map on the right depicts the spatial distribution of auto network density value itself. Darker red areas represent regions with the highest auto network density, while lighter shades indicate lower network densities. The highest densities are observed in the central and southern parts of the study area, where auto-oriented infrastructure is most concentrated. The map on the left shows the spatial distribution of the significance and coefficient values for auto network density. Areas marked with crosshatching represent locations where the relationship between auto network density and the dependent variable is not statistically significant. Regions without crosshatching indicate areas where the coefficient for auto network density is significant. The spatial variation in the coefficients suggests that the strength of the relationship between auto network density and the dependent variable is not constant across the region. The southern parts of the study area exhibit higher positive

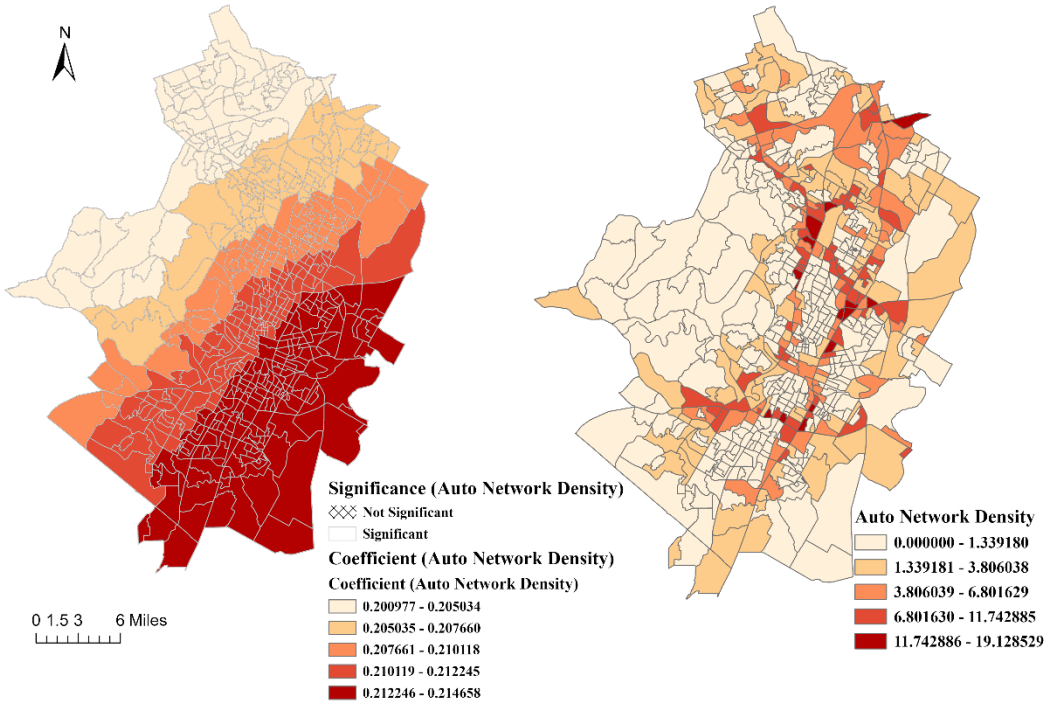
coefficients, meaning that auto network density has a higher association with pedestrian crash risk in these regions. Conversely, northern regions display weaker associations.

Figure 20 b) presents the results of the Bike Lane Density variable from the MGWR analysis. The map on the right shows the spatial distribution of bike lane density across the study area. Regions with higher bike lane density are particularly concentrated in the central Austin area. The map on the left illustrates the spatial distribution of the significance and coefficient values for bike lane density. The darkest areas in the central and southern regions are crosshatched meaning not significant. In contrast, lighter purple regions, which are in the northern part of the study area, show stronger negative associations. This suggests that the relationship between bike lane density and the dependent variable is spatially variable, with the strongest negative impacts observed in the north.

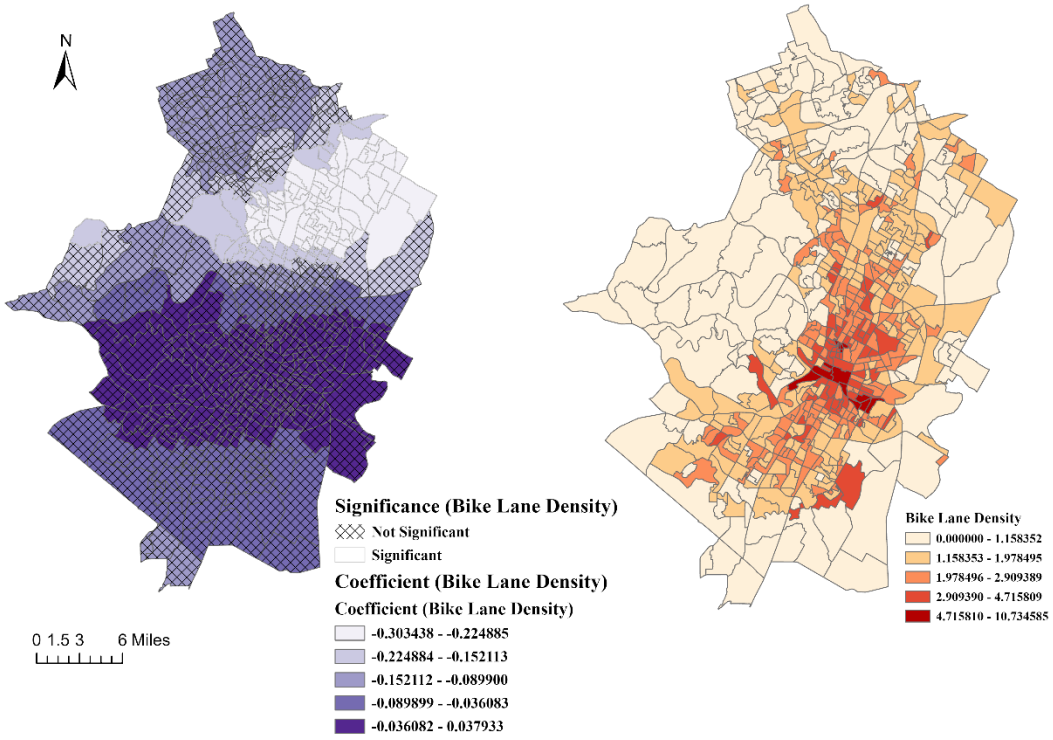
Figure 20 c) shows the results of the Employment and Household Entropy. The map on the right displays the spatial distribution of employment and household entropy values. Darker red areas represent regions with higher entropy, indicating a more diverse and balanced mix of employment types and household characteristics. The map on the left highlights the spatial distribution of significance and the coefficient values for employment and household entropy. In regions with darker shades, an increase in employment and household entropy corresponds to a stronger positive relationship with crash risk. In contrast, northern regions show weaker associations. Figure 20 d) shows the results of the Percentage of Highwage at Home Location (Pct Highwage HL) variable. The coefficients suggest that higher percentages of high-wage earners are associated with lower values of pedestrian crash risk. Conversely, lighter areas suggest a weaker or even positive relationship. The association is significant only in a small part of the study area.

The results of the Transit Frequency variable are presented in Figure 20 e). The map on the right displays the spatial distribution of transit frequency across the study area. Regions with higher transit frequency are concentrated in the central and southern parts of the study area. In regions where transit frequency is higher, the positive relationship between transit frequency and the dependent variable is less pronounced. The spatial variation in coefficients shows that northern regions, where transit frequency is lower, exhibit stronger positive associations. This suggests that in these areas, an increase in transit frequency corresponds to a larger positive impact on crash risk. In contrast, regions with higher frequency play a lesser role in influencing the dependent variable in those areas. These maps demonstrate that transit frequency has a more significant impact in regions with less frequent public transportation, particularly in the northern parts of the study area. In contrast, areas with higher transit frequency show weaker associations, emphasizing the spatial heterogeneity of transit's influence on the dependent variable.

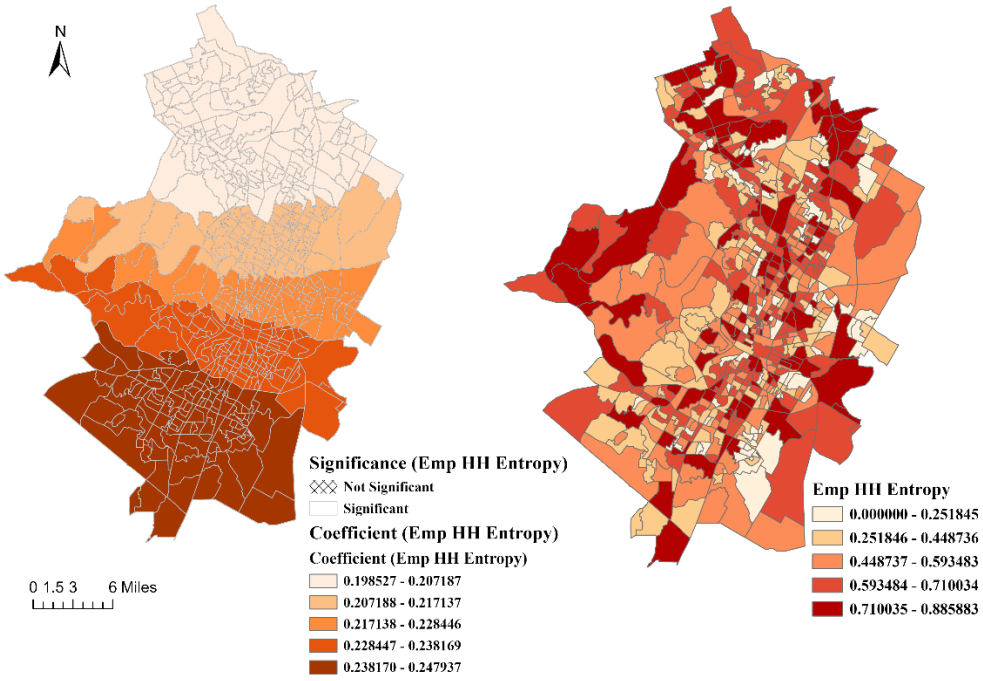
Figure 20 f) presents the results of the Percentage of Households with Two Plus Cars (Pct of Two Plus). A higher percentage of households own two or more cars is concentrated in the outliers of the study area. The map on the left shows the spatial distribution of significance and coefficient values for the Pct of Two Plus variable. The coefficient indicates stronger negative associations. The spatial variation in coefficients indicates that southern regions exhibit a weaker negative impact on crash risk. In contrast, northern areas, show a stronger impact, suggesting that two-plus car ownership plays a stronger role in those regions.



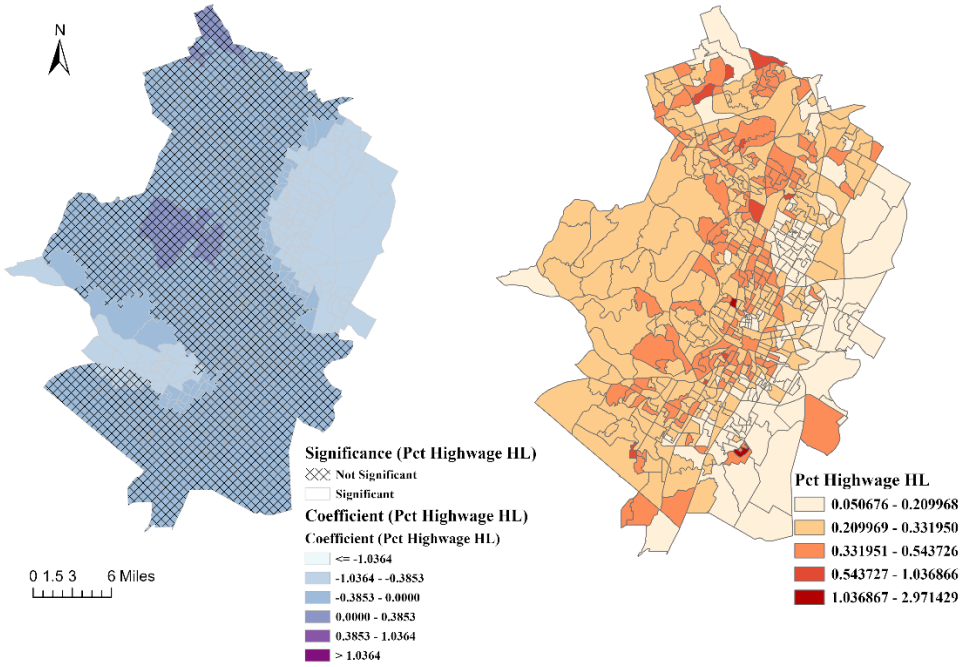
a) Auto-oriented Network Density



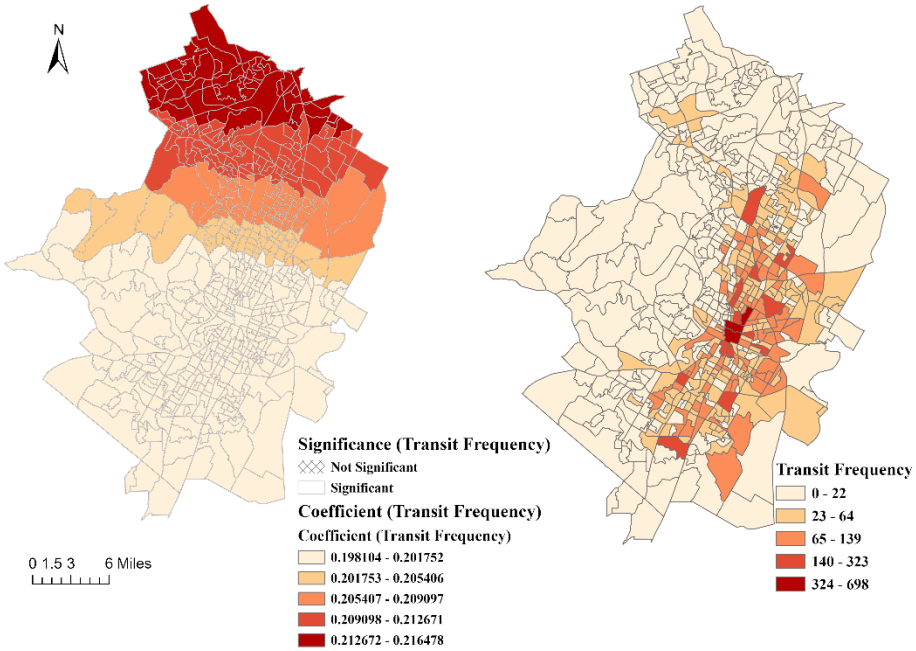
b) Bike Lane Density



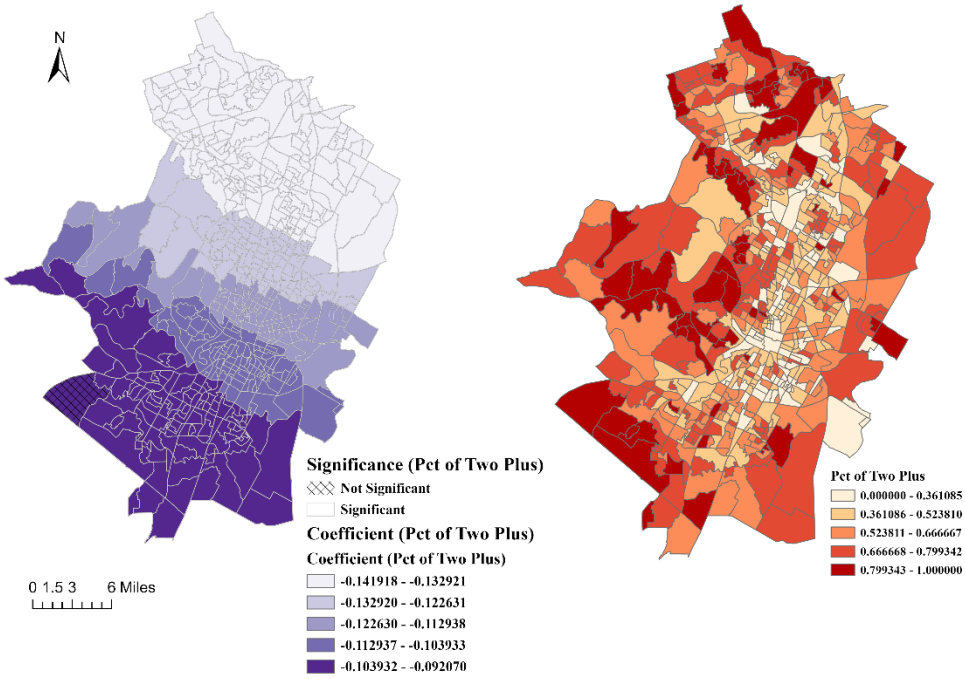
c) Employment and Household Entropy



d) Percentage of Highwage at Home Location



e) Transit Frequency



f) Percentage of Households with Two Plus Cars

Figure 20. Coefficient distribution of MGWR model

4.5. Summary

The results of the MGWR model offer crucial insights into the spatial variability of the relationships between key explanatory variables and pedestrian crash risk. By allowing the

coefficients of explanatory variables to vary across space, the MGWR model reveals localized effects that a global model, such as OLS, cannot adequately capture. Several variables, including the percentage of two-plus-car households, employment and household entropy, auto-oriented network density, and transit frequency, display stable and significant relationships with pedestrian crash risk across the study area. These findings underscore the importance of socio-demographic and infrastructural factors in shaping crash risks on a regional scale. However, other variables, such as the percentage of high-wage households and bike lane density, exhibit more localized and spatially varying effects, highlighting that certain factors may influence crash risk differently depending on their geographical context.

The negative relationship between the percentage of two-plus-car households and crash risk points to the possibility that car dependency may reduce pedestrian exposure and thus lower crash risk. On the other hand, the positive influence of employment and household entropy, auto network density, and transit frequency suggest that areas with greater land-use mix, higher traffic volumes, and more frequent transit services tend to experience heightened pedestrian crash risks. These relationships likely stem from the increased interaction between pedestrians and vehicles in denser and more transit-oriented environments.

The spatial variability observed in the effects of the percentage of high-wage households and bike lane density further reinforces the need for localized policy interventions. While higher-income locations may be associated with lower pedestrian crash risk in some areas, their impact varies significantly, requiring targeted safety measures. Similarly, the mixed effects of bike lane density indicate that the effectiveness of such infrastructure in reducing crash risks may depend on other contextual factors, such as the surrounding land use or the quality of the bike lanes themselves.

In conclusion, this study demonstrates the advantages of using the MGWR model for macro-level pedestrian crash risk analysis, revealing both stable and spatially varying relationships between key variables and pedestrian safety outcomes. Policymakers and urban planners can use these findings to develop more effective, location-specific strategies to reduce pedestrian crash risks, particularly in areas where certain factors exert strong local effects. The findings emphasize that comprehensive crash risk assessments should account for both global and local spatial patterns to better inform safety improvements in urban environments. The study has certain limitations. First, the analysis is conducted at the census block group level, and due to the Modifiable Areal Unit Problem (MAUP), the influence of factors may vary when analyzed at different spatial scales. Future research will address the MAUP issue by aggregating data across different spatial units to evaluate its impact on the findings.

5. INDIVIDUAL-LEVEL PEDESTRIAN CRASH CLUSTERING PATTERN ANALYSIS

Pedestrian crashes represent a critical traffic safety issue, often resulting in fatal outcomes and raising significant equity concerns. This study analyzed detailed records of pedestrian-involved crashes in California from 2018 to 2021, employing a novel clustering framework enhanced by an explainable AI-driven technique (CatBoost). The method significantly enhanced interpretability by effectively capturing complex non-linear relationships and interactions among features. The results indicate that pedestrian impairment and lighting conditions are pivotal in severe crash outcomes, while broader societal and demographic factors are more substantially associated with less severe cases. Non-injury pedestrian crashes tend to occur in less vulnerable, more resilient communities, whereas fatal crashes are more common in vulnerable communities with poor lighting and incomplete pedestrian infrastructure, particularly when pedestrians are under the influence of drugs or alcohol. The analysis also identified three clusters for fatal, two for injury, and two for non-injury crashes. The findings underscore the necessity for developing comprehensive safety measures that not only address situational risks but also consider the broader societal conditions. This approach advocates for integrated safety strategies that are sensitive to both the immediate context of crashes and the underlying socio-economic landscape.

5.1. Motivation

5.2. Data Preparation

This study utilizes crash data collected from the HSIS, a comprehensive multistate database comprising detailed crash data, roadway inventories, and traffic volume information across selected states (United States Department of Transportation 2024). This study covers four years, from 2018 to 2021, and focuses on California. It includes data on crash details, unit-level crash information, and roadway inventory shapefiles. The reason for using data from California is that

it offers detailed person-level information in pedestrian crash analysis, such as pedestrian actions. Additionally, it allows for a brief comparison between Texas and California. Figure 21 illustrates the data preparation process. Data processing was methodically carried out on a yearly basis. Taking 2018 as an illustrative example, each pedestrian crash for that year was merged with the corresponding crash data utilizing the crash ID as a key identifier. Subsequently, a spatial join was performed between the 2018 roadway file and the crash data, thus combining information from three distinct sources into a consolidated point-format file. The ensuing step involved extracting pedestrian-involved crash instances, specifically where the 'party_group' was marked 'Yes'. Additionally, variable selection was conducted based on the data quality and relevance to the study's aims. To gain a macro-level perspective on crash locations, the cleaned pedestrian crash points underwent intersection analysis with the Equitable Transportation Community (ETC) dataset (US Department of Transportation 2024). This dataset provides a detailed focus on transportation equity metrics, along with extensive socioeconomic and demographic data at the Census Tract level, enabling a broader understanding of the factors contributing to pedestrian crash crashes.

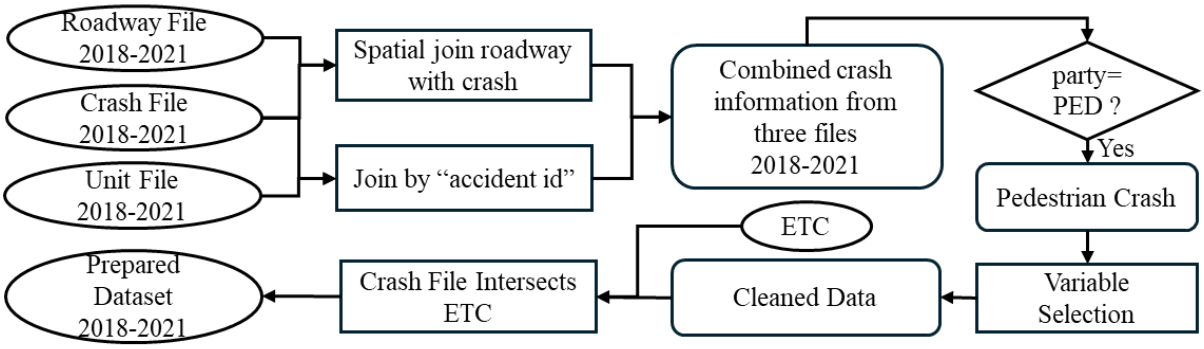


Figure 21. Data preparation process

In total, four levels of information were considered: unit level, crash level, roadway information, and zonal information. From these, 77 variables were selected to develop models

predicting the severity of pedestrian crashes. Each record represented a unique individual-level occurrence, capturing the distinct characteristics and severity outcomes for each person involved in a crash. This individual-level approach acknowledges the complexity of crash events, which often involve multiple vehicles and pedestrians, each affected to varying extents. The final cleaned dataset comprises 5,446 instances of pedestrian-involved crashes. The categories of pedestrian injury severity include three distinct classes: fatal, injury, and no injury. The distribution of these categories is as follows: injury accounts for 3,840 cases (71%), fatal for 1,227 cases (23%), and no injury for 379 cases (7%).

Previous pedestrian studies have examined variables associated with pedestrians (Baireddy, Zhou, and Jalayer 2018; Mafi, AbdelRazig, and Doczy 2018; Sasidharan, Wu, and Menendez 2015; Sun, Sun, and Shan 2019), crash characteristics (Das et al. 2019; Hossain et al. 2022), road characteristics (Batouli et al. 2020; Mafi, AbdelRazig, and Doczy 2018; Sasidharan, Wu, and Menendez 2015), socio-economic and build-environmental factors (Haddad et al. 2023; M. S. Rahman et al. 2019). This study covers all four aspects. However, pedestrian demographic information such as age and gender are not included due to the lack of available data. Table 8 illustrates the descriptive statistics of the variables used in the models. Pedestrian sobriety impairment includes ten categories, with approximately 6% of pedestrians involved found to be impaired by alcohol. Pedestrian actions include roadway, including shoulder (47%), crossing at intersection crosswalks (26%), crossing not at crosswalks (20%), not in roadway (4%), crossing at non-intersection crosswalks (2%). Location types include highways (6%), intersections (22%), and ramps (18%). Environmental justice variables, including transportation insecurity, environmental burden, health vulnerability, social vulnerability, and climate and disaster risk burden, were collected from ETC.

Table 8. Descriptive statistics of the categorical variables

Variables	No.	%
Dependent Variables		
Pedestrian Injury Severity (ped_inj_code)		
Injury	3840	70.51
Fatal	1227	22.53
NoInjury	379	6.96
Pedestrian Characteristics		
Party sobriety (impaire_1)		
Had Not Been Drinking (0%)	2782	51.08
Impairment Unknown	1354	24.86
Not Stated	374	6.87
HBD_ Under Influence	317	5.82
Not Applicable	274	5.03
HBD_ Impairment Unknown	203	3.73
Under Drug Influence	85	1.56
HBD_ Not Under Influence	52	0.95
Other Physical Impairment	4	0.07
Driver Fatigue	1	0.02
Pedestrian Action (ped_actn)		
Roadway - Include Shoulder	2591	47.58
Xing Xwalk - Intersection	1421	26.09
Xing - Not Xwalk	1091	20.03
Not in Roadway	241	4.43
Xing Xwalk - Not Intersection	101	1.85
Approach/Leave School Bus	1	0.02
Party drug or physical impairment (impairt _2)		
Not Stated	5322	97.72
Under Drug Influence	119	2.19
Other Physical Impairment	3	0.06
Driver Fatigue	2	0.04
Crash Characteristics		
Primary collision factor (cause)		
Failure to Yield	2418	44.40
Other Violations	1369	25.14
Speeding	668	12.27
Improper Turn	464	8.52
Influence of Alcohol	236	4.33
Other Than Driver	165	3.03

Variables	No.	%
Number of Vehicles (numvehs)		
1 Veh	4152	76.24
2-4 Vehs	1260	23.14
5-7 Vehs	34	0.62
Road/Location Characteristics		
Weather Condition (weather)		
Clear	4527	83.13
Cloudy	641	11.77
Raining	210	3.86
Fog	28	0.51
Snowing	23	0.42
Not Stated	7	0.13
Other	6	0.11
Wind	4	0.07
Lighting Condition (light)		
Dark - Street Light	1927	35.38
Daylight	1890	34.70
Dark - No Street Light	1412	25.93
Dusk/Dawn	178	3.27
Dark - Inoperative Street Light	27	0.50
Not Stated	11	0.20
Dark - Not Stated	1	0.02
Traffic Operation (trf_oper)		
No Controls Present	3595	66.01
Control Functioning	1831	33.62
Control Not Functioning	9	0.17
Not Stated	7	0.13
Controls Obscured	4	0.07
Roadway condition (rd_def)		
No Unusual Condition	5132	94.23
Construction - Repair Zone	152	2.79
Obstruction on Road	74	1.36
Other	31	0.57
Not Stated	22	0.40
Holes, Ruts	13	0.24
Reduced Road Width	10	0.18
Flooded	7	0.13

Variables	No.	%
Median Type (med_type)		
Paved Median	2040	37.46
Unpaved Median	1189	21.83
Striped	779	14.30
Continuous Left Turn Lane	565	10.37
Two-Way Left Turn Lane	492	9.03
Separate Structure	98	1.80
Other	85	1.56
Separate Grades	82	1.51
Railroad	79	1.45
Paved Area Occasional Traffic Lane	17	0.31
Sawtooth - Paved	6	0.11
Railroad & Bus Lanes	5	0.09
Separate Grades w/Retaining Wall	4	0.07
Reversible Peak Hour Lane(s)	3	0.06
Contains Reversible Pk Hr Ln(s)	1	0.02
Bus Lanes	1	0.02
Median Variance (med_var)		
No Variance	4446	81.64
Variable	866	15.90
Over 100’ Median & No Var.	134	2.46
Number of Lanes (no_lanes)		
2-4 Lanes	2438	44.77
6-8 Lanes	2095	38.47
10 or more	912	16.75
1 Lane	1	0.02
Land Use (rururb)		
Urbanized	4304	79.03
Rural	687	12.61
Urban	455	8.35
Terrain (terrain)		
Flat	3883	71.30
Rolling	1290	23.69
Mountainous	273	5.01
Road Surface (rdsurf)		
Dry	4920	90.34
Wet	454	8.34

Unknown	101	1.85	Loose Material	5	0.09	Snow, Icy	52	0.95
Not Stated	15	0.28	*Roadway Classification (roadwcls)*			Not Stated	18	0.33
Improper Driving	8	0.15	Freeways	3076	56.48	Slippery	2	0.04
Following Too Close	2	0.04	Multilane Roads	1746	32.06	*Highway Group (hwy_grp)*		
Crash Type (acctype)			2 Lane Roads	624	11.46	Independent Alignment- Right	4607	84.59
Auto-Pedestrian	4289	78.76	*Access Control (access)*			Undivided Highway	779	14.30
Rear-End	428	7.86	Freeway	2917	53.56	Divided Highway	60	1.10
Sideswipe	232	4.26	Conventional	2334	42.86	*Location Type (loc_typ)*		
Broadside	180	3.31	Expressway	159	2.92	Highway	3294	60.48
Hit-Object	143	2.63	One-Way City Street	36	0.66	Intersection	1187	21.80
Head-On	85	1.56	*Design Speed (desg_spd)*			Ramp	965	17.72
Other	52	0.95	60-70	3809	69.94	*Divided Roadway (divided)*		
Not-Stated	25	0.46	40-55	1270	23.32	Yes	4667	85.70
Overturn	12	0.22	25-35	367	6.74	No	779	14.30

Table 9. Descriptive statistics of the continuous variables

Variables	Variable code	Mean	STD	Min	Max
Roadway characteristics (Segment level information)					
Annual average daily traffic	aadt	102307.38	91884.36	260	719463
Average lane width	lanewid	11.86	1.46	0	25
Left shoulder width (outside)	lshl_wd2	7.87	3.38	0	33
Lef shoulder width (inside)	lshldwid	4.04	4.78	0	35
Median width	medwid	24.91	24.47	0	99
Right shoulder width (inside)	rshl_wd2	4.13	4.77	0	35
Right shoulder width (outside)	rshldwid	7.96	3.22	0	24
Travel way width (left/decreasing)	surf_wid	35.79	17.69	0	108
Environmental justice variables (Census Tract level information)					
Total Population	totPop	4439.32	2505.46	0	39373
Percent of households with no car	pctnvh	24.95	247.35	0	4902.50
Average commute time to work	avgcmm	28.72	7.40	1.2	67.43
Frequency of Transit Services per Sq Mi	trnfrq	34.05	109.68	0	1663.19
Jobs within a 45-min Drive	jb45dr	119708.03	145988.20	6.90	828440.62
Estimated Average Drive Time to Points of Interest (min)	drvpoi	63.80	1612.45	1.407	53926.23
Estimated Average Walk Time to Points of Interest (min)	wlkpoi	119.45	168.41	4.26	1255.53
Calculated average annual cost of Transportation as percent of household income	avghht	1.27	15.22	0.04	378.99
Traffic Fatalities per 100,000 people	ftltsp	10.62	30.11	0	370.98
Ozone level in the air	ozn	47.73	11.97	26.34	74.40
Particulate Matter 2.5 (PM2.5) level in the air	pm25	11.47	2.30	5.676	17.75
Diesel particulate matter level in air	dslpm	0.32	0.19	0	1.42
Air toxics cancer risk	cncrtx	30.16	9.73	0	200

Percent of tract within 1 mile of known hazardous sites	hzrdst	73.29	38.29	0	100
Percent of tract within 1 mile of known Toxics Release sites	txcrls	45.58	42.55	0	100
Percent of tract within 1 mile of known Treatment and Disposal Facilities	trtdsp	4.21	16.93	0	100
Percent of tract within 1 mile of known Risk Management Plan Sites	rskmns	23.39	33.97	0	100
Percent of houses built before 1980	ppr80h	60.55	25.01	0	100
Percent of tract within 1 mile of high volume roads	hghvlr	77.29	34.07	0	100
Percent of tract within 1 mile of railways	rlwys	45.49	42.36	0	100
Percent of tract within 5 miles of airports	arprts	77.09	37.15	0	100
Percent of tract within 3 miles of ports	prts	4.09	19.09	0	100
Percent of tract that intersects with a Watershed containing impaired water(s)	imprdw	73.54	39.76	0	100
Asthma prevalence	asthm	6.81	3.27	0	14.9
Cancer prevalence	cncr	3.58	2.04	0	13
High blood pressure prevalence	bldprs	19.15	9.51	0	49.60
Diabetes prevalence	dbts	7.12	3.96	0	23.39
Poor mental health prevalence	mntlhl	10.67	5.38	0	22.80
Percent of population with Income below 200% of poverty level	ppvrty	32.97	18.25	0	100
Percent of people age 25+ with less than a high school diploma	pndplm	18.24	14.23	0	73.8
Percent of people age 16+ unemployed	pnmply	4.28	2.69	0	19.6
Percent of total housing units that are renter-occupied	phstnr	48.33	23.39	0	100
Percent of occupied houses spend 30% or more on housing with less than 75k	phbrd7	41.03	11.44	0	100
Percent of population uninsured	pnnsrd	7.92	5.67	0	50.74
Percent of households with no internet subscription	pnntrn	13.37	9.47	0	100
GINI Index	endnql	0.49	0.82	0.1302	22.24
Percent of population 65 years or older	p65ldr	14.83	8.41	0	100
Percent of population 17 years or younger	p17yng	22.17	7.39	0	55.02
Percent of population with a disability	pdsb	11.96	5.93	0	100
Percent of population (age 5+) with limited English proficiency	plmng	9.10	8.84	0	100
Percent of total housing units that are mobile homes	pmblhm	6.23	11.89	0	100
Estimated annualized loss due to disasters	annlls	2946543.92	4795553.42	4014.13	55002605.85
Increase in number of days over 90deg by mid-century	extht	16.80	13.25	0.01	53.41
Number of days exceeding 99th percentile of precip by mid-century	extrmp	4.22	2.44	0.01	17.33
Percent change in number of days with less than 0.01 inches of precip	drghtd	3.38	2.23	-2.09	12.55
Percent of tract inundated by 0.5 sea level increase by 2100	pctnnd	0.11	0.58	0	8.44
Average Percent Land classified as Impervious Surface per Tract	mnmp	44.13	27.36	0.08	97.30

5.3. Methodology

The study begins with a prepared pedestrian crash dataset that includes 5,446 pedestrian crash records. Each record has 77 different attributes. Pedestrian injury severity is in three categories: fatal, injury, and no injury. The next phase involves hyperparameter tuning and model selection for various tree-based machine learning models. The models considered in this study include CatBoost, LightGBM, XGBoost, Random Forest, and Gradient Boosting. Through hyperparameter tuning, the study identifies the optimal settings for each model to ensure the best possible performance.

Using the tuned CatBoost model, the study employs the Tree SHAP algorithm to analyze the influence of each variable on the model's predictions. This results in a SHAP value table for each severity level - fatal, injury, and no injury - with each table containing SHAP values for the corresponding 5,446 incidents across the 77 attributes. The study then proceeds to a global explanation phase, where it examines the overall importance of variables across different injury severity levels, providing insights into which factors are most influential in fatal, injury, and no-injury crash outcomes.

Next, the study selects SHAP values by class labels, narrowing down the analysis to instances within each severity category - 1227 fatalities, 3840 injuries, and 379 no injuries - each detailed with 77 attributes. In the subsequent phase, hierarchical clustering is applied to the selected SHAP values for each severity category. This analysis reveals distinct clustering patterns, identifying three clusters within fatal incidents, and two clusters each within injury and no-injury incidents. These patterns indicate the presence of subgroups within each injury severity category, characterized by their unique variable importance profiles as determined by the Tree SHAP algorithm. Figure 22 presents the study flowchart.

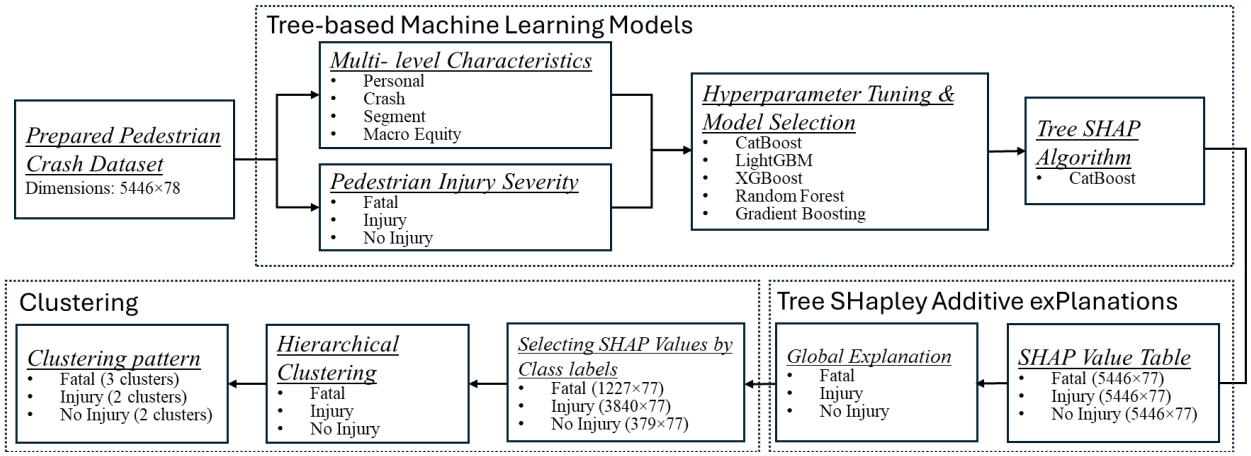


Figure 22. Study flowchart

To identify typical clustering patterns, this study conducted a hierarchical clustering analysis on the SHAP values for each severity level. Hierarchical clustering is based on a dendrogram, or cluster tree, which organizes data according to the similarity between individual data points. In hierarchical clustering, data points are grouped by merging similar clusters until all groups form one large cluster, a process known as the agglomerative strategy (F. Ding et al. 2018). Conversely, the divisive strategy starts with a large cluster that continuously splits into smaller clusters based on their heterogeneity. This study used the agglomerative strategy. To calculate the average inter-cluster distance using the average-linkage approach, the distance between each pair of observations is measured, summed, and then divided by the number of pairs in the cluster. The distance is estimated as follows:

$$d(u,v) = \sum_{tj} \frac{d(u[i], v[j])}{([u] * [v])} \tag{9}$$

where, i and j are the data instances that correspond to each cluster, and u and v are the two clusters, respectively. Rousseeuw's Silhouette approach (1987) was used to determine the number of clusters from the resulting clustered tree. The silhouette coefficient was used in this method to assess the cohesiveness measure and the distance between the generated clusters.

While cluster separation quantifies the degree to which each cluster is isolated from the others, cohesion quantifies how closely the data inside a cluster are related to one another. Equation 5 details the silhouette coefficient:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{10}$$

where, $s(i)$ is the silhouette coefficient of sample i, which is determined by comparing sample $a(i)$, which represents sample i as average distance to other samples in the same cluster, to sample $b(i)$, which represents sample i as average distance to all other clusters. The silhouette coefficient ranges from -1 to 1. A more positive value indicates a higher probability that a data point is correctly assigned to its cluster.

5.4. Results and Discussion

5.4.1. Hyperparameter Tuning and Model Selection

The first step of the pedestrian crash severity analysis involved fine-tuning the parameters and selecting the machine learning models. These models include CatBoost, LightGBM, XGBoost, Random Forest, and Gradient Boosting. For this process, the dataset was split into two subsets: 80% for training and the remaining 20% for testing. Furthermore, 20% of the training data was set aside for model validation. During the training phase, the models were validated at every three iterations. The aim of the optimization was minimizing the loss function, also referred to as the cost, which signifies the predictive capability of the model. Lower costs denote better-performing models (Khan and Ahmed 2020). Each model requires different parameters to be optimized. Grid search is a widely adopted approach for hyperparameter optimization (Chicco 2017). To enhance computational efficiency, the hyperparameter tuning was confined to parameters that significantly affect model performance. Table 3 presents the hyperparameter tuning space for each model. Parameters such as learning rate, number of estimators, criterion,

batch size, max depth, and others were fine-tuned. It is important to note that parameter tuning is not guaranteed for global optimal but local optimal.

Table 10. Hyperparameter tuning space

Models	Parameters
CatBoost	'iterations': [100, 200, 300], 'learning_rate': [0.01, 0.05, 0.1], 'depth': [4, 6, 10], 'l2_leaf_reg': [1, 3, 5, 7]
LightGBM	'num_leaves': [31, 63, 127], 'learning_rate': [0.01, 0.05, 0.1], 'n_estimators': [100, 200, 300], 'min_split_gain': [0.0, 0.1, 0.2]
XGBoost	'n_estimators': [100, 200, 300], 'learning_rate': [0.01, 0.05, 0.1], 'max_depth': [3, 5, 7] 'min_child_weight': [1, 3, 5]
Random Forest	'n_estimators': [100, 300, 500], 'criterion': ['gini', 'entropy', 'log_loss'],'max_depth': [2, 5, 7]
Gradient Boosting	'n_estimators': [100, 200, 300], 'learning_rate': [0.01, 0.05, 0.1] 'loss': ['log_loss'], 'max_depth': [2, 3, 5]

The best parameter combinations were searched using 3-fold cross-validation in the tuning space. Cross-validation ensured the findings were not dependent on a single train-test split but were validated across multiple partitions of the data. Figure 23(a) illustrates the normalized accuracy of various machine learning models over a series of search iterations. The number of iterations for each model is determined by the defined tuning space, which is why there are varying iteration counts for each model. It can be observed that the normalized accuracy for all models tends to increase with the number of iterations, suggesting that parameter tuning is effectively enhancing model performance. The XGBoost and CatBoost models show more significant improvements as iterations increase compared to other models. The shaded regions around the lines represent the confidence intervals of the measurements. In contrast, the Random Forest model shows little to no improvement in the first half of the iterations, maintaining a relatively lower accuracy throughout. Meanwhile, the LightGBM and Gradient Boosting models exhibit a more stable pattern of improvement.

Figure 23 (b) presents the log loss for the same models over search iterations, a metric where lower scores indicate better model performance. There is a downward trend in log loss values

across all models, suggesting a general improvement in model accuracy as the search progresses. Sharp increases in log loss at certain points can be interpreted as instances where the model parameters did not align well with the prediction task, potentially leading to poorer performance. Notably, CatBoost showed stability in log loss in the later iterations. Overall, CatBoost achieved the highest training accuracy and lowest training log loss.

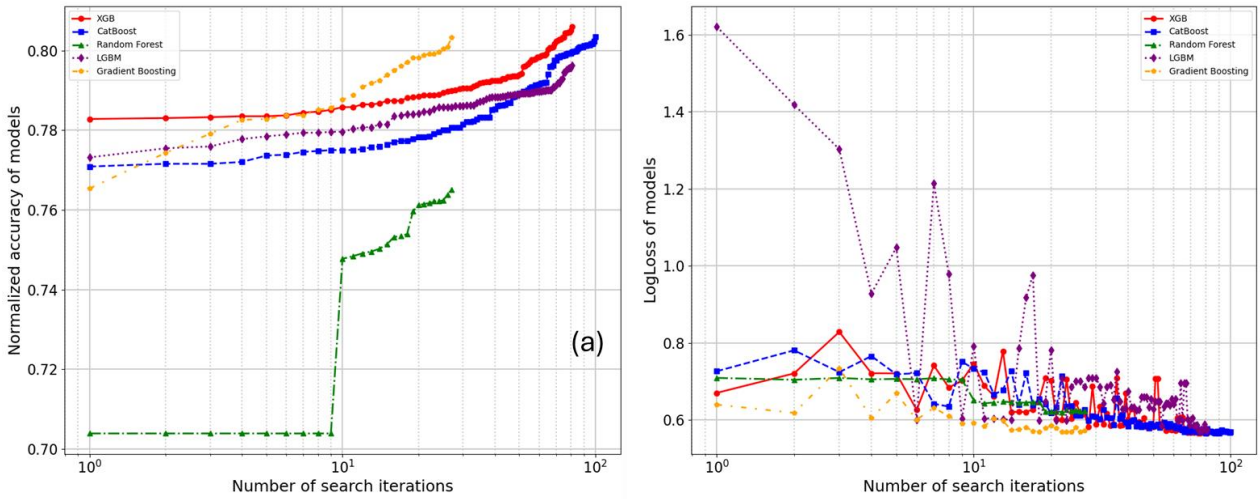


Figure 23. Model performance across search iterations: (a) normalized accuracy; (b) log loss

After assessing the training performance, it is crucial to evaluate the predictive performance. As shown in Table 11, the CatBoost model produces the highest prediction accuracy of 78.07%. This model also has the highest weighted average F1 score at 0.82, which is a harmonized metric accounting for both precision and recall. Furthermore, the CatBoost model demonstrates a commendable recall of 0.78, which emphasizes its capability to correctly identify positive instances among different classes. The use of weighted averages is particularly crucial in the presence of class imbalance, as it factors in the prevalence of each class and ensures that the performance metrics are not unduly skewed by the overrepresented class. These results establish CatBoost as the preferred model for the subsequent clustering analysis.

Table 11. Crash severity prediction performance

Model	Optimized Parameters	Accuracy	Weighted Avg		
			Precision	Recall	F1
XGBoost	'n_estimators': 100, 'min_child_weight': 1, 'max_depth': 3, 'learning_rate': 0.1	77.71%	0.84	0.78	0.8
Random Forest	'n_estimators': 300, 'max_depth': 7, 'criterion': 'entropy'	76.15%	**0.89**	0.76	0.8
CatBoost	'learning_rate': 0.1, 'l2_leaf_reg': 1, 'iterations': 300, 'depth': 4	**78.07%**	0.84	**0.78**	**0.82**
LightGBM	'num_leaves': 31, 'n_estimators': 200, 'min_split_gain': 0.1, 'learning_rate': 0.01	77.52%	0.84	0.78	0.8
Gradient Boosting	'n_estimators': 100, 'max_depth': 2, 'loss': 'log_loss', 'learning_rate': 0.1	77.89%	0.84	0.78	0.8

5.4.2. Tree SHAP Value for CatBoost Model

Utilizing tree SHAP algorithms, the study examined variable impacts on pedestrian crash severity predictions using the CatBoost model. Figure 24 presents the top 20 features by absolute mean SHAP values. The feature importance results vary for each severity level. For fatal crashes, pedestrian sobriety impairment status emerges as the most impactful feature, indicating that impairment status can significantly affect the severity of pedestrian injuries. This is likely because impaired pedestrians may have reduced awareness and attention to traffic, making them less capable of noticing approaching vehicles or accurately judging safe crossing times. Previous research corroborates this observation. For instance, Hossain et al. (2023) found that alcohol-impaired older pedestrians are more likely to be involved in fatal crashes. Batouli et al. (2020) identified pedestrian impairment as a significant factor associated with the severity of outcomes of motor vehicle crashes. Lighting conditions rank as the second most significant factor. Traffic fatalities per 100,000 people, also stand out, suggesting that fatal pedestrian crashes are closely associated with macro-level traffic fatality characteristics.

Interestingly, macro-level features such as the calculated average annual cost of transportation as a percentage of household income, PM2.5 air quality levels, total population, and the estimated average drive time to points of interest are also among the top 20 variables

associated with fatal crashes. Features like accident type and location type have considerable influence as well. The type of location is notable and potentially related to operating speeds. Several prior studies have suggested that higher speeds have more association with severe outcomes (Adanu, Dzinyela, and Agyemang 2023; Islam 2023). Moderately influential features pedestrian actions and drug or physical impairment, both having significant behavioral and condition impacts that may lead to severe outcomes. Traffic operations and the underlying cause of the crash also contribute to the model's predictions, reflecting the complexity of factors that can lead to a fatal crash.

Injury crashes show a different hierarchy of feature importance. Compared to fatal pedestrian crashes, the variables differ significantly, with more social and environmental-related variables - such as total population, percent of no-car households, GINI index, and percent of hazardous sites - presented as important variables in predicting injury crashes. The feature sobriety impairment status is the most influential. Lighting conditions and traffic fatalities also have considerable importance. Location type and cause follow, highlighting the role of crash location and causative factors. Lesser, yet notable impacts are observed for features like traffic operations and demographic elements.

Compared to crashes resulting in injury, an even greater number of social and environmental equity-related variables are identified as the top 20 important factors in predicting the occurrence of no-injury pedestrian crashes. These factors include total population, percentage of households without a car, average commute time to work, calculated average annual cost of transportation as a percentage of household income, traffic fatalities, percentage of total housing units that are renter-occupied, GINI index, and percentage of hazardous sites. The values of these factors indicate more resilient communities compared to those in the injured or fatal groups. This also

suggests that fewer crash- or road-related features are considered important in incidents resulting in no injury. Accident type is the most influential feature in predicting no-injury crashes. Pedestrian sobriety impairment still plays a role, mostly because most of the cases are not under the influence in the no-injury subset. The ‘light’ feature underscores the relevance of lighting conditions.

The SHAP value chart for crash severity shows how different features influence the likelihood of fatal, injury, and no injury outcomes. Pedestrian sobriety impairment is highly associated with fatal crashes, suggesting driver impairment is a critical risk factor, while lighting has a significant but decreasing association as the severity lessens, indicating visibility’s variable importance across severities. Accident type consistently affects all classes, especially fatal crashes, emphasizing the crash’s nature as a crucial determinant. Features like location type and cause have more influence on non-fatal crashes, hinting at the importance of location and causative factors in less severe crashes. Pedestrian action stands out for injury crashes, underlining pedestrian risks in these events. Meanwhile, total population has a notable presence in injury and no-injury crashes, suggesting demographic aspects may influence crash frequency but not necessarily result in fatalities. The figure effectively captures the varied importance of each factor across different crash severities, offering insights for developing safety measures.

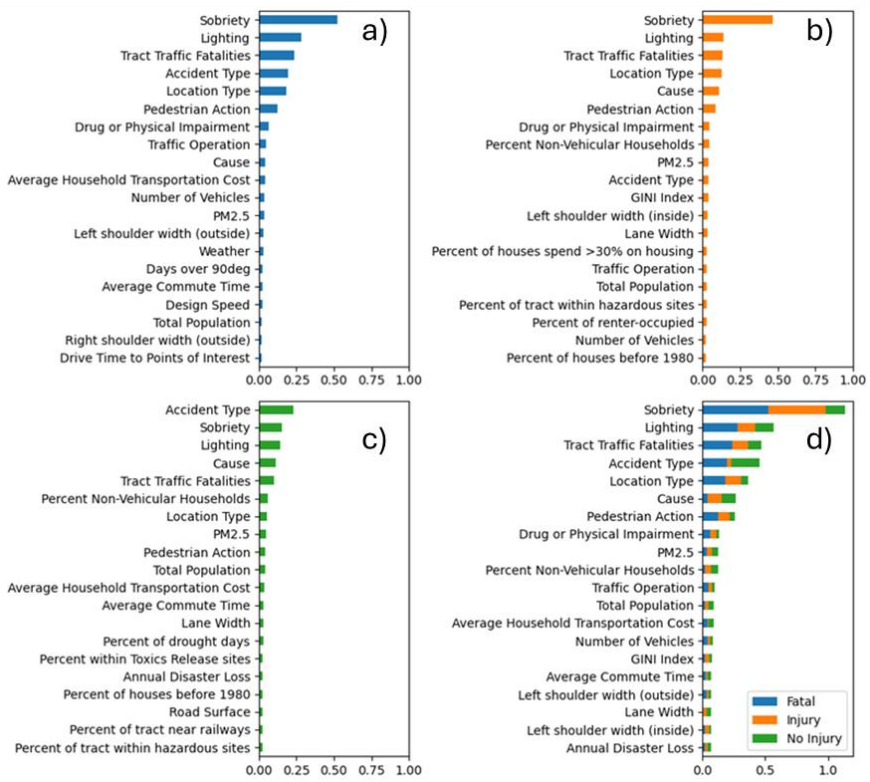


Figure 24. Global importance of top 20 variables for each crash severity level using tree SHAP Explainer: a) Fatal crash; b) Injury crashes; c) No injury crashes; d) Combined feature importance across all three classes

5.4.3. Clustering Patterns by Severity

To further identify clustering patterns across different severity levels, hierarchical clustering was applied to the SHAP values. Figure 25 presents the dendrogram for pedestrian-involved crashes at varying severities and includes corresponding silhouette scores for different numbers of clusters. The dendrograms for fatal (a), injury (b), and no injury (c) crashes illustrate the grouping of instances according to their SHAP values. On each dendrogram, the x-axis labels individual samples, while the y-axis quantifies the dissimilarity between clusters, with branch lengths signifying the point of cluster merging.

The silhouette score plot (d) presents cluster coherence over a range from 2 to 6 clusters, with scores between 0 and 1; higher scores signify well-differentiated clusters. Fatal crashes peak at three clusters, pointing to the presence of three distinct groups. Both injury and no injury

crashes show the highest silhouette scores at two clusters. For fatal crashes (a) three distinct clusters are identifiable, with one notably more separated, indicating three groups with SHAP value profiles. The injury crash dendrogram (b) shows two clusters, suggesting a less complex SHAP value profile compared to fatal crashes. The dendrogram for no injury crashes (c) displays clusters that are closely positioned, revealing a higher similarity in SHAP values. Overall, dendrograms and silhouette scores illustrate a detailed picture of the clustering within the data. The subsequent study will investigate the patterns of these clusters in depth.

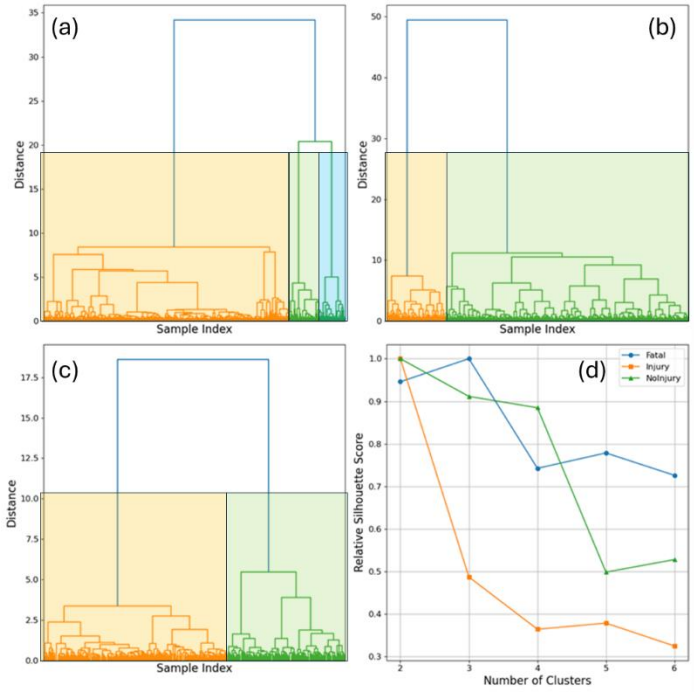


Figure 25. Dendrogram for SHAP values and silhouette scores: a) Fatal crashes; b) Injury crashes; c) No injury crashes; d) Silhouette scores for different numbers of clusters

5.4.3.1. Fatal Pedestrian Crash Clusters

For 3,840 fatal pedestrian crashes, Cluster 1 contains 1,000 samples (82%), Cluster 2 contains 132 samples (11%), and Cluster 3 contains 95 samples (8%). Figure 26 presents the clustering patterns and the representative factors contributing to fatal outcomes within each cluster. Figure 26(a) displays the mean SHAP values for the top 20 important features - those

with the highest mean absolute SHAP values -across the three clusters. b-i) zooms in on the top 8 features and provides a detailed SHAP value distribution. As shown in Figure 26(a), pedestrian sobriety impairment status ‘impair_1’ stands out in Cluster 2 with a notably negative mean SHAP value of -0.51 (blue), suggesting a significant negative association of this feature with fatal crash outcomes within this group.

In contrast, the association is positive in the other two clusters. Referring to Figure 26(b), the pedestrian sobriety impairment status shows varying impacts across clusters. For instance, the ‘Sober’ attribute, prominent in Cluster 2, presents negative SHAP values, whereas ‘Under Alcohol Influence’ and ‘Drugged’ have positive values and are major components of Cluster 1. Cluster 3, with a minor positive SHAP value of 0.27, mainly includes unstated pedestrian impairment statuses and HBD sober. Pedestrians with drug or physical impairment are positively associated with Cluster 3 but negatively associated with the other clusters, making this a significant factor contributing to fatal outcomes in Cluster 3 while not in the other two clusters. Regarding crash type, Cluster 2 has negative mean SHAP values, indicating other factors contribute to fatal outcomes in this cluster, while Clusters 1 and 3 show a moderate positive impact. As shown in Figure 6 f), ‘uto-pedestrian’ crashes have a higher positive mean SHAP value, suggesting a strong association with fatal outcomes, while other crash types are negatively associated. The following paragraphs provide a detailed analysis of each cluster and discuss potential strategies to address equity gaps.

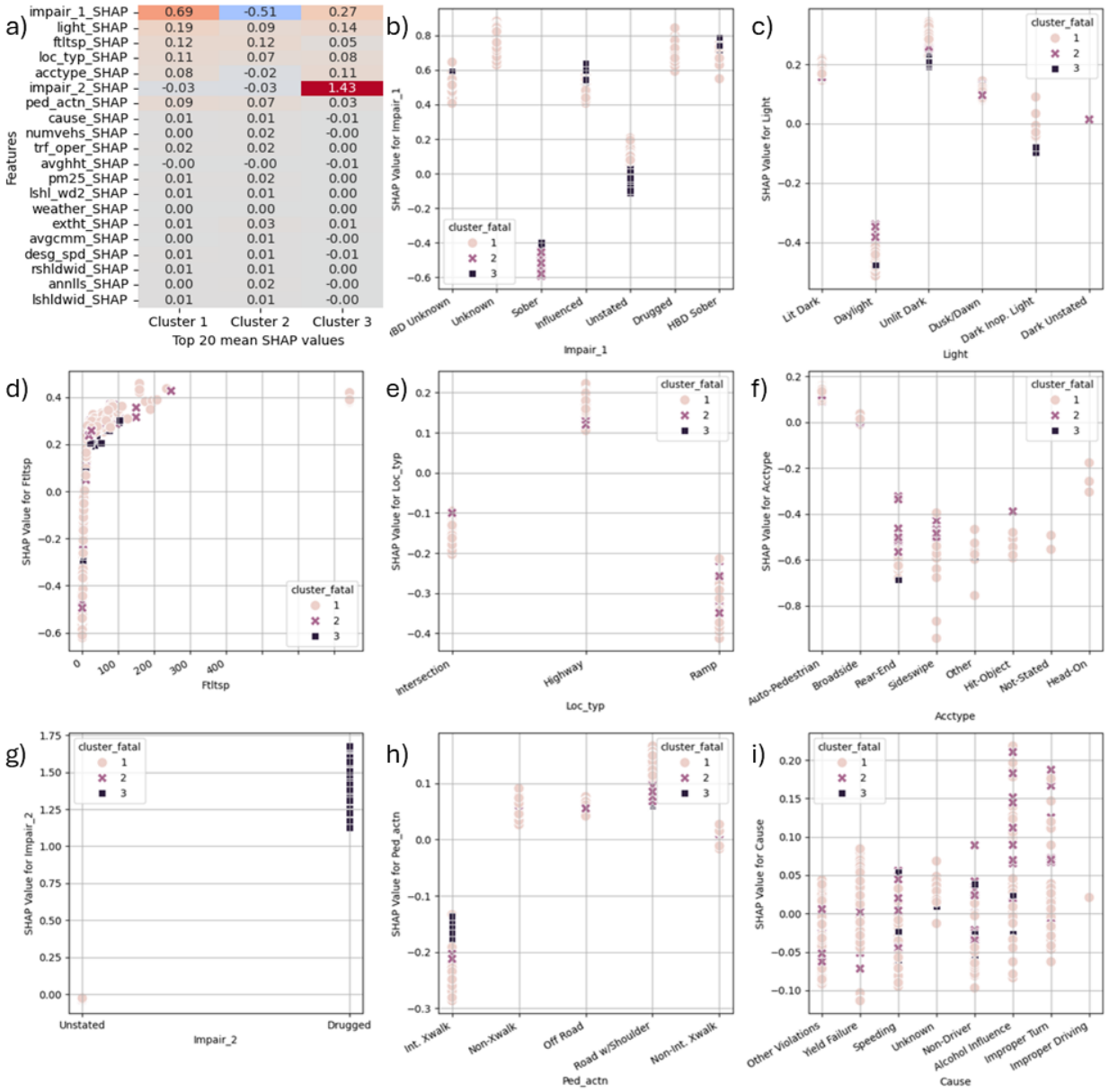


Figure 26. Cluster analysis for fatal pedestrian crashes: (a) mean SHAP values for the top 20 important features by cluster; (b-i) partial dependence plots for the top 8 important features

Fatal Cluster 1: Nighttime Highway Crashes with Poor Visibility and Pedestrian Impairment in Underserved Areas

Cluster 1 represents 82% of the fatal crashes. In this cluster, the pedestrian sobriety impairment status is largely undetermined, with influenced pedestrians accounting for 7% and drugged pedestrians for 5% of all fatal pedestrian crashes. As illustrated in Figure 26 e), most of these crashes occur on highways (69%), where vehicles typically travel at higher speeds,

reflected in common speed zones of 60-70 mph, contributing to the high risk. The light conditions are predominantly dark, with 40% of crashes occurring in unlit dark environments and 33% in lit dark conditions, indicating that poor visibility plays a significant role.

Auto-pedestrian crashes are the most frequent type within this cluster, comprising 74% of the crashes. Pedestrian actions leading to crashes often involve the road shoulder (54%) and non-crosswalk areas (21%). The leading cause of crashes is failure to yield (46%), followed by other violations (20%) and speeding (7%). This distribution suggests that driver education, law enforcement, and infrastructural changes might be necessary to address these issues. Single-vehicle crashes are the most common (63%), with incidents involving 2-4 vehicles at 18%, reflecting the isolated nature of pedestrian crashes. A lack of traffic control is evident, with 71% of crashes happening in areas without traffic controls and only 10% occurring in areas with controls. The right and left outside shoulder widths (9; 5) are larger than in the other clusters, indicating a better influence of these features.

Equity-related infrastructure, economic, and environmental factors are among the top 20 features influencing the severity of pedestrian injury in Cluster 1. The average annual cost of transportation as a percentage of household income is 1.17, and PM2.5 air quality levels are 11.64. Additionally, the area faces a growing number of extremely hot days. These factors potentially indicate the vulnerability of these communities. The average commute time is 28.73 minutes, longer than in the other clusters, indicating traffic inequity. To address the equity gap in this cluster, it is crucial to improve lighting and visibility with streetlights and enhance pedestrian infrastructure. Implementing traffic calming measures, such as speed reductions and increased traffic control with signals and crossings. Public awareness campaigns and educational programs are essential for promoting road safety. Additionally, providing public transportation subsidies

and developing affordable housing near transit can reduce commute times and transportation costs.

Fatal Cluster 2: Failing to Yield, Highway Incidents with Sober Pedestrians on Roads with Shoulders in High Traffic Fatality Areas

Cluster 2 represents 11% of the fatal crashes. Pedestrians are predominantly sober, with 11% showing no impairment, suggesting that factors other than sobriety impairment are contributing to crashes. The status of pedestrian drug or physical impairment (‘impair_2’) is largely unstated, indicating an absence of such impairments. Most crashes occur in dark conditions, either unlighted (4%) or lighted (3%), with a slightly higher proportion of crashes in daylight (2%) than in Cluster 1, suggesting that visibility issues do not exclusively account for the crashes in this cluster.

Auto-pedestrian crashes still make up a significant portion, though less than in Cluster 1, indicating diversity in crash types, evidenced by rear-end and sideswipe collisions. Highways remain the most common location for these crashes, but there are notable occurrences at intersections and ramps. Pedestrian actions leading to crashes often involve the road shoulder, similar to Cluster 1, but crashes at marked crosswalks also occur, indicating that pedestrian safety measures in these areas could be failing. Yield failure remains a leading cause of crashes, with other violations and speeding also contributing significantly. Crashes typically involve a single vehicle or up to four vehicles, reflecting a broader distribution of crash contexts. The lack of traffic control is significant, with many crashes occurring in areas without controls. Speed limits remain high in crash zones, indicating that high vehicle speeds pose a significant risk despite the sober state of pedestrians.

Economic and environmental improvements include a lower average annual cost of transportation as a percentage of household income (0.62) compared to Cluster 1. PM2.5 levels

are slightly lower, and the average commute time is marginally shorter. The traffic fatality rate is higher at 17.44, influenced by factors such as vehicle speed, pedestrian traffic volume, and the effectiveness of traffic control measures. While Cluster 2 shares some characteristics with Cluster 1, it stands out due to a greater proportion of sober pedestrians, a broader variety of crash types, and more crashes at controlled intersections. Economic, societal, and environmental conditions show subtle differences between the clusters.

Fatal Cluster 3: Non-Crosswalk, Narrow Shoulder Crashes with Alcohol or Drug-Impaired Pedestrians in Underserved Areas

Cluster 3, comprising 8% of the fatal crashes, shows distinct characteristics. This cluster is marked by a significant proportion of pedestrian crashes where individuals are under the influence, with 4% categorized as alcohol-influenced and 8% as drugged. Lighting conditions reveal many incidents occurring in lighted dark environments (5%) and unlighted dark environments (2%). Daylight and dusk/dawn crashes are less frequent but still present, indicating pedestrian risks across various lighting conditions. Non-crosswalk pedestrian actions (3%) and crashes on the road shoulder (3%) highlight the need for improvements in pedestrian infrastructure and behavior modification measures. Other violations are the leading cause of crashes, indicating the need to address various walking behaviors beyond yielding and speed control.

Most crashes involve a single vehicle, similar to Clusters 1 and 2, with a lower percentage of multi-vehicle collisions, highlighting single-vehicle impacts as a key area for intervention. A substantial portion of crashes occurred in areas without traffic controls (6%), suggesting different traffic management issues than in Cluster 1. Infrastructure in this cluster shows narrower shoulder widths, potentially reducing the safety margin for pedestrians. Socioeconomic factors indicate lower transportation costs as a percentage of household income (0.50) and a marginally

shorter average commute time (28.50). The traffic fatality rate in this cluster stands at 8.45, lower than in the other clusters, suggesting relatively lower community vulnerability. To address the equity gap in Cluster 3, it is essential to focus on improving pedestrian infrastructure and addressing behavioral issues related to alcohol and drug impairment. Building and maintaining pedestrian pathways, particularly in non-crosswalk areas and on narrow road shoulders, is crucial.

5.4.3.2. Injured Pedestrian Crash Clusters

Figure 27 visualizes the clustering patterns and the representative factors contributing to injured outcomes within each cluster. Compared to fatal pedestrian crashes, the variables differ significantly, with more social and environmental equity-related variables - such as total population, percent of no-car households, GINI index, and percent of hazardous sites - presented as important variables in predicting injury crashes. Interestingly, the pedestrian sobriety impairment status ‘sober’ has a positive value, indicating a positive association with injuries in sober pedestrians, whereas this group tends to have a negative association with fatal outcomes. Traffic fatality rates also show a mean positive association in both clusters, but this association is less pronounced compared to fatal crashes.

Injury Cluster 1: Poor Lighted, Lacking Pedestrian Infrastructure, Exacerbated by Social Disparities

Cluster 1, representing 20% of the injury subset, includes many crashes where the impairment status is not reported and pedestrians are sober. According to Figure 7(b), ‘unstated’ and ‘unknown’ impairments are associated with negative values, suggesting that most ‘unstated’ categories indicate no obvious impairment. The lighting conditions under which these crashes occur vary, with 9% happening in dark-lit environments, 7% during daylight, and 4% during dark-unlit conditions, indicating substantial risks both day and night. Highways are noted in 10%

of cases and intersections in 5%. The predominant causes of crashes - ‘yield failure’ at 8% and ‘other violations’ at 7% - point to broader issues of non-adherence to traffic laws contributing to pedestrian injuries. Traffic operation conditions vary, with 12% of injuries occurring where there are ‘no controls’ and 8% where ‘control is on’. Injuries happen on roads with shoulders (7%), at intersection crosswalks (6%), and in non-crosswalk areas (5%). Most crashes occur on dry road surfaces.

Demographic and socio-economic factors in areas where these crashes occur reveal an average total population of 4,275. Additionally, 28% of households do not own a car, suggesting a greater reliance on walking and increased pedestrian exposure to traffic risks. Traffic fatalities further underscore safety concerns, with a rate of 9, significantly lower than the fatal crash subset. The average PM2.5 level is 11, but as it reaches 12, the association with injuries becomes inverse. Sixty-three percent of the area lies within one mile of known hazardous sites, indicating a higher likelihood of crashes near potentially less pedestrian-friendly industrial zones. The housing in these areas is older, with 62% of homes built before 1980, and 51% of housing units are rented, indicating a transient population less familiar with local traffic patterns. The economic burden is evident, with 42% of occupied houses spending 30% or more on housing costs despite earning less than $75,000. The average GINI index is 0.49, positively associated with injury crashes, indicating significant income inequality correlating with disparities in traffic safety and pedestrian infrastructure investment. To address the equity gap in this cluster, it is essential to enhance lighting and visibility through the installation of street lights and reflective road markings, particularly in dark-lighted and dark-unlighted environments.

Injury Cluster 2: Sober Pedestrian Injuries Influenced by Driver Violations and Inadequate Infrastructure

Cluster 2 encompasses most of the dataset, with 80% of pedestrian injuries. In this cluster, 64% of cases involved individuals recorded as ‘sober’ at the time of the crash, with ‘influenced’ at 4%. Interestingly, ‘sober’ has a positive value opposite to the value in fatal cluster, indicating a positive association with injuries in sober pedestrians, whereas this group tends to have a negative association with fatal severity. This suggests that sobriety tends to reduce the likelihood of fatal crashes but not injuries. The lighting conditions during these crashes were predominantly ‘daylight’ at 36% and ‘lit dark’ at 26%, indicating that good visibility and other factors contribute to these crashes.

Pedestrian actions involving ‘road with shoulder’ crashes (34%) and ‘intersection crosswalk’ crashes (27%) show higher figures than in Cluster 1. Highways remain a significant risk, with 43% of injuries occurring there, similar to Cluster 1. The primary collision factors –’yield failure’ at 34%, ‘other violations’ at 18%, and ‘speeding’ at 11% - highlight that driver behavior is critical in the occurrence of these injuries. Auto-pedestrian crashes dominate at 61%, but other types such as ‘rear-end’, ‘sideswipe’, ‘broadside’, and ‘hit-object’ also contribute to the injury crashes, indicating a more diverse range of accidents. Regarding traffic operations, there is a higher incidence of pedestrian injuries in areas with ‘no controls’ (48%) and ‘control on’ (32%). To address the equity gap in this cluster, it is crucial to focus on improving infrastructure and addressing driver behavior.

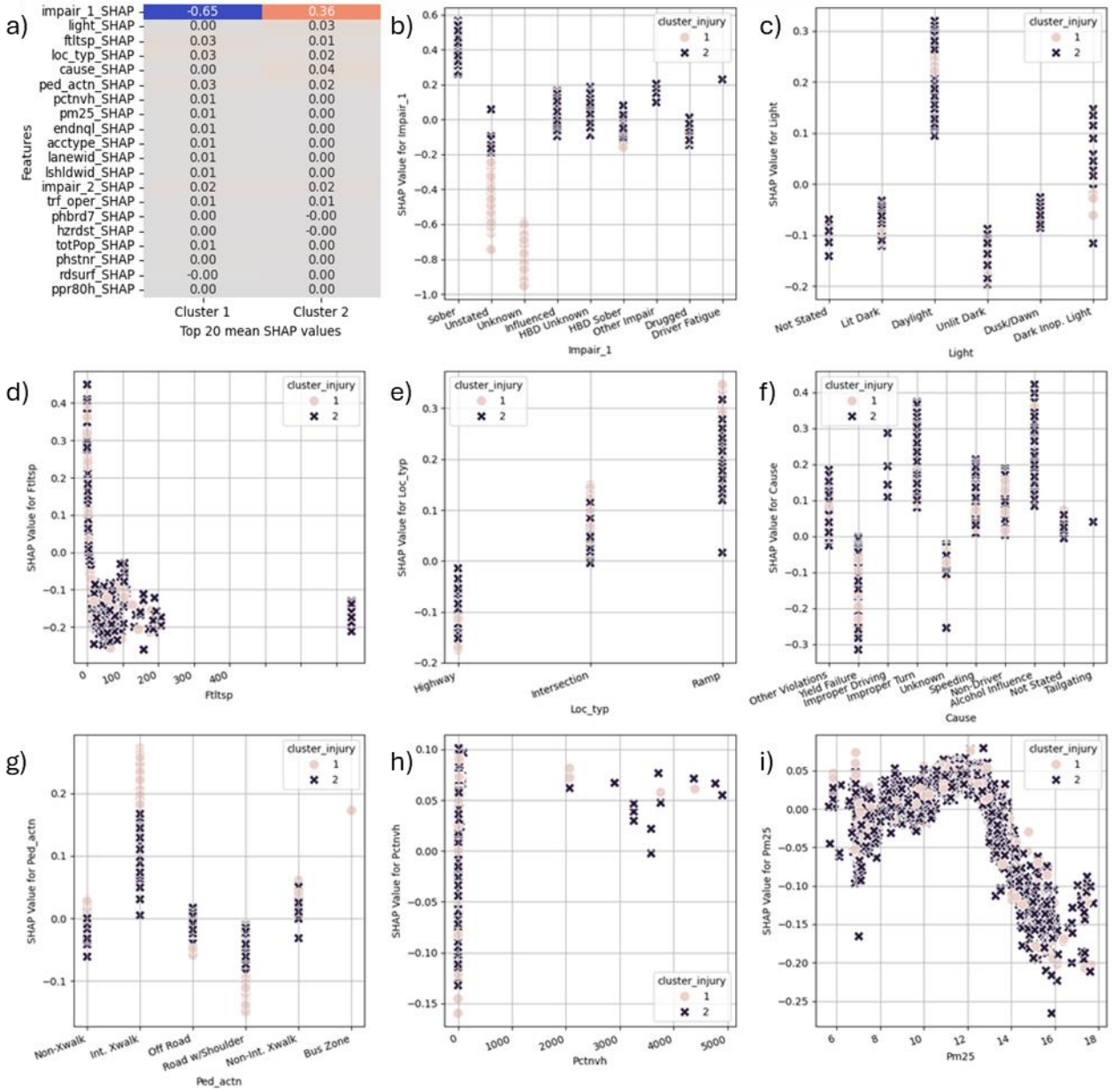


Figure 27. Cluster analysis for injured pedestrian crashes: (a) mean SHAP values for the top 20 important features by cluster; (b-i) partial dependence plots for the top 8 important features

5.4.3.3. Not Injury Pedestrian Crash Clusters

Figure 28 visualizes the clustering patterns and the representative factors contributing to not-injured crashes within each cluster. Compared to crashes that result in injury, even more social and environmental equity-related variables are presented as the top 20 important variables in predicting the occurrence of no-injury pedestrian crashes. These include total population, percent

of no-car households, average commute time to work, calculated average annual cost of transportation as a percent of household income, traffic fatalities, percent of total housing units that are renter-occupied, GINI index, and percent of hazardous sites. It also means fewer crash or road related features are presented as important features in results of no injury. It can be observed from the average values of these indicates that these communities do not present as disadvantaged as those in the fatal and injury pedestrian crash categories.

Not Injured Cluster 1: Not Injured in Prevalence of Controlled Settings in Moderately Urbanized Resilient Communities

Cluster 1 accounts for 61% of no-injury crashes. This cluster has a significant occurrence of ‘uto-pedestrian’ crashes at 58% and ‘broadside’ crashes at 3%, which, fortunately, did not result in injuries. Most of these incidents occurred during ‘daylight’ (25%) and ‘lit dark’ (20%) conditions, suggesting effective visibility and caution from both pedestrians and drivers. Impairment does not appear to be a major factor in this cluster. The dominant cause of crashes is ‘yield failure’ at 34%, followed by ‘other violations’ at 14%, highlighting the need for adherence to right-of-way rules and traffic laws. Highways are a common setting at 34%, with significant interactions at ‘intersection crosswalks’ (20%) and ‘roads with shoulders’ (27%), indicating the importance of pedestrian behavior at crosswalks in avoiding injuries.

Demographically and environmentally, these factors indicate less vulnerable communities. The average total population is 4,410, and a lower percentage (7%) of households without a car indicates less dependency on walking. With a moderate average commute time (27.62 minutes) and a relatively low annual transportation cost as a percentage of household income (0.16%), these areas suggest an efficient transportation system. A large proportion of housing units built before 1980 (62%) may indicate older infrastructure that has been adapted to modern safety

standards. Overall, it is essential to maintain and enhance the factors contributing to the resilience of these moderately urbanized communities.

Not Injured Cluster 2: No Injuries in High Density, Urban Areas with Lesser Social Disadvantages

Cluster 2 comprises 39% of no-injury crashes. ‘hit-object’ crashes are notable, representing 11% of the incidents, suggesting that pedestrian isolation facilities could potentially reduce crash injuries and offer protection. Most cases in this cluster involve pedestrians without impairment. With a significant number of crashes occurring in ‘unlit dark’ conditions (12%), as well as in ‘daylight’ (16%), the need for improved street lighting or visibility is highlighted. ‘Yield failure’ remains a significant cause at 14%, followed by ‘other violations’ at 10% and ‘speeding’ at 6%. ‘Road with Shoulder’ crashes at 26% suggest a need for better pedestrian pathways.

Demographics show a higher percentage of car-less households at 58% compared with Cluster 1. The slightly larger population (4,546) may reflect a denser urban environment. Longer average commute times (30 minutes) and a higher annual transportation cost as a percentage of household income (0.71%) indicate the need for better transportation options. A somewhat lower presence of pre-1980 housing (58%) and proximity to hazardous sites (64%) are noted. The GINI Index is higher at 0.53, suggesting greater income inequality than in Cluster 1. Policymakers may consider infrastructure investments such as better street lighting, reflective signage, and pedestrian isolation facilities to protect pedestrians from hit-object crashes. Community-specific interventions focusing on improving pedestrian pathways and ensuring safe crossing areas can be particularly effective in this high-density urban environment.

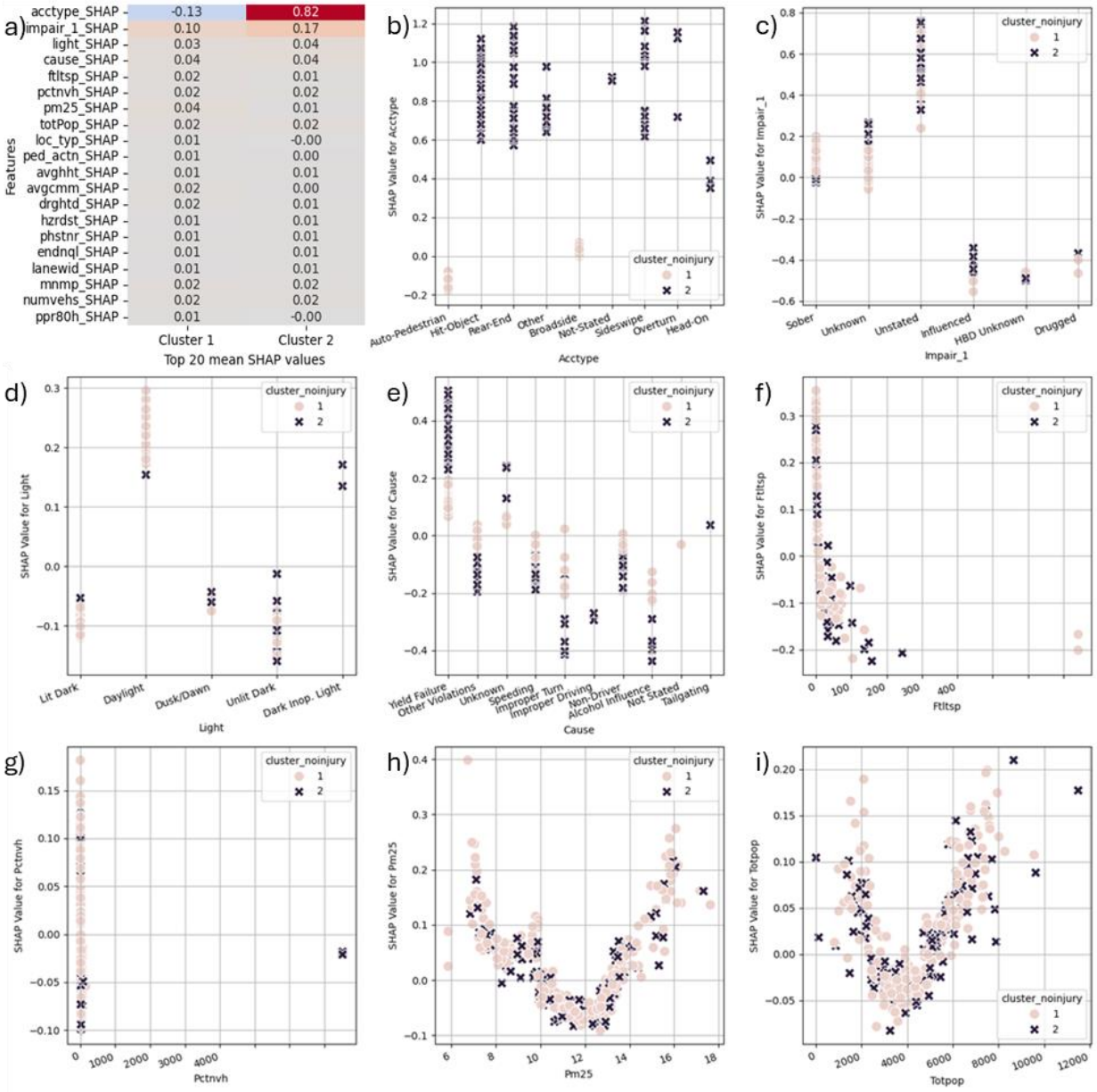


Figure 28. Cluster analysis for not injured pedestrian crashes: (a) mean SHAP values for the top 20 important features by cluster; (b-i) partial dependence plots for the top 8 important features

5.5. Summary

This chapter utilized comprehensive crash data from California, spanning from 2018 to 2021, which was gathered from HSIS. It also incorporates environmental justice features sourced from the Equitable Transportation Community, providing a rich dataset that includes extensive socioeconomic, demographic, and build-environment data at the Census Tract level. The dataset

encompasses four tiers of information: unit level, crash level, roadway details, and zonal data, from which 77 variables were chosen to develop models predicting pedestrian crash severity across three categories: fatal, injury, and no injury. The study used the Tree SHAP algorithm to explain the influence of each variable on the predictions. Hierarchical clustering is applied to these SHAP values to reveal distinct patterns within clusters.

The analysis using CatBoost model and tree SHAP values highlights distinct factors influencing the severity of pedestrian crashes. For fatal crashes, the most impactful factors include pedestrian sobriety impairment, lighting conditions, and macro-level traffic fatalities. These findings underscore the crucial role of individual pedestrian conditions and environmental factors in determining the severity of crashes. In injury crashes, the data reveals a significant influence of social and environmental variables, such as demographics and economic conditions, suggesting that broader societal factors play a key role in these less severe outcomes. For no-injury crashes, various social and environmental equity-related variables are highlighted, suggesting that communities with better walkability tend to experience less severe crashes. It is worth noting that while immediate, situational factors like pedestrian impairment and lighting are critical in severe outcomes. Broader societal and demographic influences are more influential in less severe cases. These insights can guide the development of comprehensive safety measures that address both situational risks and broader societal conditions to reduce pedestrian injuries and fatalities.

The analysis of the three fatal crash clusters provides a detailed understanding of the variables influencing pedestrian crash severities and outcomes. One cluster highlights the critical risk factors associated with nighttime highway crashes in areas with poor visibility and pedestrian impairment, suggesting the need for targeted interventions in lighting and sobriety

enforcement. Another cluster emphasizes crashes involving sober pedestrians, indicating that factors such as driver behavior, road design, and effective traffic control are significant. Meanwhile, the final cluster's focus on impaired pedestrians in vulnerable areas with inadequate infrastructure points to the necessity of addressing substance abuse and improving pedestrian pathways.

The analysis of injury clusters reveals critical insights into the factors contributing to pedestrian injuries. One cluster characterized by a mix of lighting conditions and locations lacking pedestrian infrastructure highlights the role of social disparities and infrastructure inadequacies in pedestrian injuries. A significant number of crashes occur in dimly lighted areas or where traffic controls are absent, exacerbated by socio-economic challenges such as high housing costs and income inequality. This suggests that pedestrian safety could be significantly improved through enhanced lighting, better traffic control, and targeted socio-economic interventions. The other cluster, with most injuries involving sober pedestrians, underscores the impact of driver violations and inadequate infrastructure, particularly on highways. The high incidence of crashes in daylight and controlled areas indicates that while visibility might not be a limiting factor, the design and enforcement of traffic laws, along with infrastructure improvements, are crucial.

The analysis highlights the effectiveness of existing pedestrian and traffic control measures in preventing injuries, despite varying environmental and demographic contexts. A prevalence of controlled settings in moderately urbanized, resilient communities, demonstrates that good visibility, adherence to right-of-way rules, and effective pedestrian infrastructure at intersections and roads with shoulders significantly contribute to the crashes without injury. Meanwhile, in denser urban areas with greater social challenges, the analysis underscores the importance of

addressing street lighting, enhancing pedestrian pathways, and improving transportation options to maintain safety. This study has its limitations. Pedestrian demographic information, such as age and gender, was not included due to the unavailability of this data. Future research should consider comparing different clustering methods and investigating the impact of sample imbalance on clustering outcomes.

6. CONCLUSION

This dissertation provides a comprehensive exploration of pedestrian crash risks through advanced modeling techniques, focusing on built-environment, socio-economic, and transit-related factors across multiple spatial scales. It addresses three key research questions and brings new insights to the literature on pedestrian safety.

First, regarding the factors associated with pedestrian safety, the findings reveal that auto-oriented network density consistently increases pedestrian crash risks, while pedestrian-oriented infrastructure, such as sidewalks and bike lanes, mitigates these risks. Socio-economic variables, including the percentage of zero-car households and income diversity, also play critical roles in shaping pedestrian crash patterns. These findings contribute to the existing knowledge base by highlighting the interplay between infrastructure, land use, and socio-economic factors in pedestrian safety. They emphasize the need for urban planning strategies that integrate targeted investments in pedestrian infrastructure, zoning adjustments, and transit-oriented safety interventions.

Second, the spatial variability of pedestrian crash risks at the CBG level in urban areas underscores the importance of localized planning. By employing the MGWR model, this research demonstrates how factors such as employment and household entropy, transit frequency, and bike lane density influence crash risks differently across regions. These results advance the literature by showing that global patterns often mask significant spatial heterogeneities, necessitating location-specific interventions. For example, while bike lane density generally reduces risks, its effectiveness varies depending on surrounding land use and quality.

Third, the clustering analysis of individual-level pedestrian crashes identifies distinct patterns in crash severity. Fatal crashes are influenced by pedestrian conditions, such as sobriety, and

environmental factors like lighting, while broader socio-economic and infrastructure-related variables are more significant in injury and no-injury crashes. These findings highlight the layered nature of pedestrian crash risks and contribute to the knowledge base by demonstrating how different factors interact to shape crash outcomes.

Together, these answers provide actionable insights for urban planners and policymakers. This research highlights the importance of designing safer urban environments through a multi-faceted approach that includes improving pedestrian infrastructure, optimizing transit operations, addressing socio-economic disparities, and tailoring interventions to specific locations and conditions. By integrating advanced methods such as SHAP, MGWR, and clustering analyses, this study enhances the ability to understand, predict, and address pedestrian safety challenges in urban areas.

This dissertation also identifies key limitations and areas for future research. The modifiable areal unit problem in spatial analysis, the incomplete demographic data in micro-level studies, and the constrained geographical coverage of certain datasets leave room for further refinement. Future studies should address these gaps by incorporating broader spatial scales, demographic variables like age and gender, and alternative clustering techniques. Expanding the analysis in these ways can deepen our understanding of pedestrian crash risks and lead to even more effective, equitable safety interventions.

By answering these research questions and contributing novel methodologies and findings, this dissertation provides insights into improving pedestrian safety. The integration of equity-focused insights ensures that vulnerable populations are considered in safety planning, making urban environments not only safer but also more inclusive for all.

REFERENCE

2010 AASHTO. 2010. *Highway Safety Manual*. Vol. 1. AASHTO. https://books.google.com/books?hl=zh-CN&lr=&id=M4fQiyfVRr8C&oi=fnd&pg=PP26&dq=2010+AASHTO+Highway+Safety+Manual&ots=ckqCnDOycm&sig=SB2INst-Rt34ProCbUz__sOSfTk.

Adanu, Emmanuel Kofi, Richard Dzinyela, and William Agyemang. 2023. "A Comprehensive Study of Child Pedestrian Crash Outcomes in Ghana." *Accident Analysis & Prevention* 189 (September):107146. https://doi.org/10.1016/j.aap.2023.107146.

Al-Ani, Omar, Saquib Mohammed Haroon, Doina Caragea, HM Abdul Aziz, and Eric J. Fitzsimmons. 2024. "Predicting Pedestrian Involvement in Fatal Crashes Using a TabNet Deep Learning Model." In *Proceedings of the 16th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, 19–27. IWCTS '23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3615895.3628169.

Ali, Yasir, Fizza Hussain, and Md Mazharul Haque. 2024. "Advances, Challenges, and Future Research Needs in Machine Learning-Based Crash Prediction Models: A Systematic Review." *Accident Analysis & Prevention* 194 (January):107378. https://doi.org/10.1016/j.aap.2023.107378.

Almasi, Seyed Ahmad, and Hamid Reza Behnood. 2022. "Exposure Based Geographic Analysis Mode for Estimating the Expected Pedestrian Crash Frequency in Urban Traffic Zones; Case Study of Tehran." *Accident Analysis & Prevention* 168 (April):106576. https://doi.org/10.1016/j.aap.2022.106576.

Almasi, Seyed Ahmad, Hamid Reza Behnood, and Ramin Arvin. 2021. "Pedestrian Crash Exposure Analysis Using Alternative Geographically Weighted Regression Models." *Journal of Advanced Transportation* 2021:1–13.

Ammar, Dania, Yueru Xu, Bochen Jia, and Shan Bao. 2022. "Examination of Recent Pedestrian Safety Patterns at Intersections through Crash Data Analysis." *Transportation Research Record* 2676 (12): 331–41. https://doi.org/10.1177/03611981221095513.

Anselin, Luc. 1995. "Local Indicators of Spatial Association—LISA." *Geographical Analysis* 27 (2): 93–115.

Atakishiyev, Shahin, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. 2021. "Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions." arXiv.Org. December 21, 2021. https://arxiv.org/abs/2112.11561v4.

Baireddy, Raghunandan, Huaguo Zhou, and Mohammad Jalayer. 2018. "Multiple Correspondence Analysis of Pedestrian Crashes in Rural Illinois." *Transportation Research Record* 2672 (38): 116–27. https://doi.org/10.1177/0361198118777088.

Batouli, Ghazal, Manze Guo, Bruce Janson, and Wesley Marshall. 2020. "Analysis of Pedestrian-Vehicle Crash Injury Severity Factors in Colorado 2006–2016." *Accident Analysis & Prevention* 148 (December):105782. https://doi.org/10.1016/j.aap.2020.105782.

Bernhardt, Maxwell, and Kara Kockelman. 2021. "An Analysis of Pedestrian Crash Trends and Contributing Factors in Texas." *Journal of Transport & Health* 22 (September):101090. https://doi.org/10.1016/j.jth.2021.101090.

Bo, Huang. 2018. *Comprehensive Geographic Information Systems*. Elsevier. http://www.sciencedirect.com:5070/referencework/9780128047934/comprehensive-geographic-information-systems.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

Cai, Qing, Jaeyoung Lee, Naveen Eluru, and Mohamed Abdel-Aty. 2016. "Macro-Level Pedestrian and Bicycle Crash Analysis: Incorporating Spatial Spillover Effects in Dual State Count Models." *Accident Analysis & Prevention* 93 (August):14–22. https://doi.org/10.1016/j.aap.2016.04.018.

Chang, Iljoon, Hoontae Park, Eungi Hong, Jaeduk Lee, and Namju Kwon. 2022. "Predicting Effects of Built Environment on Fatal Pedestrian Accidents at Location-Specific Level: Application of XGBoost and SHAP." *Accident Analysis & Prevention* 166:106545.

Chen, Peng, and Jiangping Zhou. 2016. "Effects of the Built Environment on Automobile-Involved Pedestrian Crash Frequency and Risk." *Journal of Transport & Health*, Built Environment, Transport & Health, 3 (4): 44856.https://doi.org/10.1016/j.jth.2016.06.008.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. San Francisco California USA: ACM. https://doi.org/10.1145/2939672.2939785.

Chicco, Davide. 2017. “Ten Quick Tips for Machine Learning in Computational Biology.” *BioData Mining* 10 (1): 1–17. https://doi.org/10.1186/s13040-017-0155-3.

Das, Subasish, Anandi Dutta, Raul Avelar, Karen Dixon, Xiaoduan Sun, and Mohammad Jalayer. 2019. “Supervised Association Rules Mining on Pedestrian Crashes in Urban Areas: Identifying Patterns for Appropriate Countermeasures.” *International Journal of Urban Sciences* 23 (1): 30–48. https://doi.org/10.1080/12265934.2018.1431146.

Das, Subasish, Reuben Tamakloe, Hamsa Zubaidi, Ihsan Obaid, and Ali Alnedawi. 2021. “Fatal Pedestrian Crashes at Intersections: Trend Mining Using Association Rules.” *Accident Analysis & Prevention* 160 (September):106306. https://doi.org/10.1016/j.aap.2021.106306.

Ding, Chuan, Peng Chen, and Junfeng Jiao. 2018. “Non-Linear Effects of the Built Environment on Automobile-Involved Pedestrian Crash Frequency: A Machine Learning Approach.” *Accident Analysis & Prevention* 112 (March):116–26. https://doi.org/10.1016/j.aap.2017.12.026.

Ding, Feng, Jian Wang, Jiaqi Ge, and Wenfeng Li. 2018. “Anomaly Detection in Large-Scale Trajectories Using Hybrid Grid-Based Hierarchical Clustering.” *Int. J. Robot. Autom* 33:5–206.

Dumbaugh, Eric, Yanmei Li, Dibakar Saha, and Wesley Marshall. 2022. “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* 16 (December):100696. https://doi.org/10.1016/j.trip.2022.100696.

Dumbaugh, Eric, Yanmei Li, Dibakar Saha, Louis Merlin, and Florida Atlantic University. 2020. “The Influence of the Built Environment on Crash Risk in Lower Income and Higher-Income Communities.” CSCRS-R11. https://rosap.ntl.bts.gov/view/dot/56824.

El-Geneidy, Ahmed, Michael Grimsrud, Rania Wasfi, Paul Tétreault, and Julien Surprenant-Legault. 2014. “New Evidence on Walking Distances to Transit Stops: Identifying Redundancies and Gaps Using Variable Service Areas.” *Transportation* 41 (1): 193–210. https://doi.org/10.1007/s11116-013-9508-z.

Erickson, Nick, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. “AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data.” arXiv. http://arxiv.org/abs/2003.06505.

ESRI. 2023. “Multiscale Geographically Weighted Regression (MGWR).” 2023. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/multiscale-geographically-weighted-regression.htm.

Geurts, Pierre, Damien Ernst, and Louis Wehenkel. 2006. “Extremely Randomized Trees.” *Machine Learning* 63 (1): 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Guo, Qiang, Pengpeng Xu, Xin Pei, S. C. Wong, and Danya Yao. 2017. “The Effect of Road Network Patterns on Pedestrian Safety: A Zone-Based Bayesian Spatial Modeling Approach.” *Accident Analysis & Prevention* 99 (February):114–24. https://doi.org/10.1016/j.aap.2016.11.002.

Haddad, Angela J., Aupal Mondal, Chandra R. Bhat, Angie Zhang, Madison C. Liao, Lisa J. Macias, Min Kyung Lee, and S. Craig Watkins. 2023. “Pedestrian Crash Frequency: Unpacking the Effects of Contributing Factors and Racial Disparities.” *Accident Analysis & Prevention* 182 (March):106954. https://doi.org/10.1016/j.aap.2023.106954.

Haleem, Kirolos, Priyanka Alluri, and Albert Gan. 2015. “Analyzing Pedestrian Crash Injury Severity at Signalized and Non-Signalized Locations.” *Accident Analysis & Prevention* 81 (August):14–23. https://doi.org/10.1016/j.aap.2015.04.025.

Hossain, Ahmed, Xiaoduan Sun, Shahrin Islam, Shah Alam, and Md Mahmud Hossain. 2023. “Identifying Roadway Departure Crash Patterns on Rural Two-Lane Highways under Different Lighting Conditions: Association Knowledge Using Data Mining Approach.” *Journal of Safety Research* 85 (June):pp-52-65. https://doi.org/10.1016/j.jsr.2023.01.006.

Hossain, Ahmed, Xiaoduan Sun, Raju Thapa, and Julius Codjoe. 2022. “Applying Association Rules Mining to Investigate Pedestrian Fatal and Injury Crash Patterns Under Different Lighting Conditions.” *Transportation Research Record: Journal of the Transportation Research Board* 2676 (6): 659–72. https://doi.org/10.1177/03611981221076120.

Islam, Mouyid. 2023. “An Exploratory Analysis of the Effects of Speed Limits on Pedestrian Injury Severities in Vehicle-Pedestrian Crashes.” *Journal of Transport & Health* 28 (January):101561. https://doi.org/10.1016/j.jth.2022.101561.

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. 2023. “Statistical Learning.” In *An Introduction to Statistical Learning*, by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor, 15–67. Springer Texts in Statistics. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-38747-0_2.

Karim, Muhammad Monjurul, Yu Li, and Ruwen Qin. 2022. “Toward Explainable Artificial Intelligence for Early Anticipation of Traffic Accidents.” *Transportation Research Record* 2676 (6): 743–55. https://doi.org/10.1177/03611981221076121.

Katanalp, Burak Yiğit, and Ezgi Eren. 2022. “GIS-Based Assessment of Pedestrian-Vehicle Accidents in Terms of Safety with Four Different ML Models.” *Journal of Transportation Safety & Security* 14 (9): 1598–1632. https://doi.org/10.1080/19439962.2021.1978022.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. “Lightgbm: A Highly Efficient Gradient Boosting Decision Tree.” *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Khan, Md Nasim, and Mohamed M. Ahmed. 2020. “Trajectory-Level Fog Detection Based on in-Vehicle Video Camera with TensorFlow Deep Learning Utilizing SHRP2 Naturalistic Driving Data.” *Accident Analysis and Prevention* 142 (105521). https://doi.org/10.1016/j.aap.2020.105521.

Khan, Md Nasim, Subasish Das, and Jinli Liu. 2024. “Predicting Pedestrian-Involved Crash Severity Using Inception-v3 Deep Learning Model.” *Accident Analysis & Prevention* 197:107457.

Kim, Karl, and Eric Y. Yamashita. 2007. “Using a K-Means Clustering Algorithm to Examine Patterns of Pedestrian Involved Crashes in Honolulu, Hawaii.” *Journal of Advanced Transportation* 41 (1): 69–89. https://doi.org/10.1002/atr.5670410106.

Kock, Ned, and Gary Lynn. 2012. “Lateral Collinearity and Misleading Results in Variance-Based SEM: An Illustration and Recommendations.” *Journal of the Association for Information Systems* 13 (7). https://doi.org/10.17705/1jais.00302.

Kullback, Solomon, and Richard A. Leibler. 1951. “On Information and Sufficiency.” *The Annals of Mathematical Statistics* 22 (1): 79–86.

Li, Xiao, Siyu Yu, Xiao Huang, Bahar Dadashova, Wencong Cui, and Zhe Zhang. 2022. “Do Underserved and Socially Vulnerable Communities Observe More Crashes? A Spatial Examination of Social Vulnerability and Crash Risks in Texas.” *Accident Analysis & Prevention* 173 (August):106721. https://doi.org/10.1016/j.aap.2022.106721.

Li, Yang, Li Song, and Wei (David) Fan. 2021. “Day-of-the-Week Variations and Temporal Instability of Factors Influencing Pedestrian Injury Severity in Pedestrian-Vehicle Crashes: A Random Parameters Logit Approach with Heterogeneity in Means and Variances.” *Analytic Methods in Accident Research* 29 (March):100152. https://doi.org/10.1016/j.amar.2020.100152.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. “Explainable Ai: A Review of Machine Learning Interpretability Methods.” *Entropy* 23 (1): 18.

Ling, Lu, Wenbo Zhang, Jie Bao, and Satish V. Ukkusuri. 2023. “Influencing Factors for Right Turn Lane Crash Frequency Based on Geographically and Temporally Weighted Regression Models.” *Journal of Safety Research* 86 (September):191–208. https://doi.org/10.1016/j.jsr.2023.05.010.

Liu, Jinli, Subasish Das, and Md Nasim Khan. 2024. “Decoding the Impacts of Contributory Factors and Addressing Social Disparities in Crash Frequency Analysis.” *Accident Analysis & Prevention* 194 (January):107375. https://doi.org/10.1016/j.aap.2023.107375.

Liu, Jinli, Subasish Das, F. Benjamin Zhan, and Md Nasim Khan. 2024. “Spatial Analysis of Geographical Disparities in Pedestrian Safety.” *Transport Policy* 156 (September):164–81. https://doi.org/10.1016/j.tranpol.2024.06.018.

Lizarazo, Cristhian, and Víctor Valencia. 2018. “Macroscopic Spatial Analysis of Pedestrian Crashes in Medellin, Colombia.” *Transportation Research Record* 2672 (31): 54–62. https://doi.org/10.1177/0361198118758639.

Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. 2019. “Consistent Individualized Feature Attribution for Tree Ensembles.” arXiv. http://arxiv.org/abs/1802.03888.

Lundberg, Scott M., and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Ma, Lu, Xuedong Yan, Chong Wei, and Jiangfeng Wang. 2016. “Modeling the Equivalent Property Damage Only Crash Rate for Road Segments Using the Hurdle Regression Framework.” *Analytic Methods in Accident Research* 11 (September):48–61. https://doi.org/10.1016/j.amar.2016.07.001.

Mafi, Somayeh, Yassir AbdelRazig, and Ryan Doczy. 2018. “Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups.” *Transportation Research Record: Journal of the Transportation Research Board* 2672 (38): 171–83. https://doi.org/10.1177/0361198118794292.

Mathew, Sonu, Srinivas S. Pulugurtha, and Sarvani Duvvuri. 2022. “Exploring the Effect of Road Network, Demographic, and Land Use Characteristics on Teen Crash Frequency Using Geographically Weighted Negative Binomial Regression.” *Accident Analysis & Prevention* 168 (April):106615. https://doi.org/10.1016/j.aap.2022.106615.

Merlin, Louis A., Chris R. Cherry, Amin Mohamadi-Hezaveh, and Eric Dumbaugh. 2020. “Residential Accessibility’s Relationships with Crash Rates per Capita.” *Journal of Transport and Land Use* 13 (1): 113–28.

Miao, Congcong, Xiang Chen, and Chuanrong Zhang. 2024. “Assessing Network-Based Traffic Crash Risk Using Prospective Space-Time Scan Statistic Method.” *Journal of Transport Geography* 119 (July):103958. https://doi.org/10.1016/j.jtrangeo.2024.103958.

MobilityData. 2024. “Mobility Database.” 2024. https://mobilitydatabase.org/.

Mokhtarimousavi, Seyedmirsajad, Jason C Anderson, Atorod Azizinamini, and Mohammed Hadi. 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): pp-100-115. https://doi.org/10.1016/j.ijtst.2020.01.001.

Mwende, Sia Isaria, Valerian Kwigizile, Jun-Seok Oh, and Ron Van Houten. 2024. “Investigating Racial and Poverty-Level Disparities Associated with Pedestrian Nighttime Crashes.” *Transportation Research Record: Journal of the Transportation Research Board*. https://doi.org/10.1177/03611981241233294.

Nasri, Mehrdad, Kayvan Aghabayk, Arsalan Esmaili, and Nirajan Shiwakoti. 2022. “Using Ordered and Unordered Logistic Regressions to Investigate Risk Factors Associated with Pedestrian Crash Injury Severity in Victoria, Australia.” *Journal of Safety Research* 81:78–90.

Noland, Robert B., Nicholas J. Klein, and Nicholas K. Tulach. 2013. “Do Lower Income Areas Have More Pedestrian Casualties?” *Accident Analysis & Prevention* 59 (October):337–45. https://doi.org/10.1016/j.aap.2013.06.009.

Oh, Jutaek, Simon Washington, and Dongmin Lee. 2010. “Property Damage Crash Equivalency Factors to Solve Crash Frequency–Severity Dilemma: Case Study on South Korean Rural Roads.” *Transportation Research Record: Journal of the Transportation Research Board* 2148 (1): 83–92. https://doi.org/10.3141/2148-10.

Osama, Ahmed, and Tarek Sayed. 2017. “Evaluating the Impact of Connectivity, Continuity, and Topography of Sidewalk Network on Pedestrian Safety.” *Accident Analysis & Prevention* 107 (October):117–25. https://doi.org/10.1016/j.aap.2017.08.001.

Papathanasopoulou, Vasileia, Ioanna Spyropoulou, Haris Perakis, Vassilis Gikas, and Eleni Andrikopoulou. 2021. “Classification of Pedestrian Behavior Using Real Trajectory Data.” In *2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 1–6. https://doi.org/10.1109/MT-ITS49943.2021.9529266.

Park, Seung-Hoon, and Min-Kyung Bae. 2020. “Exploring the Determinants of the Severity of Pedestrian Injuries by Pedestrian Age: A Case Study of Daegu Metropolitan City, South Korea.” *International Journal of Environmental Research and Public Health* 17 (7): 2358. https://doi.org/10.3390/ijerph17072358.

Patwary, A. Latif, Antora Mohsena Haque, Iman Mahdinia, and Asad J. Khattak. 2024. "Investigating Transportation Safety in Disadvantaged Communities by Integrating Crash and Environmental Justice Data." *Accident Analysis & Prevention* 194 (January):107366. https://doi.org/10.1016/j.aap.2023.107366.

Pljakić, Miloš, Dragan Jovanović, and Boško Matović. 2022. "The Influence of Traffic-Infrastructure Factors on Pedestrian Accidents at the Macro-Level: The Geographically Weighted Regression Approach." *Journal of Safety Research* 83 (December):248–59. https://doi.org/10.1016/j.jsr.2022.08.021.

Pour-Rouholamin, Mahdi, and Huaguo Zhou. 2016. "Investigating the Risk Factors Associated with Pedestrian Injury Severity in Illinois." *Journal of Safety Research* 57 (June):9–17. https://doi.org/10.1016/j.jsr.2016.03.004.

Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. "CatBoost: Unbiased Boosting with Categorical Features." *Advances in Neural Information Processing Systems* 31. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.

Rahim, Md Adilur, and Hany M. Hassan. 2021. "A Deep Learning Based Traffic Crash Severity Prediction Framework." *Accident Analysis & Prevention* 154 (May):106090. https://doi.org/10.1016/j.aap.2021.106090.

Rahman, Mashrur, Kara M. Kockelman, and Kenneth A. Perrine. 2022. "Investigating Risk Factors Associated with Pedestrian Crash Occurrence and Injury Severity in Texas." *Traffic Injury Prevention* 23 (5): 283–89. https://doi.org/10.1080/15389588.2022.2059474.

Rahman, Md Sharikur, Mohamed Abdel-Aty, Samiul Hasan, and Qing Cai. 2019. "Applying Machine Learning Approaches to Analyze the Vulnerable Road-Users' Crashes at Statewide Traffic Analysis Zones." *Journal of Safety Research* 70 (September):275–88. https://doi.org/10.1016/j.jsr.2019.04.008.

Roll, Josh, and Nathan McNeil. 2022a. "Race and Income Disparities in Pedestrian Injuries: Factors Influencing Pedestrian Safety Inequity." *Transportation Research Part D: Transport and Environment* 107 (June):103294. https://doi.org/10.1016/j.trd.2022.103294.

Rousseeuw, Peter J. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20:53–65.

Russo, Brendan J., Emmanuel James, Cristopher Y. Aguilar, and Edward J. Smaglik. 2018. “Pedestrian Behavior at Signalized Intersection Crosswalks: Observational Study of Factors Associated with Distracted Walking, Pedestrian Violations, and Walking Speed.” *Transportation Research Record* 2672 (35): 1–12. https://doi.org/10.1177/0361198118759949.

Sanders, Rebecca L., and Robert J. Schneider. 2022. “An Exploration of Pedestrian Fatalities by Race in the United States.” *Transportation Research Part D: Transport and Environment* 107 (June):103298. https://doi.org/10.1016/j.trd.2022.103298.

Sasidharan, Lekshmi, Kun-Feng Wu, and Monica Menendez. 2015. “Exploring the Application of Latent Class Cluster Analysis for Investigating Pedestrian Crash Injury Severities in Switzerland.” *Accident Analysis & Prevention* 85 (December):219–28. https://doi.org/10.1016/j.aap.2015.09.020.

Shapley, L. 1997. “7. A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307-317.” In *Classics in Game Theory*, edited by Harold William Kuhn, 69–79. Princeton University Press. https://doi.org/10.1515/9781400829156-012.

Siddiqui, Chowdhury, Mohamed Abdel-Aty, and Keechoo Choi. 2012. “Macroscopic Spatial Analysis of Pedestrian and Bicycle Crashes.” *Accident Analysis & Prevention* 45 (March):382–91. https://doi.org/10.1016/j.aap.2011.08.003.

Stipancic, Joshua, Luis Miranda-Moreno, Jillian Strauss, and Aurélie Labbe. 2020. “Pedestrian Safety at Signalized Intersections: Modelling Spatial Effects of Exposure, Geometry and Signalization on a Large Urban Network.” *Accident Analysis & Prevention* 134 (January):105265. https://doi.org/10.1016/j.aap.2019.105265.

Su, Junbiao, N. N. Sze, and Lu Bai. 2021. “A Joint Probability Model for Pedestrian Crashes at Macroscopic Level: Roles of Environment, Traffic, and Population Characteristics.” *Accident Analysis & Prevention* 150 (February):105898. https://doi.org/10.1016/j.aap.2020.105898.

Sun, Ming, Xiaoduan Sun, and Donghui Shan. 2019. “Pedestrian Crash Analysis with Latent Class Clustering Method.” *Accident Analysis & Prevention* 124 (March):50–57. https://doi.org/10.1016/j.aap.2018.12.016.

Sung, Hyungun, Sugie Lee, SangHyun Cheon, and Junho Yoon. 2022. “Pedestrian Safety in Compact and Mixed-Use Urban Environments: Evaluation of 5D Measures on Pedestrian Crashes.” *Sustainability* 14 (2): 646. https://doi.org/10.3390/su14020646.

Tasic, Ivana, Rune Elvik, and Simon Brewer. 2017. “Exploring the Safety in Numbers Effect for Vulnerable Road Users on a Macroscopic Scale.” *Accident Analysis & Prevention* 109 (December):36–46. https://doi.org/10.1016/j.aap.2017.07.029.

Turner, Shawn M., Ipek N. Sener, Michael E. Martin, L.D. White, Subasish Das, Robert C Hampshire, Michael Colety, Kay Fitzpatrick, Ravindra K. Wijesundera, and United States. Federal Highway Administration. Office of Safety. 2018. “Guide for Scalable Risk Assessment Methods for Pedestrians and Bicyclists.” FHWA-SA-18-032. https://rosap.ntl.bts.gov/view/dot/43673.

United States Department of Transportation. 2024. “Highway Safety Information System (HSIS) | FHWA.” 2024. https://highways.dot.gov/research/safety/hsis.

US Department of Transportation. 2024. “Equitable Transportation Communitys.” 2024. https://www.transportation.gov/priorities/equity/justice40/download-data.

US EPA, OP. 2014. “Smart Location Mapping.” Data and Tools. February 27, 2014. https://www.epa.gov/smartgrowth/smart-location-mapping.

Wang, Xuesong, Junguang Yang, Chris Lee, Zhuoran Ji, and Shikai You. 2016. “Macro-Level Safety Analysis of Pedestrian Crashes in Shanghai, China.” *Accident Analysis & Prevention* 96 (November):12–21. https://doi.org/10.1016/j.aap.2016.07.028.

Wang, Yiyi, and Kara M. Kockelman. 2013. “A Poisson-Lognormal Conditional-Autoregressive Model for Multivariate Spatial Analysis of Pedestrian Crash Counts across Neighborhoods.” *Accident Analysis & Prevention* 60 (November):71–84. https://doi.org/10.1016/j.aap.2013.07.030.

World Health Organization. 2024. “Road Traffic Injuries.” 2024. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

Wu, Zhenxi, Aditi Misra, and Shan Bao. 2023. “Modeling Pedestrian Injury Severity: A Case Study of Using Extreme Gradient Boosting Vs Random Forest in Feature Selection.” *Transportation Research Record: Journal of the Transportation Research Board*. https://doi.org/10.1177/03611981231170014.

Xu, Pengpeng, and Helai Huang. 2015. “Modeling Crash Spatial Heterogeneity: Random Parameter versus Geographically Weighting.” *Accident Analysis & Prevention* 75 (February):16–25. https://doi.org/10.1016/j.aap.2014.10.020.

Zafri, Niaz Mahmud, and Asif Khan. 2022. “A Spatial Regression Modeling Framework for Examining Relationships between the Built Environment and Pedestrian Crash Occurrences at Macroscopic Level: A Study in a Developing Country Context.” *Geography and Sustainability* 3 (4): 312–24.

Zuo, Ting, Heng Wei, and Na Chen. 2020. “Promote Transit via Hardening First-and-Last-Mile Accessibility: Learned from Modeling Commuters’ Transit Use.” *Transportation Research Part D: Transport and Environment* 86 (September):102446. https://doi.org/10.1016/j.trd.2020.102446.