TOWARDS SAFER PEDESTRIANS: A FRAMEWORK FOR ANALYZING AND MITIGATING PEDESTRIAN VIOLATIONS AND RELATED SAFETY ISSUES

TOWARDS SAFER PEDESTRIANS: A FRAMEWORK FOR ANALYZING AND MITIGATING PEDESTRIAN VIOLATIONS AND RELATED SAFETY ISSUES

By

Haniyeh Ghomi Rashtabadi

B.Sc., M.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

Doctor of Philosophy (2023) McMaster University

Civil Engineering Hamilton, Ontario

TITLE: Towards Safer Pedestrians: A Framework for Analyzing and Mitigating Pedestrian Violations and related Safety Issues

AUTHOR: Haniyeh Ghomi

B.Sc., M.Sc. (Iran University of Science and Technology)

SUPERVISORS: Dr. Mohamed Hussein

NUMBER OF PAGES: xv, 246

To my soulmate, Reza, for his endless love and unlimited support

&

To all the strong and powerful Iranian women fighting for their rights

Abstract

Active models of travel, particularly walking, are an integral part of the multi-modal transportation system in urban areas. Walking provides numerous benefits at the individual and community levels (e.g., health benefits, reducing traffic congestion, emissions, and energy consumption). Nevertheless, safety concerns represent a major roadblock to the optimal utilization of walking as a key mode of travel. Pedestrians are among the most Vulnerable Road Users (VRUs) who are at a higher risk of being killed or severely injured as a result of road collisions. Previous research shows that many pedestrian behaviours could increase the risk of collisions significantly. Pedestrian violations, either temporal or spatial, stand as one of the riskiest behaviours that impact pedestrian safety. However, investigating such behaviour and quantifying its impact on safety are scarce in the literature. Accordingly, this research aims at developing a comprehensive framework to analyze pedestrian violations and understand when and where they can lead to collisions. To address these goals, the research utilized historical records of collisions that involve pedestrian violations. State-of-the-art statistical models (Copula models, Bayesian Structural Equation Modelling), Machine Learning techniques (Latent Class Analysis clustering), and Deep Learning methods (Self-Organizing Map) were applied to understand the factors contributing to such collisions on the micro-level (intersection and mid-blocks) and macro-levels (traffic analysis zones) and understand the characteristics of locations that experience a high frequency of those collisions. Additionally, a novel approach (dynamic R-vine copula-based time series model) was proposed to analyze the efficiency of pedestrian safety treatments that are implemented as part of vision zero programs. This approach enables the accurate assessment of the treatments, identifying the most effective combination of treatments, and investigating the association between area characteristics and treatment combination

performance. Overall, this dissertation provides a solid understanding of pedestrian violations and safety for decision-makers, safety practitioners, and academia.

Acknowledgment

I would like to express my sincere appreciation to my supervisor, Dr. Mohamed Hussein, for his expert guidance and timely encouragement throughout my entire Ph.D. I am also thankful to him for providing me with both professional and personal support to find my path as a researcher. I would like to acknowledge his endless energy and patience, which helped me successfully complete this dissertation while facing many challenges. I would also like to thank Dr. Hussein for his ongoing mentoring, which has and will continue to help me achieve my career goals.

In addition, I would like to thank my dissertation committee members, including Dr. Saiedeh Razavi and Dr. Moataz Mohamed, who kindly agreed to serve in my Ph.D. defense committee. Their invaluable comments and guidance have helped me to achieve a solid research path towards this dissertation. I also want to thank the external examiner, Dr. Emanuele Sacchi from University of Saskatchewan, for their valuable time reviewing my dissertation, supporting my work and giving me constructive feedback for extending this study.

I appreciate the research fund provided by the Ministry of Transportation of Ontario (MTO). I also express my gratitude to the City of Hamilton and the City of Toronto for providing the data.

I would like to thank former and current members of our research group at McMaster University, including Yasmina Imad Monzer, Abdul Basith Siddiqui, Hossam El-Din Helal, Mahdi Gabaire, Abdul Razak Alozi, Mohamed Gamal Khalil, and Chao Qi for their constant support and true friendship since the very beginning.

I would like to say that no words can express my heartful gratitude to my parents and my brother, who are my backbone and my source of wisdom. Last but not least, I want to express my deep appreciation to my husband and my lifelong friend (Reza) for supporting me and giving me new dreams to pursue.

Table of Contents

List of Figures

List of Tables

Declaration of Academic Contribution

This dissertation has been prepared and written in accordance with the rules for a sandwich thesis format required by the School of Graduate Studies (SGS) at McMaster University. The sandwich thesis is a compilation of journal articles published or prepared for publication. Chapters 2 to 5 are already published as journal articles, while Chapter 6 is recently submitted for publication as journal paper. This dissertation presents the research carried out solely by Haniyeh Ghomi. Advice and guidance were provided for the whole thesis by the academic supervisor Dr. Mohamed Hussein. Information presented from outside sources, which has been used towards analysis or discussion, has been cited where appropriate; all other materials are the sole work of the author. This thesis consists of the following manuscripts in the following chapters:

Chapter 2: Ghomi, H., & Hussein, M. (2022). **An integrated text mining, literature review, and meta-analysis approach to investigate pedestrian violation behaviours**. Accident Analysis & Prevention, 173, 106712. https://doi.org/10.1016/j.aap.2022.106712

Chapter 3: Ghomi, H., & Hussein, M. (2021). **An Integrated Clustering and Copula-based Model to Assess the Impact of Intersection Characteristics on violation-related Collisions**. Accident Analysis & Prevention, 159, 106283. https://doi.org/10.1016/j.aap.2021.106283

Chapter 4: Ghomi, H., & Hussein, M. (2023). **Analyzing the Safety Consequences of Pedestrian Spatial Violation at Mid-blocks: A Bayesian Structural Equation Modelling Approach**. Transportation Research Record, p. 03611981221097964. https://doi.org/10.1177/03611981221097964

Chapter 5: Ghomi, H., & Hussein, M. (2023). **Investigating the Application of Deep Learning to Identify Pedestrian Collision-prone Zones**. Journal of Transportation Safety & Security, 1-31. https://doi.org/10.1080/19439962.2022.2164636

Chapter 6: Ghomi, H., & Hussein, M. **Moving Vision Zero Programs Forward: What Countermeasure combinations work best and where? A Dynamic Copula-based Time-series Approach**. Submitted to Accident Analysis & Prevention in January 2023.

CHAPTER 1

Introduction

1.1. Background and Motivation

Active modes of transportation, including walking and cycling, are one of the main pillars of sustainable transportation systems. Walking and cycling can be ideal modes of travel for some short and medium trips, as well as completing the first and last miles of longer trips (e.g., access public transportation stations, travel from a parking location to the final destination, etc.). Active travel modes provide numerous advantages at the individual level (e.g., health benefits) and community level (e.g., reducing traffic congestion, decreasing air pollution, and reducing energy consumption). Nevertheless, safety concerns have been one of the major roadblocks to the full utilization of non-motorized transportation as key means of travel. Given the lack of physical protection in the event of a collision with motorized road vehicles, non-motorized travellers are more likely to sustain severe consequences of collisions compared to other road users (Elvik, 2010; Prati et al., 2018). Typically, active road users are among a group of road users that are usually referred to as Vulnerable Road Users (VRUs) in the safety literature. VRUs have a higher risk of being killed or severely injured due to road collisions and may include pedestrians, cyclists, motorcyclists, and persons with disabilities (Rifaat et al., 2011; Shinar, 2012; Vanlaar et al., 2016; Yannis et al., 2020). Nevertheless, this dissertation focuses mainly on pedestrians who are placed at the top of the vulnerability pyramid, developed by the Federal Highway

Administration (FHWA, 1998). Canadian collision statistics clearly show that the consequences of road collisions are not distributed equally among all road users, with pedestrians being overrepresented in collision fatalities and serious injuries. Pedestrians accounted for 39.4% (137 records) of collisions fatalities in Canada in 2021, despite representing 11.94% of persons involved in collisions. The disparate trend is notable in the province of Ontario as well, in which pedestrians accounted for 21.64% (108 records) of collision fatalities recorded in 2021, despite representing only 4.36% (2639 records) of persons involved in collisions. Further, there was an increase of 7.69% in the proportion of fatalities and severe injuries involving pedestrians despite a reduction of 8.59% in total pedestrian-vehicle collisions in 2021 compared to 2020 (MTO, 2022). Figure 1-1 demonstrates the annual frequency of fatal road collisions in Ontario from 2010 to 2021, along with the proportion of fatalities associated with pedestrians.

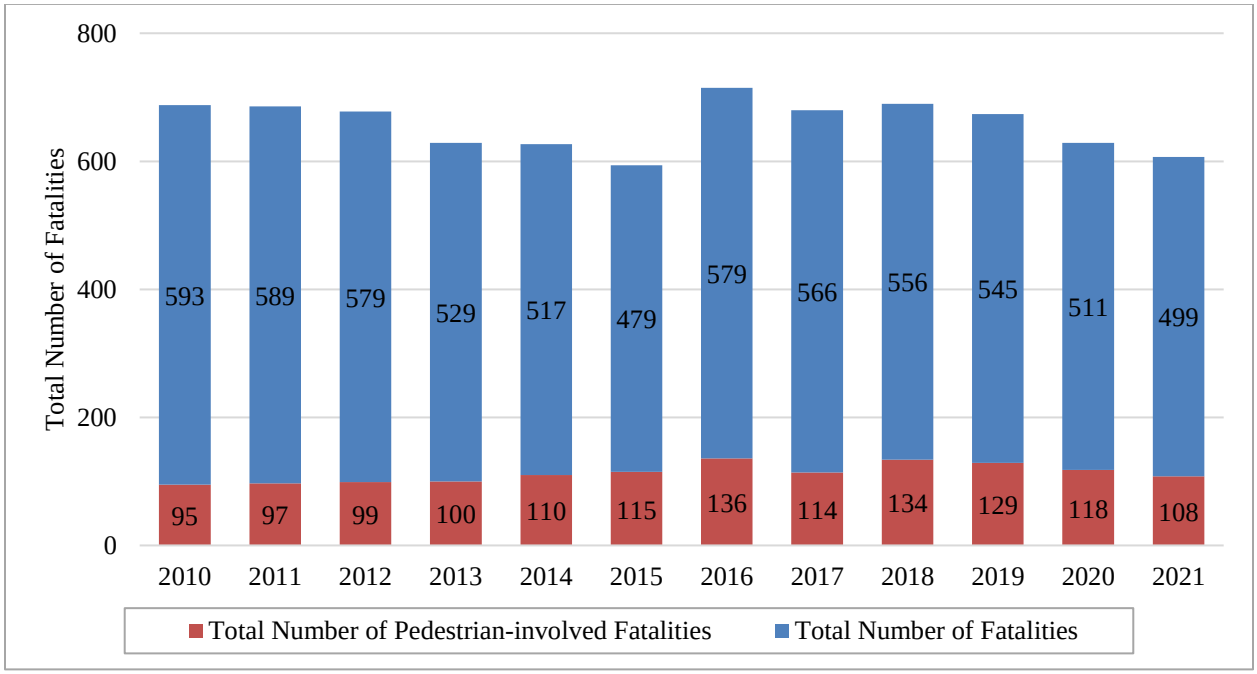


Figure 1-1 Fatal-involved collision statistics in Ontario from 2010 to 2021

In general, Figure 1-1 shows that while the frequency of collision fatalities has declined over the past decade, the ratio of pedestrian fatalities is on an uprising trend. Pedestrian fatalities accounted for 21.64% of all collision fatalities recorded in 2021 in Ontario, compared to 16.02% of the total fatalities in 2010. A closer look at road collision fatalities in Ontario would reveal that the majority of pedestrian-related collisions took place in Greater Toronto and Hamilton Area (GTHA) area. In 2021, 47.37% of total pedestrian fatalities in Ontario occurred in the GTHA area even though they were involved in only 6.48% of road collisions. These statistics emphasize the safety risks for pedestrians in the GTHA area, despite the adoption of numerous national safety plans and strategies that aim at enhancing their safety levels, such as Vision Zero programs.

In order to enhance pedestrian safety levels, transportation authorities implement a variety of interventions to mitigate fatal and serious injury collisions by managing vehicular traffic (e.g., traffic calming, speed limit reduction) and providing more pedestrian-friendly environments (e.g., sidewalk extensions and leading pedestrian intervals). However, pedestrian safety could be exacerbated by some risky behaviours that are rarely considered when designing safety interventions. Pedestrian unsafe behaviours while crossing the street are pinpointed as major contributors to increasing the risk of collisions, including traffic signal disobedience (Diependaele, 2019; Zhu et al., 2021; Wang et al., 2020; Zhang and Fricker, 2021b), jaywalking (Shiwakoti et al., 2020; Anik et al., 2021; Arhin et al., 2021), texting or conversing over the phone (Stavrinos et al., 2018; Tapiro et al., 2020), talking with a companion (Bungum et al., 2005; Thompson et al., 2013), impaired walking (Oxley et al., 2006; Reish et al., 2021), and

individual problems such as language difficulties, unfamiliarity with right-of-way rules and mental impairments (Hatfield et al., 2007). Pedestrian risky behaviours can be classified into three main categories: violating traffic rules, crossing distractedly, and crossing while intoxicated. According to the literature, violations were identified as key contributors to both the probability and the severity of pedestrian-vehicle collisions (Kim et al., 2017; Mukherjee and Mitra, 2020; Wang et al., 2019).

Pedestrian violations can be classified into two categories: temporal and spatial violations. Temporal violations occur at signalized intersections when road users traverse the intersection during a non-designated time. Spatial violations are identified when road users cross the road at a non-designated space. Spatial violations are observed in the mid-block (jaywalking) or at the intersections (crosswalk overflow). Statistics indicate that a significant proportion of pedestrian collisions that happened over the last five years (2017-2021) in the GTHA area are attributed to pedestrian violations. More importantly, while the proportion of pedestrian-vehicle collisions dropped significantly (33.91%) over the past five years, the ratio of collisions that involved pedestrian violations increased by 5.49% (Open Data Toronto, 2022; Open Data Hamilton, 2022). Such alarming statistics underscore the importance of investigating collisions that involve pedestrian violations to understand the main contributing factors to such collisions, the characteristics of the locations that encourage pedestrians to violate, and the strategies that can be implemented to mitigate violations and related collisions.

1.2. Issues and Challenges

Analyzing pedestrian violations and their implications on pedestrian safety is a challenging task due to several methodological issues and data limitations. The results of previous studies that studied this topic suffer from numerous shortcomings. For example, there is no conclusive evidence in the literature regarding the impact of various factors, such as average vehicle speed at a location, weather conditions, built-environment features, and various intersection features (such as the presence of refuge islands and countdown signals), on pedestrian violation behaviours. Moreover, previous studies provide little to no information regarding the impact of pedestrian network characteristics on pedestrian violations and subsequent safety issues. These shortcomings limit the comprehensive understanding of pedestrian violations, which restricts engineers and planners from developing appropriate mitigation strategies and designing pedestrian-friendly facilities that discourage pedestrian violations and enhance their overall safety levels.

The shortcomings of the results of previous studies can be attributed to several methodological challenges related to analyzing collision data. To start, the majority of previous studies that attempted to identify the contributing factors to pedestrian violations (or safety) relied mainly on developing traditional statistical models to model pedestrian violations (or collisions) as a function of a variety of potential contributing factors (Wang et al., 2020; Zhang and Fricker, 2021; Long et al., 2021; Pour-Rouholamin and Zhou 2016). However, the influence of the explanatory variables on violations (or collisions) may vary based on the prevailing traffic conditions and the characteristics of a location. Discovering the underlying patterns between the

different factors and pedestrian violations (or collisions) is not easily achievable through the analysis of the whole collision dataset using traditional statistical models. In fact, most of the statistical models are not capable of handling this heterogeneity issue among the explanatory variables.

Furthermore, previous studies that assessed the impact of violations on safety relied mainly on developing regression models, in which pedestrian violations were treated as an independent variable that impacts collision occurrence. This approach did not consider the personality traits of pedestrians while analyzing the violation behaviour. Some pedestrians inherently tend to take risks while crossing a road, regardless of the road characteristics and the presence of preventive countermeasures. Thus, ignoring the impact of such traits could bias the impact of violations on collision frequency and severity. From a statistical point of view, this endogeneity-biased outcome occurs due to the presence of a possible interrelationship between the independent variable in a model (i.e., pedestrian violations) and unobserved variables in the error term (i.e., the personality traits of pedestrians). Due to the impact of unobserved features, pedestrian violations could be endogenous to the occurrence and the consequence of the collisions. Accordingly, there is a need for utilizing advanced techniques to address pedestrian violation behaviour and its impact on pedestrian safety while addressing the abovementioned statistical issues.

Another important issue is related to the level of analysis in previous studies. The majority of published studies analyzed pedestrian violations at specific locations (micro-level analysis). Analyzing such hazardous behaviour on the macro level (e.g., city level) is not available in the

literature. Macro-level analysis of pedestrian violations and related safety issues can be very effective in identifying pedestrian safety issues in larger areas, understanding the characteristics of hazardous areas that experience a high frequency of collisions that involve pedestrian violations, and establishing long-term safety improvement policies. This macro-level analysis requires advanced analytical tools that can handle the complexity of the data on the macro-level, investigate the hidden relationship between a wide range of variables and pedestrian violations and safety, and recognize the unique characteristics of the hazardous areas.

Moreover, in order to mitigate pedestrian hazardous behaviours and their negative safety consequences, transportation engineers usually use a variety of mitigation strategies (safety countermeasures), ranging from specific countermeasures at specific locations, such as Lead Pedestrian Intervals (LPI), to wide-range policies, such as speed limit reduction and stricter enforcement. A crucial step to ensure the successful implementation of those safety measures is to continuously assess the impact of the adopted plans on enhancing pedestrian safety, which is important to guide future safety improvement plans and revise existing ones. Evaluating the performance of a safety intervention is typically achieved by conducting a before-and-after analysis or cross-sectional studies. However, two issues usually arise when using these two approaches. First, safety treatments can show different effects on safety over time. Some treatments can show an immediate impact on pedestrian behaviour and collisions, but over time, the impact can be reduced or even vanished. Other safety measures may not show an immediate impact, but in the long run, they can be very effective. As such, before-and-after and cross-sectional studies can yield incomplete conceptualization of the impacts of safety measures, as

they cannot easily explain the temporal trend of the different treatments. Second, the applicability of these techniques can be limited in conducting a system-wide evaluation to assess large safety initiatives. Many safety treatments are implemented as part of large system-wide strategies, such as vision zero. Accordingly, many locations (or areas) could benefit from more than one safety treatment at the same time. Therefore, pedestrian safety will be affected by more than one common error term due to the presence of interdependency between the impact of countermeasures. In this situation, evaluating the performance of each treatment in a separate manner is not accurate as it is not possible to attribute the change in collisions to one intervention. Thus, there is a need to adopt new approaches to analyze the effectiveness of safety treatments that are implemented as part of systemic initiatives, such as vision zero.

1.3. Research Objectives

The primary objective of this dissertation is to develop a better understanding of pedestrian violations and how they impact pedestrian safety. To that end, the dissertation investigates the prevalence of pedestrian-vehicle collisions that happen due to pedestrian violations, aiming to better understand their contributing factors, identify the characteristics of locations that experience a high frequency of such collisions, and assess the effectiveness of their mitigation strategies. The research utilizes several advanced statistical techniques to address the challenges discussed in the previous section. The specific objectives of the thesis are summarized as follows:

1) Understand the main contributing factors that encourage pedestrian violations and increase the risk of related collisions, on both micro and macro levels.

2) Investigate the latent relationship between pedestrian violations and safety through the analysis of pedestrian collisions that involve pedestrian violations at different elements of the urban transportation network.
3) Identify hotspot locations that experience a high frequency of collisions that involve pedestrian violations and understand their characteristics.
4) Investigate the applicability of numerous techniques in addressing the challenges associated with the data, particularly, the heterogeneity and endogeneity issues.
5) Propose an appropriate technique to evaluate the effectiveness of safety enhancement programs that target mitigating pedestrian collisions and risky behaviours.

1.4. Contributions

This dissertation provides several contributions to the current literature, summarized as follows:

- In line with the first objective of the research, the thesis presented text mining as a powerful and reliable tool for extracting information from published research. In addition, a meta-analysis framework was utilized in the thesis to develop a quantitative assessment of the factors that impact pedestrian violations and related safety issues.
- The thesis proposes the applicability of two statistical approaches to overcome the endogeneity issue of collision datasets (Objectives 2 and 4). First, a copula-based multivariate model with a joint structure was proposed to address the endogeneity issue between the unobserved explanatory variables and both the frequency and the severity of collisions that involve pedestrian violations at intersections. Second, a Bayesian Structural Equation Modeling (Bayesian SEM) technique was applied to analyze

collisions that involve pedestrian spatial violations (jaywalking). SEM technique is capable of addressing the endogeneity issue by considering unobserved (latent) variables while developing a model based on the observed explanatory variables. In other words, the main role of SEM models is to define a median variable (i.e., latent variable) to identify the hidden impacts of the observed variables on the dependent one. To the best of the authors' knowledge, this is the first research that utilizes these techniques in investigating the consequences of pedestrian violations.

- In order to understand the impact of different factors on pedestrian violations and related collisions this thesis proposed a two-staged approach that integrates a non-parametric clustering technique and a prediction model. Analyzing the impact of different explanatory variables on collisions within each cluster separately was shown to have merits as it helped to develop a better understanding of how different factors impact collision occurrence under different circumstances.
- The thesis introduces for the first time a macro-level analysis of pedestrian-vehicle collisions that happened due to pedestrian violations. The analysis was conducted using a non-parametric Deep Learning technique (i.e., Self-Organizing Map (SOM)) to identify collision-prone areas and understand the key variables that distinguish collision-prone areas from non-collision-prone ones (Objective 3). The proposed non-parametric SOM technique has superiority in handling big data and complex non-linear relationships among variables compared to the statistical and Machine Learning models.

- The research proposes a novel multivariate copula-based time series model to assess the efficiency of safety countermeasures that are implemented as part of large safety initiatives (Objective 5). The proposed approach successfully addressed the interdependency between the different countermeasures that are implemented in the same area and provided useful insights regarding the temporal trends of the different countermeasures.

1.5. Dataset

This dissertation considered two cities in the GTHA area, namely, the City of Toronto and the City of Hamilton, as the main areas for conducting the analysis. The City of Toronto is the capital city of the province of Ontario and the most populated city in Canada, with about 5,647,656 residents in 2021 (Statistics Canada, 2022). Toronto has a land area of 5,902.75 km^2 and a population density of 1050.7 people per square kilometer in 2021. Hamilton is the third largest population center in Ontario after greater Toronto and Ottawa, with an estimated 729,560 residents in 2021 (Statistics Canada, 2022). Combined, the two cities accounted for 25.18% of pedestrian fatalities in Ontario in 2021 (Open Data Toronto, 2022; Open Data Hamilton, 2022; MTO, 2022). The two cities have adopted the Vision Zero strategy, starting in 2016 in Toronto and 2019 in Hamilton. The frequency and severity of pedestrian-vehicle collisions were considered the primary data sources for this dissertation for the different studies conducted in this thesis. The collision data for the two cities were obtained from the Open Data portals of the two cities (Open Data Toronto, 2022; Open Data Hamilton, 2022). A summary of the collision data in the two cities is presented in the following sections. Multiple other data sources were

utilized to extract the different variables required for the different studies conducted in this thesis. The data sources and the extracted variables for each study will be discussed in the relevant chapters.

- **City of Hamilton collision data**

Despite an overall 10.2% reduction of pedestrian collisions in the City of Hamilton between 2017 to 2021, fatal pedestrian collisions reached a new high (9 fatalities) in 2021 (Hamilton Annual Collision Report, 2022). Pedestrians were overrepresented in collision fatalities in the city as they accounted for 56.3% of fatalities while only representing 2.61% of persons involved in collisions in 2021. Figure 1-2 shows the distribution of pedestrian-vehicle collisions that occurred in the City of Hamilton from 2017 to 2021.

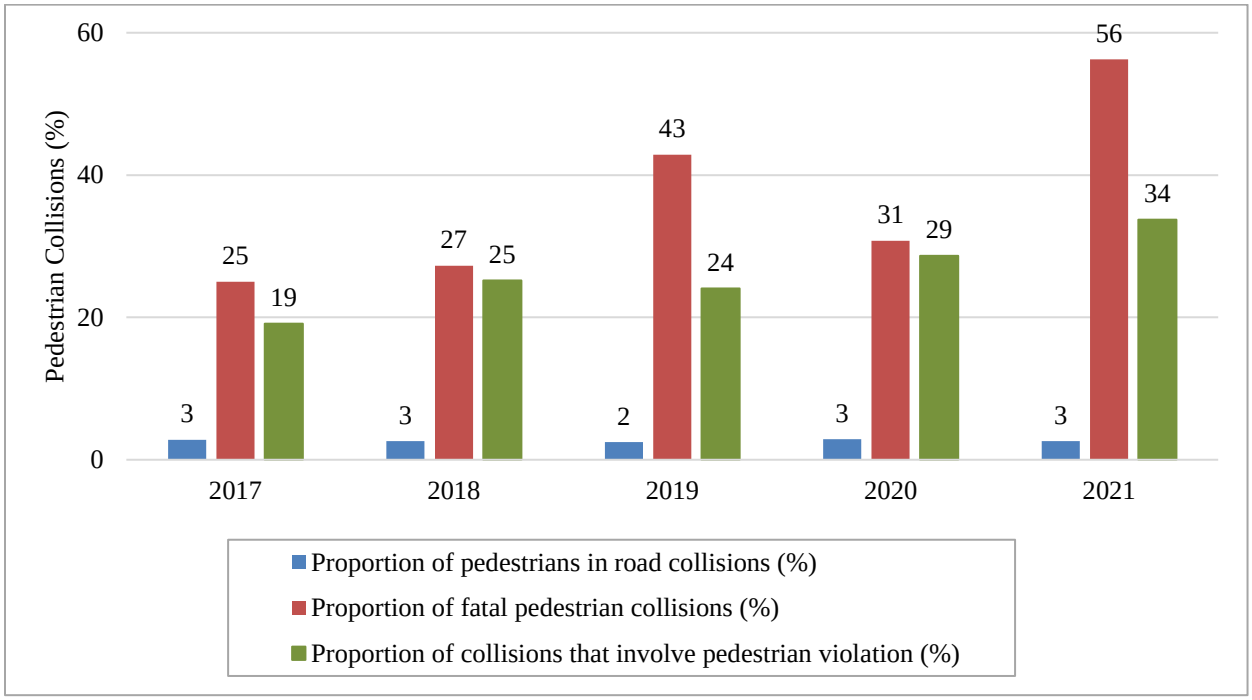


Figure 1-2 Pedestrian Collision Statistics in the City of Hamilton between 2017 and 2021

As shown in the figure, the ratio of pedestrian collisions remained stable during the past five years (around 2.7%); however, the ratio of fatal collisions continually hiked from 25% in 2017 to 56% in 2021, with the exception of 2020 (mainly due to lower traffic volume during covid lockdowns). In addition, Figure 1-2 emphasizes the negative impact of pedestrian violations on their safety. The ratio of pedestrian collisions that involved violations spiked from 19.11% in 2017 to 33.74% in 2021. The historical collision records of the City of Hamilton showed that about 90% of such collisions were serious collisions that involved either pedestrian fatalities or serious injuries (Open Data Hamilton, 2022). The spatial distribution of pedestrian collisions in Hamilton between 2017 and 2021 is presented in Figure 1-3.

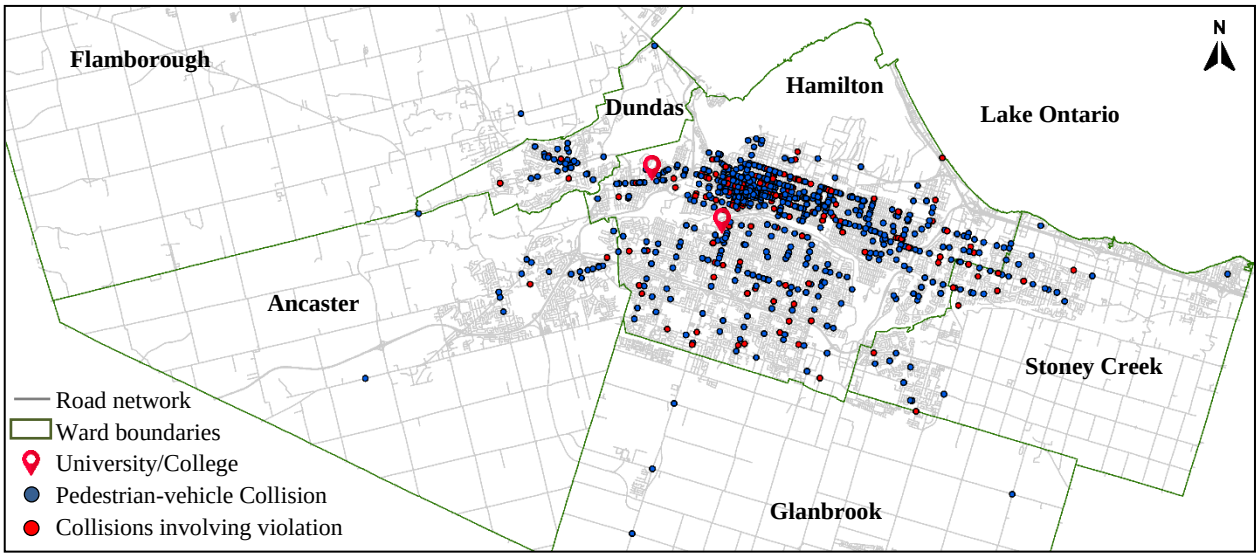


Figure 1-3 Spatial distribution of the pedestrian-vehicle collisions in Hamilton

- **City of Toronto collision data**

Figure 1-4 represents the spatial distribution of pedestrian-vehicle collisions in the city between 2017 and 2021. Overall, 1453 pedestrian-vehicle collisions were reported in the City of Toronto

during the five years considered in the analysis, resulting in 165 fatal collisions and 635 serious injury collisions. Of the 1453 collisions, 288 collisions (20%) were attributed to at least one type of pedestrian violation. Those 288 collisions involved 76 fatal collisions and 191 serious injury collisions.

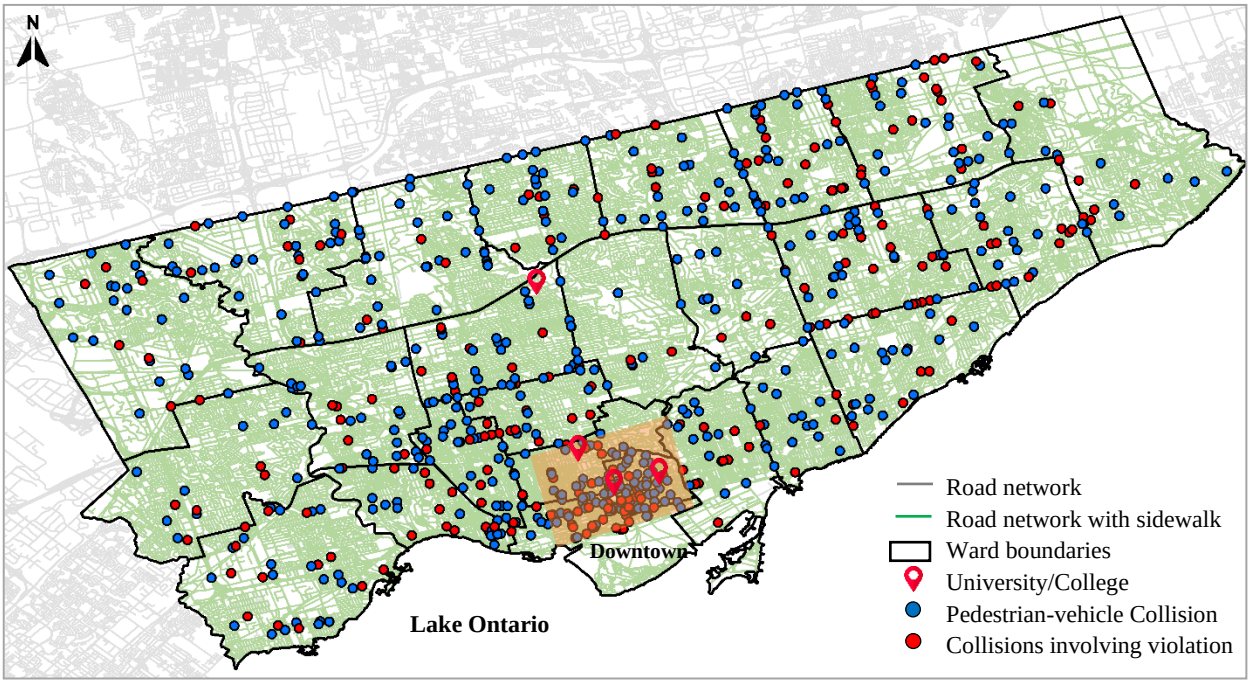


Figure 1-4 Spatial distribution of the pedestrian-vehicle collisions in Toronto

1.6. Thesis Organization

The thesis consists of seven chapters, summarized as follows:

- Chapter 1: provides the background and motivation of the work presented in this thesis, along with the research objectives, a brief description of the collision datasets used to undertake the analysis, and an overview of the thesis organization.

- Chapter 2: applies an integrated text mining, literature review, and meta-analysis approach to investigate the contributing factors to pedestrian safety and their violation behaviour, along with understanding the relationship between pedestrian violations and safety. This chapter addresses the first objective of the thesis. In this chapter, a Latent Dirichlet Allocation (LDA) Text Mining technique was applied to the title of the safety-related articles to achieve an all-rounded reference of pedestrian safety studies. This chapter also provides a four-layer framework to categorize the conducted holistic literature review based on locations that experience a high frequency of violations, the data collection and the research methods used to study such behaviour, the relationship between violations safety, and the factors that contribute to violations. Meanwhile, this chapter reviews the various strategies that can be adopted to mitigate pedestrian risky behaviours. Finally, this chapter undertakes a meta-analysis to develop a quantitative assessment of the factors that impact both pedestrian safety and violation. The results confirmed the significant impact of several factors, including the waiting time at the curbside, walking speed, the presence of bus stops and schools, and the presence of on-street parking on increasing the likelihood of pedestrian violations. On the other hand, the results did not provide conclusive evidence regarding the influence of some other factors on pedestrian violations, such as vehicle speed, the presence of refuge islands, countdown signals, pedestrian group size, and their trip purpose.
- Chapter 3: models the frequency and the severity of collisions involving pedestrian violations at intersections, aiming at identifying the contributing factors to those

collisions and understanding the safety consequences of pedestrian violations. This chapter provides a two-stage framework to address Objectives 2-4 of the dissertation. In this chapter, a Latent Class Analysis (LCA) clustering technique was first applied to divide the collisions that involve pedestrian violations into several homogeneous clusters based on the prevailing conditions of the traffic and intersection characteristics. Afterwards, a two-dimensional copula model was applied to model the frequency and severity of collisions in each cluster. The copula model was proposed to overcome the endogeneity issue between violations and the consequence of the collisions and to account for the heterogeneity among the explanatory variables. The results showed that the number of bus stops within the intersection area, the frequency of buses, and the presence of schools near the intersection are among the most influential factors that increase the frequency of collisions involving pedestrian violations.

- Chapter 4: identifies the impact of various factors on the frequency and severity of pedestrian-vehicle collisions that involve pedestrian spatial violations at mid-blocks. This chapter addresses Objectives 2-4 of the thesis by applying a Bayesian Structural Equation Modelling (SEM) framework to analyze pedestrian collisions that are attributed to spatial violations at mid-blocks in the City of Hamilton. The chapter evaluates the impact of numerous variables on violation-related collisions that were not thoroughly considered in the literature, such as pedestrian network connectivity and accessibility and a variety of location amenities and attractions. The chapter confirmed the hidden relationship between the four latent variables (namely, access to services, location vibrancy,

pedestrian network quality, and road size) and pedestrian collisions related to spatial violations. The results also showed that accessibility to services (e.g., parks, schools, bike-share stations, and bus stops) were the most influential factor on the frequency of collisions that involve spatial violation.

- Chapter 5: proposed a methodology to identify hotspot areas that promote pedestrian risky behaviours and experience a high frequency of related collisions, along with investigating the main characteristics of such areas. The second objective of the thesis was addressed in this chapter. The chapter identifies the collision-prone Traffic Analysis Zones (TAZs) that experience a high frequency of pedestrian collisions, either total collisions or those that involve pedestrian violations, in the City of Hamilton. This chapter investigates the main characteristics of such hazardous areas using Full Bayesian models with random intercepts. In addition, the chapter evaluates the applicability of the SOM model in identifying collision-prone areas and understanding the key variables that distinguish collision-prone areas from non-collision-prone ones. The results show that the SOM model identified collision-prone zones with a high accuracy that exceeded the traditional Bayesian approach, as confirmed by the results of a consistency test. The results also show that intersection density, density of bike-share stations, parking lot density, and pedestrian network directness are the most important factors in distinguishing between collision-prone and non-collision-prone zones.
- Chapter 6: provides a novel methodology for evaluating countermeasures that are installed to enhance pedestrian safety, as part of vision zero programs. This chapter

addresses the fifth objective of the dissertation. The research applied a dynamic R-vine copula-based multivariate time series modeling framework to understand the long-term efficiency of safety treatments focusing on pedestrians. In this way, a neighbourhood-level analysis was conducted using the City of Toronto's Vision Zero plans. The proposed model could address the countermeasures' co-intervention and identify the best combination of safety measures in each neighbourhood. The chapter reveals the main characteristics of neighbourhoods that receive significant benefits from each treatment over the analysis period, aiming to guide the macro-level allocation of safety treatments in the future.

- Chapter 7: provides recommendations/ideas for future research and concludes the thesis.

It is worth mentioning that chapters 2 to 5 are associated with four papers that have been published in peer-reviewed journals. Chapter 6 also represents a standalone manuscript that has been recently submitted to the Journal of Accident Analysis and Prevention and is currently under review.

1.7. References

Bai, L., & Sze, N. N. (2020). Red light running behavior of bicyclists in urban area: Effects of bicycle type and bicycle group size. Travel Behaviour and Society, 21, 226-234.

Cao, Y., Ni, Y., & Li, K. (2016). Effects of Refuge Island Settings on Pedestrian Safety Perception and Signal Violation at Signalized Intersections. 96th Annual meeting of Transportation Research Board.

Kim, M., Kho,, S. Y., & Ki, D. K. (2017). Hierarchical Ordered Model for Injury Severity of Pedestrian Crashes in South Korea. Journal of Safety Research, 61, 33–40. doi:doi.org/10.1016/j.jsr.2017.02.011

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

Long, X., Zhou, M., Zhao, H., & Song, Y. (2021). Pedestrian crossing decision during flashing green-countdown signal for urban signalized intersection. Journal of Transportation Safety & Security, 1-21.

Mukherjee, D., & Mitra, S. (2019). A comparative study of safe and unsafe signalized intersections from the view point of pedestrian behavior and perception. Accident Analysis and Prevention, 132, 105218.

Mukherjee, D., & Mitra, S. (2020). A comprehensive study on factors influencing pedestrian signal violation behaviour: Experience from Kolkata City, India. Safety Science, 124, 104610. doi:doi.org/10.1016/j.ssci.2020.104610

Ni, Y., Cao, Y., & Li, K. (2017). Pedestrians' Safety Perception at Signalized Intersections in Shanghai. Transportation Research Procedia, 25, 1955–1963. doi:doi.org/10.1016/j.trpro.2017.05.222

Osama, A., Sayed , T., & Sacchi, E. (2018). A Novel Technique to Identify Hot Zones for Active Commuters' Crashes. Transportation Research Record, 2672(38), 266-276.

Pour-Rouholamin, M., & Zhou, H. (2016). Investigating the risk factors associated with pedestrian injury severity in Illinois. Journal of Safety Research, 57, 9–17. doi:doi.org/10.1016/j.jsr.2016.03.004

Sacchi, M., Sayed, T., & El-Basyouny, K. (2015). Multivariate Full Bayesian Hot Spot Identification and Ranking New Technique. Transportation Research Record: Journal of the Transportation Research Board, 2515, 1-9.

Schleinitz, K., Petzoldt, T., Kröling, S., Gehlert, T., & Mach, S. (2019). (E-) Cyclists running the red light–The influence of bicycle type and infrastructure characteristics on red light violations. Accident Analysis & Prevention, 122, 99-107.

Statistics Canada. (2022, April). Retrieved from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm

Transport Canada. (2022, September). Retrieved from https://tc.canada.ca/en/road-transportation/statistics-data/canadian-motor-vehicle-traffic-collision-statistics-2019

Walters, C., & Ludwig, D. (1994). Calculation of Bayes posterior probability distributions for key population parameters. Canadian Journal of Fisheries and Aquatic Sciences, 51(3), 713-722.

Wang, J., Huang, H., Xu, P., Xie, S., & Wong, S. C. (2020). Random parameter probit models to analyze pedestrian red-light violations and injury severity in pedestrian–motor vehicle crashes at signalized crossings. Journal of Transportation Safety and Security, 12(6), 818-837.

Wang, K., Bhowmik, T., Yasmin, S., Zhao, S., Eluru, N., & Jackson, E. (2019). Multivariate copula temporal modeling of intersection crash consequence metrics: A joint estimation of injury severity, crash type, vehicle damage and driver error. Accident Analysis and Prevention, 125, 188-197. doi:doi.org/10.1016/j.aap.2019.01.036

World Health Organization. (2022, October). Retrieved from Road traffic injuries: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Yoneda, K., Suganuma, N., Yanase, R., & Aldibaja, M. (2019). Automated driving recognition technologies for adverse weather conditions. IATSS Research, 43, 253-262. doi:doi.org/10.1016/j.iatssr.2019.11.005

Zaki, M. H., Sayed, T., Tageldin, A., & Hussein, M. (2013). Application of Computer Vision to Diagnosis of Pedestrian Safety Issues. Transportation Research Record: Journal of the Transportation Research Board, No. 2393. doi:doi.org/10.3141/2393-09

Zhang, Y., & Fricker, J. D. (2021). Investigating temporal variations in pedestrian crossing behavior at semi-controlled crosswalks: A Bayesian multilevel modeling approach. Transportation Research Part F, 76, 92-108.

Zhu, D., Sze, N. N., Feng, Z., & Yang, Z. (2022). A two-stage safety evaluation model for the red light running behaviour of pedestrians using the game theory. Safety Science, 147, 105600.

CHAPTER 2

An Integrated Text Mining, Literature Review, and Meta-Analysis Approach to Investigate Pedestrian Violation Behaviours

The publication included in this chapter is:

Ghomi, H., & Hussein, M. (2022). An integrated text mining, literature review, and meta-analysis approach to investigate pedestrian violation behaviours. Accident Analysis & Prevention, 173, 106712. https://doi.org/10.1016/j.aap.2022.106712

The manuscript was submitted in September 2021 and accepted in May 2022. Haniyeh Ghomi is the main contributor of this manuscript. The co-author's contributions include guidance, supervision, funding, reviewing the analysis, and editing the manuscript.

2.1. Abstract

The goal of this study is to provide an overview of previous research that investigated pedestrian violation behaviour, with a focus on identifying the contributing factors of such behaviour, its impact on pedestrian safety, the mitigation strategies, the limitations of current studies, and the future research directions. To that end, the Latent Dirichlet Allocation (LDA) text mining method was applied to extract a comprehensive list of studies that were conducted during the past 21 years related to pedestrian violation behaviours. Using the extracted studies, a multi-sectional literature review was developed to provide a comprehensive understanding of the different aspects related to pedestrian violations. Afterward, a meta-analysis was undertaken, using the studies that reported quantitative results, in order to obtain the average impact of the different contributing factors on the frequency of pedestrian violations. The study found that pedestrian violations are one of the hazardous behaviours that contribute to both the frequency and severity of pedestrian-vehicle collisions. According to the literature, the waiting time at the curbside, traffic volume, walking speed, pedestrian distraction, the presence of bus stops and schools, and the presence of on-street parking are among the key factors that increase the likelihood of pedestrian violations. The study has also reviewed a wide range of strategies that can be used to mitigate violations and reduce the safety consequences of such behaviour, including simple engineering-based countermeasures, enforcement, solutions that rely on advanced in-vehicle technologies, and infrastructure connectivity features, educational programs, and public campaigns.

2.2. Introduction

Despite the strenuous efforts that are being made to enhance road safety, roadway collisions still represent a major public health problem that takes millions of lives every year. Among the different road user groups, non-motorized road users, such as pedestrians and cyclists, are considered by many to be among the most vulnerable groups, as they are at a higher risk of being killed or severely injured due to road collisions. Statistics show that pedestrians are overrepresented in fatal and severe injury of roadway crashes. Globally, pedestrians accounted for 17.2% of collision fatalities in 2019, despite representing less than 5% of persons involved in collisions (WHO, 2021). The same trend is observed on the national level as pedestrians accounted for 17.3% of collision fatalities in 2018 in Canada, despite representing only 3.4% of persons involved in collisions (Transport Canada, 2021).

Numerous studies investigated pedestrian safety issues using a wide range of techniques. The impact of many factors, such as traffic conditions, location characteristics, road network attributes, weather conditions, and other external factors on the frequency and severity of pedestrian-vehicle collisions was thoroughly assessed in the literature. Nevertheless, relatively less interest was given to study pedestrian unsafe behaviours and their impact on the overall pedestrian safety level. Previous studies showed that pedestrians who are involved in specific types of risky behaviours, such as violation, distracted walking, and walking while intoxicated, are more likely to be involved in crashes and experience more serious consequences of collisions. Among these unsafe behaviours, pedestrian violations were recognized as one of the most hazardous behaviours that influence pedestrian safety levels (Chen et al., 2011, Kim et al.,

2017; Mukherjee and Mitra, 2020; Wang et al., 2019). Pedestrian violation often refers to pedestrians' non-compliance to the rules of the road while crossing a street. Pedestrian violations can be classified into two broad categories: temporal violations and spatial violations (Sisiopiku and Akin, 2003; Zaki and Sayed, 2014; Hussein et al., 2015). Temporal violations occur when a pedestrian started to cross a signalized crosswalk during undesignated signal phases (i.e., Flash Do Not Walk or Do Not Walk). Spatial violations (i.e., jaywalking) are identified when a pedestrian crosses a road or an intersection at undesignated spaces. It is worth mentioning that the current study focused on objective safety, which refers to the actual number of collisions that occurred in a specific road segment over a specific period of time, not perceived safety, which refers to how pedestrians subjectively perceive the risk of collision in an interaction. Accordingly, the current study considered only previous research that developed an objective assessment of the impact of pedestrian violations on road safety. This includes studies that analyzed historical collisions that can be attributed to pedestrian violations. The study also considered previous research that analyzed surrogates of collisions, including, for example, pedestrian-vehicle conflicts.

A crucial step in alleviating the frequency of pedestrian violations and the consequent safety risks is to understand the contributing factors that encourage such behaviour, which enables the development of specific policies and engineering designs that mitigate such unsafe behaviour. Typically, research on pedestrian violations is conducted as part of pedestrian safety studies. This usually results in a lack of a comprehensive analysis of the factors that impact pedestrian behaviour and a scarcity of studies that provide a complete picture of such behaviour. Therefore,

it is necessary to identify to what extent the pedestrian violation behaviour has been investigated, what aspects have been analyzed, and what knowledge gaps need to be addressed in future studies. As such, the main goal of this study is to provide an overview of previous research that investigated pedestrian violation behaviour, with a focus on identifying the contributing factors of such behaviour, its impact on pedestrian safety, the utilized research methods implemented in previous research, the mitigation strategies of such behaviour, the limitations of current studies, and the potential future research directions. To that end, manual searches were conducted in the major publishers, including Elsevier, IEEE Xplore, SAGE, Scopus, Science Direct, Taylor & Francis, to extract the previous studies relevant to pedestrian violations over the past 21 years (2000-2021). The reference section of the reviewed articles was also considered as another source for finding the relevant studies. Then, the text mining technique was applied to automatically identify the relevant studies over the same period. The goal was to avoid missing any studies and provide a holistic database that covers pedestrian safety and behaviour topics.

In the second stage, a comprehensive literature review was conducted, using the identified studies, to understand where and why pedestrian violations develop and prevail, identify the utilized research methods and the data extraction approaches, assess the relationship between pedestrian violations and safety, and define the different mitigation strategies that have been proposed to mitigate pedestrian violations. Afterward, a meta-analysis framework was conducted to develop a quantitative assessment of the factors that impact pedestrian violations, based on the literature findings. Finally, knowledge gaps of the literature and the future research directions are discussed.

This study provides three main contributions: 1) the application of Text Mining to extract information of interest from unstructured huge textual databases (violation-related studies in this case); 2) conduct a robust literature review that aims at investigating pedestrian violation behaviour; 3) undertake a meta-analysis to develop a quantitative assessment of the factors that impact such behaviour.

The results of this study shall assist researchers to conduct more lucrative research in the area of pedestrian violations and put more emphasis on the underdeveloped areas. Also, the results will help transportation engineers, planners, and decision-makers to develop better design concepts to mitigate the frequency and severity of violations and enhance pedestrian safety.

The rest of the paper is organized as follows: The second section discusses the research methodology. The results of Text Mining are presented in section 3. The fourth section presents the findings of the literature review, while the fifth section addresses the results of the meta-analysis. The sixth section reviews the different strategies that have been proposed in the literature to mitigate pedestrian violations behaviour. Finally, section 7 presents the conclusions of the study and summarizes the future research directions of pedestrian violation studies.

2.3. Methodology

In order to achieve the study objectives, the Text Mining approach was developed to extract a comprehensive list of academic studies that were conducted during the past 21 years (from 2000 to 2021) related to pedestrian violation behaviours. In this regard, the Latent Dirichlet Allocation (LDA) method, which is one of the most famous Text Mining models, was applied to the title of the safety-related articles that were published in academic journals or presented at relevant

conferences. Before searching the papers, three criteria have been implemented in order to make the holistic resource database: 1) To achieve the most relevant studies, a set of valid and striking keywords related to the violation behaviour, including “violation, risky behaviours, illegal crossing, and safety perception” were distinguished to be utilized along with more specific words like “spatial and temporal violation, jaywalking, red-light violation, intersection, and mid-block areas”. Table 1 provides the full list of keywords that utilized to extract the studies; 2) The study focused on peer-reviewed academic papers that published in English-language scientific journals and presented at international conferences. Therefore, other studies (such as industrial reports published in commercial magazines and newsletters) that involved pedestrian violation analysis were removed from the further analysis; 3) The main focus of the current study is to investigate pedestrian violation behaviour from an engineering perspective, with an attempt to help transportation engineers and transportation planners to mitigate the frequency of the violation behaviours. Therefore, the studies that discussed the role of habits, cultures, and social norms were not considered for review.

Table 2-1 Full list of the keywords

violation	temporal violation	distraction	Pedestrian crossing
risky behaviours	jaywalking	Pedestrian behaviour	crosswalk
illegal crossing	red-light violation	Crossing behaviour	mid-block areas
safety perception	intersection	Crossing decisions	spatial violation

Using the extracted studies, a multi-sectional literature review was developed to provide a comprehensive overview of the different studied aspects related to pedestrian violation behaviour in the literature, including the locations in which pedestrian violations are common, the prevalent data collection methods in violation studies, the major techniques employed to investigate pedestrian violations, the main contributing factors that encourage pedestrians to violate, and the relationship between pedestrian violations and both the frequency and the severity of the related collisions. Subsequently, a range of mitigation strategies that have been proposed in the literature to reduce the frequency of violations and alleviate the safety consequences of such behaviour was investigated. Afterward, a meta-analysis was conducted, using the studies that reported quantitative results only, to develop an assessment of the average impact of the different contributing factors on the frequency of pedestrian violations. The following subsections provide a brief description of the LDA method and meta-analysis framework utilized in this study.

2.1.1 LDA Modeling

Text Mining refers to the process of exploiting beneficial information from a mixture of uncorrelated large textual data. The integration of Text Mining and Natural Language Processing (NLP) produced a new generation of probability-based Text Mining methods, named topic models. The main idea behind these models is that each topic in a document demonstrates the probability distribution over words in that specific document. LDA is a Bayesian-based topic model that was developed to extract topics from discrete datasets (Blei et al., 2003). The LDA method is developed based on a three-level hierarchical Bayesian model, which provides more accurate results compared to the distributional semantics models (such as the LSI model). In

addition, the LDA model is characterized by its modularity and extensibility, which enables the incorporation of more complicated models to enhance the accuracy of the text mining. Considering a series of n documents that each document consists of m words, a corpus could be shown as S_{ij} where i={1,…,n}, j={1,…,m}. LDA predicts the probability of a corpus as Equation (2-1):

$$P(S|\alpha\ ,\beta) = \prod_{s=1}^{N} \int P\ (\mu_i|\alpha\ )\left(\prod_{m=1}^{M_s} \sum_{z_{im}} P(z_{im}|\mu_i)P(S_{im}|z_{im},\beta)\right) d\mu_i \qquad (2\text{-}1)$$

where μ_i is the topic distribution for document i, ω_k represents the word distribution in the k^{th} topic, z_{im} shows the topic assigned to the m^{th} word in document i, and α and β are the parameters of the Dirichlet prior distribution for each topic and each word per topic, respectively. The current study considered a symmetric prior distribution in order to develop LDA model, as shown in Equation (2-2).

$$f(x,\alpha) = \frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K} \prod_{i=1}^{K} x_i^{\alpha-1} \qquad (2\text{-}2)$$

In a symmetric distribution α is equal to 1 and K is the dimension of the Dirichlet distribution.

Several methods were proposed to estimate the distribution of topic (μ) and word (ω). One of the well-known techniques is the Gibbs sampling method that estimates the probability of a value in each topic, which was dedicated to every word (Geman and Geman, 1984).

The effective sample size was equal to the number of topics (3048 records). In order to generate the posterior predictive checks, a posterior predictive distribution was simulated based on the observed variables and the predicted outcomes. Using “Tidy” package in R Studio software, there was a significant association between the predicted fitted model and the actual observed dataset. Moreover, the two cross-validation indexes, including Leave-One-Out cross-validation

Information Criterion (LOOIC) and Watanabe–Akaike Information Criterion (WAIC) were developed. The result showed the values of -1324.27 and -7804.91 for the two indexes, respectively.

Finally, two additional metrics, namely the "perplexity criteria" and the "topic coherence score" were reported to assess the LDA model performance. In language modeling, perplexity is a measure of the model accuracy (i.e., how well a model predicts a sample). It can be calculated according to Equations (2-3):

$$Perplexity = exp\left\{-\frac{\sum_{i=1}^{n} logp(f_i)}{\sum_{i=1}^{n} m_i}\right\} \qquad (2\text{-}3)$$

where n is the number of documents, f_i is the words in document i, m_i is the number of words in document i. The "perplexity criteria" was found to be 984.38 for the developed LDA model.

The topic coherence score is a measure of how interpretable the topics are to humans. Topics are represented as the top N words with the highest probability of belonging to that particular topic. The coherence score measures how similar these words are to each other. There are different measures of topic coherence, including, for example, the CV coherence score, the UMass coherence score, and the UCI coherence score. In this study, we reported the UMass coherence score. The UMass coherence score for the developed LDA model was found to be 0.5695.

2.1.2 Meta-Analysis Technique

The meta-analysis technique enables researchers to statistically combine the results of separate studies, which provides an opportunity to define the most important factors that impact a dependent variable and understand the overall impact of these factors when no conclusive results

have been reported in the literature (Rosenthal and DiMatteo, 2001). Initially, 137 studies that investigated the different factors that impact pedestrians' violation behaviour was reviewed. The reviewed studies were filtered to keep only the studies that provided a quantitative assessment of the impact of the factors under investigation. In addition, only studies that reported descriptive statistics of the results, including the sample size, mean, and standard deviation, were considered in the meta-analysis. Also, it should be noted that in order for a factor to be considered in the meta-analysis, it must appear in at least three different quantitative studies, which is the minimum standard of the meta-analysis concept (Rosenthal and DiMatteo, 2001). In total, 65 studies satisfied the aforementioned conditions and were therefore considered in the analysis.

The meta-analysis was conducted using the Comprehensive Meta-Analysis V3 software through two scenarios. The initial scenario was to conduct a unified meta-analysis framework to investigate the impact of the contributing factors on pedestrian violations, regardless of the type of violations. While the idea behind the second scenario was to split the studies into two broad categories based on the type of violations: spatial violation and temporal violation. Then, conduct a meta-analysis to assess the impact of the different factors on each of the two types of violations. The main issue of the latter scenario was that several variables were only investigated in one or two studies, which violates the minimum standard of the meta-analysis concept (Rosenthal and DiMatteo, 2001). However, such an assumption was ignored in order to assess the possibility of the first scenario. Therefore, two meta-analysis frameworks were employed to evaluate the impact of the factors on both temporal and spatial violations separately. According to the results, it was found that the impact of all factors on both types of violations displayed

similar trends (although the exact impact expressed in terms of the odds ratio varies). This means that factors that discourage pedestrians from violations (e.g., higher vehicle speeds and the presence of heavy vehicles) have the same impact on both spatial and temporal violations. Similarly, factors that encourage pedestrians to violate (e.g., lower traffic volume and the presence of refuge islands) have the same impact on both spatial and temporal violations. Although conducting a separate meta-analysis for each type of violation will enable better insight, analyzing a few studies may result in inaccurate insight. Therefore, the results of the first scenario will be discussed in this study.

For each variable, a meta-analysis was conducted in order to get an average impact of its impact on pedestrian violations. In the meta-analysis, the dependent variable is typically the outcome of the variable under investigation, which can be represented by many indicators (such as the odds ratio (OR), the relative risk (RR), probability (P), or the correlation coefficient (r)). In this study, the odds ratio indicator was used as the dependent variable. For the studies that did not report the odds ratios directly, the presented indicator was converted to the odds ratios in order to have a unified assessment indicator in all studies. For the studies that assessed the impact of a variable on the probability of violations, the odds ratio was calculated according to Equation (2-4) as follows:

$$OR = [P/(1-P)] \quad (2\text{-}4)$$

For studies that presented the relative risk of violation, the odds ratio was calculated according to Equation (2-5) as follows:

$$OR = [\frac{RR\ (1-P)}{1-(P*RR)}] \quad (2\text{-}5)$$

Also, the odds ratio can be calculated if a study reported the Correlation Coefficient between a variable and pedestrian violations, using Equation (2-6) below:

$$log(OR) = [\frac{2*r}{\sqrt{1-r^2}} * \frac{\pi}{\sqrt{3}}] \quad (2\text{-}6)$$

Once the odds ratio was estimated for all variables, the meta-analysis was conducted using the third version of the Comprehensive Meta-Analysis software to obtain an average estimate of the odds ratio of the variable under investigation among the different studies.

A random-effects modeling technique was utilized in this study, as it provides more flexibility to handle the studies with lower weights compared to the fixed effect modeling. Also, the 95% confidence level was considered for confidence interval calculations and the estimation of the variable's significance. The present technique involves the allocation of specific weights to individual studies based on their respective precision or standard error. In this study, a modified random-effects model was employed, which incorporates a penalty term for studies with lower weights. This penalty term was established in relation to the sample size of the studies, where studies with larger sample sizes are granted higher weights.

2.4. Meta-Analysis Technique

The major publishers including Elsevier, IEEE Xplore, SAGE, Scopus, Science Direct, Taylor & Francis were considered as the sources to search for relevant papers and provide a unique database related to pedestrian violation behaviour. In total, 3048 studies related to violations and safety over the past 21 years were found. The RStudio software was utilized to apply the Text Mining algorithm on the titles of the extracted articles, using several statistical packages. The

process of data clearance is the main step in Text Mining to address the linguistic noise issue. This problem refers to several types of linguistic diversities like letter case (e.g., pedestrian and PEDESTRIAN), the structure of words (e.g., study and studies), spelling (e.g., behaviour and behaviour), common words (e.g., the, to, in), and special symbols (e.g., punctuation). The procedure of data cleaning was adopted from Pahwa et al., (2018) in six steps to reduce the impact of linguistic noise on Text Mining analysis: 1) a code was assigned to each word in the title of the studies, 2) all the capital words were transferred to lowercase, 3) all the prefix and suffix words were removed to convert the words to their room format, 4) the spell checker package was activated in the R software to convert the words to a unit spelling format, 5) the common stop words were eliminated based on the stop word listed defined in NLP, and 6) the dataset was cleaned from any special characters, numbers, punctuations, symbols, extra white spaces, etc. After the implementation of the above-mentioned filters, 26,513 raw words detected before clearance were converted to 3,726 more specific words. Figure 2-1 shows the word clouds scheme for the initial raw words in the database and the clean one.

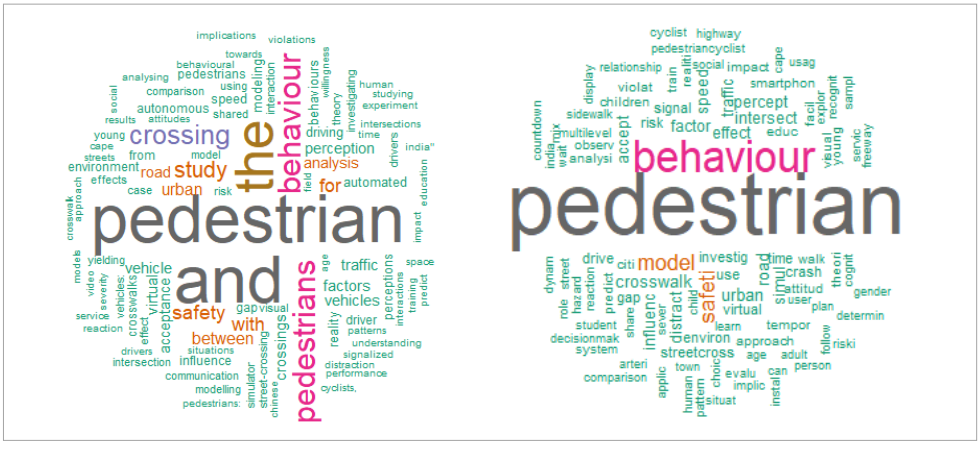


Figure 2-1 Generated word clouds for raw (left) and cleaned (right) words

In the concept of word clouds, the size of a word demonstrates the frequency of the word that is repeated in the text. It can be seen that in the raw database, the frequency of the common words (e.g., and, the, for, with) is relatively high. Also, the two words (pedestrian and pedestrians) were considered as two separate words. The two versions written types of behaviour and behaviour were considered separately as well. Then, the LDA model was applied to the remaining 3,726 words to extract the studies related to pedestrian violation behaviour. In order to find the optimal number of topics, a preliminary perplexity analysis was conducted based on the approach developed by Blei et al., (2003). In this approach, the LDA model will be examined based on several pre-defined topics to find the best performance of the model along with the optimum number of topics. In this study, the lowest rate of perplexity was gained through the definition of 6 topics. Figure 2-2 shows the top 5 frequent words that are repeated in each topic.

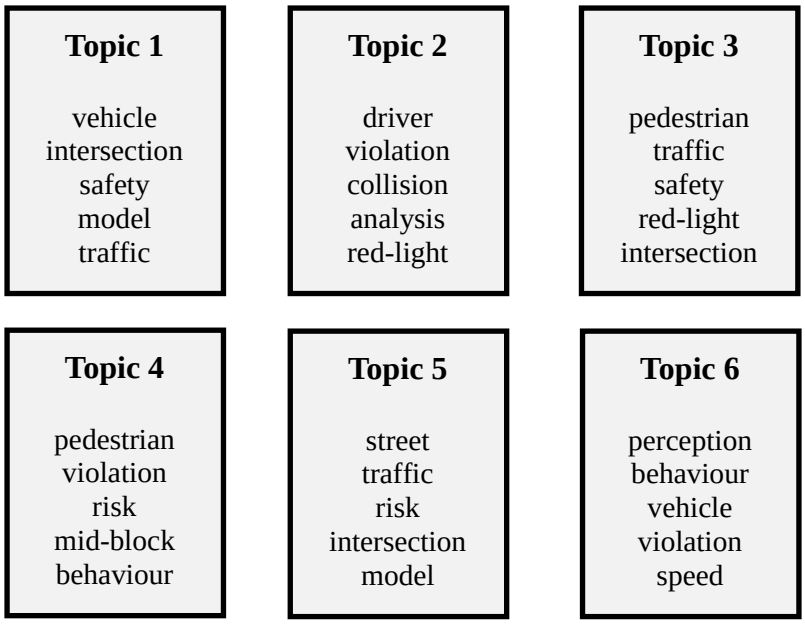


Figure 2-2 Top 5 frequent words in extracted 6 topics

Based on the frequent words, the process of topic identification will be started to assign a unique topic to each category. For example, the words “vehicle, intersection, safety, model, and traffic” in Topic 1 are most probably related to the statistical models developed to predict pedestrian-vehicle collisions that happen at intersections. The papers dedicated to Topics 3 and 4 are the only important studies related to the scope of the current study. In total, 137 studies that belonged to Topics 3 and 4 were considered as the main studies that will be reviewed in the study.

2.5. Literature Review

This section summarizes the findings of a multi-sectional literature review that was conducted to fully understand the pedestrian violation behaviour and its safety consequences. The first subsection focuses on evaluating where in the transportation network pedestrian violations were studied. The data collection approaches that were adopted in previous studies to analyze pedestrian violations are highlighted in the second subsection. The third subsection addresses the different research methods adopted to investigate pedestrian violations. The fourth subsection summarizes the main contributing factors that influence pedestrian violations. Finally, the fifth subsection provides a summary of the findings of previous studies that assessed the relationship between pedestrian violations and safety.

2.5.1 Study Locations

The majority of previous studies (68% of the reviewed studies) analyzed pedestrian violations, either temporal or spatial, at intersections. Intersections are believed to be the most hazardous

locations for pedestrians in the transportation network due to the high frequency of interactions between pedestrians and vehicles (Brosseau et al., 2013; Cao et al., 2016). Of the studies that analyzed pedestrian violations at intersections, 85% of the studies were conducted at signalized intersections (e.g., Fu and Zou, 2016; Hussein et al., 2015; Mukherjee and Mitra, 2019), while only 15% of the studies investigated pedestrian violations at unsignalized intersections (e.g., Aghabayk et al., 2021).

Nevertheless, some researchers shed light on the importance of studying pedestrian illegal crossings at mid-blocks (e.g., Sisiopiku and Akin, 2003; Tarko and Azam, 2011; Demiroz, et al., 2015; Mukherjee and Mitra; 2020). Roughly, 28% of the reviewed studies investigated pedestrian spatial violations at mid-block areas. Previous studies demonstrated that pedestrian violations at mid-blocks can be more dangerous and lead to more serious collisions, mainly due to the higher speed of vehicles and the fact that drivers do not expect to encounter pedestrians at these locations (e.g., Toran Pour et al., 2017; Papić et al., 2020).

Finally, a few studies investigated pedestrian violations at other locations such as parking lots or private driveways. Kim et al., (2008) showed that pedestrian violations could be risky in parking lots, especially since the pedestrian paths and the right of way are not usually well-defined in such locations. Sanchez (2009) demonstrated that walking diagonally between aisles in parking lots (defined as violations in the study) was the main reason for conflicts between vehicles and pedestrians. Kim and Ulfarsson (2019) indicated that the unclear definition of pedestrian access and pedestrian-unfriendly location of walkways in parking lots makes them violation-prone locations that are hazardous and hard to navigate through, especially for older pedestrians.

2.5.2 Data Collection Method

Various datasets were utilized in previous studies to analyze pedestrian violations, including collision data, surveys, video data, and data extracted from virtual reality and driving simulators. Analyzing historical collision data that involved pedestrian violations may be considered as the most accessible approach to assess such behaviour and evaluate its contributing factors. For example, Miranda-Moreno et al., (2011) analyzed pedestrian-vehicle collisions in Montreal, Canada to evaluate the impact of built-environment factors on pedestrian violations at signalized intersections. Mukherjee and Mitra (2020) integrated historical collision records with survey data to understand the relationship between temporal violations and the severity of the pedestrian-vehicle collisions at 55 signalized intersections in Kolkata, India. Ghomi and Hussein (2021) analyzed eight years of collision data in the City of Hamilton, Canada to identify the impact of the violation behaviour on the severity of collisions at intersections.

Another popular approach to study pedestrian violations in the literature was to analyze people's responses to survey questions. For example, Deb et al., (2017) distributed a 6-point Likert scale self-assessment questionnaire among a sample of 500 workers in the U.S. in order to investigate their behaviours while crossing the street. Sisiopiku and Akin (2003) surveyed university students on the Michigan University campus and found that more than 70% of the students reported that they cross a major street on campus at an undesignated location (jaywalking). Chu et al., (2004) designed a survey that aimed at assessing the impact of block sizes on pedestrian jaywalking behaviour. Ren et al., (2011) analyzed pedestrian responses to survey questions along with video data collected at 26 intersections to develop an understanding of the relationship

between pedestrian violations and the crosswalk length. Ni et al., (2017) administrated a survey that was conducted with pedestrians at 32 crosswalks in Shanghai, China to understand the impact of crosswalk characteristics on pedestrian violations. Dommes et al., (2015) analyzed survey responses to assess the relationship between on-street parking and temporal violations in Lille, France. Liu and Tung (2014) surveyed 32 pedestrians to analyze the association between vehicle speed and pedestrian violation at a signalized intersection in Yunlin, Taiwan.

Moreover, video data were advocated as a reliable data source to analyze pedestrian violation behaviour in many recent studies. For example, Zaki et al., (2013) analyzed video data collected at a major signalized intersection in Vancouver, Canada, using computer vision techniques, to analyze the crossing behaviour of pedestrians and investigate the relationship between violations and safety. Hediyeh et al., (2014) utilized computer vision to study the impact of walking speed on temporal violations, using video data collected at a signalized intersection in California. The impact of available traffic gap and waiting time on pedestrian violation was investigated through analyzing collected video data in various studies (e.g., Pawar and Patil, 2016; Russo et al., 2018; Brosseau et al., 2013). Cao et al., (2016) assessed the impact of wider medians on the probability of temporal violations in Shanghai, China using collected video data at a busy signalized intersection. Mukherjee and Mitra (2020) analyzed video data collected at 55 signalized intersections in Kolkata, India to assess the impact of on-street parking on pedestrian violation behaviour. The impact of pedestrian traits, such as walking speed, on pedestrian violations was also investigated using video data in many studies (e.g., Guo et al., 2016; Goh et al., 2012). The impact of pedestrian distraction, mainly by surfing on mobile and texting while walking, on

violations was also investigated using video data collected at 100 selected crosswalks in Cluj County, Romania (Hamann et al., 2017).

Additionally, a few studies relied on virtual reality and simulators to understand the impact of multiple factors on pedestrian violations, including for example the number of lanes (Petritsch et al., 2005), and presence of refuge islands (Ling et al., 2013). Table 2-2 provides a summary of the data collection methods including the number of samples and the investigated location.

Table 2-2 Summary of the data collection methods

Data collection method	Sample reference	Number of samples	Venue of interest
Historical collision records	Miranda-Moreno et al., (2011)	signalized intersections	Montreal, Canada
	Mukherjee and Mitra (2020)	55 signalized intersections	Kolkata, India
	Ghomi and Hussein (2021)	intersections (2010-2017)	Hamilton, Canada
Survey	Deb et al., (2017)	425 workers	United States
	Sisiopiku and Akin (2003)	711 university students	Michigan University campus
	Chu et al., (2004)	86 respondents	Tampa Bay area, Florida
	Ren et al., (2011)	26 intersections	3 cities (Nanjing, Wuhan, and Shizuishan), China
	Ni et al., (2017)	32 crosswalks	Shanghai, China
	Dommes et al., (2015)	422 adult pedestrians	Lille, France
	Liu and Tung (2014)	32 pedestrians	Yunlin, Taiwan
Video data	Zaki et al., (2013)	a major signalized intersection	Vancouver, Canada
	Hediyeh et al., (2014)	a signalized intersection	California, United States
	Pawar and Patil (2016)	two uncontrolled mid-block crossings	Kolhapur and Mumbai, Maharashtra
	Russo et al., (2018)	4 signalized intersections	New York and Arizona, United States
	Brosseau et al., (2013)	13 intersections	Montreal, Canada
	Cao et al., (2016)	a signalized intersection	Shanghai, China

	Guo et al., (2016)	12 unsignalized mid-block pedestrian crosswalks	Nanning, China
	Goh et al., (2012)	4 signalised and non-signalised crosswalks	Kuala Lumpur, Malaysia
	Hamann et al., 2017	100 crosswalks	Cluj County, Romania
Simulation	Petritsch et al., (2005)	3 major intersections	Florida, United States
	Ling et al., (2013)	12 crosswalks	China

2.5.3 Analysis Method

Previous studies adopted a wide range of methods to investigate the violation behaviour of pedestrians, as summarized below. It should be noted that some studies employed more than one technique to conduct the analysis.

2.5.3.1 Classical Statistical Modeling

Statistical modeling is one of the most applied techniques for analyzing pedestrian violation behaviour (Goh et al., 2012; Li and Fernie, 2010). Previous studies adopted a wide range of statistical models to analyze collision data, survey responses, or other data sources, to assess the impact of different factors on pedestrian violations and establish the relationship between violation and collisions. Several studies relied on simple linear regression models to identify the significant factors that impact pedestrian violations (e.g., Bernhoft and Carstensen, 2008; Bian et al., 2009; Ling et al., 2013). Petritsch et al., (2005) investigated the impact of several significant factors on the frequency of temporal violations through the implementation of stepwise regression analysis. The logistic regression model was also commonly utilized to study pedestrian violations. For example, Chen et al., (2017) investigated the impact of waiting time on pedestrian temporal violation through the implementation of a binary logit model to analyze

video data collected from 13 signalized intersections in Suzhou, China. The same model was utilized in several other studies (Koh and Wong, 2014; Pawar and Patil, 2016; Brosseau et al., 2013; Liu and Tung, 2014; Kang et al., 2013; Ni et al., 2017; and Zhang et al., 2016). Generalized linear models were also implemented in some studies (e.g., Mukherjee and Mitra, 2020). Miranda-Moreno et al., (2011) developed negative binomial and log-linear regression models to investigate the effect of built-environment factors on the violation behaviour of pedestrians at signalized intersections in Montreal, Canada. The discrete choice model was also implemented to extract the main contributing factors associated with pedestrian decision-making in several studies (e.g., Papadimitriou, 2012). Wang et al., (2020) developed a random parameter probit model to investigate the main contributing factors on pedestrian temporal violations in Hong Kong, based on historical collision records (2010-2012).

2.5.3.2 Advanced Statistical Models

Advanced multivariate regression models were also applied in violation-related studies. Ghomi and Hussein (2021) applied a copula-based model to study the factors that impact pedestrian violations at intersections in Hamilton, Canada. Other studies developed Structural Equation Modeling (SEM) to analyze safety perception and subjective norms of the pedestrian while crossing the street (e.g., Mo and Mo, 2017; Zhou et al., 2016; Kummeneje and Rundmo, 2019). Zhang and Fricker (2021b) utilized a Bayesian multilevel logistic model to analyze pedestrian jaywalking at a semi-controlled road segment in Indiana, United States.

2.5.3.3 Risk-based Methods

Few studies that employed risk-based models in order to quantify the probability of a pedestrian engaging in a violation-related collision are found in the literature. The hazard-based duration method was applied to study the impact of pedestrian group size on the probability of violation in Guo et al., (2011). King et al., (2009) developed a relative risk ratio to investigate the relationship between pedestrian violations and safety. Long et al., (2021) developed a decision model to assess the impact of two risk functions (cost and time) on pedestrian attitude while crossing the crosswalk during the Flash-Do-Not-Walk phase at a signalized intersection in China.

2.5.3.4 Cross-Sectional and Before-and-After Analysis

The use of cross-sectional analysis and time-series (before-and-after) studies to investigate the change in the violation behaviour following changes to the intersection design or signal timing was quite popular in the literature. Evaluating the impact of pedestrian countdown signals on the violation behaviours was examined in many previous studies, using cross-sectional analysis (e.g., Lipovac et al., 2013; Fu and Zou, 2016) or before-and-after study (e.g., Arhin and Noel, 2007). For example, Schattler et al., (2007) conducted a before and after evaluation of the behaviour of pedestrians after installing countdown signals at 13 intersections in Illinois. The study showed that the rate of pedestrians' temporal violations decreased significantly after the installation of the countdown signals. Cross-sectional studies were also conducted to assess the impact of other factors on pedestrian violations, such as the block size (Oakes et al., 2007) and pavement marking at crosswalks (Guo et al., 2016).

2.5.3.5 Machine Learning Algorithms

Clustering Machine Learning algorithms, which refer to the process of grouping similar correlated events in a set of homogenous clusters, were among the adopted methods in the literature to analyze pedestrian violations. For example, Papadimitriou et al., (2013) employed Principal Component Analysis (PCA) and a Two-Step clustering algorithm to investigate violators' attitudes in 19 European countries. Sasidharan et al., (2015) utilized a Latent Class Clustering technique to investigate the impact of pedestrians' temporal violation on the severity of collisions in Switzerland. Ghomi and Hussein (2021) developed a Latent Class model to identify the underlying contributing factors to pedestrian violations at 759 intersections in Hamilton, Canada.

Classification Machine Learning algorithms was another category that was utilized in previous studies to investigate violation behaviours. For example, Lyons et al., (2001) investigated the impact of gap acceptance on pedestrian temporal violations. The study implemented a traditional artificial neural network to analyze the observed illegal crossings at several signalized mid-blocks in Wales, UK. Zhang, et al., (2020) applied a Recurrent Neural Network model to investigate the characteristics of spatial violators at a signalized intersection adjacent to the University of Central Florida.

Meanwhile, several studies attempted to predict the impact of many contributing factors on pedestrian unsafe behaviours using predictive Machine Learning models. For example, Kadali et al., (2014) developed an Artificial Neural Network model to predict the impact of pedestrians' waiting time on their gap acceptance behaviour. In another study, Anik et al., (2021) developed

an Artificial Neural Network model to predict the probability of pedestrians' jaywalking at several mid-blocks in Dhaka, Bangladesh.

2.5.3.6 Microsimulation Models

Few studies that relied on microsimulation models to investigate pedestrian unsafe behaviours are found in the literature. For example, Yang et al., (2006) developed a microsimulation model to investigate the impact of gap acceptance on pedestrian temporal violations near two universities located in Xi'an, China. The study found that the higher frequency of pedestrian violations is directly impacted by pedestrian group size and the speed of the approaching vehicles. In another study, Ibitoye et al., (2021) developed a VISSIM-based microsimulation model to analyze the impact of many factors, including pedestrian spatial violations on the frequency of pedestrian-vehicle conflicts. The study found that longer waiting time acts as a motivator to spatial violations.

In summary, various methods and techniques were used in the literature to study pedestrian violation behaviour as summarized in Table 2-2. Meanwhile, a comprehensive review was conducted on the studies developed to investigate the frequency and severity of pedestrian-vehicle collisions. As shown in the table, both categories of studies employed similar techniques, with traditional statistical models being the predominant technique employed to study pedestrian safety and their crossing behaviour.

Table 2-3 Summary of the methods adopted to investigate pedestrian violations and their safety

Approach	Sample reference		
	Violation	**Collision**	**Collision involved violation**
Classical Statistical models	**Statistical tests:** Goh et al., (2012); Li and Fernie (2010) **Simple linear regression:** Bernhoft and Carstensen (2008); Bian et al., (2009); Ling et al., (2013) **Stepwise regression model:** Petritsch et al., (2005) **Generalized linear models:** Papadimitriou (2012); Koh and Wong (2014); Pawar and Patil (2016); Brosseau et al., (2013); Liu and Tung (2014); Kang et al., (2013); Ni et al., (2017); Zhang et al., (2016); Chen et al., (2017); Wang et al., (2020)	**Generalized linear models:** Miranda-Moreno et al., (2011); Haleem et al., (2015); Moudon et al., (2011); Tefft (2013); Zegeer et al, (2004); Shankar et al., (2003); Aidoo et al., (2013); Ulfarsson et al., (2010)	**Simple linear regression:** Loukaitou-Sideris et al., (2007) **Generalized linear models:** Mukherjee and Mitra (2020); Pour-Rouholamin and Zhou (2016); Kim et al., (2008); Mukherjee and Mitra (2019)
Advanced Statistical models	**Bayesian multilevel logistic model:** Zhang and Fricker (2021b) **Copula-based model:** Ghomi and Hussein (2021) **Structural Equation Modeling:** Mo and Mo (2017); Zhou et al. (2016); Kummeneje and Rundmo (2019)	**Hierarchical logistic model:** Forbes (2015) **Hierarchical Bayesian random effects:** Song et al., (2020) **Structural Equation Modeling:** Al-Mahameed et al., (2019); Sheykhfard et al, (2021b)	**Hierarchical logistic model:** Kim et al., (2017) **Copula-based model:** Wang et al., (2019)

Risk-based methods	**Hazard-based duration method:** Guo et al., (2011) **Risk decision model:** Long et al., (2021)	**Hazard-based duration method:** Haque et al., (2021); Al Kaabi et al., (2012)	**Relative risk ratio:** King et al., (2009)
Cross-sectional and Before-and-After analysis	**Cross-sectional analysis:** Lipovac et al., (2013); Fu and Zou (2016); Oakes et al., (2007) **Before-and-after analysis:** Arhin and Noel (2007); Schattler et al., (2007)	**Cross-sectional analysis:** Guo et al., (2016); Amoh-Gyimah et al., (2016) **Before-and-after analysis:** Dommes et al., (2012); King et al., (2003); Nie and Zhou (2016)	
Machine Learning algorithms	**Unsupervised learning:** Papadimitriou et al., (2013); Ghomi and Hussein (2021); Sasidharan et al., (2015) **Supervised classification methods**: Lyons et al., (2001); Zhang, et al., (2020) **Supervised regression models:** Kadali et al., (2014); Anik et al., (2021)	**Unsupervised learning:** Mohamed et al., (2013); Kaplan and Prato (2013) **Supervised classification methods**: Casali et al., (2021); Pineda-Jaramillo (2020); Das et al., (2020) **Supervised regression models:** Ding et al., (2018); Ka et al., (2019); Li et al., (2020)	**Unsupervised learning:** Sasidharan et al., (2015)
Microsimulation models	Yang et al., (2006); Ibitoye et al., (2021)	Zaki et al., (2013); Hussein et al., (2015); Waizman et al., (2015);	Puscar et al., (2018)

Moreover, the last column in Table 2-3 shows the studies that investigated pedestrian-vehicle collisions that involved pedestrian unsafe behaviours. As shown in the table, limited studies were conducted in this area, which shows the need for future research to study the impact of violations on pedestrian safety in more detail.

2.5.4 Contributing Factors to Pedestrian Violations

According to the literature, the contributing factors to pedestrian violations at a specific location can be divided into six categories, including traffic-related factors, location-specific factors, pedestrian-related factors, environmental and external factors, built-environment factors, and socio-economic factors. A brief discussion of the six categories is provided as follows:

2.5.4.1 Traffic-related Factors

Most of the previous studies showed that the frequency of pedestrian violations decreases significantly as roads become more congested. The studies explained that the unavailability of adequate gaps between the vehicles in congested conditions discourages pedestrians from crossing illegally (Hamed, 2001; Yagil, 2000; Pawar and Patil 2016; Yoneda et al., 2019; Zhu et al., 2021a). Meanwhile, other studies showed that the probability of gap acceptance is dependent on the pedestrian age and the waiting time before crossing (Brewer et al., 2006; Nassr et al., 2017; Zhuang and Wu, 2011). Oxley et al., (2005) studied the relationship between the available gap on a one-way street and the spatial violation behaviour in three age groups (30-45, 60-69, and older than 75 years old). The results demonstrated that pedestrians between 60-69 years old were the least probable group to jaywalk unless a significantly large gap between approaching vehicles is available. Surprisingly, pedestrians over 75 years old accepted riskier gaps in more

than 70% of the cases, which raised questions regarding the visual abilities, the visual processing speed, and the reaction time of senior pedestrians.

Moreover, previous studies assessed the impact of vehicle speed on the frequency of pedestrian violations. The majority of studies treated vehicle speed as a categorical variable (i.e., low speed, high speed.) rather than a continuous one. The majority of the studies identified the average vehicle speed as another traffic-related factor that contributes to pedestrian violation (Kadali et al., 2014; Zhang and Fricker, 2021a).

Several studies showed that the relationship between violation and speed vary among pedestrians of different age groups (Lobjois and Cavallo, 2007; Liu and Tung 2014). On the other hand, other studies did not find a correlation between vehicle speed and violation behaviour of pedestrians in any age group (Alexander et al., 2002; Oxley et al., 2005).

Additionally, previous studies investigated the impact of the type of vehicles on the frequency of pedestrian violations. In summary, studies showed that pedestrians preferred to wait for a longer time on the curbside and did not initiate a risky crossing when encountering heavy and large vehicles (Hamed 2001; Zhuang and Wu 2011; Zhu et al., 2021a).

Furthermore, previous studies were consistent regarding the impact of parked vehicles near the crosswalk on pedestrian violation behaviours. Studies showed that the frequency of pedestrian violations increases at locations where on-street parking is provided near crosswalks, mainly due to the lack of visibility of approaching vehicles (Dommes et al., 2015; Jahandideh et al., 2017). Mukherjee and Mitra (2020) found that parked vehicles and other obstacles near pedestrian crosswalks increase the probability of jaywalking significantly.

2.5.4.2 Location-specific Features

According to the literature, location-specific factors that have an impact on the violation behaviour of pedestrians include the presence of a central refuge island, signal timing, the presence of countdown pedestrian signal, the number of roadway lanes, and intersection/roadway width.

While many studies reported significant safety benefits for central refuge islands (e.g., Aidoo et al., 2013; Pour-Rouholamin and Zhou, 2016), the majority of the previous studies showed that the presence of central refuge islands encourages pedestrians to accept riskier crossing behaviour. Ishaque and Noland (2008) found that the compliance of pedestrians with traffic signals would drop drastically with the presence of central refuge islands at locations where pedestrians must wait for a long time to cross the street. The idea is that pedestrians do not have to find an adequate gap in both directions, since they can wait in the median and wait for an adequate gap in the other direction (Li and Ferine, 2010). The impact of medians' width on the probability of temporal violations was also investigated by Cao et al., (2016) and showed that the likelihood of temporal violations increased by 15% for each 1% increase in width of the central medians. However, the analysis conducted by Xu et al., (2013) revealed the negative impact of pedestrian infrastructure at intersections, like medians, on the frequency of violations.

Previous research presented much evidence regarding the strong correlation between pedestrian temporal violations and signal timing. In fact, proper signal design leads to a reduction of pedestrian delay which has a positive impact on reducing the frequency of pedestrian violations (Rosenbloom 2009; Tiwari et al., 2007; Guo et al., 2011). Also, other studies demonstrated that

extending the signal clearing time (pedestrian walk phase) reduces the frequency of temporal violations significantly (Brosseau et al., 2013, Ren et al., 2011).

Moreover, the impact of the countdown signal at signalized intersections on pedestrian behaviour was controversial in the literature. Several studies showed the safety benefits of the countdown signals and pedestrian behaviour improvements (Arhin and Noel, 2007). According to the study, when pedestrians know the remaining time in the signal phase, they can adjust their walking speed so that they can complete the crossing during the pedestrian clearance time without creating any serious conflicts with vehicles moving in the next phase. However, another study (Ni et al., 2017) showed that the presence of countdown signals promotes pedestrian risky behaviour. The study demonstrated that showing the remaining time to the "Do Not Walk" phase leads to more pedestrians being impatient so that they prefer to cross illegally than waiting for more than a whole cycle to cross legally.

Furthermore, previous studies agree that the frequency of pedestrian violation events drops significantly at locations with a higher number of lanes (Zhang et al., 2018; Ghomi and Hussein, 2021; Zhang et al., 2019). Different studies provided different explanations for these findings. Petritsch et al. (2005) reported that as the number of lanes increases, pedestrian perception regarding the intersection changes to be riskier and unsafe to cross. Thus, pedestrians prefer to comply with the pedestrian signal and only cross the intersection during designated phases. Furthermore, locations with more traffic lanes usually existed at major roads that carry higher traffic volume, which reduces the frequency of pedestrian violations. Nevertheless, Ren et al., (2011) did not find a significant correlation between the number of lanes and the frequency of violations.

Several studies also showed that longer crosswalks experience a higher frequency of spatial and temporal violations (e.g., Cambon de Lavalette, et al., 2009). Cao et al. (2017) analyzed video data collected at seven signalized crosswalks in the city of Changchun, China to investigate the key contributing factors to the pedestrian spatial violation. The results of the study showed that the length of the crosswalk is strongly correlated with the frequency of pedestrians' spatial violation. However, few studies did not observe a significant relationship between crosswalk length and pedestrian violation behaviour. Ren et al., (2011) analyzed pedestrian behaviours at signalized intersections with various crosswalk lengths ranging from 8 and 23 meters and did not find any significant correlation between crosswalk length and temporal violation behaviour.

2.5.4.3 Pedestrian-related Factors

Among the different groups of factors that impact the violation behaviour of pedestrians, pedestrian-related factors are by far the most important factors that impact this behaviour. Pedestrians respond differently to traffic and signal parameters depending on their characteristics, such as age, gender, and walking speed. Previous studies investigated the influence of seven different pedestrian-related factors on pedestrian violation behaviour; namely, pedestrian desired speed, age, gender, group size, waiting time to cross, the purpose of the trip, and distraction. It should be noted that while some studies (e.g., Cœugnet et al., 2019) showed that the culture of a society and the social norms may impact some pedestrian behaviours, such as violations, the current study did not attempt to analyze the impact of such factors on the violation behaviour of pedestrians.

Most previous studies showed that pedestrians who are involved in violation events are significantly faster than higher those who cross legally. This suggests that pedestrians who are

younger and physically stronger who have higher walking speeds are more likely to engage in illegal crossings (Hediyeh et al., 2014; Hussein and Sayed, 2015a, Zhu et al., 2022). The results of the study by Goh et al., (2012) showed that the average speed of spatial violators at unsignalized crosswalks was 1.1 times higher than the speed of non-violators. Also, Guo et al., (2016) found that the speed of pedestrians who started to cross the intersection at the end of the "Walk" phase was much higher than the speed of those who started crossing at the beginning of the phase.

The vast majority of studies also showed that younger pedestrians are more likely to be involved in violation events (Hamed, 2001; Brosseau et al., 2013; Ren et al., 2011; Yagil, 2000). Likewise, Dommes et al., (2017) and Sucha et al., (2017) showed that older pedestrians are more likely to accept longer waiting times and only cross the intersection during designated phases. However, few studies found age to be an insignificant factor that does not impact the violation behaviour at all (Ni et al., 2017; Ren et al., 2011).

Additionally, most of the findings of previous studies were consistent regarding the higher probability of men being involved in violation events (Brosseau et al., 2013; Guo et al., 2011; Dommes et al., 2017; Díaz, 2002). Hamed (2001) showed that men's involvement in spatial violation events is 2.6 times higher than women's engagement in such risky behaviour at undivided mid-block locations. The study also showed that men are 1.4 and 3.1 times more likely to be involved in temporal violation than women when they start crossing at the curbside and the central refuge island, respectively. However, other studies conclude opposite results or demonstrated an insignificant relationship between gender difference and risky crossing

behaviours (Tom and Granié, 2011; Holland and Hill, 2010; Ren et al., 2011; Elliott, 2004; Ni et al., 2017).

Moreover, it is well established in the literature that pedestrians who walk in groups are more likely to be involved in risky violation scenarios compared to individual pedestrians (Brosseau et al., 2013; Hamed, 2001; Ren et al., 2011, Zhu et al., 2021b). Guo et al., (2011) confirmed this finding as they observed that the waiting time of pedestrians walking alone at signalized intersections is 3.6 times higher than the waiting time of pedestrians moving in groups. The study explained this as pedestrians walking in groups follow the first impatient group member that decides to cross illegally instead of waiting for the dedicated signal phase. Zhang and Fricker (2021a) showed that the probability of spatial violation is higher if other pedestrians already started to cross the street illegally.

Regarding the waiting time, previous studies agree that both actual waiting time (i.e., the time between the arrival of a pedestrian at the crosswalk and the time the pedestrian starts to cross the crosswalk) and maximum waiting time (i.e., the time between the arrival of a pedestrian at the crosswalk and the end time of the red signal) at the signalized intersection are by far the most significant factors that impact pedestrian's temporal violation decisions. The majority of previous studies demonstrated that longer actual waiting time at the curbside led to a higher probability of pedestrian violations. Several studies have also shown that pedestrians who waited longer to cross tend to accept shorter traffic gaps between oncoming vehicles, which increases the risk of collision. For example, Koh and Wong (2014) analyzed pedestrian crossing behaviour at seven signalized intersections in Singapore using Logistic regression. The study reported that the average accepted gap was much shorter for violators compared to non-violators. In another

study, Russo et al., (2018) developed an ordinal regression model to analyze the behaviour of approximately 3000 pedestrians at four signalized intersections in New York and Arizona. The results of the study showed that the longer the actual waiting time, the higher the frequency of temporal violations is. Tiwari et al., (2007) evaluated pedestrian crossing behaviours at seven signalized intersections in Delhi, India. The study revealed that as the maximum waiting time of pedestrians increases, due to prolonged red phases, the probability of pedestrian temporal violation increases as well. In another study, Brosseau et al., (2013) applied a logistic regression model to investigate the impact of waiting time on pedestrian temporal violation behaviour at thirteen signalized intersections in Montreal, Canada. The results showed a strong direct relationship between the pedestrian temporal violation and the maximum waiting time. Guo et al., (2011) yielded the same conclusion in their study that investigated the temporal violation behaviour at seven crosswalks in China.

Previous studies also showed that trip purpose has a significant impact on pedestrians' violation decisions. Pedestrians are more likely to jaywalk or cross the road during an undesignated phase when they are in a rush, which means that people heading to work or school are more likely to take higher risks while crossing compared to pedestrians who are walking for leisure (Guo et al., 2011). The study conducted by Hamed (2001) demonstrated that pedestrians travelling for a non-work trip can wait at the curbside of undivided roads 1.8 times longer than those who are travelling for a work trip before they decide to cross the road illegally. On divided roads, pedestrians travelling for a non-work trip can wait up to 3 times longer than pedestrians heading to work before they engage in a risky crossing.

Moreover, several studies investigated the impact of pedestrian distraction on the probability of violation behaviours (Nasar and Troyer, 2013). Surfing on mobile and texting while walking was identified as the major causes of distraction among violators (Hamann et al., 2017; Byington and Schwebel, 2013; Aghabayk et al., 2021). Deb et al., (2017) indicated walking in groups as another source of distraction that increases the likelihood of violations, as pedestrians engage in conversation with each other and pay less attention to their surroundings.

2.5.4.4 Environmental and External Factors

Weather conditions, illumination, and time of the day were the three environmental factors that were investigated extensively in previous studies.

As for the weather conditions, a positive association between adverse weather conditions and the frequency of pedestrian violations was reported in many studies (Sisiopiku and Akin, 2003; Yang and Li, 2005). Most drivers reduce their speed and pay more attention to the road during adverse weather conditions; however, pedestrians show riskier behaviours during harsh weather conditions. Li and Fernie (2010) showed that the pedestrian compliance rate with the pedestrian signal in clear weather conditions was 2.3 times higher than pedestrian compliance during harsh weather. Waiting for a long time to cross an intersection is very challenging during extremely cold weather, snowstorms, or thunderstorms. Accordingly, many pedestrians may choose to cross an intersection at an undesignated phase or space just to reduce their waiting time or to reach their destination faster. This behaviour is very risky, especially with the poor visibility and lack of friction available to drivers, which increase the risk of a collision significantly.

Previous studies also reported a direct relationship between the frequency of pedestrian violations and road lighting. Zhang, et al., (2016) demonstrated that a lower frequency of pedestrian temporal violations was observed at locations with no illumination.

The impact of the time of the day on pedestrian violation behaviour was also tested in the literature. Wang et al., (2011) collected video data at five intersections in Beijing, China to study pedestrian crossing behaviour. The results showed that the rate of violation among pedestrians during peak hours is extremely higher than off-peak hours. Zhang et al., (2016) indicated that the probability of pedestrian temporal violation was lower during the day compared to night when the road illumination is not adequate.

2.5.4.5 Built-Environment Factors

The literature addressed the impact of four factors related to the built-environment characteristics on pedestrian violations, namely land use, the presence of bus stops, the presence of schools, and block size.

Residential zones were identified as one of the common land use types that promote pedestrian violations (Schneider et al., 2009). Pulugurtha and Repaka (2008) considered both population density and residential land use as the contributing factors to pedestrian violations. Other studies identified commercial land use and open spaces as land use types that attract more pedestrian risky activities (Miranda-Moreno et al., 2011).

Transit-related factors, such as the presence of bus stops and the frequency of buses, were also investigated in several published studies. Previous research showed that pedestrians would accept riskier crossings in order to avoid missing a bus that is about to depart its stop (Chu et al., 2004; Balk, et al., 2014). These violations are usually accompanied by a high crossing speed and

lack of attention to traffic (Miranda-Moreno et al., 2011). Zaki et al. (2013) indicated that 67% of the spatial violations that occurred at a busy intersection in Vancouver, Canada was attributed to pedestrians trying to catch buses at one bus stop, located at the southwest corner of the intersection. However, Mukherjee and Mitra (2019) did not find any impact of the presence of bus stops at an intersection on pedestrian spatial violation behaviour. Other studies found that there is a positive relationship between the higher number of bus stations within a predefined buffer and the frequency of pedestrian violations. The significant buffer value was 100 feet in (Pulugurtha and Repaka, 2008), and 50 meters in (Ghomi and Hussein, 2021). Ghomi and Hussein (2021) demonstrated the positive impact of the frequency of buses on both the frequency of violations and the severity of collisions that happened due to violations. As the frequency of buses increases, pedestrians do not feel the pressure to catch a stopping bus as the waiting time for the next bus would be shorter.

Moreover, the literature agreed that intersections located near schools usually experience a high frequency of pedestrian violations (Miranda-Moreno et al., 2011; Mukherjee and Mitra, 2020). The students at elementary schools were identified as the predominant temporal violators, especially in the morning as they are rushing to go to school on time (Mukherjee and Mitra, 2020). Ghomi and Hussein (2021) investigated the impact of the school size on the frequency of violations and found that the likelihood of violations increased with the presence of large schools in the intersection area.

Previous studies also showed that longer block size in residential areas was one of the factors that promote jaywalking. Chu et al., (2004) studied jaywalking behavior at 48 blocks in Florida. The study relied on surveying pedestrians and conducting observational studies to assess

pedestrian jaywalking behavior. The study concluded that the larger block size increases the probability of pedestrian spatial violations, particularly, when major bus stops are present. Oakes et al. (2007) investigated pedestrian behaviour in Minneapolis, Minnesota, and found that longer block sizes increased the probability of jaywalking in residential areas by 40%.

2.5.4.6 Socio-Economic Factors

It is commonly acknowledged that the behaviour of pedestrians in an area is strongly impacted by a variety of socioeconomic factors. Hamed (2001) found that people who own a private vehicle have a lower likelihood to be involved in risky crossing maneuvers while walking. The results of the study showed that the likelihood of violations among pedestrians without access to private cars is 2.4 times higher than it is among pedestrians who own at least one private car. The study also noted that individuals who live near major divided streets are more likely to have riskier behaviour compared to people who live on local streets. Zaki et al., (2013) analyzed pedestrian behaviour at a major signalized intersection in Vancouver, Canada, and identified several socioeconomic factors as the main contributors to the high-risk behaviour of pedestrians at the intersection, including poverty and drug use. McIlroy et al., (2019) developed a questionnaire to investigate the impact of income level on pedestrian unsafe behaviours. The study analyzed the responses of around 3500 pedestrians across six economically distinct countries. According to the results, the average scores of violations and lack of safety awareness were significantly higher in low-income communities. Useche et al., (2020) employed a Walking Behaviour Questionnaire (WBQ) method to demonstrate the impact of walkability on both temporal and spatial violations in Spain. The study utilized a cross-sectional method to provide adequate sample size of pedestrians. According to the results, a strong inverse relationship was

found between the walkability score and the frequency of the violations. Esmaili et al., (2021) investigated the impact of pedestrian demographics (age and gender), as well as income level on collisions that involved pedestrian violation. The principal component analysis was conducted on the responses of 520 questionnaires distributed in the city of Mashhad, Iran. According to the results, young male with low income were the dominant group among violators.

2.5.5 Relationship Between Pedestrian Violation and Their Safety

The negative impact of pedestrians' violations on their safety is well established in the literature. There is a unanimous agreement among researchers that the frequency of pedestrian violations is strongly correlated with the frequency and severity of pedestrian-vehicle collisions. For example, King et al., (2009) analyzed pedestrian crossing behaviours and police-recorded collisions at six signalized intersections and the surrounding mid-blocks in Brisbane, Australia to investigate the relationship between pedestrian unsafe behaviours and their safety level. The results revealed that pedestrian temporal and spatial violations could increase the collision rate by up to 8 times. Puscar et al., (2018) investigated the relationship between the violation behaviour of road users at the right turn channels and the pedestrian-vehicle conflicts. The study showed that pedestrian spatial violations are the main contributor to pedestrian-vehicle conflicts at the studied locations. Loukaitou-Sideris et al., (2007) analyzed the collisions that occurred in urban crosswalks in Los Angeles and found that pedestrian jaywalking is one of the main contributors to pedestrian-vehicle collisions at the analyzed locations.

As for the collision severity, Mukherjee and Mitra (2019) analyzed the relationship between pedestrian temporal violations and the severity of collisions at 24 signalized intersections in Kolkata, India. The study concluded that there is a direct relationship between the frequency of

pedestrian temporal violations and the frequency of fatal collisions. In a consequent larger study, Mukherjee and Mitra (2020) implemented a negative binomial model to analyze collision records at 55 intersections in Kolkata, India. Similarly, the results showed a significant direct relationship between the frequency of temporal violations and the frequency of fatal collisions. The study proposed that the rate of temporal violations at intersections can be used as a surrogate index to identify hazardous intersections. Wang et al., (2019) investigated the impact of "crossing on the red" on the severity of collisions in Hong Kong. The results showed that two age groups (pedestrians aged 11 years old and younger and those who are over 66 years old) are more likely to be involved in severe injuries as a result of temporal violations. Sasidharan et al., (2015) investigated the impact of pedestrian temporal violations on the severity level of collisions via analyzing police records (2009-2012) in Switzerland. The study employed a Latent Class Clustering method aims at providing homogenous subsets of collision dataset and developed a binary logit model in each cluster, separately. According to the results, pedestrian unsafe behaviours demonstrate a strong direct impact on the collisions resulted in fatalities and severe injuries. Kim et al., (2017) utilized the hierarchical order technique to analyze more than 137 thousand pedestrian collisions in South Korea between 2011 and 2013. The results showed that spatial violation of pedestrians at mid-blocks locations and the temporal violation of drivers (red light running) were the main contributing factors to severe injury pedestrian-vehicle collisions. Pour-Rouholamin and Zhou (2016) reported a reduction of fatal collisions by 12% for pedestrians who cross the roadway at the dedicated crosswalks compared to jaywalkers who cross the roadway at undesignated areas. Kim et al., (2008) concluded that the likelihood of fatality would decrease by 16% for pedestrians crossing the roadway at the dedicated crosswalks.

In summary, the violation behaviour of pedestrians, either temporally or spatially, was identified by many studies as a risky behaviour that is strongly contributing to the frequency and severity of pedestrian collisions. According to the extensive studies conducted on pedestrian- safety, the main contributing factors impacting both the frequency and severity of pedestrian collisions were extracted. Table 2-4 summarizes the factors that are identified as remarkable contributors to pedestrian safety in the literature, as well as the expected impact of these factors on both the frequency of violations and pedestrian safety.

Table 2-4 Summary of the contributing factors to pedestrian violations

Category	Factor	Impact on Violation	Impact on Safety
Traffic-related factors	Higher traffic volume	+	-
	Higher vehicle speed	+	*
	Higher percentage of heavy vehicles	+	-
Location-specific factors	Higher number of lanes	+	-
	Longer and wider crosswalk	+	+
	Presence of central refuge islands	-	+
	Presence of countdown signals	-	*
	Presence of traffic signals	-	-
Pedestrian-related factors	Being young	-	+
	Being male	*	+
	Higher walking speed		+
	Larger group size		+
	Work/School trip purpose		+
	Longer waiting time before crossing		+
Environmental factors	Adverse weather conditions	*	+
	Lack of illumination	+	*
Built-environment factors	land use that attracts pedestrian activities	*	+
	Presence of schools and bus stops	+	+
	Larger block size		+

Positive (+) sign indicates that the factor is positively associated with the frequency of violations and collisions.
Negative (-) indicates that the factor is positively associated with the frequency of violations and collisions.

Star (*) sign shows that the literature was not conclusive regarding the impact of the factor on the frequency of pedestrian violation and/or collisions.

Based on the table, there is a consistency between the impact of several factors on both the frequency of violations and pedestrian safety, including presence of longer and wider crosswalk, presence of traffic signals, and presence of intersection amenities (like schools and bus stops). However, the majority of the contributing factors demonstrated various impacts on pedestrian violations and their safety. Higher traffic volume, presence of heavy vehicles and higher number of lanes increase the frequency and/or the severity of pedestrian-vehicle collisions. while such situations act as a warning to increase pedestrian awareness and discourage them from violating. According to the literature, there is a direct positive relationship between several factors (e.g., higher vehicle speed, lack of countdown signals, and lack of illumination) and pedestrian safety; however, the previous studies were not conclusive regarding the impact of these factors on the frequency of violations. The inverse result was found for the other three factors (being male, adverse weather condition, and crowded land uses).

2.6. Results of Meta-Analysis

As discussed earlier, the study utilized a meta-analysis framework to develop a quantitative assessment of the factors that impact pedestrian violations, based on the literature findings. The impact of the different factors on pedestrian violations is expressed in terms of the odds ratio (OR) and the corresponding confidence intervals. The reported odds ratio represents the average impact of each factor on the pedestrian violation, as expressed in Equation (2-7):

$$\text{Impact} = OR - 1 \quad (2\text{-}7)$$

An odds ratio greater than (1) represents a positive association between this factor and the frequency of pedestrian violations, while an odds ratio that is less than 1 indicates a negative association. The larger the odds ratio, the greater the impact of the factor on the pedestrian violations.

Additionally, the reported confidence intervals for each factor are vital in understanding whether there is an agreement among previous studies regarding the impact of this factor on pedestrian violations or not. Generally, a confidence interval that does not include (1) indicates an agreement between studies regarding the impact of a factor. A confidence interval that includes (1) indicates that the previous research was inconclusive regarding the impact of a factor, with some studies reporting a positive association between this factor and the frequency of violation, and other studies reporting an opposite trend. The results of the meta-analysis are shown in Table 2-5 below.

Table 2-5 Results of a meta-analysis

Category	Factor	Number of studies	Odd ratio	Confidence interval		P-value
				Lower	Upper	
Traffic-related factors	Lower traffic volume	12	1.22	1.14	1.3	0.00
	Higher vehicle speed (vph)	9	0.98	0.92	1.10	0.00
	Presence of heavy vehicles	7	0.95	0.89	0.99	0.03
	Presence of on-street parking	4	1.18	1.01	1.29	0.02
Location-specific factors	Lower number of lanes	16	1.34	1.21	1.49	0.00
	Presence of central refuge islands	13	1.04	0.97	1.12	0.02
	Presence of traffic signals	19	1.03	0.95	1.12	0.00
	Presence of countdown signals	6	0.98	0.93	1.18	0.04
	Longer and wider crosswalk	10	1.02	1.01	1.08	0.03
Pedestrian-	Age (Being young)	21	1.16	0.96	1.31	0.00

related factors	Gender (Being male)	17	1.01	0.93	1.09	0.03
	Higher walking speed (m/sec)	11	1.08	1.01	1.18	0.00
	Larger group size	8	1.01	0.95	1.08	0.00
	Work/School trip purpose	9	1.06	0.97	1.32	0.04
	Longer waiting time before crossing	16	1.23	1.04	1.32	0.03
	Distracted pedestrian	9	1.12	1.07	1.21	0.00
Environmental factors	Adverse weather conditions	11	1.03	0.98	1.07	0.01
	walking during peak hour (evening)	5	1.07	0.98	1.18	0.00
Built-environment factors	land use types that attract pedestrian activities (residential and commercial)	9	1.33	1.24	1.43	0.00
	Presence of schools and bus stops	5	1.16	1.02	1.27	0.02

As shown in the table, twenty factors were considered in the meta-analysis. All these factors were quantitatively assessed in more than three published studies. Based on the results reported in Table 2-4, the waiting time at the curbsides, traffic volume, walking speed, pedestrian distraction, number of lanes, land use types that attract pedestrian activities, the presence of bus stops and schools, and the presence of on-street parking are the key factors that increase the likelihood of pedestrian violations.

According to the meta-analysis, an agreement among previous studies that as pedestrians wait for a long time to cross at signalized intersections, they are more likely they will cross the intersection illegally. Waiting time is by far the most influential factor on pedestrian temporal violation behaviour. Pedestrian walking speed is another significant factor that was shown to have a positive association with the frequency of violations. Pedestrians with higher walking speeds usually trust their abilities to finish their crossing safely before approaching vehicles to arrive at the crosswalk, which encourages them to violate.

Distracted pedestrians are also more likely to violate (OR = 1.12, with a confidence interval between 1.07 and 1.21), as they may not be aware of the surrounding risks in many situations.

Also, the presence of on-street parking was shown to be a significant factor that increases both the frequency of violations and the severity of the potential consequences. On-street parking limits the sight distance available to pedestrians, so they may choose to jaywalk or cross the road during undesignated times, thinking that there are no approaching vehicles. Meanwhile, parked vehicles also limit the drivers' sight distance so that they may not have enough time to react to a violating pedestrian, which increases the risk of collision.

Moreover, there is an agreement among previous studies that land use types that attract more pedestrian activities, including the residential and commercial land use types, experience a higher frequency of pedestrian violations. The presence of bus stops and schools in the vicinity of a location is directly associated with a higher frequency of pedestrian violations. Also, the frequency of unsafe behaviours is more common among younger pedestrians and those who walk in groups.

The meta-analysis also indicates that larger intersections and road segments with a higher number of lanes experience fewer pedestrian violations. Crowded and larger locations discourage pedestrians from violations as they perceive such behaviour as dangerous behaviour in such an environment. The meta-analysis also shows that that the frequency of pedestrian violations decreases significantly as roads become more congested (i.e., traffic volume increases), which can be explained by the unavailability of adequate gaps between vehicles in congested conditions, which discourage pedestrians from crossing illegally.

Table 2-4 also shows that increasing the percentage of heavy vehicles decreases the frequency of pedestrian violations significantly (OR = 0.95, with a confidence interval between 0.89 and

0.99). This can be explained by the severe consequences that pedestrians foresee in the event of a collision with a large vehicle, which discourages them from violations.

Nevertheless, the meta-analysis showed that previous studies were inconclusive regarding the impact of the average vehicle speed, type of traffic control devices, the presence of refuge islands, pedestrian attributes (age and gender), and the time of the day on pedestrian violation behaviour. According to Table 2-4, pedestrian gender, group size, average vehicle speed, longer crosswalk, adverse weather condition, and the type of traffic control device have on average almost no impact on the violation behaviour, with some studies reporting a positive association with the frequency of violations and others reporting a negative association. Younger pedestrians and those who are walking to get to work/school are generally more likely to violate, although few studies showed otherwise. As well, the meta-analysis results showed that the frequency of violations increases in the evening (OR = 1.07), although some limited research reported otherwise, as can be understood from the reported confidence interval (0.98-1.18).

2.7. Mitigation Strategies

In order to mitigate pedestrian violations (or reduce their frequency), different mitigation strategies have been proposed and tested in the literature. Mitigation strategies focus on reducing the frequency of violations and prevent their serious safety consequences, either in the short term or in the long term. Short-term strategies usually involve the use of a variety of engineering countermeasures that aim at reducing the frequency of violations and/or enhancing enforcement. Some short-term strategies also focus on the early detection of violators and warning both violators and approaching vehicles regarding the potential risks in order to enhance the overall

safety level. Long-term solutions usually involve a combination of educational problems and public campaigns that aim at changing pedestrian behaviour in the long term. A brief discussion of the mitigation strategies is provided as follows:

2.7.1 Engineering-based Mitigation Strategies

Many previous studies investigated the efficiency of a variety of countermeasures in reducing the frequency of pedestrian violations. For example, Sisiopiku and Akin (2003) investigated the impact of two types of physical barriers (i.e., vegetation and concrete wall) and warning signs on pedestrian behaviours in Michigan. Based on the results, barriers and signs reduced 65% and 34% of the frequency of the spatial violations, respectively. Vasudevan et al., (2011) compared the behaviour of pedestrians before and after the installation of pedestrian call buttons at a mid-lock location and two intersections in Nevada. The study concluded that this countermeasure helped to decrease the frequency of pedestrian risky crossings significantly. Arhin et al., (2021) evaluated the impact of a newly designed right-of-way sign on the frequency of pedestrian-vehicle conflicts at 32 uncontrolled crosswalks. The results of the study demonstrated a 73% reduction in pedestrian jaywalking. Zhang and Fricker (2021a) investigated the right-of-way and the frequency of conflicts between pedestrians and vehicles in a semi-controlled crosswalk located near to Purdue University campus, Indianapolis. According to the results, the likelihood of conflict is independent of road users' speed. However, the distance between approaching vehicle to the crosswalk could increase the probability of conflict. In addition to the countermeasures that were assessed in previous studies, the results of the meta-analysis conducted in this study suggest that developing proper signal timing that minimizes pedestrian waiting time at signalized intersections, eliminating on-street parking at locations that experience

a high frequency of violations, and working on short- and long-term solutions to reduce pedestrian distraction may be very beneficial in reducing both the frequency of violations and the risk of related collisions.

2.7.2 Enforcement

Enforcement is another approach that is being adopted by many jurisdictions to eliminate pedestrian risky crossing behaviours, including temporal and spatial violations. Some studies investigated the efficiency of this strategy in reducing the frequency of pedestrian violations. For example, Savolainen et al., (2011) evaluated the impact of enforcement programs that targeted pedestrian violations in Detroit, United States. The results of the study showed a 17.1% and 7.8% reduction in the rate of pedestrian violations during and after the enforcement campaigns, respectively. Li et al., (2021) evaluated the impact of law enforcement cameras on pedestrian behaviour, including crossing speed, waiting time, and gap acceptance at an uncontrolled crosswalk in Nanjing, China. The results showed that the installation of a camera increased the probability of conflicts between pedestrians and vehicles; however, the severity of collisions due to pedestrian spatial violations dropped significantly. Muley et al., (2021) assessed the impact of a modified enforcement program of the Ministry of Interior on pedestrian unsafe behaviours in Qatar. The results showed that the new program increased the level of awareness, safety perception, and adaptation among pedestrians. However, the literature lacks information regarding the long-term effect of enforcement and the optimal allocation of resources to achieve an overall acceptable level of reduction in the frequency of violations.

2.7.3 Educational Programs and Public Campaigns

As a long-term strategy, several jurisdictions considered adopting educational programs and public campaigns that aim at increasing public awareness of the serious consequences of reckless crossing practices. The impact of such programs has been evaluated in many studies in literature. For example, Shiwakoti et al., (2020) evaluated the impact of a campaign in Melbourne, Australia, in which, the city installed posters that portrays the consequences of pedestrian jaywalking at two signalized intersections. The study found a statistically significant reduction in pedestrian spatial violation rates of 9% following the posters installation. Twisk et al., (2014) developed a before-and-after study in order to evaluate the impact of five short-term educational programs on pedestrian behaviours in Michigan. Based on the results, red-light violation had the highest rate of reduction. In another study, Zhang et al., (2013) developed a pilot educational program on the campus of the University of South Florida and found that students' perception regarding the right-of-way was improved significantly.

2.7.4 Technology-based Strategies

With the recent advances in connectivity and connected vehicle applications, several technology-based strategies are being promoted as potential solutions to mitigate pedestrian violations and reduce the severity of the consequences of such behaviour. These technology-based solutions depend mainly on the automated detection of violators and warning drivers and/or the violating pedestrians of the potential hazard. For example, Wu et al., (2014) designed a system, known as "802.11p", based on Dedicated Short-Range Communications (DSRC) that issues several warnings to the driver regarding the presence of pedestrians in a potential hazard, including violating and distracted pedestrians. Harding et al., (2014) developed a smartphone application

that detects the jaywalking pedestrians, based on their crossing location, and communicates with approaching vehicles, via DSRC, to warn drivers regarding the potential hazard. Anaya et al., (2014) introduced a system known as "V2ProVu" to warn violating pedestrians with potential risks. The system sends audible messages to the pedestrians' smartphones to warn pedestrians regarding illegal crossings and probable serious interactions with approaching vehicles. In another study, Rahman et al., (2019) developed a DSRC-based device equipped with a camera that is installed on the windshield of the vehicle. The system uses time-to-collision measurement as an indicator to alert the driver if a violator pedestrian is detected. Khosravi et al., (2018) developed a system called (Smart Walk Assistant) that could identify pedestrians who enter the crosswalk in the red interval (temporal violators). The system could alert the drivers regarding a dangerous situation using the intersection connectivity features (Roadside Unit and Wi-Fi).

In summary, a few studies provided quantitative assessments of the efficiency of the proposed mitigation strategies. Therefore, further studies are needed to evaluate the different strategies based on their efficiency and identify the appropriate conditions for adopting those proposed solutions.

2.8. Conclusions and Future Directions

This study provides a holistic review of pedestrian violation behaviour in order to develop a solid understanding of the factors that contribute to violations, locations that experience a high frequency of violations, the data collection and the research methods used to study such behaviour, the relationship between violations safety, and the different strategies that can be adopted to mitigate this behaviour. The study utilized a Text Mining method to identify all

related studies in the past 21 years. The study also conducted a meta-analysis to assess the impact of the different contributing factors on the frequency of pedestrian violations. The study found that pedestrian violation is one of the hazardous behaviours that contribute to both the frequency and severity of pedestrian-vehicle collisions. Previous research investigated the effect of a wide range of factors on pedestrian violations. According to the literature, there is a consensus regarding the positive association between the frequency of pedestrian violations and many factors, including longer waiting time before crossing, the presence of schools and bus stops on-street parking, long crosswalks, long blocks, and pedestrian distraction. Also, crowded locations that have high traffic volume, a high percentage of heavy vehicles, and a higher number of lanes usually experience a lower frequency of pedestrian violations. On the other hand, the literature did not provide conclusive evidence regarding the impact of many factors on pedestrian violations, including vehicle speed, the presence of refuge islands, the presence of traffic signals, countdown signals, pedestrian attributes (gender, age, and group size), trip purpose, weather conditions, and time of the day. Previous studies have also assessed a wide range of strategies that can mitigate violations and reduce the safety consequences of such behaviour. The mitigation strategies ranged from simple engineering-based countermeasures, such as physical barriers, pedestrian call buttons, and warning signs to the use of advanced technologies in mitigating violations and the automated detection of violators.

Based on the findings of the study, several future research directions can be proposed, summarized as follows:

- As for the data collection methods, there is a need to rely on new sources of data other than historical collision records to enhance/complement our understanding of the

violation behaviour. It is commonly acknowledged that collision data suffers from many shortcomings. Specifically, with pedestrian violations, collision data do not provide a complete picture of pedestrians' actions at the time of the collision, which impacts the accuracy of the results. Analyzing video data that captures the natural walking behaviour of pedestrians is a promising approach to investigate pedestrian violations. Video data enables to conduct of micro-level analyses of the violation behaviour and develop models that evaluate the frequency of violations as a function of a variety of factors. Video data also enable the establishment of the relationship between violation and safety, expressed using traffic conflicts, which is a reliable surrogate measure of safety that mitigates many issues associated with collision data. Other data collection methods, such as the cell phone or probe data, can also provide useful information regarding pedestrian walking behaviour in general and their violation behaviour in particular. Such data collection methods were promoted as promising approaches to investigate road users' safety in the literature (e.g., Rose 2006, Kummala 2002, Fukushima 2009). Adopting these data collection methods could provide more detailed and accurate data related to pedestrian violation behaviours. For example, cell phones continuously capture the real-time pedestrian location, so factors like the exact waiting time at the curbside before a pedestrian illegally crosses a signalized intersection can be accurately measured and modeled based on pedestrian traits, location characteristics, and weather conditions. This could assist planners in enhancing the design of pedestrian signals to reduce the frequency of pedestrian temporal violations at signalized intersections. Spatial violations can be tracked and analyzed for individual pedestrians over a large space and time frame.

This would provide an opportunity to analyze such behaviour and assess the impact of individual traits on violation decisions. Also, pedestrian speed profiles and trajectories can be analyzed during the identified violation events to check if pedestrians undertook sudden evasive actions at the time of the violation (e.g., pedestrians suddenly stop, speed up, change direction, etc.), which can be used to analyze the safety implications of violation events.

Aside from cellphone data, Naturalistic driving dataset (NDD) could be considered as another data collection method. Although some previous studies used the Naturalistic driving datasets (NDD) to analyze pedestrian behaviour safety (e.g., Sheykhfard et al., 2021a, Rasch et al., 2020), this data source has not been widely used to investigate pedestrian violation behaviour One of the studies that explored the use of NDD to investigate pedestrian behaviour is Wang et al., (2018), which used the NDD to investigate children's violation behaviour near two primary schools in Nantong, China. The study showed that younger children are less likely to violate traffic rules compared to older children. Although the NDD dataset is a promising data source that can facilitate conducting in-depth analyses of pedestrian violations, it suffers from several limitations that need to be considered by future studies that may rely on it, including, for example, authenticity and validity of the results, high initial cost, and simulation errors.

- Regarding the research methods adopted to study pedestrian violations, different emerging approaches show promising results in developing a better understanding of the violation behaviour, including Machine Learning techniques. Machine Learning algorithms have the power to capture the underlying relationships between the different

explanatory variables and pedestrian violations. The major advantage of Machine Learning techniques in comparison with traditional statistical models is that their results are independent of any prior assumption related to the variables (i.e., the linearity, the standardization, the normality of the predictors, and the presence of missing values), which is a common issue in dealing with collision datasets. Moreover, Machine learning models are well-suited to handle independent variables that are impacted by a large number of factors, which is the case for pedestrian violations. It is also acknowledged that Machine learning models are flexible, which increases their ability to handle overfitting and tendency for reduced bias. Since Machine Learning models do not require predefined model forms, their ability to capture the non-linear behaviours of independent variables is higher than traditional statistical models. Such flexibility facilitates the transferability of Machine Learning models, as they are capable of being generalized across models trained with different techniques or ensembles taking collective decisions. In addition, some Machine Learning techniques, such as unsupervised clustering techniques and supervised decision tree prediction models, can be useful in studying complex behaviour such as pedestrian violations, as they are capable of investigating the change of the impact of the explanatory variable with the changes in location, traffic, and pedestrian characteristics. Nevertheless, one of the disadvantages of some Machine Learning techniques, such as artificial neural networks, is that they act as a black box, which makes it difficult to interpret the model results and study the impact of the influencing factors on pedestrian violations.

- With respect to the mitigation strategies, the literature lacks quantitative assessments of the efficiency of many of the mitigation strategies addressed in this study. Future studies should assess the short-term and the long-term impact of the different mitigation strategies in terms of how successful they are in reducing the frequency of violations and mitigating severe collisions that occur due to violations. Besides, more research is needed to investigate the optimal use of advanced in-vehicle technologies and infrastructure connectivity features in mitigating violations and the early detection of violating pedestrians. Future research also should investigate the best medium to communicate with both drivers and pedestrians to warn them regarding potential hazards without creating new sources of hazard.
- Future studies that focus on assessing the mitigation strategies will need to conduct a quantitative assessment of the efficiency of different strategies. Microsimulation models that are developed based on a solid understanding of the violation behavior can be a perfect tool to conduct such quantitative assessments, compare the efficiency of different strategies at specific locations, and select the optimal strategies that work best under prevailing traffic and geometric conditions. However, in order to ensure the reliability of the results, it is crucial to rely on microsimulation models that are developed using a proper modeling approach, mimics actual pedestrian behaviour, and are calibrated with a proper methodology that ensures that the model results are realistic and accurate (Hussein and Sayed, 2015b; Hussein and Sayed, 2017; Hussein and Sayed 2019).
- As highlighted in the study, the literature was inconclusive regarding the impact of many factors on pedestrian violations. There is a need to consider advanced analysis techniques

and using reliable data sources to overcome this issue in future research. Also, macro-level analysis that assesses the impact of socio-economic factors, land use, and household characteristics on pedestrian behaviour and safety can provide crucial insights for planners and engineers to design convenient and safe transportation networks. Future studies also are encouraged to investigate the relationship between pedestrian distraction and violations. Pedestrian distraction has been identified as a potential contributor to pedestrian violations; however, very limited studies were undertaken to quantify the impact of distraction on pedestrian violations and the associated collisions.

- Regarding the relationship between violations and safety, it is essential to develop micro-level and macro-level analyses of the pedestrian-vehicle collisions that involve pedestrian violation. Such analyses are important to understand the factors that increase the severity of the safety consequences of pedestrian violations and determine the impact of the location characteristics and the neighborhood/zone traits on the frequency and severity of collisions that involve pedestrian violations.

2.9. Reference

Aghabayk, K., Esmailpour, J., Jafari, A., & Shiwakoti, N. (2021). Observational-based study to explore pedestrian crossing behaviors at signalized and unsignalized crosswalks. *Accident Analysis and Prevention, 151*, 105990.

Aidoo, E. N., Amoh-Gyimah, R., & Ackaah, W. (2013). The effect of road and environmental characteristics on pedestrian hit-and-run accidents in Ghana. *Accident Analysis and Prevention, 53*, 23-27.

Al Kaabi, A., Dissanayake, D., & Bird, R. (2012). Response time of highway traffic accidents in Abu Dhabi: Investigation with hazard-based duration models. *Transportation Research Record, 2278, 1*, 95-103.

Alexander, J., Barham, P., & Black, I. (2002). Factors influencing the probability of an incident at a junction: results from an interactive driving simulator. *Accident Analysis and Prevention , 34*, 779–792.

Al-Mahameed, F., Qin, X., Schneider, R. J., & Shaon, M. R. (2019). Analyzing pedestrian and bicyclist crashes at the corridor level: structural equation modeling approach. *Transportation research record 2673, 7*, 308-318.

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

Anaya, J. J., Merdrignac, P., Shagdar, O., Nashashibi, F., & Naranjo, J. E. (2014). Vehicle to Pedestrian Communications for Protection of Vulnerable Road Users. *IEEE Intelligent Vehicles Symposium* . Michigan, USA.

Anik, M., Hossain, M., & Habib, M. (2021). Investigation of pedestrian jaywalking behaviour at mid-block locations using artificial neural networks. *Safety science, 144*, 105448.

Arhin, S. A., & Neol, E. C. (2007). Impact of Countdown Pedestrian Signals on Pedestrian Behavior and Perception of Intersection Safety in the District of Columbia. *IEEE Intelligent Transportation Systems Conference ITSC.* doi:10.1109/ITSC.2007.4357761

Arhin, S. A., Gatiba, A., Anderson, M., Manandhar, B., Ribbisso, M., & Acheampong, E. (2021). Effectiveness of modified pedestrian crossing signs in an urban area. *Journal of Traffic and Transportation Engineering (English Edition)* .

Balk, S. A., Bertola, M. A., Shurbutt, J., & Do, A. (2014). *Human Factors Assessment of Pedestrian Roadway Crossing Behavior.* Washington, DC.: Federal Highway Administration.

Bernhoft, I. M., & Carstensen, G. (2008). Preferences and behaviour of pedestrians and cyclists by age and gender. *Transportation Research Part F:Traffic Psychology and Behaviour, 11*(2), 83-95.

Bian, Y., Ma, J., Rong, J., Wang, W., & Lu, J. (2009). Pedestrians' level of service at signalized intersections in China. *Transportation Research Record, 2114*, 83-89.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Brewer, M., Fitzpatrick, K., Whitacre, J., & Lord, D. (2006). Exploration of Pedestrian Gap Acceptance Behaviour at Selected Locations. *Transportation Research Record, 1982*, 132-140.

Brosseau, M., Zangenehpour, S., Saunier, N., & Miranda-Moreno, L. (2013). The impact of waiting time and other factors on dangerous pedestrian crossings and violations at signalized intersections: A case study in Montreal. , pp. *Transportation Research Part F*.

Byington, K., & Schwebel, D. (2013). Effects of mobile internet use on college student pedestrian injury risk. *Accident Analysis and Prevention, 51*, 78–83. doi:doi.org/10.1016/j.aap.2012.11.001

Cambon de Lavalette, B., Tijus, C., Poitrenaud, S., Leproux, C., Bergeron, J., & Thouez, J. P. (2009). Pedestrian crossing decision-making: A situational and behavioral approach. *Safety Science, 47*, 1248–1253.

Cantillo, V., Arellana, J., & Rolong, M. (2015). Modelling pedestrian crossing behaviour in urban roads: a latent variable approach. *Transportation Research Part F: Traffic Psychology and Behaviour, 32*, 56-67.

Cao, N. B., Qu, Z. W., Song, X. M., Zhao, L. Y., Bai , Q. W., & Luo, R. Q. (2017). Modeling the Variation in the Trajectory of Crosswalk Overflow Violation Pedestrians in China and Countermeasure. *Mathematical Problems in Engineering*. doi:doi.org/10.1155/2017/3483809

Cao, Y., Ni, Y., & Li, K. (2016). Effects of Refuge Island Settings on Pedestrian Safety Perception and Signal Violation at Signalized Intersections. *96th Annual meeting of Transportation Research Board.*

Casali, M., Malchiodi, D., Spada, C., Zanaboni, A. M., Cotroneo, R., Furci, D., . . . Blandino, A. (2021). A pilot study for investigating the feasibility of supervised machine learning approaches for the

classification of pedestrians struck by vehicles. *Journal of Forensic and Legal Medicine, 84*, 102256.

Chen, J., Shi, J. J., Li, X. L., & Zhao, Q. (2011). Pedestrian Behavior and Traffic Violations at Signalized Intersections. *11th International Conference of Chinese Transportation Professionals (ICCTP)* (pp. 1103-1110). Americal Society of Civil Engineers.

Chen, S., Xing, J., & Cao, Y. (2017). The impact of waiting time on pedestrian violations at signalized intersections. *Civil Engineering and Urban Planning: An International Journal (CiVEJ), 4*(2). doi:10.5121/civej.2017.4201

Chu, X. H., Guttenplan, M., & Maltes, M. R. (2004). Why People Cross Where They Do: The Role of Street Environment . *Transportation Research Record, 1878*, 3-10.

Cœugnet, S., Cahour, B., & Kraiem, S. (2019). Risk-taking, emotions and socio-cognitive dynamics of pedestrian street-crossing decision-making in the city. *Transportation Research Part F, 65*, 141-157.

Das, S., Le, M., & Dai, B. (2020). Application of machine learning tools in classifying pedestrian crash types: A case study. *Transportation Safety and Environment, 2*(2), 106-119.

Das, S., Manski, C. F., & Manuszak, M. (2005). Walk or wait? An empirical analysis of street crossing decisions. *Journal of Applied Econometrics, 20*(4), 445-577.

Deb, S., Strawderman, L., DuBien, J., Smith, B., Carruth, D. W., & Garrison, T. M. (2017). Evaluating pedestrian behavior at crosswalks: Validation of a pedestrian behavior questionnaire for the U.S. population. *Accident Analysis and Prevention, 106*, 191-201.

Demiroz, Y. I., Onelcin, P., & Alver, Y. (2015). Illegal road crossing behavior of pedestrians at overpass locations: factors affecting gap acceptance, crossing times and overpass use. *Accident Analysis and Prevention, 80*, 220-228.

Díaz, E. M. (2002). Theory of planned behavior and pedestrians' intentions to violate traffic regulations. *Transportation Research Part F: Traffic Psychology and Behaviour, 5*(3), 169-175.

Ding, C., Chen, P., & Jiao, J. (2018). Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: A machine learning approach. *Accident Analysis and Prevention, 112*, 116-126.

Dommes , A., Cavallo, V., Vienne, F., & Aillerie, I. (2012). Age-related differences in street-crossing safety before and after training of older pedestrians. *Accident Analysis and Prevention, 42*, 42-47.

Dommes, A., Granié, M. A., Cloutier, M. S., Coquelet, C., & Huguenin-Richard, F. (2015). Red light violations by adult pedestrians and other safety-related behaviors at signalized crosswalks. *Accident Analysis and Prevention, 80*, 67-75.

Elliott, M. A. (2004). *The attitudes and behaviour of adolescent road users: an application of the theory of planned behaviour.* TRL Report 601.

Eluru, N., Bhat, C. R., & Hensher, D. A. (2008). A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis and Prevention, 40*(3), 1033–1054.

Esmaili, A., Aghabayk, K., Parishad, N., & Stephens, A. N. (2021). Investigating the interaction between pedestrian behaviors and crashes through validation of a pedestrian behavior questionnaire (PBQ). *Accident Analysis and Prevention, 153*, 106050.

Forbes, J. H. (2015). Pedestrian injury severity levels in the Halifax Regional Municipality, Nova Scotia, Canada: hierarchical ordered probit modeling approach. . *Transportation Research Record 2519*, 172-178.

Fu, L., & Zou, N. (2016). The influence of pedestrian countdown signals on children's crossing behavior at school intersections. *Accident Analysis and Prevention, 94*, 73-79.

Fukushima, M. (2009). Progress of an ITS Field Operational Test for Traffic Safety and Congestion in Japan. *International Journal of ITS Researc, 7*(2).

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 721-741.

Ghomi, H., & Hussein, M. (2021). An integrated clustering and copula-based model to assess the impact of intersection characteristics on violation-related collisions. *Accident Analysis and Prevention, 159*, 106283.

Goh, B. H., Subramaniam, K., Wai, Y. T., & Mohamed, A. A. (2012). PEDESTRIAN CROSSING SPEED: THE CASE OF MALAYSIA. *International Journal for Traffic and Transport Engineering, 2*(4), 323-332.

Guo, H., Gao, Z., Yang, X., & Jiang, X. (2011). Modeling pedestrian violation behavior at signalized crosswalks in China: A hazards-based duration approach. *Traffic Injury Prevention*, 96–103.

Guo, Y., Liu, P., Liang, Q., & Wang, W. (2016). Effects of parallelogram-shaped pavement markings on vehicle speed and safety of pedestrian crosswalks on urban roads in China. *Accident Analysis and Prevention, 95*, 438–447.

Haleem, K., Alluri, P., & Gan, A. (2015). Haleem, K., Abdel-Aty, M., 2010. Examining traffic crash injury severity at unsignalized intersections. J. Saf. Res. 41, 347–357. *Accident Analysis and Prevention, 81*, 14-23.

Hamann, C., Dulf, D., Baragan-Andrada, E., Price, M., & Peek-Asa, C. (2017). Contributors to pedestrian distraction and risky behaviours during road crossings in Romania. *Injury Prevention, 0*, 1-7.

Hamed, M. M. (2001). Analysis of pedestrians' behavior at pedestrian crossings. *Safety Science, 38*, 63-82.

Haque, M. M., Oviedo-Trespalacios, O., Sharma, A., & Zheng, Z. (2021). Examining the driver-pedestrian interaction at pedestrian crossings in the connected environment: A Hazard-based duration modelling approach. *Transportation Research Part A, 150*, 33-48.

Harding, J., Powell, G., Yoon, R., Fikentscher, J., Doyle, C., Sade, D., . . . Wang, J. (2014). *Vehicle-to-Vehicle Communications: Readiness of V2V Technology for Application* . U.S. Department of Transportation.

Hediyeh, H., Sayed, T., & Zaki, M, M. (2014). Automated analysis of pedestrian crossing speed behavior at scramble-phase signalized intersections using computer vision techniques. *International Journal of Sustainable Transport, 8*(5), 382– 397. doi:doi.org/10.1080/15568318.2012.708098

Holland, C., & Hill, R. (2010). Gender differences in factors predicting unsafe crossing decisions in adult pedestrians across the lifespan: a simulation study. *Accident Analysis and Prevention, 42*(4), 1097-1106.

Hussein, M., & Sayed, T. (2015). A unidirectional agent based pedestrian microscopic model. *Canadian Journal of Civil Engineering, 42*, 1114–1124.

Hussein, M., & Sayed, T. (2015). Microscopic Pedestrian Interaction Behavior Analysis Using Gait Parameters. *Transportation Research Record, 2519*, 28–38.

Hussein, M., & Sayed, T. (2017). A bi-directional agent-based pedestrian microscopic model. *Transportmetrica A: Transport Science, 13*(4), 326-355.

Hussein, M., & Sayed, T. (2019). Validation of an agent-based microscopic pedestrian simulation model in a crowded pedestrian walking environment. *Transportation planning and technology, 42*(1), 1-22.

Hussein, M., Sayed, T., Reyad, P., & Kim, L. (2015). Automated pedestrian safety analysis at a signalized intersection in New York City: Automated data extraction for safety diagnosis and behavioral study. *Transportation Research Record: Journal of the Transportation Research Board, 2519*(1), 17-27.

Ibitoye, B. A., Ibrahim, N. A., Adeyemoh, N. A., Kazeem, M. B., & Daudu, M. (2021). Impact of Road Users Behaviour on Intersection Performanceusing Vissim Micro-Simulation. *Journal of Architecture and Civil Engineering, 6*(7), 39-45.

Ishaque, M. M., & Noland, R. B. (2008). Behavioural issues in pedestrian speed choice and street crossing behaviour: a review. *Transport Reviews , 28*(1), 61–85.

Jahandideh, Z., Mirbaha, B., & Rassafi, A. A. (2017). Modeling the Risk Intensity of Crossing Pedestrians in Intersections Based on Selected Critical Time to Collision. *Transportation Research Board.* Washington D.C.

Ka, D., Lee, D., Kim, S., & Yeo, H. (2019). Study on the Framework of Intersection Pedestrian Collision Warning System Considering Pedestrian Characteristics. *Transportation Research Record, 2673*(5), 747-758.

Kadali, B. R., Rathi, N., & Perumal, V. (2014). Evaluation of pedestrian mid-block road crossing behaviour using artificial neural network. *Journal of traffic and transportation engineering (English edition), 1*(2), 111-119.

Kang, L., Xiong, Y., & Mannering, F. L. (2013). Statistical analysis of pedestrian perceptions of sidewalk level of service in the presence of bicycles. *Transportation Research Part A: Policy and Practice, 55*, 10-21.

Kaplan, S., & Prato, C. G. (2013). Cyclist-motorist crash patterns in Denmark: A latent class clustering approach. *Traffic Injury Prevention, 14*, 725–733.

Karl, K., Brunner, M., & Yamashita, E. (2008). Modeling fault among accident-Involved pedestrians and motorists in Hawaii. *Accident Analysis & Prevention , 40*(6), 2043-2049.

Khosravi, S., Beak, B., Head, K. L., & Saleem, F. (2018). Assistive System to Improve Pedestrians' Safety and Mobility in a Connected Vehicle Technology Environment. *Transportation Research Record: Journal of the Transportation Research Board, 2672*(19), 145-156.

Kim, J. K., Ulfarsson, G. F., Shankar, V. N., & Kim, S. (2008). Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis. *Accident Analysis and Prevention, 40*, 1695-1702.

Kim, M., Kho,, S. Y., & Ki, D. K. (2017). Hierarchical Ordered Model for Injury Severity of Pedestrian Crashes in South Korea. *Journal of Safety Research, 61*, 33–40.

Kim, S., & Ulfarsson, G. F. (2019). Traffic safety in an aging society: Analysis of older pedestrian crashes. *Journal of Transportation Safety & Security, 11*(3), 323-332.

King, M. J., Soole, D., & Ghafourian, A. (2009). Illegal pedestrian crossing at signalised intersections: incidence and relative risk. *Accident Analysis and prevention, 41*(3), 485–490.

King, M. R., Carnegie, J. A., & Ewing, R. (2003). Pedestrian Safety Through a Raised Median and Redesigned Intersections. *Transportation Research Record 1828, 1*, 56-66.

Koh, P. P., & Wong, Y. D. (2014). Gap acceptance of violators at signalised pedestrian crossings. *Accident Analysis and Prevention, 62*, 178–185.

Kummala, J. (2002). *Travel Time Service Utilising Mobile Phones.* (Helsinki: Finish Road Administration): Report No. 55/2002.

Kummeneje, A. M., & Rundmo, T. (2019). Risk perception, worry, and pedestrian behaviour in the Norwegian population. *Accident Analysis and Prevention, 133*, 105294.

Li, H., Zhang, Z., Sze, N. N., Hu, H., & Ding, H. (2021). Safety effects of law enforcement cameras at non-signalized crosswalks: A case study in China. *Accident Analysis and Prevention*, 106124.

Li, J., Yao, L., Xu, X., Cheng, B., & Ren, J. (2020). Deep reinforcement learning for pedestrian collision avoidance and human-machine cooperative driving. *Information Sciences, 532*, 110-124.

Li, Y., & Fernie, G. (2010). Pedestrian behavior and safety on a two-stage crossing with a center refuge island and the effect of winter weather on pedestrian compliance rate. *Accident Analysis and Prevention, 42*, 1156–1163.

Ling, Z., Ni, Y., Cherry, C., & Li, K. (2013). 73. , 2013. Pedestrian Level of Service at Signalized Intersections in China Using Contingent field survey and Pedestrian Crossing Video Simulation. *Transportation Research Record*.

Lipovac, K., Vujanic, M., Marie, B., & Nesic, M. (2013). Pedestrian behavior at signalized pedestrian crossings. *Journal of Transportation Engineering, 139*(2), 165-172.

Liu, Y. C., & Tung, Y. C. (2014). Risk analysis of pedestrians' road-crossing decisions: Effects of age, time gap, time of day, and vehicle speed. *Safety Science, 63* , 77–82.

Lobjois, R., & Cavallo, V. (2007). Age-related differences in street-crossing decisions: The effects of vehicle speed and time constraints on gap selection in an estimation task. *Accident Analysis and Prevention, 39*, 934–943.

Long, X., Zhou, M., Zhao, H., & Song, Y. (2021). Pedestrian crossing decision during flashing green-countdown signal for urban signalized intersection. *Journal of Transportation Safety & Security*, 1-21.

Loukaitou-Sideris, A., Liggett, R., & Sung, H. G. (2007). Death on the Crosswalk A Study of Pedestrian-Automobile Collisions in Los Angeles. *Journal of Planning Education and Research, 26*(3), 338–351.

Lyons, G., Hunt, J., & McLeod, F. (2001). A neural network model for enhanced operation of midblock signalled pedestrian crossings. *European journal of operational research, 129*(2), 346-354.

Ma, J., Song, W., Fang, Z., Lo, S., & Liao, G. (2010). Experimental study on microscopic moving characteristics of pedestrians in built corridor based on digital image processing . *Building and Environment , 45*, 2160-2169.

McIlroy, R. C., Plant, K. L., Jikyong, U., Hoài, N. V., Bunyasi, B., Kokwaro, G. O., . . . Stanton, N. A. (2019). Vulnerable road users in low-, middle-, and high-income countries: Validation of a Pedestrian Behaviour Questionnaire. *Accident Analysis and Prevention, 131*, 80-94.

Miranda-Moreno, L. F., Morency, P., & El-Geneidy, A. M. (2011). The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections. *Accident Analysis and Prevention, 43*, 1624–1634.

Mo, Z. G., & Mo, L. G. (2017). Study of pedestrian crossing behavior at signalized intersection based on structural equation model. *Advanced in Transportation Studies (ATS), 2*, 45-54.

Mohamed, M. G., Saunier, N., Miranda-Moreno, L. F., & Ukkusuri, S. V. (2013). A clustering regression approach: A comprehensive injury severity analysis of pedestrian-vehicle crashes in New York, US and Montreal, Canada. *Safety science, 54*, 27–37.

Moudon, A. V., Lin, L., Jiao, J., Hurvitz, , P., & Reeves, P. (2011). The Risk of Pedestrian Injury and Fatality in Collisions with Motor Vehicles, a Social Ecological Study of State Routes and City Streets in King County, Washington. *Accident Analysis and Prevention, 43*(1), 11-24.

Mukherjee, D., & Mitra, S. (2019). A comparative study of safe and unsafe signalized intersections from the view point of pedestrian behavior and perception. *Accident Analysis and Prevention, 132*.

Mukherjee, D., & Mitra, S. (2020). A comprehensive study on factors influencing pedestrian signal violation behaviour: Experience from Kolkata City, India. *Safety Science, 124*, 104610.

Muley, D., Kharbeche, M., Ghonim, O., Madkoor, A., & Mohamed, Y. (2021). Does Pedestrian Penalty Affect Pedestrian Behavior? A Case of State of Qatar. *The 12th International Conference on Ambient Systems, Networks and Technologies (ANT)* (pp. 234-241). Warsaw, Poland: Procedia Computer Science 184.

Nasar, J., & Troyer, D. (2013). Pedestrian injuries due to mobile phone use in public places. *Accident Analysis and Prevention, 57*, 91–95. doi:doi.org/10.1016/j.aap.2013.03.021

Nassr, M. M., Zulkiple, A., Albargi, W. A., & Khalifa, N. A. (2017). Modeling pedestrian gap crossing index under mixed traffic condition. *Journal of Safety Research, 63*, 91-98.

Ni, Y., Cao, Y., & Li, K. (2017). Pedestrians' Safety Perception at Signalized Intersections in Shanghai. *Transportation Research Procedia, 25*, 1955–1963.

Nie, B., & Zhou, Q. (2016). Can new passenger cars reduce pedestrian lower extremity injury? A review of geometrical changes of front-end design before and after regulatory efforts. *Traffic Injury Prevention, 17*(7), 712-719.

Nilsen, P. (2004). What makes community based injury prevention work? In search of evidence of effectiveness. *Injury Prevention, 10*, 268–274.

Oakes , J. M., Forsyth, A., & Schmitz, K. H. (2007). The effects of neighborhood density and street connectivity on walking behavior: the Twin Cities walking study. *Epidemiologic Perspectives and Innovations, 4*(16). doi:doi.org/10.1186/1742-5573-4-16

Oxley, J. A., Ihsen, E., Fildes, B. N., Charlton, J. L., & Days, R. H. (2005). Crossing roads safely: An experimental study of age differences in gap selection by pedestrians. *Accident Analysis and Prevention, 37*(5).

Pahwa, B., Taruna, S., & Kasliwal, N. (2018). Sentiment Analysis- Strategy for Text Pre-Processing. *International Journal of Computer Applications, 180*(34), 15-18.

Papadimitriou, E. (2012). Theory and Models of Pedestrian Crossing Behaviour along Urban Trips. *Transportation Research Part F: Traffic Psychology and Behaviour, 15*, 75-94.

Papadimitriou, E., Theofilatos, A., & Yannis, G. (2013). Patterns of pedestrian attitudes, perceptions and behaviour in Europe. *Safety science, 53*, 114-122.

Papić, Z., Jović, A., Simeunović, M., Saulić, N., & Lazarević, M. (2020). Underestimation tendencies of vehicle speed by pedestrians when crossing unmarked roadway. *Accident Analysis and Prevention, 143*, 105586.

Pawar, D. S., & Patil, G. R. (2016). Critical gap estimation for pedestrians at uncontrolled mid-block crossings on high-speed arterials. *Safety Science, 86*, 295-303.

Petritsch, T. A., Landis, B. W., McLeod, P. S., Huang, H. F., Challa, S., & Guttenplan, M. (2005). Level-of-Service Model for Pedestrians at Signalized Intersections. *Transportation Research Record, 1939*, 55-62.

Pineda-Jaramillo , J. D. (2020). A Shallow Neural Network approach for identifying the leading causes associated to pedestrian deaths in Medellín. *Journal of Transport & Health, 19*, 100912.

Poudel-Tandukar, K., Nakahara, S., Ichikawa, M., Poudel, K. C., & Jimba, M. (2007). Risk perception, road behavior, and pedestrian injury among adolescent students in Kathmandu, Nepal. *Injury Prevention, 13*(4), 258-263.

Pour-Rouholamin, M., & Zhou, H. (2016). Investigating the risk factors associated with pedestrian injury severity in Illinois. *Journal of Safety Research, 57*, 9–17.

Pulugurtha, S. S., & Repaka, S. R. (2008). Assessment of Models to Measure Pedestrian Activity at Signalized Intersections. *Transportation Research Record, 2073*, 39-48.

Puscar, F., Sayed, T., Bigazzi, A., & Zaki, M. (2018). Multimodal Safety Assessment of an Urban Intersection by Video Analysis of Bicycle, Pedestrian, and Motor Vehicle Traffic Conflicts and Violations. *Transportation Research Board.* Washington, D.C.

Rahman, M., Islam, M., Calhoun, J., & Chowdhury, M. (2019). Real-Time Pedestrian Detection Approach with an Efficient Data Communication Bandwidth Strategy. *Transportation Research Record*.

Rasch, A., Panero, G., Boda, C. N., & Dozza, M. (2020). How do drivers overtake pedestrians? Evidence from field test and naturalistic driving data. *Accident Analysis and Prevention, 139*, 105494.

Rasouli, A., & Kotseruba, I. (2022). Intend-Wait-Cross: Towards Modeling Realistic Pedestrian Crossing Behavior .

Ren, G., Zhou, Z., Wang, W., Zhang, Y., & Wang, W. (2011). Crossing Behaviors of Pedestrians at Signalized Intersections. *Transportation Research Record, 2264*, 65–73.

Rose, G. (2006). Mobile Phones as Traffic Probes: Practices, Prospects and Issues. *Transport Reviews, 26*(3), 275-291.

Rosenbloom, T. (2009). Crossing at a red light: Behaviour of individuals and groups. *Transportation Research Part F*, 389–394.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59–82.

Russo, B. J., James, E., Aguilar, C. Y., & Smaglik, E. J. (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.

Sanchez, P. E. (2009). *Pedestrian and Traffic Safety in Parking Lots at SNL/NM: Audit Background Report (No. SAND2009-1630).* California: Sandia National Laboratories.

Sasidharan, L., Wu, K. F., & Menendez, M. (2015). Exploring the Application of Latent Class Cluster Analysis for Investigating Pedestrian Crash Injury Severities in Switzerland. *Accident Analysis and Prevention, 85*, 219–228.

Savolainen, P. T., Gates, T. G., & Datta, T. K. (2011). Implementation of Targeted Pedestrian Traffic Enforcement Programs in an Urban Environment . *Transportation Research Record: Journal of the Transportation Research Board, 2265*(1), 137-145.

Schattler, K. L., Wakim, J. G., Datta, T. K., & McAvoy, D. (2007). Evaluation of pedestrian and driver behaviors at countdown pedestrian signals in Peoria, Illinois. *Transportation Research Record 2002*, 98–106.

Schneider, R. J., Arnold, L. S., & Ragland, D. R. (2009). Pilot Model for Estimating Pedestrian Intersection Crossing Volumes. *Transportation Research Record, 2140*, 13-26.

Shankar, V. N., Ulfarsson, G. F., Pendyala, R. M., & Nebergall, M. B. (2003). Modeling crashes involving pedestrians and motorized traffic. *Safety Science, 41*, 627–640.

Sheykhfard, A., Haghighi, F., Papadimitriou, E., & Van Gelder, P. (2021). Review and assessment of different perspectives of vehicle-pedestrian conflicts and crashes: Passive and active analysis approaches. *Journal of Traffic and Transportation Engineering (English Edition), 8*(5), 681-702.

Sheykhfard, A., Haghighi, F., Nordfjærn, T., & Soltaninejad, M. (2021). Structural equation modelling of potential risk factors for pedestrian accidents in rural and urban roads. *International Journal of Injury Control and Safety Promotion, 28*(1), 46-57.

Shiwakoti, N., Tay, R., & Stasinopoulos, P. (2020). Development, testing, and evaluation of road safety poster to reduce jaywalking behavior at intersections. *Cognition, Technology & Work, 22*, 389-397.

Sisiopiku, V. P., & Akin, D. (2003). Pedestrian behaviors at and perceptions towards various pedestrian facilities: an examination based on observation and survey data. *Transportation Research Part F: Traffic Psychology and Behaviour, 6*(4), 249–274.

Song, L., Li, Y., Fan, W., & Wu, P. (2020). Modeling pedestrian-injury severities in pedestrian-vehicle crashes considering spatiotemporal patterns: Insights from different hierarchical Bayesian random-effects models. *Analytic Methods in Accident Research, 28*, 100137.

Stevenson, M. (2006). Building safer environments: injury, safety, and our surroundings. *Injury Prevention, 12*, 312–315.

Sucha, M., Dostal, D., & Risser, R. (2017). Pedestrian-driver communication and decision strategies at marked crossings. *Accident Analysis and Prevention, 102*, 41-50.

Tarko, A., & Azam, M. S. (2011). Pedestrian injury analysis with consideration of the selectivity bias in linked police-hospital data. *Accident Analysis and Prevention, 43*(5), 1689-1695.

Tefft, B. C. (2013). Impact speed and a pedestrian's risk of severe injury or death. *Accident Analysis and Prevention, 50*, 871–878.

Terwilliger, J., Glazer, M., Schmidt, H., Domeyer, J., Toyoda, H., Mehler, B., . . . Fridman, L. (2019). Dynamics of Pedestrian Crossing Decisions Based on Vehicle Trajectories in Large-Scale Simulated and Real-World Data.

Tiwari, G., Bangdiwala, S., Saraswat, A., & Gaurav, S. (2007). Survival analysis: Pedestrian risk exposure at signalized intersections. *Transportation Research Part F*, 77–89.

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

Toran Pour, A., Moridpour, S., Tay, R., & Rajabifard, A. (2017). Modelling pedestrian crash severity at mid-blocks. *Transportmetrica A: Transport Science, 13*(3), 273-297.

Twisk, D. A., Vlakveld, W. P., Commandeur, J. J., Shope, J. T., & Kok, G. (2014). Five road safety education programmes for young adolescent pedestrians and cyclists: A multi-programme evaluation in a field setting. *Accident Analysis and Prevention*, 55-61.

Ulfarsson, G., Kim, S., & Booth, K. (2010). Analyzing fault in pedestrian–motor vehicle crashes in North Carolina. *Accident Analysis and Prevention, 42*, 1805–1813.

Useche, S. A., Alonso, F., & Montoro, L. (2020). Validation of the walking behavior questionnaire (WBQ): a tool for measuring risky and safe walking under a behavioral perspective. *Journal of Transport & Health, 18*, 100899.

Vasudevan, V., Pulugurtha, S. S., Nambisan, S. S., & Dangeti, M. R. (2011). Effectiveness of Signal-Based Countermeasures for Pedestrian Safety. *Transportation Research Record No. 2264*, 44–53.

Waizman, G., Shoval, S., & Benenson, I. (2015). Micro-Simulation Model for Assessing the Risk of Vehicle–Pedestrian Road Accidents. *Journal of Intelligent Transportation Systems, 19*(1), 63-77.

Wang, H., Tan, D., Schwebel, D. C., Shi, L., & Miao, L. (2018). Effect of age on children's pedestrian behaviour: Results from an observational study. *Transportation Research Part F: Traffic Psychology and Behaviour, 58*, 556-565.

Wang, J., Huang, H., Xu, P., Xie, S., & Wong, S. C. (2020). Random parameter probit models to analyze pedestrian red-light violations and injury severity in pedestrian–motor vehicle crashes at signalized crossings. *Journal of Transportation Safety and Security, 12*(6), 818-837.

Wang, K., Bhowmik, T., Yasmin, S., Zhao, S., Eluru, N., & Jackson, E. (2019). Multivariate copula temporal modeling of intersection crash consequence metrics: A joint estimation of injury severity, crash type, vehicle damage and driver error. *Accident Analysis and Prevention, 125*, 188-197.

Wang, W., Guo, H., Gao, Z., & Bubb, H. (2011). Individual differences of pedestrian behaviour in midblock crosswalk and intersection. *International Journal of Crashworthiness, 16*(1), 1-9.

World Health Organization. (2021, July). Retrieved from Road traffic injuries: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Wu, X., Miucic, R., Yang, S., Al-Stouhi, S., Misener, J., Bai, S., & Chan, W. (2014). Cars Talk to Phones: A DSRC Based Vehicle-Pedestrian Safety System. *Vehicular Technology Conference (VTC Fall)*. Vancouver, BC, Canada.

Xu, Y., Li, Y., & Zhang, F. (2013). Pedestrians' intention to jaywalk: Automatic or planned? A study based on a dual-process model in China. *Accident Analysis and Prevention, 50*, 811-819. doi:doi.org/10.1016/j.aap.2012.07.007

Yagil, D. (2000). Beliefs, motives and situational factors related to pedestrians' self-reported behavior at signal-controlled crossings. *Transportation Research Part F , 3*, 1-13.

Yang, J. G., & Li, Q. F. (2005). Estimating pedestrian delays at signalised intersections in developing cities by Monte Carlo method. *Mathematics and Computers in Simulation, 68*(4), 329 -337. doi:doi.org/10.1016/j.matcom.2005.01.017

Yang, J., Deng, W., wang, J., Li, Q., & Wang, Z. (2006). Modeling pedestriansÕroad crossing behavior in trafficsystem micro-simulation in China. *Transportation Research Part A, 40*, 280-290.

Yoneda, K., Suganuma, N., Yanase, R., & Aldibaja, M. (2019). Automated driving recognition technologies for adverse weather conditions. *IATSS Research, 43*, 253-262.

Zajac, S., & Ivan, J. (2003). Factors influencing injury severity of motor vehicle–crossing pedestrian crashes in rural Connecticut. *Accident Analysis and Prevention, 35*, 369–379.

Zaki, M. H., & Sayed, T. (2014). Automated Analysis of Pedestrians' Nonconforming Behavior and Data Collection at an Urban Crossing. *Transportation Research Record: Journal of the Transportation Research Board, 2443*, 123-133.

Zaki, M. H., Sayed, T., Tageldin, A., & Hussein, M. (2013). Application of Computer Vision to Diagnosis of Pedestrian Safety Issues. *Transportation Research Record: Journal of the Transportation Research Board, No. 2393.*

Zegeer, C. V., Esse, C. T., Stewart, J. R., Huang, H., & Lagerwey, P. (2004). Safety analysis of marked versus unmarked crosswalks in 30 cities. *ITE Journal (Institute of Transportation Engineers), 74*(1), 34–41.

Zhang, C., Chen, F., & Wei, Y. (2019). Evaluation of pedestrian crossing behavior and safety at uncontrolled mid-block crosswalks with different numbers of lanes in China. *Accident Analysis and Prevention, 123*, 263-273.

Zhang, C., Zhou, B., Qiu, T. Z., & Liu, S. (2018). Pedestrian crossing behaviors at uncontrolled multi-lane mid-block crosswalks in developing world. *Journal of Safety Research, 64*, 145-154.

Zhang, G., Tan, Y., & Jou, R. C. (2016). Factors influencing traffic signal violations by car drivers, cyclists, and pedestrians: A case study from Guangdong, China. *Transportation Research Part F: Traffic Psychology and Behaviour, 42*(1), 205-216.

Zhang, S., Abdel-Aty, M., Yuan, J., & Li, P. (2020). Prediction of Pedestrian Crossing Intentions at Intersections Based on Long Short-Term Memory Recurrent Neural Network. *Transportation Research Record, 2674*(4), 57–65.

Zhang, Y., & Fricker, J. D. (2021). Incorporating conflict risks in pedestrian-motorist interactions: A game theoretical approach. *Accident Analysis and Prevention, 159*, 106254.

Zhang, Y., & Fricker, J. D. (2021). Investigating temporal variations in pedestrian crossing behavior at semi-controlled crosswalks: A Bayesian multilevel modeling approach. *Transportation Research Part F, 76*, 92-108.

Zhang, Y., Gawade, M., Lin, P. S., & McPherson, T. (2013). Educational Campaign for Improving Pedestrian Safety: A University Campus Study. *13th COTA International Conference of Transportation Professionals (CICTP 2013)* (pp. 2756 – 2766). Procedia - Social and Behavioral Sciences 96.

Zhou, H., Romero, S. B., & Qin, X. (2016). An extension of the theory of planned behavior to predict pedestrians' violating crossing behavior using structural equation modeling. *Accident Analysis and Prevention, 95*, 417-424.

Zhu, D., Sze, N. N., & Bai, L. (2021). Roles of personal and environmental factors in the red lightrunning propensity of pedestrian: Case study at the urbancrosswalks. *Transportation Research Part F, 76*, 47-58.

Zhu, D., Sze, N. N., & Feng, Z. (2021). The trade-off between safety and time in the red light running behaviors of pedestrians: A random regret minimization approach. *Accident Analysis and Prevention, 158*, 106214.

Zhu, D., Sze, N. N., Feng, Z., & Yang, Z. (2022). A two-stage safety evaluation model for the red light running behaviour of pedestrians using the game theory. *Safety Science, 147*, 105600.

Zhuang, X., & Wu, C. (2011). Pedestrians' crossing behaviors and safety at unmarked roadway in China. *Accident Analysis & Prevention, 43*(6), 1927–1936.

CHAPTER 3

An Integrated Clustering and Copula-based Model to Assess the Impact of Intersection Characteristics on violation-related Collisions

The publication included in this chapter is:

Ghomi, H., & Hussein, M. (2021). An Integrated Clustering and Copula-based Model to Assess the Impact of Intersection Characteristics on violation-related Collisions. Accident Analysis & Prevention, 159, 106283. https://doi.org/10.1016/j.aap.2021.106283

The manuscript was submitted in May 2021 and accepted in June 2021. Haniyeh Ghomi is the main contributor of this manuscript. The co-author's contributions include guidance, supervision, funding, reviewing the analysis, and editing the manuscript.

3.1 Abstract

The main goal of this study is to investigate the impact of a variety of factors on the frequency and the severity of pedestrian-vehicle collisions that involve pedestrian violations. To that end, the collision dataset of the City of Hamilton between 2010 and 2017 was reviewed to filter out pedestrian collisions that involved pedestrian violations. A Latent Class Analysis (LCA) method was applied to divide the dataset into a set of homogeneous clusters, based on traffic and intersection characteristics. A copula-based multivariate model was then developed for each cluster in order to study the impact of the different factors on collisions under the prevailing conditions of each cluster. The results showed that the number of bus stops within the intersection area is directly associated with the frequency and the severity of collisions involving pedestrian violations. A reduction in collisions was observed with the increase in the frequency of buses at intersections that are located along main transit routes. Moreover, the presence of schools near the intersection tends to increase the frequency of collisions involving pedestrian violations, especially at large intersections. The results also revealed that the presence of central refuge islands, despite their overall safety benefits, increases the likelihood of collisions involving pedestrian violations in large intersections. The results of this study provide valuable insights for a better understanding of the safety consequences of pedestrian violations. Such understanding assists engineers and planners to design intersections that reduce the frequency of pedestrian violations and mitigate their negative safety consequences.

3.2 Introduction

Pedestrian-vehicle collisions are traumatic events that take millions of lives worldwide every year and impose drastic costs to societies. Statistics show that pedestrians are overrepresented in collision fatalities. For example, pedestrians accounted for 17.3% of collision fatalities in 2018 in Canada, despite representing only 3.4% of persons involved in collisions (Transport Canada, 2021). Numerous studies attempted to investigate the factors that contribute to the frequency and severity of pedestrian-vehicle collisions. The impact of a variety of factors was thoroughly investigated, including traffic-related factors (e.g., vehicle speed and traffic volume), location-specific characteristics (e.g., number of lanes and type of traffic control device), road network characteristics (e.g., intersection density), and environmental and external factors (e.g., illumination and weather condition).

Nevertheless, less interest was given to study pedestrian unsafe behaviours and their impact on the overall pedestrian safety level. Several behaviours have been identified in the literature as risky behaviours that may increase the risk of collisions. Among those behaviours, pedestrian violations were identified as a key hazardous behaviour that increases both the probability and the severity of pedestrian-vehicle collisions (Kim et al., 2017; Mukherjee and Mitra, 2020; Wang et al., 2019). Locally, historical collision records of the City of Hamilton, Ontario showed that about 20% of pedestrian-vehicle collisions that occurred at intersections between (2010-2017) were mainly attributed to pedestrian violations, among which, 90% were serious collisions that involved either pedestrian fatalities or serious injuries. Nevertheless, a very limited number of studies that investigated pedestrian violations and their impact on pedestrian-vehicle collisions are found in the literature. Therefore, there is a need to investigate collisions that involve

pedestrian violations in order to identify the factors that impact their occurrence and understand the relationship between the violation behaviour and the severity of such collisions.

In order to investigate collisions that involve pedestrian violations, two issues related to the analysis techniques usually arise. First, the majority of the studies that investigated pedestrian safety relied on a variety of statistical modeling techniques, either on the micro or the macro-level. In such models, historical collision records are used to model collision frequency (and severity) as a function of a variety of factors. However, the influence of these explanatory variables may vary based on the prevailing traffic conditions and the characteristics of the collision location. Discovering the underlying patterns between the different factors and pedestrian collisions is not easily achievable through the analysis of the whole collision dataset using traditional statistical models. Given the heterogeneous nature of pedestrian violations, a more robust approach is needed to investigate the association between such behaviour and collisions.

A prevalent approach to mitigate this issue is to divide the collision dataset into several subsections (clusters), based on the prevailing conditions of the collision location, traffic, and pedestrian characteristics. Analyzing the impact of different explanatory variables, on collisions that involve pedestrian violations within each cluster separately may help to develop a better understanding of how different factors impact collision occurrence under different circumstances. In this regard, Machine Learning clustering techniques can be applied to divide the collision dataset into a set of clusters prior to the development of any statistical models. This would provide an opportunity to study the occurrence of collisions that involve pedestrian

violations in detail and understand the consequences of such collisions, in terms of collision severity.

Second, the majority of studies that investigated the safety impacts of pedestrian violations considered the decision of violation as an explanatory variable that impacts collision frequency and/or severity, without considering the personality traits of pedestrians. Some people inherently tend to take risks while crossing a road, regardless of the road characteristics and the presence of preventive countermeasures. Thus, ignoring the impact of such traits could bias the impact of violations on collision severity. From a statistical point of view, this endogeneity-biased outcome occurs due to the presence of possible interrelationship between the independent variable in a model (i.e., violation) and unobserved variables in the error term (i.e., the personality traits of pedestrians). Due to the impact of unobserved features, violations could be endogenous to the consequence of the collisions. Copula-based multivariate models are one of the most common techniques that could address the endogeneity bias through a joint structure. Based on the definition, the copula is a function that attempts to tie multiple multivariate distributions to the uniform marginal function of each distribution. In other words, the main role of the copula is to connect the dependency between several factors through producing a multivariate distribution (Eluru et al., 2010). Given the nature of pedestrian violations, the copula-based multivariate model is considered an appropriate statistical modeling technique to investigate the safety consequences of pedestrian violations.

Therefore, this paper aims at combining a Machine Learning clustering technique and a copula-based multivariate model to investigate the impact of a variety of contributing factors on both violations and severity of collisions that involve pedestrian violations. Historical collision

records of the City of Hamilton, Ontario between 2010 to 2017 were obtained, and pedestrian collisions that occur at intersections in the city were filtered out. A wide range of factors that may impact pedestrian violations and collisions were obtained from various sources, including vehicles and pedestrian exposure variables, intersection-specific factors (such as traffic control type, intersection size, presence of central refuge islands), and factors related to the amenities in the vicinity of the intersections (mainly bus stops and school zones).

In the first stage, the study utilized the Latent Class Analysis (LCA) clustering algorithm to divide the collision dataset into a set of homogeneous clusters. Several studies recommended the implementation of the LCA technique to study pedestrian-vehicle collisions (Kaplan and Prato, 2013; Mohamed et al., 2013; Zhao et al., 2019). However, this technique has not been utilized to investigate the consequences of pedestrian violations on the frequency and the severity of the collisions and the different factors that impact such collisions, which is a key contribution of this study. In the second stage, a two-dimensional copula-based multivariate model was applied in order to investigate the impact of the different factors on both the frequency and severity of collisions involving violations simultaneously in each cluster.

Therefore, the study provides two methodological contributions to the literature: 1) highlighting the significance of using a clustering algorithm (such as the LCA) to develop a better understanding of the impact of the different factors on the collisions that involve pedestrian violations; and 2) the use of the copula-based model to mitigate the endogeneity-biased outcome that may occur in such analyses.

The results of this study provide valuable insights for a better understanding of the consequences of pedestrian violations and the different factors that affect the frequency and severity of such

collisions. Such understanding assists engineers and planners to design intersections that reduce the frequency of pedestrian violations and mitigate their negative safety consequences. The rest of the paper is organized as follows: The second section provides a summary of the literature review. The third section addresses the methodology of the study. The details of the data collection are presented in section 4. The results of the study are presented and discussed in section 5. Finally, the sixth section summarizes the conclusions and the recommendations of the study.

3.3 Literature review

The following subsections focus on reviewing the previous studies in three key areas. The first area provides a summary related to the main contributing factors to pedestrian-vehicle collisions. The second area is related to defining the different factors that impact pedestrian violation behaviour. Finally, the third area addresses the relationship between violation behaviour and pedestrian safety.

3.3.1 Pedestrian Safety

Extensive research can be found in the literature that attempted to investigate the contributing factors to both frequency and severity of pedestrian-vehicle collisions. The majority of the previous studies recognized the significant impact of several traffic-related factors, such as traffic volume, vehicle speed, and the presence of heavy vehicles on pedestrian-involved collisions (Mohamed et al., 2013; Zhao et al., 2013). Besides, various studies demonstrated the impact of location-specific characteristics on pedestrian safety, including the number of lanes (Pour-Rouholamin and Zhou, 2016; Sasidharan et al., 2015), the presence of central refuge

islands (Ulfarsson et al., 2010; Aidoo et al., 2013) and intersection amenities (Miranda-Moreno et al., 2011; Ding et al., 2018). Moreover, pedestrian traits and behaviours, such as age, gender, violation, and distraction were identified as key contributing factors to pedestrian safety (Zaki et al., 2013; Hussein et al., 2015; Haleem et al., 2015; Amoh-Gyimah et al., 2016). Finally, several environmental-related factors, including weather conditions, illumination, and time of collision showed a direct impact on the pedestrian-vehicle collisions (Moudon et al., 2011; Forbes 2015).

3.3.2 Pedestrian violation behaviours

Relatively, less interest was given to develop a solid understanding of pedestrian unsafe behaviours, such as violations, and study the consequences of such behaviour on the frequency and severity of pedestrian-vehicle collisions. The main focus of those studies was to identify the key contributing factors that impact pedestrian violations. According to the literature, pedestrian violations have been tied with traffic-related factors (Zhu et al., 2021; Yoneda et al., 2019; Nassr et al., 2017), intersection-specific characteristics (Ishaque and Noland, 2008; Cao et al., 2016), built environment factors (Miranda-Moreno et al., 2011; Oakes et al., 2007), pedestrian attitudes (Zareharofteh et al., 2021; Aghabayk et al., 2021; Russo et al., 2018), and environmental features (Zhu et al., 2021; Wsang et al., 2011; Zhang et al., 2016). Previous studies adopted various methods to study the impact of those factors on pedestrian violations, including traditional statistical models (Pawar and Patil, 2016; Chen et al., 2017; Ni et al., 2017), cross-sectional and time-series analysis (Fu and Zou, 2016; Guo et al., 2016), surveys (Chu et al., 2004; Ren et al., 2011), Machine learning techniques (Papadimitriou et al., 2013; Zhang et al., 2020), and micro-simulation analysis (Zaki et al., 2013; Hediyeh et al., 2014).

Although the literature has investigated the impact of a wide range of variables on pedestrian violations, there are still many inconsistencies among the results of the previous studies, particularly for variables like the presence of central island, number of lanes, and the presence of bus stops within the intersection area. For example, the presence of central refuge islands at intersections was shown to have a positive association with the frequency of pedestrian violations in many studies. Li and Ferine (2010) found that only 13% of pedestrians started to cross the intersection during the Walk phase at intersections with a central island in Vancouver, Canada. Cao et al., (2016) reported that the likelihood of temporal violations increased by 15% for each 1% increase in width of the central medians, based on a video-based study in Shanghai, China. On the other hand, other studies reported completely opposite findings. For example, Xu et al., (2013) showed a negative association between pedestrian infrastructure at intersections, like medians, and the frequency of pedestrian violations, based on a study that was conducted in Beijing, China.

Most previous studies found a negative association between the number of lanes and the likelihood of pedestrian violations. For example, Ma et al., (2020) analyzed pedestrian spatial violations (i.e., jaywalking) at three signalized intersections in China, using a Bayesian modeling framework. The study found that the probability of jaywalking decreased at locations with a higher number of lanes. However, some studies reached a different conclusion. For example, Ren et al., (2011) analyzed pedestrian behaviours at 26 signalized intersections in three major cities in China. The study did not find a significant correlation between the number of lanes and the frequency of violations.

Also, most previous studies showed that the presence of bus stops within the intersection area would increase the frequency of pedestrian violations significantly. For example, Zaki et al., (2013) applied a computer vision technique to investigate pedestrian behaviour at a major signalized intersection in Vancouver, Canada. The study found that 67% of the spatial violations that occurred at the intersection are attributed to pedestrians trying to catch buses at one bus stop, located at the southwest corner of the intersection. However, the results of Mukherjee and Mitra (2019) did not support this hypothesis. The study analyzed the historical collision records at 24 intersections in Kolkata, India between 2011 and 2016. The study did not find any impact of the presence of bus stops at an intersection on pedestrian spatial violation behaviour.

Regarding the impact of the presence of schools on the violation behaviour, previous studies seem to agree that intersections that are located near schools usually have a high pedestrian violation. For example, Mukherjee and Mitra (2020) analyzed pedestrian behaviour at 55 signalized intersections in Kolkata, India. The results showed that temporal violations increased significantly at intersections that are located near elementary schools. The study reported that students are the predominant temporal violators at these intersections, especially in the morning as they are rushing to go to school on time.

3.3.3 Relationship between pedestrian violation and their safety

Furthermore, although a few studies highlighted the importance of mitigating pedestrian violations to enhance pedestrian safety (e.g., Harding et al., 2014; Wu et al., 2014), the association between pedestrian violations and safety has been understudied in the literature. Wang et al., (2019) investigated the impact of “crossing on red” on the severity of collisions in Hong Kong. The results showed that two age groups (pedestrians aged 11 years old and younger

and those who are over 66 years old) are more likely to be involved in severe injuries as a result of temporal violations. In another study, Mukherjee and Mitra (2020) implemented a negative binomial model to analyze collision records at 55 intersections in Kolkata, India. The results showed a significant direct relationship between the frequency of temporal violations and the frequency of fatal collisions. The study proposed that the rate of temporal violations at intersections can be used as a surrogate index to identify hazardous intersections.

In summary, it can be seen that the relation between the violation behaviour of pedestrians and their safety level is not well established. Therefore, there is a need to investigate the relationship between pedestrian violations and the frequency and severity of pedestrian in more detail. Moreover, previous studies were not conclusive regarding the impact of several factors on pedestrian violations. Therefore, the impact of these factors on the consequences of pedestrian violations cannot be established. A thorough investigation of the impact of these factors on pedestrian violations and their safety consequences using a technique that accounts for the possibility that these factors may contribute to pedestrian violations differently under different traffic and environmental conditions would help to mitigate the inconsistency of the results found in the literature.

3.4 Methodology

In order to achieve the study objectives, historical collision records of the City of Hamilton between 2010 and 201۷ were obtained. The analysis focused on pedestrian-vehicle collisions that occurred at intersections only. The impact of explanatory variables on the risk of violation and the severity of collisions due to violation behaviour was evaluated in two different approaches.

In the first approach, the copula-based multivariate model was applied to the whole dataset to investigate the contributing factors impact on both violation and severity. In the second approach, a two-staged analysis was developed. First, the LCA method was applied to divide the collision dataset into homogeneous clusters, based on the intersection and traffic characteristics. Then, a two-dimensional copula-based model was developed to assess the potential interrelationships between pedestrian violation and the severity of collisions in each cluster. Finally, the performance of the two approaches is compared based on the AIC criteria. This comparison highlighted the importance of the implementation of the LCA technique. The following sections provide a brief description of the LCA method and the copula-based model.

3.4.1 LCA

Despite the widespread implementation of the traditional clustering techniques, such as nearest neighborhood and K-means in the transportation safety literature (e.g., Anderson, 2009; Maji et al., 2018), they suffer from several limitations, mainly regarding dealing with the outliers and the missing values (Shu, 2020). To overcome these limitations, a new generation of clustering algorithms, known as model-based clustering is being promoted. The LCA method is one of the most common model-based clustering techniques that can be developed for databases with categorical or ordinal dependent variables. As such, these models are considered appropriate for evaluating collision data, which are ordinal in nature (Kaplan and Prato, 2013; Mohamed et al., 2013; Zhao et al., 2019).

The LCA method is an unsupervised clustering technique that has several benefits compared to the traditional clustering techniques (such as nearest neighborhood and K-means):

- The LCA method has no prior assumption regarding the linearity, the standardization, and the normality of the predictors.
- The LCA method deals with missing data in a more sensible way through a missing at random assumption; however, the traditional techniques have no strong assumption for the missing values.
- Since the LCA method is a model-based clustering technique, it is more capable of handling a mixture of independent variables (e.g., numeric, and categorical) compared to other clustering techniques.

The main purpose of LCA algorithms is to cluster the raw database into several subsets, by maximizing the homogeneity within each subset while minimizing it between the different subsets simultaneously (Zhao et al., 2019). The LCA method employs the binomial finite mixture model to predict the probability of independent variable allocation to the clusters, based on Equation (3-1):

$$\pi_j^i = \sum_x \sum_{i=1,2,\dots} \sum_{j=1,2,..} \pi_x^X . \pi_j^{i|X} \qquad (3\text{-}1)$$

where X is the hidden class, π_x^X is the size of class x, and $\pi_j^{i|X}$ is the probability of category j of a variable i to be allocated to a hidden class x. Several statistical criteria can be used to assess the goodness of fit of the developed models and select the optimal number of clusters, including the Akaike information criterion (AIC) and Bayesian information criterion (BIC). In this study, the AIC was used to assess the best number of clusters.

3.4.2 Copula-based multivariate model

In this study, a two-dimensional copula approach is developed in order to model the frequency of collisions involving pedestrian violations and collision severity simultaneously, with error terms ε_i and σ_i, respectively. The joint distribution of the i^{th} level of violation and j^{th} severity level for collision i can be shown as Equation (3-2):

$$Pr(m_i = i, n_i = j) = Pr\{[\alpha_{i-1} - zx_i < \varepsilon_i < \alpha_i - zx_i], [\beta_{j-1} - qx_i < \sigma_i < \beta_j - qx_i]\} \quad (3\text{-}2)$$

where m_i and n_i are violation and severity indicators for collision i, α_i and β_j are thresholds related to the dependent variables, x_i is the vector of explanatory variables, z and q are the parameter coefficients.

Considering S_1, S_2,…, S_q as Q random variables with uniform distribution, Q-dimensional copula model with θ indicator can be expressed as Equation (3-3):

$$A_\theta(s_1, s_2, \dots, s_q) = Pr\ (S_1 < s_1\ , S_2 < s_2\ , \dots.,\ S_q < s_q) \quad (3\text{-}3)$$

In the copula-based multivariate models, the maximum likelihood concept is applied to estimate the parameter coefficients (Rana et al., 2010). To address the heterogeneity among the variables, the copula model estimates an association parameter (θ) as a function of independent variables (Wang et al., 2019) shown as Equation (3-4):

$$A_{\theta_s} = f(\alpha x_s) \quad (3\text{-}4)$$

where θ_s is the association parameter for collision s, f is the function of copula structure, and α is the coefficients vector for the copula parameters. According to the acceptable range of the dependency parameter, various types of f will use to estimate θ. The functional form of the Frank model, which has been developed in the study is calculated as $A_{\theta_s} = exp(\alpha x_s)$.

3.5 Data

The following section presents a brief description of the collision dataset and other supporting data that were collected to undertake the study. The pedestrian-vehicle collision records that occurred at intersections in the City of Hamilton, Ontario from 2010 to 2017 represent the main source of data in this study. In total, 1453 pedestrian-vehicle collisions were reported at 759 intersections in the city during the eight years considered in the analysis. The dataset includes a description of the action of both pedestrian and vehicle at the time of the collision. Pedestrian actions include whether pedestrians involving in the collision were crossing normally, crossing outside the designated crosswalks (spatial violations), crossing during a Do-Not-Walk phase (temporal violations), which enable to identify of the collisions that involve pedestrian violations. A total of 288 collisions (20% of total collisions) were attributed to pedestrian violations in the collision dataset. The 288 collisions resulted in seven fatalities and 252 injuries.

Moreover, five other data sources were utilized in order to extract the independent variables required for the analysis, including Canadian 2016 census data, Hamilton Street Railway (HSR) transit route dataset, Hamilton School Board dataset, Hamilton Open Data website of the City of Hamilton (Open Hamilton, 2021), and Geospatial Datacenter of McMaster University. ArcMap 10.7.1. was utilized to integrate the information of different sources, which enables the development of the required models. First, location-specific characteristics, including intersection size, number of lanes, type of traffic control device implanted at the collision location, and whether the intersection is divided or undivided were extracted from the Hamilton Open GIS map for each intersection in the collision dataset. Second, two exposure parameters, namely, Average Annual Daily Traffic (AADT) and Pedestrian Kilometer Travelled (PKT), were

used to account for road users' exposure in the analysis. The AADT at each intersection was provided by the City of Hamilton to account for vehicle exposure. Unfortunately, a direct measure for pedestrian exposure at each intersection (i.e., pedestrian volume) was not available. In order to overcome this issue, the study utilized the PKT as a surrogate measure of pedestrian exposure. The City of Hamilton was divided into 191 tracts, based on the 2016 Canadian census data (Statistics Canada, 2021). The PKT in each tract was calculated according to the methodology reported in (Nordback et al., 2017), as presented in Equation (3-5).

$$PKT = \sum_{i=1}^{n} Pedestrians_i \times L_i \times 365 \qquad (3\text{-}5)$$

Where $(Pedestrians)_i$ is the total number of walking trips conducted by pedestrians in tract (i) and L_i is the length of road segment (i) in the tract. To determine the total number of walking trips, the dominant mode of transportation in each tract (such as walking, private car, public transit, etc.) was extracted from the census data. PKT in each tract was then used as a measure of pedestrian exposure at all intersections located in this tract.

Third, transit-related parameters, including the number of bus stops and the frequency of buses at each intersection were extracted from the Hamilton Street Railway (HSR) dataset and the Geospatial Datacenter of McMaster University. The exact location of all bus stops in Hamilton was geocoded in ArcMap software. Then, a buffer with a pre-defined radius was generated around the center of each intersection to obtain the number of bus stops that exist within the intersection area. Previous studies considered various buffer sizes when studying the impact of bus stops on pedestrian safety and behaviour, ranging from 5 to 150 meters (Miranda-Moreno et al., 2011; Pulugurtha and Repake, 2008; Schneider et al., 2009). As such, an initial sensitivity analysis was conducted, in which, four different buffer values were used to develop the copula-

based models, including 15, 30, 50, and 100 meters. The sensitivity analysis showed that a 50-meter buffer resulted in the best performance of the copula-based model, as it yields the lowest AIC value. Hence, a 50-meter buffer radius was used to obtain the number of bus stops within each intersection area. Moreover, in order to obtain the frequency of buses within the intersection area, the schedules of 34 bus routes that operate in the City of Hamilton were obtained from the HSR website. Since the frequency of the buses varies significantly between the days of the week and the time of the day, it was decided to consider the bus frequency in the morning and evening peak-hours during weekdays (7:00 – 9:00 AM and 5:00 – 7:00 PM, respectively), as these times experience the highest frequency of buses. For each intersection, the frequency of the buses at all stops within the predefined buffer (50 meters) was used as a measure for bus exposure in the developed models.

Finally, the impact of the presence of schools within the intersection area on collisions involving pedestrian violations was also investigated. To that end, the locations of the educational institutions in the City of Hamilton were extracted from the Open Hamilton dashboard.

A 300-meter radius buffer was generated in ArcMap software to obtain the number of schools in the vicinity of each intersection. This buffer was also selected based on a preliminary sensitivity analysis, in which three different values were examined (100 m, 300m, and 400 m). The 300-meter radius showed the best performance of the developed models, as expressed by the AIC value. The number of students in each school was obtained from the Hamilton School Board website. This enables exploring the impact of both the number and size of schools within the intersection area. A descriptive summary of the factors extracted for each intersection in the study is presented in Table 3-1.

Table 3-1 Descriptive Summary of the Variables

Variable	Mean	Std. Dev.	Min.	Max.
Number of bus stops within the intersection area	2	2.6	0	16
Frequency of buses within the intersection area	38.65	54.61	12	446
Number of schools within the intersection area	1	0.91	0	4
School size within the intersection area	141.87	300.02	67	1515
Number of lanes	556 collisions (less than three lanes per direction) 897 collisions (at larger intersections)			
Traffic control device	971 collisions occurred at signalized intersections, while 482 collisions occurred at unsignalized intersections			
Presence of refuge island	1396 collisions (no central refuge islands) 57 collisions occurred (with central refuge islands)			
Log (AADT)	10.44	0.45	4.75	14.25
Log (PKT)	2.56	0.39	1.18	3.34

3.6 Results and Discussion

The LCA model was applied to classify the collision dataset into a set of homogenous clusters. The LCA was implemented using *mclust* and *poLCA* statistical packages that are available in RStudio 1.2.5042 software. Different model outputs were assessed using the AIC value, and the model with the minimum AIC value was selected. The optimal results involved classifying the dataset into three latent clusters, as shown in Table 3-2, with an AIC value of 12428.91.

As shown in the table, the first cluster includes 234 intersections and accounted for 40.3% of the total collisions that are recorded in the dataset. Intersections in this cluster experienced the lowest percentage of collisions involved pedestrian violations (10.31%). The majority of intersections in this cluster were large, signalized intersections. 8.2% of intersections in this cluster have central refuge islands, the highest percentage in all clusters. Intersections in this cluster have moderate exposure to transit buses, with an average of 38.96 buses during the peak hours of the day. Most intersections in this cluster also do not have any schools within 300

meters of the intersections. The intersections in this cluster experienced the highest exposure to traffic and are located in tracts with the lowest exposure to pedestrians.

Table 3-2 Results of LCA Clustering

Factors	Cluster 1	Cluster 2	Cluster 3
Total number of collisions	585 (40.3%)	503 (34.6%)	365 (25.1%)
Total number of intersections	234	393	124
Percentage of collisions involving pedestrian violations	10.31%	22.20%	28.60%
The dominant number of bus stops (percentage of intersections)	one or two (47.2%)	zero (69%)	more than three (53.2%)
The average number of daily buses during peak hours	38.96	11.7	75.3
Percentage of intersections with at least one school	7.52%	41.4%	100.00%
The average number of students	10.45	172.47	322.78
Percentage of intersections with more than three lanes	73.80%	32.71%	74.2%
Percentage of intersections equipped with traffic signals	90.60%	16.50%	98.10%
Percentage of intersections equipped with central refuge islands	8.20%	0.60%	1.60%
Average Log (AADT)	10.64	10.21	10.48
Average Log (PKT)	2.44	2.57	2.74
Lowest			Highest

Cluster 2 includes 393 intersections that experienced 34.6% of the total collisions. Almost 22% of the collisions that occurred in this cluster involved some sort of pedestrian violations. The intersections in this cluster experience the lowest exposure to buses due to the lack of bus stops at 70% of the intersections. About 40% of the intersections have at least one school within 300 meters of the intersections, with an average school size of 172 students. Most of the intersections in this cluster are small to medium unsignalized intersections. Intersections in this cluster have the lowest exposure to vehicles among the three clusters.

Finally, cluster 3 includes 124 intersections that experienced about 25.1% of the total collisions. The intersections in this cluster experienced the highest percentage of collisions that involved pedestrian violations (28.6%) among the three clusters. These intersections have the highest exposure to transit buses, with an average of 75 buses during the peak hours of weekdays. As well, all intersections in this cluster have at least one school within 300 meters of the intersection, with an average school size of 323 students; the highest across all clusters. The majority of intersections in this cluster are large, signalized intersections that are located in tracts with the highest exposure to pedestrians.

Three copula-based multivariate models were developed to investigate the impact of the different factors on the frequency and severity of collisions that involve pedestrian violations under the prevailing conditions of each cluster. The copula-based models considered two binary indicators as the dependent variables. The first binary dependent variable is called “Violation”, which indicates whether the collision involves a pedestrian violation or not. This variable takes a value of 1 if the collision involves a pedestrian violation (which is the case for 288 collisions in the dataset) and 0 otherwise (which is the case for the remaining 1165 collisions in the dataset). The second binary dependent variable is called “Fatality”, which indicates whether the collision that involves a pedestrian violation resulted in a fatality or not. This variable takes a value of 1 if the collision that involves pedestrian violation resulted in a fatality (7 collisions in the dataset) and 0 otherwise. The parameter estimates of the copula-based multivariate models are reported in Table 3-3. A brief discussion of the results shown in Table 3-3 is presented as follows:

Table 3-3 Results of the copula-based multivariate model

Factors	Cluster 1				Cluster 2				Cluster 3			
	V*	Sig.	F**	Sig.	V*	Sig.	F**	Sig.	V*	Sig.	F**	Sig.
Number of bus stops (one or two)	**0.332**	**0. 042**	**0.317**	**0.006**	**0.101**	**0.000**	**1.226**	**0.009**	**0.039**	**0.014**	**1.357**	**0.007**
Number of bus stops (more than three)	**1.33**	**0.029**	**1.094**	**0.026**	**0.402**	**0.034**	**0.296**	**0.027**	**18.314**	**0.036**	**0.172**	**0.032**
Frequency of buses	**-0.002**	**0.000**	**-0.002**	**0.009**	**0.004**	**0.009**	**-0.02**	**0.041**	**-0.003**	**0.006**	**-0.001**	**0.050**
Presence of schools	**0.001**	**0.031**	**3.421**	**0.021**	**0.659**	**0.025**	**0.074**	**0.012**	**-**		**-**	
School size	**0.02**	**0.001**	**15.855**	**0.000**	**0.525**	**0.041**	**0.113**	**0.000**	**-**		**-**	
Higher number of lanes (more than three lanes)	**-0.346**	**0.029**	**-0.382**	**0.041**	**0.665**	**0.008**	0.086	0.874	**-0.903**	**0.009**	**-10.367**	**0.000**
Presence of refuge island	**0.15**	**0.005**	**0.214**	**0.021**	**-0.006**	**0.004**	0.791	0.916	**0.711**	**0.016**	**0.721**	**0.012**
Presence of traffic signals	**-1.733**	**0.050**	**-0.1**	**0.046**	**-0.894**	**0.050**	**-0.188**	**0.013**	**-0.457**	**0.030**	**-2.371**	**0.037**
Log(AADT)	**0.091**	**0.041**	**0.47**	**0.001**	**0.13**	**0.017**	**0.147**	**0.001**	**0.154**	**0.005**	**0.629**	**0.021**
Log(PKT)	**0.003**	**0.025**	**0.33**	**0.022**	0.071	0.451	0.118	0.776	**0.225**	**0.022**	**0.607**	**0.000**

* V is the indicator of collisions involving pedestrian violations.

** F is the fatal collisions that involve pedestrian violations.

- **Number of bus stops within intersection area**

The number of bus stops within the intersection area is directly associated with the collisions involving a pedestrian violation in all clusters. This may be explained by the increase in the frequency of violations at intersections with more bus stops, as many pedestrians may accept riskier crossing behaviour to catch a bus at the intersection which was observed in many studies in the literature (e.g., Chu et al., 2004; Pulugurtha and Sambhara, 2011; Zaki et al., 2013). The impact of the higher number of bus stops on the collisions involving violations is most notable in cluster 3. This cluster is characterized by a high frequency of buses in peak hours and the presence of large schools in the intersections' area, which lead to an expected increase in the frequency of violations.

Results also showed that the number of bus stops within the intersection area is directly associated with the fatal collisions that involve pedestrian violations in all clusters. However, the highest impact of this factor on fatal collisions was noticed in cluster 3, which mainly includes large intersections that have high exposure to traffic. This indicates that if a collision that involves violations occurred in this cluster, there is a higher chance that this collision is severe, as the large size of the intersection and the high traffic volume increases the risk of pedestrian fatality.

- **Frequency of buses**

Results showed an inverse relationship between the frequency of buses in the peak hours and both total collisions and fatal collisions involving pedestrian violations in all clusters, with only one exception, that is the collisions involving pedestrian violations in cluster 2. As the frequency

of buses increase, pedestrians do not feel the pressure to catch a stopping bus as the waiting time for the next bus would be shorter. This leads to an expected reduction of violators and consequently, a reduction in the severity of collisions involving violations. As for the second cluster, the majority of intersections in this cluster are not well-served by busses. On average, just 11 buses stop by an intersection during the six peak hours of the day (less than 2 buses per peak hour). Under such poor transit service, intersections that are served by more buses, which still represent a relatively low frequency, may experience a high frequency of violations, as passengers accept riskier crossing behaviour to avoid the long waiting time to the next bus. This may explain the direct association between the violations and the frequency of buses in this cluster.

- **Presence of schools and school size**

The presence of schools near intersections and the school size were found to have a significant positive impact on the probability of violations in clusters 1 and 2. Previous studies showed that the presence of schools near intersections increases the frequency of risky behaviours at intersections, which are more common among younger pedestrians and those who walk in groups (Miranda-Moreno et al., 2011; Mukherjee and Mitra, 2020). Consequently, the collisions that involve pedestrian violations increase with the presence of large schools in the intersection area. As for the fatal collisions, a positive association was also found between the two factors and the fatal collisions. The impact on the presence of schools and school size on fatal collisions was more notable in cluster 1, due to the impact of the high traffic volume and the large intersection size on increasing the severity of collisions as discussed earlier. As for the third cluster, all

intersections in this cluster have at least one school within the intersections' area, and consequently, investigating the impact of the presence of schools and school size on collisions that involve violations was not possible in this cluster.

- **The number of lanes**

Table 3-3 shows that large intersections experience fewer collisions related to pedestrian violations in clusters 1 and 3. Previous studies (e.g., Petritsch et al., 2005; Mukherjee and Mitra, 2019) showed that larger intersections discourage pedestrians from violations as they perceive such a behaviour as a dangerous behaviour in this environment. Cluster 2 represents an exception since larger intersections in this cluster are expected to have more collisions related to pedestrian violations. The majority of the intersections in the second cluster are unsignalized intersections that connect minor roads with low traffic volume. In such settings, more people may accept riskier crossing behaviour and violate, compared to signalized major intersections. As such, larger intersections in such environments may experience more collisions involving violations due to the increased exposure to violators and the complexity of larger intersections for pedestrian crossing.

Table 3-3 also shows that increasing the size of the intersection is associated with a reduction in the probability of fatal collisions that involve pedestrian violations (clusters 1 and 3). An opposite result was found in cluster 2, in which larger intersections are associated with higher fatal collisions; however, the parameter estimated for this factor was not statistically significant. It is worth mentioning that a very strong negative association between the number of lanes and fatal collisions was found in the third cluster, as shown in Table 3-3. The explanation of such a

trend is not obvious and may be attributed to some factors that have not been investigated in the study. All intersections in this cluster have at least one school within a close distance from the intersection, and these schools are much larger, in terms of the number of students than their counterparts in other clusters. Thus, intersections in these clusters may be located in reduced speed zones that reduce the average vehicle speed. The intersections in this cluster also experience the largest frequency in buses in peak hours, which is usually associated with a reduction in the average vehicle speed (Yoneda et al., 2019). The lower operating speed may contribute to reducing the severity of collisions at locations in this cluster. Unfortunately, the operating speed data is not available to validate this hypothesis.

- **Refuge island**

Results of the copula model indicate that the collisions involve pedestrian violations increases at intersections equipped with central refuge islands in clusters 1 and 3. Also, the results showed a positive association between the presence of central refuge islands and fatal collisions across all clusters. While many studies reported significant safety benefits for central refuge islands (e.g., Aidoo et al., 2013; Pour-Rouholamin and Zhou, 2016), previous studies showed that central refuge islands encourage pedestrian to accept riskier crossing behaviour (e.g., Das et al., 2005; Hamed, 2001; Ishaque and Noland, 2008). The increase in the frequency of violations at intersections with central refuge islands may explain the increased risk of collisions that involve pedestrian violations. The impact of this factor on total and fatal collisions involving violations is most notable in cluster 3, mainly due to the impact of intersection size and high traffic volume. The negative coefficient reported in cluster 2 suggests that central refuge islands may actually be

associated with a lower probability of violations at smaller minor intersections since this cluster includes the largest percentage of small intersections and the largest percentage of intersections that are not served by transit buses. The results suggest that large major intersections that have central refuge islands need to be equipped with other countermeasures that aim at reducing the risk of pedestrian violations in order to maximize the positive safety impacts of central refuge islands.

- **Traffic signals**

According to the copula model results, signalized intersections are associated with a lower risk of violations and fatal collisions involving pedestrian violations in all clusters. Based on the parameters presented in Table 3-3, traffic signals were shown to be most beneficial in mitigating fatal collisions that involve pedestrian violations in cluster 3, which includes the largest percentage of large and major intersections that are heavily served by transit buses and provide access to schools.

- **Traffic and pedestrian exposure**

As expected, a direct relationship was observed between the risk of violations and fatal collisions involving pedestrian violation, and both pedestrian and vehicle exposure. The highest impact of exposure was found in the third cluster, which contains major intersections that have frequent bus service and large schools within the intersections' area.

3.7 Investigating the significance of LCA

In order to assess the significance of the clustering techniques in providing a better understanding of the impact of the studied factors on collisions that involve pedestrian violations, the copula-based model was implemented on the whole dataset (without clustering). The parameter estimates of the copula-based multivariate model for the whole collision dataset (without clustering) are presented in Table 3-4. Based on the results presented in Table 3-3 (with clustering) and Table 3-4 (without clustering), several remarkable differences were found and discussed below:

Table 3-4 Results of the copula-based multivariate model on the whole dataset

Factors	All Dataset			
	V*	Sig.	V*	Sig.
Number of bus stops (one or two)	**0.110**	**0.015**	**0.110**	**0.015**
Number of bus stops (more than three)	**2.18**	**0.112**	**2.18**	**0.112**
Frequency of buses	**0.003**	**0.000**	**0.003**	**0.000**
Presence of schools	**0.001**	**0.042**	**0.001**	**0.042**
School size	**0.380**	**0.001**	**0.380**	**0.001**
Higher number of lanes (more than three lanes)	**-0.697**	**0.007**	**-0.697**	**0.007**
Presence of refuge island	0.513	0.179	0.513	0.179
Presence of traffic signals	**-0.548**	**0.050**	**-0.548**	**0.050**
Log(AADT)	**0.065**	**0.004**	**0.065**	**0.004**
Log(PKT)	**0.004**	**0.017**	**0.004**	**0.017**

* V is the indicator of collisions involving pedestrian violations.
** F is the fatal collisions that involve pedestrian violations.

- The high number of bus stops at an intersection (more than three), did not have a significant impact on both the frequency and the severity of collisions that involve

pedestrian violations when the copula-based model was applied to the whole dataset. However, the impact of this factor was notable and statistically significant when the model was applied after clustering the dataset, as shown in Table 3-3.

- The impact of the lower number of bus stops at an intersection on the fatal collisions that involve pedestrian violation was consistent when the model was applied with and without clustering the dataset. However, the impact of this factor varies significantly among the three clusters as it shows a different influence on collisions based on the prevailing traffic conditions and location characteristics, which was not easily observed by applying the model on the whole dataset without clustering.
- The frequency of buses showed a direct relationship with the frequency of collisions that involve pedestrian violations when the model was applied to the whole dataset. However, when the model was applied after clustering the collision data, a difference in the impact of this factor was observed between the second cluster (positive association) and clusters 1 and 3 (negative association), as discussed earlier.
- Implementation of the copula-based model on the whole dataset showed that there is no significant relationship between the presence of the school in the vicinity of an intersection and the severity of collisions due to violations. However, the LCA technique addressed this misleading result by demonstrating the significant positive impact of this factor on the consequences of the pedestrian violations in clusters 1 and 2. The same trend can be found for the impact of the school size on both the frequency and severity of collisions that occurred due to pedestrian violations. The impact of school size was not

significant when the model was applied for the whole collision dataset, while the influence becomes obvious when the data were clustered using the LCA technique.

- According to Table 3-4, the presence of the central refuge islands did not have a statistically significant impact on collisions that involved pedestrian violations when the Copula-based model was applied to the whole collision records. However, the implementation of the LCA to cluster the data prior to modeling collisions revealed vital information regarding how the impact of this factor varies among clusters based on intersection size, as shown in Table 3-3.
- Analyzing the whole dataset failed to capture the different impact of the number of lanes on the frequency of collisions at intersections located in cluster 2. As discussed earlier, the number of lanes was positively associated with the frequency of collisions that involve pedestrian violations in the second cluster, which is likely attributed to the traffic control device at most of the intersections.

3.8 Conclusion

In this study, an integrated Machine Learning unsupervised clustering algorithm and a copula-based multivariate model were applied to analyze pedestrian-vehicle collisions that involve pedestrian violations. The analysis was conducted using the historical collision records of the City of Hamilton, Ontario from 2010 to 2017. The goal was to investigate the main contributing factors that impact both the frequency and the severity of collisions involving pedestrian violations. The main findings of the study were:

- There is a strong positive relationship between the presence of bus stops and schools in the vicinity of an intersection and the frequency of collisions that involve violations. Such amenities usually generate a high frequency of pedestrian violations, and consequently, more collisions that involve violations are observed.
- A high frequency of buses in peak hours is usually associated with a lower frequency of total and fatal collisions that involve pedestrian violations. This may be explained by the fact that pedestrians do not feel the pressure to catch a stopping bus when they know that the waiting time for the next bus would be short.
- More collisions that involve violations are observed in smaller intersections since pedestrians are discouraged from illegal crossing in large and crowded intersections.
- Larger intersections that are well-served by transit buses and have schools within a close distance of the intersection tend to have a lower rate of fatal collisions that involve violations. The operating speed is potentially lower at these intersections due to the higher frequency of buses in the peak hours and the presence of schools. However, the lack of operating speed data at the studied location data does not enable the validation of this hypothesis.
- The presence of central refuge islands increases the likelihood of collisions resulted from pedestrian violations in large intersections.

According to the above-mentioned findings, the study provides several recommendations that aim at mitigating the safety consequences of pedestrian violations, summarized as follows:

- More care should be given to the transit service operational parameters (particularly, the bus frequency) and location of bus stops at major intersections. Along with the importance of the proper design for transit service quality, it plays a significant role in mitigating pedestrian violations and reducing the frequency of collisions involving violations.
- The design of the walking infrastructure at intersections that are located near school zones or along major bus routes needs to be properly designed particularly for reducing the likelihood of pedestrian violations, as these locations usually observe a higher frequency of collisions involving violations.
- Large and major intersections that have central refuge islands need to be equipped with other countermeasures that aim at reducing the frequency of pedestrian violations to maximize the safety benefits of refuge islands.

Moreover, several research directions can be recommended for future studies, including:

- Future studies should analyze more datasets from different cities to investigate the impact of culture and behavioural differences on the results.
- Conduct micro-level analysis of pedestrian violations at different intersections to develop a better understanding of such a behaviour and its safety consequences.
- Investigating the impact of the location of bus stops on pedestrian violations and safety is of great importance, particularly at large and major intersections.
- There is a need to investigate the impact of the actual operating speed of the vehicles on the severity of collisions that involve pedestrian collisions.

- The study used the PKT in each tract as a surrogate measure of pedestrian exposure at collision locations. However, more precise measures for pedestrian exposure can be explored, including collecting extra survey data or implementation of activity-based algorithms to estimate the pedestrian exposure at an intersection, as discussed in Xie et al. (2018) and Li et al. (2020).
- Future studies should continue to explore the impact of other contributing factors on collisions that involve pedestrian violations.

3.9 References

Aghabayk, K., Esmailpour, J., Jafari, A., & Shiwakoti, N. (2021). Observational-based study to explore pedestrian crossing behaviors at signalized and unsignalized crosswalks. *Accident Analysis and Prevention, 151*, 105990.

Aidoo, E. N., Amoh-Gyimah, R., & Ackaah, W. (2013). The effect of road and environmental characteristics on pedestrian hit-and-run accidents in Ghana. *Accident Analysis and Prevention, 53*, 23-27.

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

Anaya, J. J., Merdrignac, P., Shagdar, O., Nashashibi, F., & Naranjo, J. E. (2014). Vehicle to Pedestrian Communications for Protection of Vulnerable Road Users. *IEEE Intelligent Vehicles Symposium* . Michigan, USA.

Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis and Prevention, 41*, 359–364.

Cao, Y., Ni, Y., & Li, K. (2016). Effects of Refuge Island Settings on Pedestrian Safety Perception and Signal Violation at Signalized Intersections. *96th Annual meeting of Transportation Research Board.*

Chen, S., Xing, J., & Cao, Y. (2017). The impact of waiting time on pedestrian violations at signalized intersections. *Civil Engineering and Urban Planning: An International Journal (CiVEJ), 4*(2).

Chu, X. H., Guttenplan, M., & Maltes, M. R. (2004). Why People Cross Where They Do: The Role of Street Environment . *Transportation Research Record, 1878*, 3-10.

Das, S., Manski, C. F., & Manuszak, M. D. (2005). Walk or wait? An empirical analysis of street crossing decisions. *Journal of Applied Econometrics, 20*(4), 529–548.

Ding, C., Chen, P., & Jiao, J. (2018). Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: A machine learning approach. *Accident Analysis and Prevention, 112*, 116-126.

Domenichini, L., Branzi, V., & Smorti, M. (2019). Influence of drivers' psychological risk profiles on the effectiveness of traffic calming measures. *Accident Analysis and Prevention, 123*, 243-255.

Eluru, N., Paleti, R., Pendyala, R. M., & Bhat, C. R. (2010). Modeling Injury Severity of Multiple Occupants of Vehicles: Copula-Based Multivariate Approach. *Transportation Research Record: Journal of the Transportation Research Board, 2165*, 1-11.

Forbes, J. H. (2015). Pedestrian injury severity levels in the Halifax Regional Municipality, Nova Scotia, Canada: hierarchical ordered probit modeling approach. . *Transportation Research Record 2519*, 172-178.

Fu, L., & Zou, N. (2016). The influence of pedestrian countdown signals on children's crossing behavior at school intersections. *Accident Analysis and Prevention, 94*, 73-79.

Garder, P. (1989). Pedestrian Safety At Traffic Signals: A Study Carried Out With The Help Of A Traffic Conflicts Technique. *Accident Analysis and Prevention, 21*(5), 435-444.

Guo, Y., Liu, P., Liang, Q., & Wang, W. (2016). Effects of parallelogram-shaped pavement markings on vehicle speed and safety of pedestrian crosswalks on urban roads in China. *Accident Analysis and Prevention, 95*, 438–447.

Haleem, K., Alluri, P., & Gan, A. (2015). Haleem, K., Abdel-Aty, M., 2010. Examining traffic crash injury severity at unsignalized intersections. J. Saf. Res. 41, 347–357. *Accident Analysis and Prevention, 81*, 14-23.

Hamed, M. M. (2001). Analysis of pedestrians' behavior at pedestrian crossings. *Safety Science, 38*, 63-82.

Harding, J., Powell, G., Yoon, R., Fikentscher, J., Doyle, C., Sade, D., . . . Wang, J. (2014). *Vehicle-to-Vehicle Communications: Readiness of V2V Technology for Application* . U.S. Department of Transportation.

Hediyeh, H., Sayed, T., & Zaki, M, M. (2014). Automated analysis of pedestrian crossing speed behavior at scramble-phase signalized intersections using computer vision techniques. *International Journal of Sustainable Transport, 8*(5), 382– 397.

Hussein, M., Sayed, T., Reyad, P., & Kim, L. (2015). Automated Pedestrian Safety Analysis at a Signalized Intersection in New York City: Automated Data Extraction for Safety Diagnosis and Behavioral Study. *Transportation Research Record Journal of the Transportation Research Board, 2519*, 17-27.

Ishaque, M. M., & Noland, R. B. (2008). Behavioural issues in pedestrian speed choice and street crossing behaviour: a review. *Transport Reviews , 28*(1), 61–85.

Kaplan, S., & Prato, C. G. (2013). Cyclist-motorist crash patterns in Denmark: A latent class clustering approach. *Traffic Injury Prevention, 14*, 725–733.

Khosravi, S., Beak, B., Head, K. L., & Saleem, F. (2018). Assistive System to Improve Pedestrians' Safety and Mobility in a Connected Vehicle Technology Environment. *Transportation Research Record: Journal of the Transportation Research Board, 2672*(19), 145-156.

Kim, J. K., Ulfarsson, G. F., Shankar, V. N., & Kim, S. (2008). Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis. *Accident Analysis and Prevention, 40*, 1695-1702.

Kim, M., Kho,, S. Y., & Ki, D. K. (2017). Hierarchical Ordered Model for Injury Severity of Pedestrian Crashes in South Korea. *Journal of Safety Research, 61*, 33–40.

LI, H., Wu, D., Graham, D. J., & Sze, N. N. (2020). Comparison of exposure in pedestrian crash analyses: A study based on zonal origin-destination survey data. *Safety science, 131*, 104926.

Li, Y., Fernie, G. (2010). Pedestrian behavior and safety on a two-stage crossing with a center refuge island and the effect of winter weather on pedestrian compliance rate. *Accident Analysis and Prevention, 42*, 156–163.

Ma, Y., Lu, S., & Zhang, Y. (2020). Analysis on Illegal Crossing Behavior of Pedestrians at Signalized Intersections Based on Bayesian Network. *Journal of Advanced Transportation*.

Maji, A., Velaga, N. R., & Urie, Y. (2017). Hierarchical clustering analysis framework of mutually exclusive crash causation parameters for regional road safety strategies. *International Journal of Injury Control and Safety Promotion, 25*(3), 257-271.

Miranda-Moreno, L. F., Morency, P., & El-Geneidy, A. M. (2011). The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections. *Accident Analysis and Prevention, 43*, 1624–1634.

Mohamed, M. G., Saunier, N., Miranda-Moreno, L. F., & Ukkusuri, S. V. (2013). A clustering regression approach: a comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. *Safety Science, 54*, 27–37.

Moudon, A. V., Lin, L., Jiao, J., Hurvitz, , P., & Reeves, P. (2011). The Risk of Pedestrian Injury and Fatality in Collisions with Motor Vehicles, a Social Ecological Study of State Routes and City Streets in King County, Washington. *Accident Analysis and Prevention, 43*(1), 11-24.

Mukherjee, D., & Mitra, S. (2019). A comparative study of safe and unsafe signalized intersections from the view point of pedestrian behavior and perception. *Accident Analysis and Prevention, 132*.

Mukherjee, D., & Mitra, S. (2020). A comprehensive study on factors influencing pedestrian signal violation behaviour: Experience from Kolkata City, India. *Safety Science, 124*, 104610.

Nassr, M. M., Zulkiple, A., Albargi, W. A., & Khalifa, N. A. (2017). Modeling pedestrian gap crossing index under mixed traffic condition. *Journal of Safety Research, 63*, 91-98.

Ni, Y., Cao, Y., & Li, K. (2017). Pedestrians' Safety Perception at Signalized Intersections in Shanghai. *Transportation Research Procedia, 25*, 1955–1963.

Nordback, K., Sellinger, M., & Phillips, T. (2017). *Estimating Walking and Bicycling at the State Level*. Portland, OR: National Institute for Transportation and Communities (NITC).

Oakes , J. M., Forsyth, A., & Schmitz, K. H. (2007). The effects of neighborhood density and street connectivity on walking behavior: the Twin Cities walking study. *Epidemiologic Perspectives and Innovations, 4*(16).

Open Hamilton. (2021, April). Retrieved from https://open.hamilton.ca/

Osama, A., & Sayed, T. (2017). Evaluating the impact of connectivity, continuity, and topography of sidewalk network on pedestrian safety. *Accident Analysis and Prevention, 107*, 117-125.

Papadimitriou, E., Theofilatos, A., & Yannis, G. (2013). Patterns of pedestrian attitudes, perceptions and behaviour in Europe. *Safety science, 53*, 114-122.

Pawar, D. S., & Patil, G. R. (2016). Critical gap estimation for pedestrians at uncontrolled mid-block crossings on high-speed arterials. *Safety Science, 86*, 295-303.

Petritsch, T. A., Landis, B. W., McLeod, P. S., Huang, H. F., Challa, S., & Guttenplan, M. (2005). Level-of-Service Model for Pedestrians at Signalized Intersections. *Transportation Research Record, 1939*, 55-62.

Pour-Rouholamin, M., & Zhou, H. (2016). Investigating the risk factors associated with pedestrian injury severity in Illinois. *Journal of Safety Research, 57*, 9–17.

Pulugurtha, S., & Sambhara, V. R. (2011). Pedestrian crash estimation models for signalized intersections. *Accident Analysis and Prevention, 43*(1), 439–446.

Rahman, M., Islam, M., Calhoun, J., & Chowdhury, M. (2019). Real-Time Pedestrian Detection Approach with an Efficient Data Communication Bandwidth Strategy. *Transportation Research Record*.

Rana, T. A., Sikder, S., & Pinjari, A. R. (2010). Copula-Based Method for Addressing Endogeneity in Models of Severity of Traffic Crash Injuries: Application to Two-Vehicle Crashes. *Transportation Research Record: Journal of the Transportation Research Board, 2147*, 75-87.

Ren, G., Zhou, Z., Wang, W., Zhang, Y., & Wang, W. (2011). Crossing behaviors of pedestrians at signalized intersections. *Transportation Research Record, No. 2264*, 65-73.

Russo, B. J., James, E., Aguilar, C. Y., & Smaglik, E. J. (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.

Sasidharan, L., Wu, K. F., & Menendez, M. (2015). Exploring the Application of Latent Class Cluster Analysis for Investigating Pedestrian Crash Injury Severities in Switzerland. *Accident Analysis and Prevention, 219–28*, 85. doi:doi.org/10.1016/j.aap.2015.09.020

Savolainen, P. T., Gates, T. G., & Datta, T. K. (2011). Implementation of Targeted Pedestrian Traffic Enforcement Programs in an Urban Environment . *Transportation Research Record: Journal of the Transportation Research Board, 2265*(1), 137-145.

Schneider, R. J., Arnold, L. S., & Ragland, D. R. (2009). Pilot Model for Estimating Pedestrian Intersection Crossing Volumes. *Transportation Research Record, 2140*, 13-26.

Shiwakoti, N., Tay, R., & Stasinopoulos, P. (2020). Development, testing, and evaluation of road safety poster to reduce jaywalking behavior at intersections. *Cognition, Technology & Work, 22*, 389-397.

Shu, X. (2020). *Knowledge Discovery in the Social Sciences: A Data Mining Approach.* Oakland, California: University of California Press.

Statistics Canada. (2021, April). Retrieved from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm

Transport Canada. (2021, April). Retrieved from https://tc.canada.ca/en/canadian-motor-vehicle-traffic-collision-statistics-2018

Ulfarsson, G. F., Kim, S., & Booth, K. (2010). Analyzing fault in pedestrian–motor vehicle crashes in North Carolina. *Accident Analysis and Prevention, 42*, 1805–1813.

Vasudevan, V., Pulugurtha, S. S., Nambisan, S. S., & Dangeti, M. R. (2011). Effectiveness of Signal-Based Countermeasures for Pedestrian Safety. *Transportation Research Record No. 2264*, 44–53.

Wang, K., Bhowmik, T., Yasmin, S., Zhao, S., Eluru, N., & Jackson, E. (2019). Multivariate copula temporal modeling of intersection crash consequence metrics: A joint estimation of injury severity, crash type, vehicle damage and driver error. *Accident Analysis and Prevention, 125*, 188-197.

Wang, W., Guo, H., Gao, Z., & Bubb, H. (2011). Individual differences of pedestrian behaviour in midblock crosswalk and intersection. *International Journal of Crashworthiness, 16*(1), 1-9.

Wu, X., Miucic, R., Yang, S., Al-Stouhi, S., Misener, J., Bai, S., & Chan, W. (2014). Cars Talk to Phones: A DSRC Based Vehicle-Pedestrian Safety System. *Vehicular Technology Conference (VTC Fall)*. Vancouver, BC, Canada.

Xie, S. Q., Dong, N., Wong, S. C., Huang, H., & Xu, P. (2018). Bayesian approach to model pedestrian crashes at signalized intersections with measurement errors in exposure. *Accident Analysis & Prevention, 121*, 285-294.

Xu, Y., Li, Y., & Zhang, F. (2013). Pedestrians' intention to jaywalk: Automatic or planned? A study based on a dual-process model in China. *Accident Analysis and Prevention, 50*, 811-819.

Yoneda, K., Suganuma, N., Yanase, R., & Aldibaja, M. (2019). Automated driving recognition technologies for adverse weather conditions. *IATSS Research, 43*, 253-262.

Zaki, M. H., Sayed, T., Tageldin, A., & Hussein, M. (2013). Application of Computer Vision to Diagnosis of Pedestrian Safety Issues. *Transportation Research Record: Journal of the Transportation Research Board, No. 2393.*

Zareharofteh, F., Hidarnia, A., Morowatisharifabad, M., & Eslami, M. (2021). Unsafe behaviours in Iranian adult pedestrians. *Journal of Transport & Health, 21*, 101058.

Zhang, G., Tan, Y., & Jou, R. C. (2016). Factors influencing traffic signal violations by car drivers, cyclists, and pedestrians: A case study from Guangdong, China. *Transportation Research Part F: Traffic Psychology and Behaviour, 42*(1), 205-216.

Zhang, S., Abdel-Aty, M., Yuan, J., & Li, P. (57–65). Prediction of Pedestrian Crossing Intentions at Intersections Based on Long Short-Term Memory Recurrent Neural Network. *Transportation Research Record, 2674*(4), 2020.

Zhao, H., Yang, G., Zhu, F., Jin, X., Begeman, P., Yin, Z., . . . Wang, Z. (2013). An investigation on the head injuries of adult pedestrians by passenger cars in China. *Traffic Injury Prevention, 14*, 712–717.

Zhao, Sh., Iranitalab, A., Khattak, A.J. . (2019). A clustering approach to injury severity in pedestrian-train crashes at highway-rail grade crossings. *Journal of Transportation Safety and Security, 11*(3), 305-322.

Zhu, D., Sze, N. N., & Bai, L. (2021). Roles of personal and environmental factors in the red lightrunning propensity of pedestrian: Case study at the urbancrosswalks. *Transportation Research Part F, 76*, 47-58.

CHAPTER 4

Analyzing the Safety Consequences of Pedestrian Spatial Violation at Mid-blocks: A Bayesian Structural Equation Modelling Approach

The publication included in this chapter is:

Ghomi, H., & Hussein, M. (2021). Analyzing the Safety Consequences of Pedestrian Spatial Violation at Mid-blocks: A Bayesian Structural Equation Modelling Approach. Transportation Research Record, p. 03611981221097964. https://doi.org/10.1177/0361198122109796

The manuscript was submitted in August 2021 and accepted in June 2022. Haniyeh Ghomi is the main contributor of this manuscript. The co-author's contributions include guidance, supervision, funding, reviewing the analysis, and editing the manuscript.

4.1 Abstract

The main goal of this study is to understand the impact of a variety of factors on the frequency and severity of pedestrian-vehicle collisions that involve pedestrian spatial violations at mid-blocks. To that end, the historical collision records of the City of Hamilton between 2010 and 2017 were obtained, and collisions that occur at mid-blocks were filtered out. A Bayesian Structural Equation Modelling (SEM) framework was developed to investigate the impact of a wide range of factors on such collisions. First, a classical SEM was developed to group the different factors into sets of latent variables. Four latent variables were defined, including access to services, location vibrancy, pedestrian network quality, and road size. Then, the Bayesian SEM was implemented to investigate the relationship between the latent variables and collisions. The results showed that access to services (e.g., parks, schools, bike-share stations, and bus stops) were the most influential factor on the frequency of collisions that involve spatial violation, followed by the pedestrian network quality. Pedestrian network quality and road size were found to be the most influential factors on the severity of collisions. The location of bike-share stations, pedestrian network connectivity, exposure to walkers, and the number of lanes were the four observed variables that explained the highest percent of the variance in each latent group, respectively. The results of this study should assist engineers and planners to develop better design concepts to mitigate collisions that are caused by pedestrian spatial violations in urban areas.

4.2 Introduction

Promoting non-motorized modes of transportation, such as walking and biking, has become a central objective to many transportation agencies around the world. Numerous policies and design concepts have been introduced to encourage active modes of travel, aiming at promoting sustainable communities and reducing single-occupancy vehicle trips. Nevertheless, safety concerns have been one of the major roadblocks for the full utilization of active travel modes as key modes of travel in many North American cities. Many transportation safety professionals consider pedestrians and cyclists to be among the most vulnerable road user groups who have a higher risk of being killed or severely injured as a result of road collisions. Historical collision data clearly show that pedestrians and cyclists are overrepresented in collision fatalities and serious injuries. For example, pedestrians accounted for 17.3% of collision fatalities in 2018 in Canada, despite representing only 3.4% of persons involved in collisions (Transport Canada, 2021).

While pedestrian safety has been investigated extensively in the literature, pedestrian behaviour and its impact on their safety have been relatively understudied. Previous studies showed that pedestrian-unfriendly design of the urban road networks, lack of effective pedestrian facilities, and inadequate prior education of pedestrians promote many risky pedestrian behaviours that impact the overall road safety level, such as spatial violation (Soathong et al., 2021).

Crossing the street at undesignated spaces (spatial violations) has become a common way to cross streets in big cities. No surprise, such behaviour was shown to be a major contributor to increasing the frequency and severity of pedestrian-vehicle collisions. Historical collision

records of the City of Hamilton, Ontario showed that about 35% of pedestrian-vehicle collisions that occurred at mid-block locations between (2010-2017) were mainly attributed to pedestrian violations. Data also shows that 91.4% of those collisions were serious collisions that involved either pedestrian fatalities or serious injuries.

In this study, spatial violations in mid-blocks are defined in two cases: Pedestrian crossings outside a designated mid-block crosswalk that exist within 30 meters of the crossing location, and 2) Crossings in mid-blocks with no close-by marked crosswalks where pedestrians did not yield the right-of-way to vehicles. The two cases were identified after a careful review of pedestrian crossing laws in the province of Ontario and some major cities in the province (specifically, the City of Toronto). The Highway Traffic Act of the province of Ontario, Chapter H.8, Section 144(22) stated that "Where portions of a roadway are marked for pedestrian use, no pedestrian shall cross the roadway except within a portion so marked" (Traffic Act, 1990). The law does not stipulate how far from the nearest intersection one must be in order to legally cross mid-block. Some cities, such as the City of Toronto, follow police advice to generally use 30 meters from the nearest intersection as a 'rule of thumb' (City of Toronto, 2022). This means that if a pedestrian crosses the street at unmarked crosswalks while they are within 30 meters of an intersection, it is considered a legal offense. The study followed this concept and considered the crossings that occur outside the crosswalk, but within 30 meters of the crosswalk to be a spatial violation. Also, the Municipal Code of the City of Toronto (Section 950-300B) states that "No person shall, except where traffic control signals are in operations, or where traffic is being controlled by a police officer, or at a pedestrian crossover, proceed so as not to yield the right-of-

way to vehicles and streetcars on the roadway" (City of Toronto, 2022). This means that pedestrian crossings in the mid-blocks with no close-by marked crosswalk can still be illegal if pedestrians do not yield the right-of-way to vehicles. Based on that concept, The study considered pedestrian mid-block crossings where no close-by marked crosswalks exist to be spatial violations if the pedestrians did not provide the right-of-way to vehicles.

Pedestrian spatial violations are typically influenced by a variety of contributing factors, such as social norms and habits, road network characteristics, traffic conditions, and built environment characteristics, among other factors. Previous studies investigated the impact of a wide range of factors on pedestrian spatial violations and attempted, to some extent, to assess the impact of such behaviour on pedestrian-vehicle collisions. However, the impact of many factors, such as pedestrian network characteristics (network connectivity and accessibility) and location amenities, on the frequency and severity of collisions that involve pedestrian violations still requires further investigation. Moreover, the majority of previous studies assessed the impact of spatial violations on safety through advanced regression models, which consider the spatial violation decision as an independent variable that impacts collision occurrence. These models did not consider the personality traits of pedestrians while analyzing the spatial violations. Some pedestrians inherently tend to take risks while crossing a road, regardless of the road characteristics and the presence of preventive countermeasures. Thus, ignoring the impact of such traits could bias the impact of violations on collision frequency and severity. From a statistical point of view, this endogeneity-biased outcome occurs due to the presence of possible interrelationship between the independent variable in a model (i.e., spatial violation) and

unobserved variables in the error term (i.e., the personality traits of pedestrians). Due to the impact of unobserved features, spatial violations could be endogenous to the consequence of the collisions.

Bayesian Structural Equation Modeling (SEM) is one of the most common techniques that are capable of addressing the aforementioned endogeneity bias, by considering unobserved (latent) variables while developing a model based on the observed explanatory variables. In other words, the main role of SEM models is to define a median variable (i.e., latent variable) to identify the hidden impacts of the observed variables on the dependent one (Joreskog, 1973). Given the potential endogeneity bias of pedestrian spatial violations, the Bayesian SEM is considered an appropriate statistical technique to investigate the safety consequences of pedestrian spatial violations, in terms of the frequency and severity of the resulted collisions. Although several studies recommended the implementation of the classical SEM method to study road safety (Lee et al., 2008; Kim et al., 2011), this technique has not been utilized to investigate the safety consequences of pedestrian spatial violations.

The main objective of the study is to analyze pedestrian collisions at mid-block locations that are attributed to spatial violations. The goal is to understand the impact of a variety of factors on the frequency and severity of pedestrian-vehicle collisions that involve pedestrian spatial violations. To that end, historical collision records of the City of Hamilton, Ontario between 2010 to 2017 were obtained, and pedestrian collisions that occur at mid-block locations were filtered out. A wide range of factors that may impact pedestrian collisions that involve spatial violation was obtained from various sources, including exposure parameters, location-specific characteristics

(such as the number of lanes, presence of central refuge islands, and road surface condition), the amenities and attractions that exist in the vicinity of the collision location (such as parking lots, bus stops, schools, convenience stores, and parks), pedestrian network features (such as connectivity, block size), and land-use at the collision location. A Bayesian SEM framework was developed to investigate the impact of the considered variables on both the frequency and the severity of collisions that involve pedestrian spatial violations at mid-block areas. The results of the model were analyzed to understand the impact of the different factors on the violation-related collisions, along with identifying the relative importance of each factor in influencing the frequency and the severity of collisions.

This study provides two main contributions: 1) the application of the Bayesian SEM to assess the safety consequences of pedestrian spatial violation and identifying the contributing factors that affect the frequency and severity of collisions that involve pedestrian spatial violations; and 2) the study investigated the impact of numerous variables on violation-related collisions that were not thoroughly considered in previous studies, such as pedestrian network connectivity and accessibility, and a variety of location amenities and attractions. The results of this study provide valuable insights for a better understanding of the factors that encourage pedestrians to spatial violation and increase the risk of collisions, along with the impact of road and pedestrian network characteristics on pedestrian behaviour and safety. Such understanding assists transportation engineers and planners to develop better design concepts to mitigate the frequency and severity of collisions that are caused by pedestrian spatial violations in urban areas. The rest of the paper is organized as follows: The following section provides a summary of the literature

review. Afterward, the research methodology is documented, followed by a summary of the data collection and processing. Next, the results of the study are presented, and a brief discussion of the results is provided. Finally, the last section of the paper presents the conclusions and the recommendations of the study.

4.3 Literature Review

The literature review focused on reviewing previous studies in two key areas: 1) understanding the contributing factors to pedestrian spatial violations at mid-blocks and the safety consequences of such behaviour and 2) investigating the applications of SEM models in pedestrian safety research. The following subsections provide a summary of the findings of the literature review in the two areas.

4.3.1 Pedestrian Spatial Violations

Many pedestrians engage in spatial violations while crossing to save time and reduce the walking distance (Turner et al., 2019). Nevertheless, pedestrians' decisions to whether violate or not vary significantly depending on many factors. Waiting time to cross and the available gap between vehicles have been identified as important factors that influence pedestrians' decision to violate (Yoneda et al., 2019; Zhang et al., 2016). Vehicle speed was also identified as another traffic-related factor that contributes to pedestrian violation in many studies (Papić et al., 2020; Kadali et al., 2020).

Many studies also showed that road characteristics play an important role in pedestrian decisions to violate. For example, pedestrians were shown to be more eager to cross the street without the

right of way at mid-block locations equipped with central refugee islands (Cao et al., 2016; Pour-Rouholamin and Zhou, 2016). The large block size was also found to be an important factor that increases the probability of spatial violation (Oakes et al., 2007). The presence of bus stops at a location increases the frequency of spatial violations, especially at times where buses are waiting at the bus stop (Zaki et al., 2013). Considering pedestrian traits, while some studies showed that men are more likely to engage in high-risk situations and end up in collisions that happened due to spatial violation (Abdullah et al., 2021; Useche et al., 2021), other studies found that gender has no significant influence on such behaviour (Holland and Hill, 2010; Tom and Granié, 2011). Younger pedestrians were shown to be more likely to violate compared to older pedestrians in many studies (Tom and Granié, 2011). Previous studies also highlighted the role of habits, social norms, and past experiences on pedestrian violation behaviour (Rankavat and Tiwari, 2020; Papadimitriou et al., 2017).

Moreover, spatial violations were shown to be an important contributor to the frequency and severity of pedestrian collisions. For example, Hussein et al., (2015) analyzed the association between pedestrian violation and the frequency of pedestrian-vehicle conflicts at a signalized intersection in New York City. The study identified pedestrian violations as the main contributors to pedestrian-vehicle conflicts and showed that 18% of the pedestrians tend to cross the street in a non-designated space. Kim et al., (2017) utilized the hierarchical order technique to analyze more than 137,400 pedestrian collisions in South Korea between 2011 and 2013. The results showed that spatial violations at mid-blocks and temporal violation of drivers (red light running) were the main contributing factors to severe injury collisions. Pour-Rouholamin and

Zhou (2016) reported that pedestrians who cross the roadway at the dedicated crosswalks are 12% less likely to engage in a fatal collision compared to violation. In another study, Ghomi and Hussein (2021) applied an integrated clustering and copula-based model to investigate the impact of pedestrian violations at intersections on both the frequency and the severity of collisions in the City of Hamilton, Ontario. The study showed a strong association between pedestrian violations at intersections and the frequency and severity of pedestrian-vehicle collisions, especially at intersections that have multiple bus stops and schools in the vicinity of the intersection.

4.3.2 SEM Applications

The application of SEM in the Transportation field is most popular in Transportation Planning and travel behaviours (Fillone et al., 2005). Several studies showed the merits of SEM in road safety applications, especially for identifying the contributing factors of the frequency (Kim et al., 2011) and severity (Turner et al., 2019) of motor-vehicle collisions. Other studies utilized SEM to develop a safety risk index in urban areas (Schorr et al., 2014) and evaluate the unsafe behaviour and drivers' aggression (Hassan and Abdel-Aty, 2011). SEM has also gained recent popularity in investigating pedestrian collisions. Al-Mahameed et al., (2019) defined road network characteristics, exposure, and social status as the main influential latent factors on the frequency of collisions that involve pedestrians and cyclists. Sheykhfard et al., (2021) demonstrated that road characteristics were the most important latent factors that impact the frequency of pedestrian collisions. Other studies developed SEM to analyze survey data to evaluate the safety perception and subjective norms of a pedestrian while crossing the streets (Soathong et al., 2021). As can be seen in the literature, the application of SEM in pedestrian

safety studies is still limited. There are almost no studies that applied SEM to assess the contributing factors of collisions that involve pedestrian spatial violations.

4.4 Methodology

In order to achieve the study objectives, historical collision records of the City of Hamilton between 2010 and 201٧ were obtained. The analysis focused on pedestrian-vehicle collisions that involved pedestrian spatial violations at mid-block locations. The Bayesian SEM approach was adopted to evaluate the underlying impact of the explanatory variables on the frequency and the severity of collisions that involve spatial violations. The analysis was carried out in a two-stage procedure. In the first stage, the relationship between the manifest variables and the latent ones was calibrated by developing a classical SEM. In the second stage, a Bayesian SEM was applied to investigate the impact of the latent variables on both the frequency and severity of collisions that involve pedestrian spatial violations. The study considered a wide range of explanatory variables along with two independent variables (the frequency of spatial violations and the severity of collisions that happened due to violations). A brief description of the proposed technique is addressed as follows:

4.4.1 Bayesian SEM

SEM is a multivariate statistical technique that assesses the interrelated dependency among observed variables and unobserved (latent) variables, through the incorporation of regression models, factor analysis, path analysis, and analysis of variance, simultaneously. SEM approaches consist of two main components. The first component, known as the measurement model,

describes the association between the observed variables (independent variables) and the latent factors. The second component, known as the latent model, explains the relationship between endogenous and exogenous latent variables by presenting the direction and effectiveness between the variable (Bollen, 1989). The latent model is developed based on Equation (4-1):

$$A = aX + yY + \tau \qquad (4\text{-}1)$$

where X and Y are vectors of the endogenous and exogenous latent variables, respectively; a and b are the coefficients of the latent variables, and τ is the vector of errors. Also, the measurement models for the exogenous and endogenous variables follow the formulas of Equations (4-2) and (4-3), respectively:

$$L = b_i X + \varepsilon \qquad (4\text{-}2)$$

$$M = a_i Y + \epsilon \qquad (4\text{-}3)$$

where L and M are vectors of the observed exogenous and endogenous variables, b_i and a_i are the coefficient matrices of the latent exogenous and endogenous variable i of the observed variables, and ε and ϵ are the error terms (Bollen, 1989). Figure 4-1 presents a simplified general structure of an SEM model.

Generally, SEM has several benefits compared to the typical statistical models. The SEM method is capable of estimating multiple relationships among variables at the same time. This approach can evaluate the performance of the unobserved/latent variable while predicting the dependent factors based on a series of manifest variables. Moreover, SEM can estimate the error term for each of the observed variables in the measurement part of the model. Finally, the SEM is capable of overcoming the multicollinearity issue among the variables.

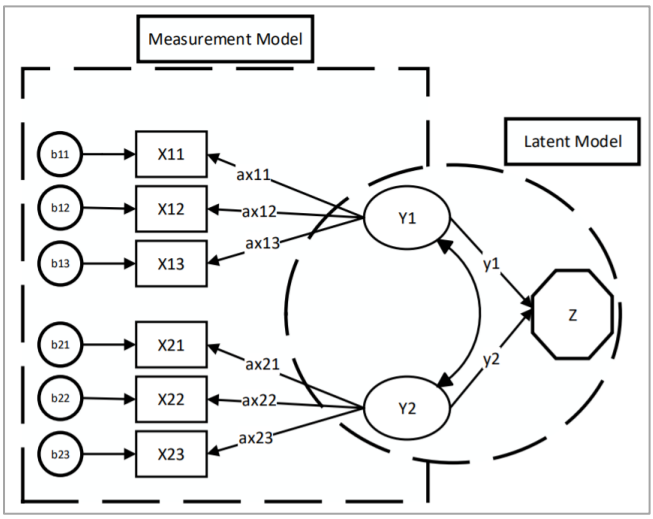


Figure 4-1 SEM structure

The Maximum Likelihood (ML) method is the most common estimation approach for classical SEM techniques. However, the ML estimator is not able to present the best performance while dealing with residual correlation, cross-loadings, and the absence of multivariate normality, which leads to biased factor loadings. To overcome these issues, the third generation of the SEM was integrated with Bayesian models. In Bayesian SEM, the uncertainties are considered in the predictive model and the requirement for the normal distributions is released (Mwangi and Wanjoya, 2016). Bayesian SEM employs Gibbs sampler from a Markov Chain Monte Carlo (MCMC) simulation method to predict the posterior distribution of the latent variables. However, the prior distribution of the variables needs to be determined first. Since there is no sufficient information regarding the prior distribution of the variables, the normal distribution with zero mean and a very large variance (e.g., 1000) is considered as an acceptable prior distribution of the parameters.

In order to ensure the model convergence, three independent Markov chains were run for 10,000 iterations for each parameter. The first 5,000 iterations in each chain were treated as burn-in samples that are not considered in the calculations. The convergence of each parameter was checked using the Proportional scale reduction (PSR), which examines the between- and within-chain variation. PSR value that is close to 1 typically indicates that the model has converged (Asparouhov and Muthen, 2010).

4.5 Data collection and processing

Pedestrian-vehicle collisions that occurred at mid-block locations in the City of Hamilton, Ontario from 2010 to 2017 represent the main source of data in this study. In total, 617 pedestrian-vehicle collisions were reported at mid-block locations (19,728 sections) in the city during the eight years considered in the analysis. The collision dataset provided by the City of Hamilton specifies the exact location of the collision, which was used to determine whether the location occurred outside a close-by marked crosswalk or not. If the collision occurred outside a marked crossway that exists within 30 meters of the collision location, it is defined as a collision that involves pedestrian spatial violation. The collision dataset also provides information regarding the pedestrian action before the collision, including whether or not the pedestrian yielded the right-of-way to vehicles. Collisions that involve pedestrians who did not yield the right-of-way to vehicles were also defined as collisions that involve pedestrian spatial violation. A total of 214 collisions (34.7% of total collisions) were attributed to pedestrian spatial violations at mid-blocks, resulted in 11 fatalities and 192 injuries.

Moreover, in order to determine the potential contributing factors to collisions that involved spatial violations, a thorough review of the literature was first conducted to identify the factors that promote pedestrian spatial violations at mid-blocks. According to the literature, various factors were identified as contributors to the spatial violation behaviour, including road user exposure, road network characteristics (mainly block size and road class), location-specific factors (such as the number of lanes and the presence of central refuge islands), built-environment factors (mainly, bus stops and schools), and land use. The literature also provided little discussion regarding the impact of pedestrian network characteristics (mainly directness and connectivity) on the violation behaviour. Consequently, it was decided to investigate the potential impact of those pedestrian network indicators on violation-related collisions. Afterwards, a list of additional potential contributing factors was identified based on a preliminary analysis of the spatial distribution of the violation-related collisions and the correlation between the location of collisions and those factors. Those additional factors included several location amenities and attractions (namely, bike share stations, playgrounds, parking lots, convenience stores, recreational trails, and restaurants), other location-specific factors (such as illumination, traffic composition), and the distance between collision location and the nearest intersection. Finally, the correlation between the selected factors was investigated to avoid utilizing highly correlated factors in the developed model. Below, a brief description of the selected factors, their calculation details, and the correlation analysis that was conducted to select the final list of factors is provided.

First, the collision dataset provided useful information regarding each collision, including weather conditions at the time of the collision, illumination, road surface condition, type of vehicles involved in the collision, number of lanes, road class, Average Annual Daily Traffic (AADT), and whether the road segment is divided or undivided. This enables the direct extraction of those factors for each collision in the data set.

Additionally, the study utilized seven other data sources to extract the rest of the potential contributing factors, including Esri ArcGIS online website, Open Street Map website, Canadian 2016 census data, Hamilton Street Railway (HSR) transit route dataset, Hamilton School Board dataset, Hamilton Open Data website of the City of Hamilton, and Geospatial Datacenter of McMaster University (Open Street Map, 2022; McMaster University, 2021; Statistics Canada, 2021; Open Data Hamilton, Esri, 2021). ArcMap 10.7.1. was utilized to merge the information of different sources, which enables the development of the required models.

As for the pedestrian network accessibility indicators, two indicators were used, including pedestrian network connectivity at the collision location and pedestrian route directness at the collision location. Road network characteristics at the collision area included block size, road class, and distance between the collision location to the nearest intersection. The class of the road at which the collision occurs was provided in the collision dataset. To estimate the other four parameters, the transportation network of the City of Hamilton was converted to a set of nodes and links, where the links represent the road segments, and the nodes represent the intersections. The geo-coded road network of the City of Hamilton was extracted from the Open Street Map website (2022). The block size was measured as the direct distance between two adjacent

intersections. The length of the road between the location of the collision and the nearest intersection was considered as the distance to the nearest intersection. The ratio of intersections to the summation of intersections and dead-end streets within a radius of 400 m from the collision location was considered as an indicator for pedestrian network connectivity at the collision location. Finally, the sidewalk layer was mapped on the road network to estimate the Pedestrian Route Directness within a radius of 400 m from the collision location. This factor indicates the degree of the sidewalk's orientation and is calculated as Equation (4-4).

$$Directness = \frac{Walking\ distance\ (route\ distance)}{Straight-line\ distance\ (geodetic\ distance)} \qquad (4\text{-}4)$$

Regarding the land-use, parcel-based land-use data of the City of Hamilton was obtained from the Geospatial Datacenter of McMaster University (2021) and merged with collision layer in ArcMap software. Then, the "Intersect" function was utilized to divide a mid-block segment between adjacent parcels if it crossed the boundary of the parcel. In order to extract the dominant land use, a 400-meter buffer was generated from each collision location. The study considered three common categories of land-use: residential, commercial, and institutional/office land-use.

As for the exposure parameters, the AADT was used as a direct exposure measure for traffic. The AADT at each collision location was available in the collision dataset. Unfortunately, a direct measure for pedestrian exposure at each collision location (i.e., pedestrian volume) was not available. In order to overcome this issue, the study utilized the number of walking trips as a surrogate measure of pedestrian exposure. The City of Hamilton was divided into 191 tracts, based on the 2016 Canadian census data (2021) and the dominant mode of transportation in each tract was calculated in each tract. The census data layer was joined to the mid-block layer in

ArcMap software in order to distribute the mid-blocks within the tracts. Then, the “Intersect” function was utilized to divide a mid-block segment between adjacent tracts if it crossed the boundary of the tract. Finally, the total number of walking trips that was overlaid on a 400 m buffer generated around the center of the collision location (i) was counted and considered as the total number of walking trips for collision (i). It should be noted that the 400-meter buffer that was used to determine the abovementioned factors was selected based on a preliminary sensitivity analysis, in which different buffer sizes (50 – 1000 m) were tested and the buffer that was associated with the best performance of the developed SEM was selected.

Regarding the location/collision-specific factors, six variables were considered, namely, the number of lanes, the presence of central refuge islands, illumination at the collision location, the type of vehicle involved in a collision, the road surface conditions, and the weather conditions at the time of the collision. These six factors were provided in the collision dataset and were used directly in the analysis.

Finally, the study considered the impact of a variety of amenities and attractions in the collision area, including, the number of schools and bus stops within the collision location area, and the presence of trails, playgrounds, parks, restaurants, parking lots, bike-share stations, and convenience stores near the collision location. The number of bus stops was extracted from the Hamilton Street Railway (HSR) dataset and geocoded in ArcMap software. Then, a buffer with a pre-defined radius was generated around the center of the mid-block at which the collision occurs to obtain the number of bus stops that exist within the collision area. Previous studies considered various buffer sizes when studying the impact of bus stops on pedestrian safety,

ranging from 5 to 300 meters (Ghomi and Hussein, 2021; Miranda-Moreno et al., 2011). Based on the results of a preliminary sensitivity analysis, a 200-meter buffer showed the highest accuracy of the developed SEM model. Hence, a 200-meter buffer was used to estimate the number of bus stops within each collision location area.

The locations of schools in the City of Hamilton were extracted from the Open Hamilton dashboard (2021). A 400-meter buffer was generated from the center of the mid-block at which the collision occurs to determine the number of schools existing near the collision location. The buffer size was also selected based on a preliminary sensitivity analysis, in which different buffer sizes were tested and the buffer that was associated with the best performance of the developed SEM was selected. The number of the rest of the amenities and attractions (trails, playgrounds, parks, restaurants, parking lots, bike-share stations, and convenience stores) at each collision location was extracted from the Open Data website of the City of Hamilton (2021) and Esri ArcGIS online website (2021), using a buffer of 400-meters from the center of the mid-block at which the collision occurred.

After extracting all the factors, the Spearman correlation matrix was developed to study the potential correlation between them. The correlation between the two variables was considered significant if the correlation coefficient was higher than 0.7. In this regard, a significant correlation was found between 1) weather condition and road surface and 2) the presence of trails and the proportion of parks. Consequently, weather conditions and the number of parks were eliminated from the dataset, leaving 23 factors as potential contributors to the collisions

that involve pedestrian spatial violation at mid-block locations. A descriptive summary of the 23 factors is presented in Table 4-1.

Table 4-1 Descriptive Summary of the Variables

Category	Variable	Mean	Std. Dev.	Min.	Max.
Exposure parameters	LOG (AADT)	3.83	0.58	1.34	4.73
	Log (Walkers)	1.75	0.35	1	2.43
Pedestrian network accessibility indicators	Directness	0.69	0.18	0.12	0.98
	Connectivity	0.31	0.06	0	0.43
Road network characteristics	Block size (meters)	181.49	188.62	9.27	2072.82
	Distance to intersection (meters)	473.75	582.35	10	4874.52
	Road Class	1= arterial (62.9%), 2= local (37.1%)			
Location/collision-specific factors	Number of lanes	3.16	1.17	1	6
	Road surface	1= dry (80.7%), 2= otherwise (19.3%)			
	Refuge islands	1= yes (2.4%), 2= no (97.6%)			
	Type of vehicles	1= light (72.8%), 2= heavy (27.2%)			
	Illumination	1= yes (28%), 2= no (72%)			
Location Amenities and attractions	Number of bus stops	7.34	10.69	0	17
	Number of schools	1.66	1.62	0	7
	bike-share stations	1= yes (45.5%), 2= no (54.5%)			
	Playgrounds	1= yes (51.1%), 2= no (48.9%)			
	Parking lots	1= yes (36.1%), 2= no (63.9%)			
	Trails	1= yes (22.5%), 2= no (77.5%)			
	Restaurant	1= yes (43.1%), 2= no (56.9%)			
	Convenience store	1= yes (47%), 2= no (53%)			
Land-use Factors	Residential land-use	1= yes (42.1%), 2= no (57.9%)			
	Commercial land-use	1= yes (26.9%), 2= no (73.1%)			
	Institutional land-use	1= yes (14.6%), 2= no (85.4%)			

4.6 Results and Discussion

The model development process started with forming a measurement model with six latent variables. The initial model was developed to evaluate the presence of causal effects among

latent variables and assess the multicollinearity issue. The model demonstrated a high χ^2 value of 2305.88 with 171 degrees of freedom, which indicated a poor fit with the data. The model did not converge due to the presence of several negative values in both error variance and covariances. Also, the relationship between observed variables and the corresponding latent variable was not statistically significant in many of the categories.

In order to overcome these issues, the insignificant observed variables (illumination, type of vehicle, and road class) were removed from the model. Also, two latent variables were dropped out and their significant parameters were merged with other potential latent variables. The modified structure of the measurement model included four latent variables: access to services, location vibrancy, pedestrian network quality, and road size. The "access to services" latent variable included eight observed factors (the presence of playground, restaurant, bike-share stations, parking lots, trails, convenience stores, and the number of bus stops and school). Five variables, including AADT, walkers, and commercial/ residential/ institutional land-uses were classified as "location vibrancy" latent variables. "Pedestrian network quality" latent variable included four observed variables (block size, pedestrian network connectivity indicators, pedestrian route directness, and the distance between the collision location and the nearest intersection). The "road size" included road surface condition, number of lanes, and the presence of refuge islands at the collision location. The model connected the four exogenous latent variables with the two endogenous variables, collisions and fatal collisions that involved pedestrian spatial violations.

The proposed model with four latent variables demonstrated significant results for all input variables. The value of χ^2 was 1426.72, with 221 degrees of freedom. The value of the two error indicators, Standardized Root Mean Square Residual (SRMR) and Root Mean Square Error of Approximation (RMSEA) were 0.0407 and 0.0192, which were lower than the acceptable threshold (0.05).

Once the model structure was set, it was imposed in the Bayesian SEM model to estimate the relationship between latent variables and endogenous variables. The RStudio software was utilized to develop the Bayesian SEM method using the "*blavaan*" statistical package. Figure 4-2 shows the graphical structure of the Bayesian SEM model and the group of manifest variables utilized for each of the latent variables.

The values on the arrows show the coefficients of the variable while the values in parentheses indicate the squared correlation coefficient (R^2), which express the percentage of the variance that was explained by that observed variable (at 95% confidence level).

The results of the Bayesian SEM are presented in Table 4-2. The table presents the parameter estimates that show the impact of the observed variables on the four extracted latent variables, along with the influence of the four latent variables on the two endogenous variables (the frequency of collisions and fatal collisions that involve pedestrian spatial violations).

Based on the results, access to services demonstrated the highest impact on pedestrian collisions that involve spatial violations. This latent variable reflects the aggregation of the amenities and attractions at a mid-block location. As the value of this variable increases (i.e., more access to services is available), more pedestrians are attracted to use these facilities, and the probability of

spatial violation increases significantly, which results in increasing the frequency of collisions that involve pedestrian spatial violations. The coefficients of the manifest factors related to access to services were all positive and significant at a 95% confidence level.

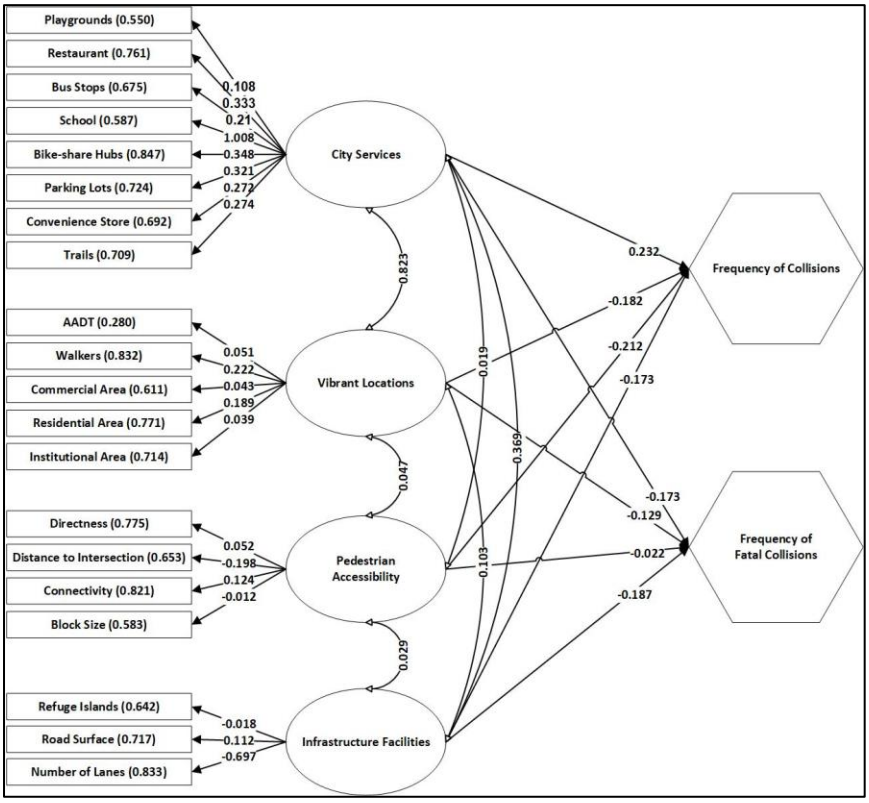


Figure 4-2 Graphical results of the Bayesian SEM model

Furthermore, it should be noted that in the SEM technique, the proportion of variance that is explained by each observed variable is equal to the square of the correlation coefficient (R^2). Based on the last column of Table 4-2, the presence of bike-share stations was found to be the observed variable that explains the highest percent of the variance for the "access to services" latent variable. The City of Hamilton has an efficient bike-sharing system (So Bi bike-share) that

has stations distributed all over the city. Bike-share stations are important attractions for pedestrians as they walk to get to the bike and use it as the main mode to get to the destination or to switch mode (mainly, from or to transit buses) and use the bike in a part of their trips. It should be noted that 81% of the bike-share stations are located within 100 m of a bus stop, which makes bike-share a convenient transportation option that can be integrated with transit. As such, it is expected to have many violations occur at these locations, especially in situations like when a cyclist tries to park the bike and cross the street to catch an approaching bus. Also, parking lots, restaurants, and trail entrances were found to be significant variables within the "access to services" latent variable. Based on these findings, special attention is required when selecting the location of the bike-share stations, especially those that are close to bus stops. As well, preventive measures are needed to reduce the frequency of pedestrian spatial violation near bus stops, bike-share stations, parking lots, restaurants, and trail entrances.

Nevertheless, a negative association was found between the frequency of fatalities and the "access to services" latent variable. One of the reasons that can explain this finding is that locations with amenities like schools, playgrounds, access to trails are usually located in reduced speed zones, and the vehicle operating speeds are typically low. Consequently, the severity of collisions is expected to be lower. Unfortunately, the distribution of the vehicle operating speeds at collision locations is not available to test this hypothesis. Another possible reason for this finding is that drivers usually pay more attention to violating pedestrians at locations with such amenities, which reduces the severity of potential collisions.

Table 4-2 Result of Bayesian SEM method

Latent Variables	Observed Variables	Estimate	Std. Dev.	PSR	R_squared
Coefficients for observed variables					
Access to services	Playgrounds	0.108	0.022	1.001	0.55
	Restaurant	0.333	0.019	1.005	0.761
	Bus Stops	0.21	0.021	1	0.675
	School	1.008	0.065	1.004	0.587
	Bike-share hubs	0.348	0.019	1.001	0.847
	Parking Lots	0.321	0.019	1.007	0.724
	Convenience Store	0.272	0.021	1.001	0.692
	Trails	0.274	0.017	1.001	0.709
Location vibrancy	AADT	0.051	0.027	0.999	0.28
	Walkers	0.222	0.101	1.001	0.832
	Commercial area	0.043	0.028	1	0.611
	Residential area	0.189	0.0163	1.006	0.771
	Institutional area	0.039	0.023	0.999	0.714
Pedestrian network quality	Directness	0.052	0.029	1.026	0.775
	Distance to Intersection	-0.198	0.124	1.03	0.653
	Connectivity	0.124	0.041	1	0.821
	Block Size	-0.012	0.021	1.01	0.583
Road size	Refuge Islands	-0.018	0.009	1	0.642
	Road Surface	0.112	0.015	0.999	0.717
	Lanes	-0.697	0.284	1.004	0.833
Regression weights of the latent variables					
Frequency of collisions that involve spatial violations	Location Amenities and Attraction	0.232	1.196	1.05	
	Exposure	-0.182	1.123	1.04	
	Road/pedestrian Network characteristics	-0.212	0.282	1	
	Location/collision-specific factors	-0.111	0.493	1.057	
Frequency of Fatal collisions that involve spatial violations	Location Amenities and Attraction	-0.173	0.676	1.029	
	Exposure	-0.129	0.636	1.024	
	Road/pedestrian Network characteristics	-0.022	0.136	1.003	
	Location/collision-specific factors	-0.187	0.293	1.05	

The “pedestrian network quality” latent variable was found to be the second most influential factor on the frequency of collisions that involve violation. Since the four variables that constitute this latent variable demonstrate a better level of accessibility and pedestrian convenience in the road network, the rate of pedestrians’ conformity will increase as they experience a more pedestrian-friendly environment (i.e., with the increase of the value of this latent variable). Subsequently, both the frequency and severity of collisions that involve pedestrian spatial violations will decrease, as can be observed from the negative sign of the latent variable coefficients. Pedestrian Network Connectivity was found to be the main factor associated with this latent variable, based on the percentage of the variance explained. Based on the results, locations with poor pedestrian network connectivity and large block size require countermeasures that mitigate pedestrian violation. When planning new developments, block size, the connectivity of the pedestrian network, and ensuring that pedestrians can access their desired destination in the shortest possible distance are essential measures to mitigate violation and related collisions.

Moreover, the “location vibrancy” latent variable was found to have a significant but negative impact on both the frequency of total collisions and fatal collisions that involve pedestrian spatial violations at mid-block (although the impact of exposure on the severity of collisions was not statistically significant) locations. This finding was expected since higher exposure to traffic (higher AADT) discourages pedestrians from spatial violation, as was reported in many previous studies. Also, higher pedestrian exposure and land-uses that attract more pedestrians increase the awareness of the drivers of pedestrians, which reduces the risk of collisions.

Finally, the “road size” latent variable showed an indirect significant impact on both the frequency and the severity of collisions attributed to violation. Based on the definitions of the observed variables of this latent variable, the results indicate that the presence of refuge islands, dry surface conditions, and the lower number of lanes at a location will increase the frequency of both total and fatal collisions that involve pedestrian spatial violations. Previous research showed that the presence of refuge islands increases the probability of spatial violation (Cao et al., 2016; Pour-Rouholamin and Zhou, 2016). Consequently, the presence of refuge islands will increase the frequency of collisions due to violations. Similarly, previous research showed that pedestrians are discouraged from violation in adverse weather conditions and as the number of lanes increases (Ghomi and Hussein, 2021), which explains the higher frequency of violation-related collisions in dry weather conditions and at roads with a lower number of lanes.

4.7 Conclusion

In this study, a Bayesian SEM model was developed to analyze pedestrian collisions that are attributed to spatial violations at mid-blocks. Pedestrian-vehicle collisions that occurred in the City of Hamilton, Ontario from 2010 to 2017 were the main source of data for this study. The SEM model aimed at investigating the interrelationship between a variety of factors, categorized in four latent variable groups, and two endogenous dependent variables (the frequency and the severity of collisions that involve spatial violations). The four latent variable groups included: access to services (e.g., parking lots, schools, bus stops, trails, restaurants, among others), location vibrancy, road size (e.g., number of lanes and the presence of refuge islands), and pedestrian network quality, such as pedestrian network connectivity and block size.

The results showed a significant impact of the access to services on the frequency of violation-related collisions, particularly, bike-share stations, trail access points, restaurants, and parking lots. More collisions were observed at locations with bike-share stations that are located near bus stops, which highlights the significance of the proper selection of bike-share stations and applying appropriate countermeasures at such locations to mitigate pedestrian spatial violation. Lack of pedestrian network connectivity and large block size were found to be highly correlated with the frequency and the severity of pedestrian collisions that involved spatial violations at mid-blocks. Accordingly, locations with poor pedestrian network connectivity and large block size require countermeasures that reduce the frequency of spatial violation. Additionally, block size, the connectivity of the pedestrian network, and ensuring that pedestrians can access their desired destination in the shortest possible distance are essential measures to consider when planning new areas. Finally, violation-related collisions were found to be more likely to happen at locations that have refuge islands and a low number of lanes.

Nevertheless, the study is subject to several limitations that should be addressed in future studies.

- The study utilized an estimate of the number of walkers at collision locations as a surrogate measure of pedestrian exposure. While this is a commonly used surrogate measure for pedestrian exposure in the safety literature, more precise measures for pedestrian exposure can be explored, including collecting extra survey data or applying activity-based models to estimate the pedestrian volume at a location accurately.
- The estimated number of walkers at collision locations in this study is based on Canadian census data that are only available at the tract level. Although this method has some benefits,

it suffers from a major drawback. Specifically, many mid-blocks that are located in the same tract will be assigned similar numbers of walking trips regardless of other road characteristics. Specifically, the coarse-grained pedestrian volume estimates used in this study may introduce error into the parameter estimates for roadway-level variables. Therefore, there is a need to consider pedestrian volume at a fine-grained, street-by-street level in future studies.

- It is essential for future studies to include the vehicle operating speed distribution in the analysis as it can provide an explanation for the impact of many factors on collision severity and enhance the accuracy of the results.
- The current study utilized one dataset from one city. Future studies should analyze more datasets from different cities to investigate the impact of culture and behavioural differences on the results. It should be noted that the "spatial violations" referred to in this study may occur within a different legal context for pedestrian crossings than in other studies in the literature. However, investigating the nuanced differences in the legality of mid-block crossing in the various jurisdictions examined in previous studies is beyond the scope of this study.
- Despite the well-established safety benefits of refuge islands (Oakes et al., 2007; Aidoo et al., 2013; Ulfarsson et al., 2010), the results of the model suggest that the presence of refuge islands may be associated with an increase in the frequency of collisions that involve violations. This result may be an artifact of the coarse measure used to represent pedestrian exposure, but it could potentially be due to the increase in the frequency of spatial violations at locations with refuge islands. Accordingly, it may be valuable to apply some mitigation

measures at locations with refuge islands that aim at reducing the frequency of spatial violations in order to avoid having a high frequency of these violation-related collisions.

- Finally, road networks characteristics seem to have a significant impact on spatial violation behaviour and consequent safety issues. Future studies should conduct a more detailed analysis of the pedestrian network indicators and evaluate the impact of micro-scale characteristics related to the built environment factors on the results.

4.8 Reference

Abdullah, M., Dias, C., & Oguchi, T. (2021). Road Crossing at Unmarked Mid-Block Locations: Exploring Pedestrians' Perception and Behavior. Shiraz,Iran: Springer.

Aidoo, E. N., Amoh-Gyimah, R., & Ackaah, W. (2013). The effect of road and environmental characteristics on pedestrian hit-and-run accidents in Ghana. Accident Analysis and Prevention, 53, 23-27.

Al-Mahameed, F. J., Qin, X., Schneider, R. J., & Shaon, M. R. (2019). Analyzing Pedestrian and Bicyclist Crashes at the Corridor Level: Structural Equation Modeling Approach. Transportation Research Record, 2673(7), 308-318.

Asparouhov, T., & Muthen, B. (2010). Bayesian Analysis Using Mplus: Technical Implementation.

OpenStreetMap. (2021, June). Retrieved from https://www.openstreetmap.org/#map=10/43.2607/-79.9352

Bollen, K. A. (1989). Structural Equations with Latent Variables. New York: Wiley.

Cao, Y., Ni, Y., & Li, K. (2016). Effects of Refuge Island Settings on Pedestrian Safety Perception and Signal Violation at Signalized Intersections. 96th Annual meeting of Transportation Research Board.

Esri . (July, 2021). Retrieved from ArcGIS Online: https://www.esri.com/en-us/arcgis/products/arcgis-online/overview

Fillone, A. M., Montalbo, C. M., & Tiglao, N. C. (2005). Assessing Urban Travel: a Structural Equations Modeling (SEM) Approach. Eastern Asia Society for Transportation Studies, 5, 1050-1064.

Ghomi, H., & Hussein, M. (2021). An integrated clustering and copula-based model to assess the impact of intersection characteristics on violation-related collisions. Accident Analysis and Prevention, 159, 106283.

Hassan, H. M., & Abdel-Aty, M. A. (2011). Analysis of drivers' behavior under reduced visibility conditions using a structural equation modeling approach. Transportation Research Part F: Traffic Psychology and Behaviour, 14(6), 614-625.

Holland, C., & Hill, R. (2010). Gender differences in factors predicting unsafe crossing decisions in adult pedestrians across the lifespan: a simulation study. Accident Analysis and Prevention, 42(4), 1097-1106.

Hussein, M., Sayed, T., Reyad, P., & Kim, L. (2015). Automated pedestrian safety analysis at a signalized intersection in New York City: Automated data extraction for safety diagnosis and behavioral study. Transportation Research Record: Journal of the Transportation Research Board, 2519(1), 17-27.

Joreskog, K. G. (1973). Analysis of covariance structures. In Multivariate Analysis-third edition (pp. 263-285). New York: Academic Press.

Kim, M., Kho,, S. Y., & Ki, D. K. (2017). Hierarchical Ordered Model for Injury Severity of Pedestrian Crashes in South Korea. Journal of Safety Research, 61, 33–40.

Kadali, B. R., & Vedagiri, P. (2020). Role of number of traffic lanes on pedestrian gap acceptance and risk taking behaviour at uncontrolled crosswalk locations. Journal of Transport & Health, 19, 100950.

Kim, K., Pant, P., & Yamashita, E. (2011). Measuring Influence of Accessibility on Accident Severity with Structural Equation Modeling. Transportation Research Record: Journal of the Transportation Research Board, 2236, 1-10.

Lee, J., Chung,, J., & Son, B. (2008). Analysis of Traffic Accident Size for Korean Highway Using Structural Equation Models. Accident Analysis and Prevention, 40, 1955–1963.

Library, G. D. (2021, April). McMaster University. Retrieved from https://library.mcmaster.ca/collections/geospatial-data

Miranda-Moreno, L. F., Morency, P., & El-Geneidy, A. M. (2011). The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections. Accident Analysis and Prevention, 43, 1624–1634.

Mwangi, M. E., & Wanjoya, A. (2016). Bayesian Structural Equation Modeling: A Business Culture Application in Kenya. Science Journal of Applied Mathematics and Statistics, 4(2), 37-42.

Oakes , J. M., Forsyth, A., & Schmitz, K. H. (2007). The effects of neighborhood density and street connectivity on walking behavior: the Twin Cities walking study. Epidemiologic Perspectives and Innovations, 4(16).

Open Hamilton. (2021, April). Retrieved from https://open.hamilton.ca/

Papadimitriou, E., Lassarre, S., & Yannis, G. (2017). Human factors of pedestrian walking and crossing behaviour. Transportation Research Procedia, 25, 2002-2015.

Papić, Z., Jović, A., Simeunović, M., Saulić, N., & Lazarević, M. (2020). Underestimation tendencies of vehicle speed by pedestrians when crossing unmarked roadway. Accident Analysis and Prevention, 143, 105586.

Pour-Rouholamin, M., & Zhou, H. (2016). Investigating the risk factors associated with pedestrian injury severity in Illinois. Journal of Safety Research, 57, 9–17.

Rankavat, S., & Tiwari, G. (2020). Infuence of actual and perceived risks in selecting crossing facilities by pedestrians. Travel Behaviour and Society, 21, 1-9.

Sheykhfard, A., Haghighi, F., Nordfjærn, T., & Soltaninejad, M. (2021). Structural equation modelling of potential risk factors for pedestrian accidents in rural and urban roads. International Journal of Injury Control and Safety Promotion, 28(1), 46-57.

Schorr, J. P., & Hamdar, S. H. (2014). Safety Propensity Index for Signalized and Unsignalized Intersections: Exploration and Assessment. Accident Analysis and Prevention, 71, 93-105.

Soathong, A., Chowdhury, S., Wilson, D., & Ranjitkar, P. (2021). Investigating the motivation for pedestrians' risky crossing behaviour at urban mid-block road sections. Travel Behaviour and Society, 22, 155-165.

Statistics Canada. (2021, April). Retrieved from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm

TRAFFIC ACT, R.S.O. Chapter H.8. (1990), https://www.ontario.ca/laws/statute/90h08#BK246.

The City of Toronto. (2022, February). Retrieved from https://www.toronto.ca/311/knowledgebase/kb/docs/articles/transportation-services/district-transportation-services/traffic-operations/rules-for-crossing-the-street-jaywalking-pedestrian-traffic-signals.html

Transport Canada. (2021, April). Retrieved from https://tc.canada.ca/en/canadian-motor-vehicle-traffic-collision-statistics-2018

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

Turner, S., Smith, M., & Tse, I. (2019). Understanding Vulnerable Road User Crash Risk (On Auckland's High-Risk Routes). Transportation Group 2019 Conference, (pp. 3-6). Te Papa, New Zealand.

Ulfarsson, G., Kim, S., & Booth, K. (2010). Analyzing fault in pedestrian–motor vehicle crashes in North Carolina. Accident Analysis and Prevention, 42, 1805–1813.

Useche, S. A., Hezaveh, A. M., Llamazares, F. J., & Cherry, C. 2. (2021). Not gendered but diferent from each other? A structural equation model for explaining risky road behaviors of female and male pedestrian. Accident Analysis and Prevention, 150, 105942.

Yoneda, K., Suganuma, N., Yanase, R., & Aldibaja, M. (2019). Automated driving recognition technologies for adverse weather conditions. IATSS Research, 43, 253-262.

Zaki, M. H., Sayed, T., Tageldin, A., & Hussein, M. (2013). Application of Computer Vision to Diagnosis of Pedestrian Safety Issues. Transportation Research Record: Journal of the Transportation Research Board, No. 2393.

Zhang, G., Tan, Y., & Jou, R. C. (2016). Factors influencing traffic signal violations by car drivers, cyclists, and pedestrians: A case study from Guangdong, China. Transportation Research Part F: Traffic Psychology and Behaviour, 42(1), 205-216.

CHAPTER 5

Investigating the Application of Deep Learning to Identify Pedestrian Collision-prone Zones

The publication included in this chapter is:

Ghomi, H., & Hussein, M. (2023). Investigating the Application of Deep Learning to Identify Pedestrian Collision-prone Zones. Journal of Transportation Safety & Security. https://doi.org/10.1080/19439962.2022.2164636

The manuscript was submitted in January 2022 and accepted in December 2022. Haniyeh Ghomi is the main contributor of this manuscript. The co-author's contributions include guidance, supervision, funding, reviewing the analysis, and editing the manuscript.

5.1 Abstract

The main objective of this study is to understand the factors that contribute to the frequency of both the total pedestrian-vehicle collisions and collisions that involve pedestrian violations and identify collision-prone areas. The two Full Bayes (FB) macro-level models were applied to historical collision records of the City of Hamilton to identify the collision-prone zones and the key factors that contribute to collision occurrence in TAZs. Finally, a self-organizing map (SOM) deep learning model was developed to identify collision-prone zones for the two collision classes. The results showed that the SOM model identified collision-prone zones with a high accuracy that exceeded the traditional Bayesian approach, based on the developed consistency test. As for the total collisions, the SOM model revealed that intersection density is the most important factor in distinguishing between collision-prone and non-collision-prone zones, followed by the pedestrian network directness and the proportion of residential land uses. As for the collisions that involved pedestrian violations, intersection density was also found to be the most important factor, followed by the density of bike-share stations and parking lots in a TAZ. The results of this study could aid planners in designing pedestrian-friendly networks and develop specific recommendations to enhance safety in unsafe zones.

5.2 Introduction

The importance of investigating pedestrian collisions and developing engineering solutions to enhance pedestrian safety cannot be overstated. Historical collision data clearly show that pedestrians are overrepresented in collision fatalities and serious injuries. According to Transport Canada, pedestrians represented 17.8% of total fatalities in 2019 in Canada, while they

comprised only 3.3% of persons involved in collisions (Transport Canada, 2021). Extensive research can be found in the literature that studied pedestrian-vehicle collisions and investigated the contributing factors to both the frequency and the severity of collisions. The majority of the previous studies were undertaken at the micro-level to assess pedestrian safety issues at specific locations (e.g., Mohamed et al., 2013; Pour-Rouholamin and Zhou, 2016; Ding et al., 2018). Less interest was given to analyzing pedestrian safety on the macro-level (e.g., city-level), even though macro-level pedestrian safety analyses can be very effective in identifying pedestrian safety issues in larger areas, understanding the characteristics of unsafe areas, and establishing long-term safety improvement policies. Furthermore, previous macro-level pedestrian safety studies considered the total pedestrian-vehicle collisions as the main source of data to conduct the analysis. While the total pedestrian-vehicle collisions are useful in understanding pedestrian safety trends, they do not provide a comprehensive conceptualization of pedestrian safety issues. For example, several pedestrian behaviours have been identified in the literature as risky behaviours that may increase the risk of collisions, including pedestrian violations, distractions, and impaired walking. Among those behaviours, pedestrian violations were identified as a key hazardous behaviour that increases both the probability and the severity of collisions (Kim et al., 2017; Mukherjee and Mitra, 2020; Wang et al., 2019; Ghomi and Hussein, 2021). Yet, macro-level studies rarely analyze collisions that involve such risky behaviours. Thus, it is extremely beneficial for macro-level pedestrian safety studies to investigate collisions that involve pedestrian unsafe behaviours in addition to the total collisions in order to define the

characteristics of areas that promote such risky behaviours and contribute to the related collisions.

Moreover, the literature review shows that the Bayesian techniques, either Empirical Bayes (EB) or Full Bayes (FB) models, are the most prevalent approaches to model collision occurrence at the macro-level and identify collision-prone areas. Although Bayesian models are powerful statistical tools that have been successfully used in many transportation safety applications, they usually come up with a prohibitive computational cost, especially when dealing with a high-dimensional dataset (Smith, 1991). Specifically, Full Bayes models require defining a prior distribution of the model parameters that are updated to develop a posterior distribution of those parameters. While the posterior distributions of the parameters are heavily affected by the selected prior distributions, there is no absolute way to select the most appropriate prior distributions for the model parameters (Wang, 2004). In addition, considering a probability distribution function for model parameters is not always the most adequate approach to account for the uncertainty of the parameters, as discussed in detail in Walters and Ludwig (1994).

One of the techniques that have promising potential to identify hotspot areas and identify contributing factors to collisions at the macro-level is Deep Learning. Although the Deep Learning technique has attracted tremendous attention in several fields, it is still underutilized in pedestrian safety studies. The Deep Learning technique, as a non-parametric method, has superiority in handling big data and complex non-linear relationships among variables compared to the statistical and Machine Learning models (Ma et al., 2015 a, b). Meanwhile, unsupervised algorithms of the Deep Learning technique could directly distinguish between collision-prone

and non-collision-prone areas instead of the two-step process in the Bayesian methods (i.e., developing a collision prediction model then identifying the hotspot areas). A few studies employed the common methods of Deep Learning to predict vehicular collision risk (Bao et al., 2019; Cai et al., 2019; Hollander et al., 2021). However, this concept has not been widely used in macro-level pedestrian safety studies.

The main objective of this study is to conduct a macro-level pedestrian safety analysis using both the total pedestrian-vehicle collisions and pedestrian-vehicle collisions that involve pedestrian violations. The goal is to understand the factors that contribute to the frequency of the two classes of collisions, identify collision-prone zones that experience a high frequency of collisions, and understand their characteristics. In addition, the study evaluates the applicability of the Deep Learning technique in identifying collision-prone areas and understanding the key variables that distinguish collision-prone areas from non-collision-prone ones. Such analysis would provide a better understanding of pedestrian safety on the macro-level and aid engineers and planners in developing specific planning recommendations to enhance safety in unsafe areas and areas that have serious problems with pedestrian violations.

In order to achieve the study objectives, a macro-level pedestrian safety analysis was conducted in the City of Hamilton, Ontario. The City of Hamilton was divided into 236 Traffic Analysis Zones (TAZs). Pedestrian-vehicle collisions that occurred in the city between 2010 and 2017 were obtained from the City of Hamilton and aggregated to the TAZs level. The collision database included information regarding the action of pedestrians who were involved in collisions, which enabled the extraction of collisions that involved pedestrian violations. Several

factors that are expected to impact pedestrian safety on the macro-level were extracted for each zone, including traffic-related factors, pedestrian network connectivity indicators, pedestrian route directness indicators, built-environment factors, attraction and amenities in each zone, land use, and socio-economic factors. Two separated datasets were created to identify the collision-prone zones based on both the total collisions and the collisions that involved pedestrian violations. First, the FB approach was applied to develop two macro-level collision prediction models for the two databases. Collision-prone zones were then identified, and the key factors that contribute to collision occurrence in those zones were extracted. Afterwards, the Deep Learning technique was utilized to automatically identify hotspots for the two datasets and identify the key variables that distinguish between collision-prone and non-collision-prone areas. Standard Deep Learning algorithms (such as convolutional neural networks) are not capable of capturing uncertainty in the model, which is a necessary step for the Deep Learning practitioner (Krzywinski and Altman, 2013). Therefore, a Self-Organizing Map (SOM) technique, which is one of the unsupervised advanced versions of Artificial Neural Networks (ANN), was utilized.

The paper provided several contributions to literature. First, the study analyzed pedestrian safety on the macro-level based on the collisions that involved pedestrian violations, which helped to develop a better understanding of the factors that contribute to such collisions. The study considered the impact of a wide range of factors that were not considered in previous studies on the safety of pedestrians on the macro level, including built-environment factors, socio-economic indicators, and attractions and amenity-related features in TAZs. The paper also applied the SOM Deep Learning model to identify the collision-prone zones based on total and violation-

related pedestrian collisions and highlighted the potential of such a technique in pedestrian safety applications. Finally, the study identifies and ranks the variables that can be used to distinguish between collision-prone and non-collision-prone zones on a city level.

5.3 Literature Review

5.3.1 Identification of collision-prone locations

The first attempts to identify the collision-prone locations for pedestrians relied mainly on the frequency of the collisions and the rate of the collision occurrence (e.g., Deacon et al., 1997; Barker and Baguley, 2001). To overcome the issue of the Regression-to-the-Mean (RTM) that is associated with the use of the collision frequency and rate, the EB and FB models were utilized to identify the collision-prone locations, either on the micro-level in the majority of studies or on the macro-level in a few studies. On the micro-level, Sacchi et al., (2015) proposed a multivariate FB approach to identify the collision-prone locations for pedestrians in the City of Vancouver. The study analyzed pedestrian-vehicle collisions at 137 signalized intersections in the City of Vancouver to identify the collision-prone intersections. El-Basyouny and Sayed (2009) analyzed vehicular collisions at 99 signalized intersections in the City of Edmonton, Alberta, using the Multivariate Poisson-lognormal (MVPLN) model to identify the hotspot locations. In another study, El-Basyouny and Sayed (2013) applied the same technique to identify collision-prone locations among 236 signalized intersections in the Greater Vancouver Area.

On the macro-level, Osama et al., (2018) utilized the FB technique to identify the collision-prone zones for pedestrians and cyclists in the City of Vancouver according to the historical collision

records. The study identified walk trips and Bicycle Kilometers Travelled (BKT) as the key contributing factors for active road user collisions. Lee et al., (2017) developed a mixed-effects negative binomial model to investigate the hotspot locations for both pedestrians and bicyclists in Florida. The study utilized three-year historical collision records (2010-2012) that occurred at 8350 major intersections. According to the results, several factors were identified as the key contributing factors to pedestrian and bike collisions, including population density, age, the proportion of trips made by public transit, and walking trips.

In summary, macro-level analysis of pedestrian safety is limited in the literature. Previous studies relied mainly on Bayesian techniques to develop macro-level collision prediction models that are used in the identification and ranking of collision-prone locations. The application of techniques like Machine learning and Deep Learning in macro-level pedestrian safety studies has not been effectively investigated in the literature. In addition, the impact of many factors on the macro-level safety of pedestrians is still understudied, including network connectivity features (e.g., network coverage, complexity), built-environment factors (e.g., signal density and congestion of bus stops), and the location amenities, such as parks, convenience stores, and parking lots.

5.3.2 Applications of Deep Learning in safety studies

The application of the Deep Learning technique in the transportation field is most popular in Transportation Planning to predict speed (e.g., Ma et al., 2017; Peng and Xu, 2021), traffic flow (such as Wu et al., 2018; Lv et al., 2015), travel time (e.g., Hou and Edara, 2018; Bhandari and Parc, 2022); and travel mode inference (e.g., Dabiri and Heaslip, 2018). Several studies showed

the merits of Deep Learning techniques in vehicular safety applications. For example, Bao et al., (2019) utilized the Deep learning concept to estimate the collision risk in Manhattan, New York City, in 2015. The study implemented a spatiotemporal convolutional model on three different sizes of citywide grids. According to the study, the proposed model was more accurate than econometric and Machine Learning models. Cai et al., (2019) applied a Convolutional Neural Network (CNN) to investigate the relationship between high-resolution collision records and collision prediction in Florida. The proposed methodology increased the accuracy of the collision prediction compared to the results obtained by two statistical models (spatial Poisson-lognormal model and negative binomial model) and a conventional ANN. Hollander et al., (2021) applied the CNN technique to investigate the relationship between road safety and a variety of factors, including built-environment, land use, street characteristics, and specific transportation policies in two Canadian cities, namely, Toronto and Montreal.

As can be seen in the literature, the Deep Learning technique is a promising tool that can lead to a better understanding of road safety issues and incorporate multiple features to enhance the accuracy of collision prediction. Nevertheless, there are almost no applications of the Deep learning technique in the literature related to the macro-level assessment of pedestrian safety and the identification of collision-prone zones for pedestrians.

5.4 Methodology

The study considered two approaches to identify the collision-prone zones, namely, the Full Bayesian approach and the self-organizing map (SOM) unsupervised Deep Learning technique. The two techniques were used to identify collision-prone locations based on the total pedestrian-

vehicle collisions in TAZs and the collisions that involved pedestrian violations. The following sections provide a brief description of the FB approach, the identification of collision-prone zones, and the SOM model.

5.4.1 Full Bayes macro-level collision prediction models

Bayesian methods are widely considered an effective statistical approach to develop collision prediction models as they take into consideration the stochastic nature of the collision data. Bayesian models treat the predicted collision frequency as a random variable that is associated with a specific probability distribution. The probability distribution of the predicted collision frequency is typically obtained in two stages. In the first stage, the prior distribution of the collision frequency is determined. In the second stage, the Bayes theorem is applied to update the prior distribution into a posterior distribution based on the observed collision (Heckerman 1999). Recently, the FB approach has gained popularity as a powerful methodology for a wide range of safety applications, including the development of collision prediction models. In this approach, a Poisson-Lognormal model is developed, in which collision frequency at the zonal level is modeled as a dependent variable. Consider Y_i as the number of collisions at zone i, it is assumed that Y_i follows the Poisson distribution with an expected rate (λ_i), according to Equation (5-1):

$$Y_i \mid \lambda_i \sim Poisson\ (\lambda_i) \qquad (5\text{-}1)$$

where λ_i itself is presumed to be a random variable that can be expressed according to Equation (5-2) as follows:

$$Ln(\lambda_i) = \alpha_i + \alpha_1 Ln(VKT_i) + \alpha_2 Ln(PKT_i) + \sum_i \beta_i X_i + u_i + \varepsilon_{ij} \quad (5\text{-}2)$$

where:

- α_i is the model intercept at zone i
- VKT_i is the average vehicle kilometer travelled at zone i
- PKT_i is the average pedestrian kilometer travelled at zone i
- X_i is a vector of covariates (independent variables)
- β_i is the coefficient for covariate X_i
- u_i is the heterogeneity parameter
- ε_{ij} is the random effect parameter modeled with a lognormal distribution ($\varepsilon_{ij} \sim Normal\ (0, \delta_\varepsilon^2)$) as the prior distribution

As mentioned, the total collisions and violation-involved collisions were integrated at the zonal level. Each zone has various unobserved factors (e.g., geometric and environmental factors) that can impact the occurrence of pedestrian violations and collisions. To account for this issue, the literature suggests introducing an additional variance component in the model by allowing the model intercept α to vary between different zones instead of using a fixed intercept for all zones (El-Basyouny and Sayed, 2012; Hussein et al., 2020). The random intercept (α_i) is developed according to Equation (5-3) as follows:

$$\alpha_i = \alpha_0 + \rho_i \quad , \quad \rho_i \sim Normal\ (0, \delta_i^2) \qquad (5\text{-}3)$$

Where δ_i^2 is the additional variance component in the model that accounts for the intra-zonal variation.

The Stata-MP16 statistical software was utilized to determine the parameter coefficients of the model. To start, the prior distribution of the parameters should be determined first. Based on the literature, it is common to use the diffused normal distributions (with zero mean and large

variance) as a prior distribution for the regression parameters to account for the lack of information about the prior distribution. In this study, the whole set of parameters (i.e., α_i and β_i) was assumed as non-informative parameters with a prior distribution that follows the Normal Distribution with a zero mean and large variance, i.e., Normal (0, 10^3). The posterior distribution of the parameters was obtained using the Markov Chain Monte Carlo (MCMC) simulation technique.

To ensure the model convergence, two independent Markov chains were run for 40,000 iterations for each parameter. The first 5,000 iterations in each chain are treated as burn-in samples that are not considered in the calculations. The convergence of each parameter was checked using the ratio of the Monte Carlo Standard Error (MCSE) to the standard deviation, in which values lower than 5% indicate convergence. Gibbs sampling algorithm was utilized to develop the Bayesian model, which is an appropriate sampling technique dealing with probabilistic models with discrete dependent variables so that such conditional probability can be estimated. Moreover, in order to assess the goodness of fit of the developed models, the Deviance Information Criterion (DIC) and Watanabe–Akaike Information Criterion (WAIC) was used (Spiegelhalter et al., 2002; Watanabe, 2010). Both DIC and WAIC are considered a generalization of the Akaike Information Criterion (AIC) and are typically used to assess the relative quality of statistical models for a given set of data. Generally, a model with a lower value of DIC/WAIC represents a better fit.

5.4.2 Identification of collision-prone (hotspot) zones

According to the literature, several approaches can be used to identify the collision-prone zones in the FB context, including Posterior Poisson Mean (PM), Potential for Safety Improvement (PSI), Median Rank of Posterior Distribution of Poisson Mean, and Observed Crash Counts. The current study identified the collision-prone zones for both total collisions and violation-related collisions using the PSI approach, following the methodology discussed in (Lan and Persaud, 2011). Each zone's PSI value is determined by subtracting the expected collision frequency (γ_i) from the long-term mean of collision counts for each zone (λ_i), as shown in Equation (5-4).

$$PSI_i = \gamma_i - \lambda_i \qquad (5\text{-}4)$$

The expected collision frequency (γ_i) is calculated as Equation (5-5):

$$\gamma_i = \frac{\lambda_i(K+Y_i)}{K+\lambda_i} \qquad (5\text{-}5)$$

Where K is the overdispersion parameter.

5.4.3 Self-Organizing Map (SOM)

The SOM method is one of unsupervised Deep Learning techniques that have some advantages over the common algorithms, such as ANN. The main concept of the SOM model is to break the high-dimensional space of the input into several regular low-dimensional subsets (usually a two-dimension subset) through the implementation of a non-linear procedure combined with visualized clustering technique. A typical SOM model consists of two layers, named input layer and the output layer. The input layer is linked to each variable in the dataset. While the output layer builds a two-dimensional array of neurons. The main role of the output layer is to

demonstrate the distribution of the variables as units of the grid. Figure 5-1 shows the scheme of an n-dimensional SOM model.

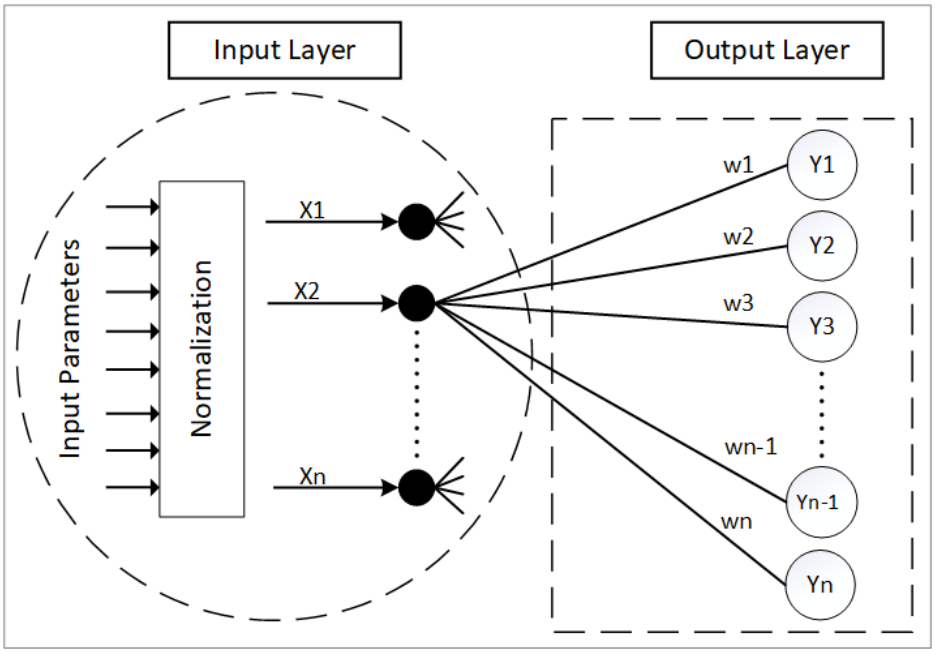


Figure 5-1 A scheme of n-dimensional SOM model

The procedure of dimension reduction and clustering is achieved in three consecutive steps, namely, normalization, training, and information extraction. During the first step, the entire variables will be normalized in order to handle the variation of the variables' scale. In the second step, the input vectors will be clustered into two classes based on the distance between the weight vector that is assigned to the variables and the input vector. Finally, the extracted map layer will be post-processed in order to extract the clusters and visualize outputs.

Similar to most ANN methods, the SOM operates in two stages, namely training and mapping. In the first stage, the model will use the training subset of the dataset as an input space in order to

generate a two-dimensional representation of the data (named as training space). In the second stage, the remaining dataset will be added to the generated map to validate the results (named as map space).

The map space consists of components called nodes or neurons, which are arranged as a hexagonal or rectangular grid with two dimensions. Each node in the map space is associated with a weight vector, which is the position of the node in the input space. While nodes in the map space stay fixed, training consists in moving weight vectors toward the input data (reducing the Euclidean distance) without spoiling the topology induced from the map space. After training, the map can be used to classify additional observations for the input space by finding the node with the closest weight vector (smallest distance metric) to the input space vector. Equation (5-6) shows an updated neuron v with a weight vector $W_v(r)$:

$$W_v(r+1) = W_v(r) + \emptyset(u, v, r).\partial(r).(A(n) - W_v(r))) \qquad (5\text{-}6)$$

where r shows the steps, n is an index to demonstrate the training sample, u is the index of the best matching unit for the input vector $A(n)$, $\partial(r)$ is the learning coefficient that decreases monotonically, and $\emptyset(u, v, r)$ is the neighborhood function that gives the distance between the neuron u and the neuron v in step r (Kohonen, 2013).

In this study, a 10-fold cross-validation technique was developed to evaluate the performance of the model. After exhausting the ten iterations, the average results of the model in the ten iterations are considered as the model output. To evaluate the performance of the SOM model, three main statistical criteria are used; namely, Correctly Classified Rate (CCR), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). According to the literature, CCR is

typically used as an index that measures the overall accuracy of a model. The MAE calculates the average magnitude of the errors in a set of predictions, without considering their direction. Finally, RMSE is used as it is a scale-dependent index that is recognized as an appropriate measure of accuracy for numerical predictions.

Once the SOM model is implemented and assessed, it provides useful information regarding the factors that contribute to distinguishing between the two classes (i.e., collision-prone zones and non-collision-prone zones), as will be shown in the result section.

Finally, two methods were applied to check the consistency between the collision-prone zones identified by the proposed SOM model and the PSI method. First, a within-method consistency test was developed to investigate the accuracy of the "truly" identified hotspots and safe zones in each method. According to (Sacchi et al., 2015), the mean of the collision records over five years can be considered as the True Poisson Mean (TPM) for each zone. The top 10% of the zones with the highest TPM are assumed as truly identified collision-prone zones (Sacchi et al., 2015). To that end, the False Identification is calculated according to Equation (5-7) as follows:

$$\text{False Identification} = \frac{\text{False Positive TAZs+False Negative TAZs}}{\text{n}} \qquad (5\text{-}7)$$

Where n is the total number of zones (236 in this study). Second, a between-methods consistency test was applied to investigate the similarity of the hotspots extracted from the two methods (SOM vs. PSI). Such test for the top 10% of the hotspots is developed based on Equation (5-8) as follows:

$$Test = \{H_1, H_2, \dots, H_n\}_{i,SOM} \cap \{H_1, H_2, \dots, H_n\}_{i,PSI} \qquad (5\text{-}8)$$

Where n represents the total number of zones located in that dataset and H_i is the i^{th} ranked zone recognized as a collision-prone zone.

5.5 Data

The City of Hamilton was divided into 236 TAZs. The frequency of pedestrian-vehicle collisions that occurred in each TAZ between 2010 and 2017 was considered the main source of data in this study. In total, 2089 pedestrian-vehicle collisions were reported resulting in 45 fatal collisions (2.15%), and 1859 injury collisions (88.99%). A total of 509 collisions (24.4% of total collisions) involved at least one type of pedestrian violation were identified. Of the 509 collisions that involved pedestrian violations, 17 were fatal collisions (3.34%), and 451 were injury collisions (88.61%).

Six other data sources were utilized to extract a list of variables that are expected to contribute to collision occurrence in each TAZ. The variables extracted from these data sources can be categorized into seven broad categories: road user's exposure variables, pedestrian network connectivity indicators, pedestrian route directness indicators, built-environment factors, land use variables, attraction and amenities in TAZ, and socio-economic variables. ArcMap 10.7.1. was utilized to integrate the information of different sources, which enables to conduct of the required analysis.

5.5.1 List of contributing factors

Two exposure parameters, namely, Vehicle Kilometer Travelled (VKT) and Pedestrian Kilometer Travelled (PKT) were utilized to account for road users' exposure in the analysis. The

VKT in each zone was calculated according to the methodology reported in (Nordback et al., 2017), as presented in Equation (5-9).

$$VKT_i = \sum_{j=1}^{n} AADT_{ji} \times L_{ji} \times 365 \qquad \text{i=1,2…,236} \qquad (5\text{-}9)$$

Where $AADT_{ji}$ is the Average Annual Daily Traffic of road j in TAZ (i) and L_{ji} is the length of road segments (j) in TAZ (i). The AADT for each road segment was extracted from the Hamilton Open Data website of the City of Hamilton (Open Hamilton, 2021). The historical trends of AADT have been reviewed, and no significant changes to the AADT were observed between 2010-2017. Therefore, the average value of VKT over the eight years was considered as the traffic exposure in each zone.

Moreover, (PKT) was considered a surrogate of pedestrian exposure in each zone. The 2016 Canadian census data (Statistics Canada, 2021) provided the total number of walking trips and population density in the tracts dedicated to the City of Hamilton. The census dataset was integrated into the zone map layer in ArcMap software to calculate the frequency of walkers in each TAZ. Whereas a high number of zones consist of more than one tract, the weighted average of walkers was estimated in each zone, as presented in Equation (5-10).

$$Weighted\ Average\ of\ Walkers = \frac{\sum_{i=1}^{n}(proportion\ of\ walkers)_i * (Population\ density)_i}{Population\ density} \qquad (5\text{-}10)$$

Then, the PKT in each zone was calculated based on the weighted average of walkers and the length of road segments in each TAZ (L_i), as shown in Equation (5-11) below.

$$PKT = \sum_{i=1}^{n}(Weighted\ Average\ of\ Walkers)_i \times L_i \times 365 \qquad (5\text{-}11)$$

- **Pedestrian network connectivity Pedestrian**

Four variables were used to measure the pedestrian network connectivity, including intersection density, network density, degree of network coverage, and complexity. In order to estimate these parameters, it was necessary to convert the transportation network of the City of Hamilton to a set of nodes and links, where the links represent the sidewalks and the nodes represent the intersections. The geo-coded road network of the City of Hamilton was extracted from the Open Street Map website (OpenStreetMap, 2021) and integrated into the zone map in order to distribute the links and nodes among the different TAZs. The "Intersect" function in ArcMap software was utilized to divide a link between adjacent zones if it crossed the boundary of the zone.

Once links and nodes in each TAZ are determined, the four connectivity parameters can be calculated. First, the intersection density index, which indicates the proportion of intersections (regardless of their types) to the sidewalk network in each zone, can be calculated according to the expression shown in Equation (5-12).

$$(Intersection\ density)_i = \frac{(Total\ number\ of\ intersections)_i}{(TAZ\ area)_i} \quad i=1,2\ldots,236 \qquad (5\text{-}12)$$

Second, the network density index, which represents the proportion of sidewalks in each zone, can be evaluated as shown in Equation (5-13).

$$(Network\ density)_i = \frac{(Total\ length\ of\ sidewalk\ links\ in\ a\ TAZ)_i}{(TAZ\ area)_i} \quad i=1,2\ldots,236 \qquad (5\text{-}13)$$

Third, the degree of network coverage, which represents the percentage of the road network covered with sidewalks, can be determined according to Equation (5-14).

$$(Degree\ of\ Network\ Coverage)_i = \frac{(Number\ of\ Sidewalk\ Links\ in\ TAZ)_i}{(Number\ of\ Street\ Links\ in\ TAZ)_i} \quad i=1,2\ldots,236 \qquad (5\text{-}14)$$

Finally, the zonal complexity, which represents the ratio between the number of sidewalks to the number of intersections in a TAZ, can be calculated based according to Equation (5-15) as follows:

$$(Complexity)_i = \frac{(Number\ of\ Sidewalk\ Links\ in\ TAZ)_i}{(Number\ of\ Nodes\ in\ TAZ)_i} \quad i=1,2\ldots,236 \qquad (5\text{-}15)$$

- **Pedestrian route directness**

Three graph measures were considered to represent the directness of the pedestrian sidewalk network, including the average edge length, average length per vertex, and linearity. The first two measures represent the continuity of sidewalks in a TAZ, while the linearity shows the degree of orientation of the sidewalks. Equations (5-16, 5-17, 5-18) to show the mathematical representation of the three parameters consecutively.

$$(Average\ edge\ length)_i = \frac{(Total\ length\ of\ the\ zonal\ sidewalk\ network)_i}{(Number\ of\ links\ in\ TAZ)_i} \quad i=1,2\ldots,236 \qquad (5\text{-}16)$$

$$(Average\ length\ per\ vertex)_i = \frac{(Total\ length\ of\ a\ zonal\ network)_i}{(Number\ of\ Street\ links\ in\ TAZ)_i} \quad i=1,2\ldots,236 \qquad (5\text{-}17)$$

$$(Linearity)_i = \frac{(Walking\ distance\ (route\ distance))_i}{(Straight\text{-}line\ distance\ (geodetic\ distance))_i} \quad i=1,2\ldots,236 \qquad (5\text{-}18)$$

- **Built-environment factors**

This study considered two built-environment factors in the analysis, namely, signal density and bus stop density. The signal density indicates the number of traffic signals in a TAZ per unit area. The Transportation Data Management System of the City of Hamilton provided the exact

location of traffic signals in the city as of 2019 (Public Hamilton, 2021). Thus, the number of signals in each TAZ can be easily calculated, and consequently, the signal density in each TAZ can be estimated.

The bus stop density indicates the number of bus stops in a TAZ per unit area. The exact location of all bus stops in the City of Hamilton was extracted from the Hamilton Street Railway (HSR) dataset and geocoded in ArcMap software. Thus, the total number of bus stops in each TAZ was calculated, and consequently, the bus stop density in each TAZ can be obtained by dividing the number of bus stops in a TAZ by its area.

- **Land use factors**

In order to assess the impact of land use on pedestrian safety, three land uses were considered, those are residential use, commercial use, and institutional/office use. The proportion of the TAZ areas dedicated to each of these three categories was obtained from the Geospatial Datacenter of McMaster University (McMaster University 2021). The ratio of the area dedicated to each land use to the total TAZ area was used in the analysis as a measure of the different land uses in each TAZ.

- **Attractions and amenity-related factors**

The density of a variety of amenities and attractions was calculated in each TAZ. Six amenities were considered in this category, including playgrounds, parks, restaurants, parking lots, bike-share stations, and convenience stores. The proportion of parks in each TAZ was divided by the relevant TAZ area in order to calculate the density of parks per unit TAZ area. For the rest of the

amenities, the number of each amenity in a TAZ was divided by the TAZ area to estimate the amenity density.

- **Socio-economic factors**

Population, labor force, household, and the number of jobs were the four socio-economic indicators considered in this study. These indicators were extracted from the Canadian 2016 census data. The census dataset was integrated into the zone map layer in ArcMap software to calculate each socio-economic indicator in each TAZ. Whereas a high number of zones consist of more than one tract, the weighted average of each indicator in each TAZ was estimated in a similar way that the walking trips were distributed among TAZs in Equation (5-13).

5.5.2 Final Variables list

After extracting the variables for each TAZ, the Spearman correlation matrix was developed to examine the potential correlation between these variables. Based on Moore et al., (2013), the correlation between the two variables is considered significant if the correlation coefficient is higher than 0.7. In this regard, a significant correlation was found between serval pairs of variables, including network density and intersection density, network density and residential density, average length per vertex and linearity, population and labor force, population and household, and labor force and number of jobs. In order to avoid using correlated variables in the developed models, four variables were eliminated that are network density, average length per vertex, population, and labor force. Based on that decision, a total of 20 variables were left as

potential contributors to the pedestrian-vehicle collisions in the different TAZs of the City of Hamilton. A descriptive summary of the 20 variables is presented in Table 5-1.

Table 5-1 Descriptive Summary of the Variables

Category	Variable	Mean	Std. Dev.	Min.	Max.
Pedestrian Connectivity	Intersection Density (Intersection/m^2)	0.06	0.06	0	0.28
	Degree of network coverage	1.93	1.83	0	22.29
	Complexity	5.88	11.03	0	156
Pedestrian Route Directness	Average edge length	75.16	54.20	0	618.3
	Linearity	0.79	0.32	0	1
Built Environment	Signal Density (signal/m^2)	0.002	0.004	0	0.035
	Bus Stop Density (stop/m^2)	0.009	0.022	0	0.282
Socio-economic	Household (per 1000 people)	5.87	2.66	0.685	16.16
	Job (per 1000 people)	6.59	4.57	0.43	23.48
Land use	Residential Density (%)	29.14	21.39	0	68.56
	Commercial Density (%)	4.58	9.37	0	83.69
	Institutional/Office Density (%)	5.89	9.88	0	84.52
Exposure	Log (VKT)	4.6	6.8	1	5.79
	Log (PKT)	15.7	44.1	0.008	70.54
Amenities and Attractions	Convenience Store Density (store/m^2)	0.0005	0.001	0	0.014
	Playgrounds Density (playground/m^2)	0.001	0.002	0	0.014
	Proportion of Parks (%)	0.001	0.001	0	0.011
	Restaurant Density (restaurant/m2)	0.0006	0.002	0	0.016
	Bike-share stations Density (station/m^2)	0.0007	0.002	0	0.014
	Parking lots Density (lot/m^2)	0.0003	0.001	0	0.008

5.6 Results and Discussion

5.6.1 FB macro-level collision prediction models

Four macro-level pedestrian collision prediction models were developed, including 1) Model M1, a macro-level FB with a fixed intercept for the total collisions; 2) Model M2, a macro-level FB with a fixed intercept for the collisions that involve pedestrian violations; 3) Model M3, a macro-level FB with a random intercept for the total collisions; and 4) Model M4, a macro-level FB with a random intercept for the collisions that involve pedestrian violations. Table 5-2 shows the estimated coefficients for the four models. As shown in the table, the models with random intercepts (models M3 and M4) outperformed the models with fixed intercepts (models M1 and M2), based on the value of the DIC associated with each model. Accordingly, the study relied on the two models (M3 and M4) to investigate the impact of the different factors on collisions and the identification of collision-prone zones.

- **Contributing factors to total pedestrian collisions**

According to the results of the FB models, several factors were positively associated with a higher frequency of pedestrian collisions in TAZs. Among the investigated factors, the two exposure parameters (VKT and PKT), pedestrian connectivity indicators (i.e., intersection density and network complexity), signalized intersection density, land use, socio-economic factors, sidewalk linearity, and amenities located in TAZ showed a direct impact on the frequency of pedestrian-vehicle collisions. However, some factors, including the degree of network coverage, the average edge length, and bus stop density, demonstrated a negative association with the frequency of pedestrian collisions on the macro-level.

Table 5-2 Results of the FB collision prediction models

Category	Variables	Fixed Intercept								Random Intercept							
		Model M1 (Total collisions)				Model M2 (Collisions involving violations)				Model M3 (Total collisions)				Model M4 (Collisions involving violations)			
		Mean	Std. Dev.	MCSE	Sig.	Mean	Std. Dev.	MCSE	Sig.	Mean	Std. Dev.	MCSE	Sig.	Mean	Std. Dev.	MCSE	Sig.
Pedestrian Connectivity	Intersection Density	7.86	1.92	0.001	0.0005	21.22	1.44	0.002	0.0014	7.76	2.09	0.001	0.0005	22.35	1.39	0.002	0.0014
	Degree of network coverage	-0.59	2.13	0.001	0.0005	0.12	1.32	0.002	0.0015	-0.43	0.04	0.001	0.0250	0.14	0.12	0.002	0.0167
	Complexity	3.97	3.52	0.002	0.0006	7.89	2.26	0.002	0.0009	3.82	0.05	0	0.0000	7.16	0.1	0.002	0.0200
Pedestrian Route Directness	Average edge length	-0.04	4.4	0.001	0.0002	0.76	1.57	0.001	0.0006	-0.04	0.07	0.002	0.0286	0.81	0.17	0.002	0.0118
	Linearity	0.68	1.93	0.001	0.0005	2.44	8.11	0.002	0.0002	0.62	0.11	0.003	0.0273	2.18	0.09	0.001	0.0111
Built Environment	Signal Density	0.07	0.1	0.001	0.0100	-0.08	0.05	0.001	0.0200	0.05	0.06	0.001	0.0167	-0.06	0.07	0	0.0000
	Bus Stop Density	-0.05	0.06	0	0.0000	0.12	0.13	0.002	0.0154	-0.03	0.09	0.001	0.0111	0.81	0.13	0.002	0.0154
Amenities and Attractions	Playgrounds	0.11	2.81	0.001	0.0004	0.06	7.22	0.001	0.0001	0.15	0.06	0	0.0000	0.12	0.13	0.002	0.0154
	Parking lots	0.02	2.45	0.002	0.0008	0.02	1.05	0.001	0.0010	0.02	0.09	0.001	0.0111	0.16	0.06	0.001	0.0167
	Restaurant Density	0.02	3.02	0.002	0.0007	0.01	1.16	0.002	0.0017	0.03	0.07	0.001	0.0143	0.01	0.15	0.002	0.0133
	Bike-share stations	0.26	2.1	0.002	0.0010	0.03	1.23	0.001	0.0008	0.31	0.09	0.002	0.0222	0.81	0.13	0.002	0.0154
	Convenience Store	0.03	7.83	0.002	0.0003	0.02	1.35	0.001	0.0007	0.01	0.05	0.001	0.0200	0.08	0.11	0.001	0.0091
	Proportion of Parks	0.08	0.06	0.011	0.1833	0.67	0.04	0.007	0.1750	0.04	0.07	0.001	0.0143	0.71	0.05	0	0.0000
Socio-economic	Household	0.26	2.1	0.003	0.0014	0.03	1.23	0.001	0.0008	0.37	0.1	0.001	0.0100	0.08	0.05	0.001	0.0200
	Job	0.11	2.81	0.002	0.0007	0.06	7.22	0.002	0.0003	0.08	0.06	0.001	0.0167	0.19	0.12	0.002	0.0167
Land use	Residential Density	0.02	0.07	0.001	0.0143	0.11	0.14	0.002	0.0143	0.12	0.05	0.001	0.0200	0.04	0.14	0.001	0.0071
	Commercial Density	0.02	3.02	0.001	0.0003	0.01	1.16	0.001	0.0009	0.13	0.04	0.001	0.0250	0.07	0.14	0.001	0.0071
	Institutional/Office Density	0.03	7.83	0.001	0.0001	0.02	1.35	0.001	0.0007	0.12	0.07	0.001	0.0143	0.02	0.09	0.001	0.0111
Exposure	Log (VKT)	0.06	0.1	0.001	0.0100	0.12	0.09	0.002	0.0222	0.02	0.09	0.002	0.0222	0.09	0.07	0.001	0.0143
	Log (PKT)	0.23	0.07	0.001	0.0143	0.06	0.06	0.001	0.0167	0.13	0.05	0.001	0.0200	0.04	0.11	0.001	0.0091
DIC		1078.14				846.36				930.52				699.52			
WAIC		-10482				-7622.16				-11213				-7804.91			

The negative association between the degree of network coverage and pedestrian collisions indicates that pedestrian safety increases as the sidewalk network covers a higher percentage of the road network. This finding confirms the results reported in some previous studies (e.g., Osama and Sayed, 2017; Yu, 2015). Meanwhile, the inverse impact of bus station density on the frequency of pedestrian collisions may be attributed to many factors, such as the lower average speed at locations with frequent bus service (Miranda-Moreno et al., 2011) and the relatively well-developed pedestrian infrastructure in these locations.

- **Contributing factors to collisions involving violations**

As shown in Table 5-2, there is a direct positive relationship between the frequency of the collisions that involve pedestrian violations and almost all the investigated variables, except the signalized intersection density. Such negative impact indicates that the presence of signalized intersections controls pedestrian traffic and reduces the likelihood of jaywalking, which in turn, reduces the frequency of collisions that result due to violations. According to the results of Model 4, shown in Table 5-2, the two pedestrian connectivity indicators (i.e., intersection density and network complexity) showed a similar impact on both total pedestrian collisions and collisions that involve pedestrian violations. However, the impact of the two factors on collisions that involve pedestrian violations is much higher than their impact on total pedestrian collisions. This indicates that the complexity of traffic movements at intersections may lead to serious consequences of pedestrian violations and also shed light on the importance of developing countermeasures to mitigate pedestrian violations at intersections. The explanation of the positive impact of the degree of network coverage on the frequency of collisions that involve

pedestrian violations is not straightforward. It may be the case that TAZs with a higher network coverage reflect the presence of attractions that generate higher activities in a TAZ (e.g., commercial and recreational attractions). Such attractions would necessitate a higher network coverage, but at the same time, they may increase the likelihood of pedestrian violations and the frequency of collisions that involve violations consequently. Nevertheless, further analysis is required to investigate such a relationship in more detail.

5.6.2 Identification of collision-prone zones

Collision-prone zones for both the total pedestrian collisions (based on model M3) and the collisions that involve pedestrian violations (based on model M4) were identified according to the methodology presented in section 3. Overall, 31 collision-prone zones were identified with respect to total pedestrian-vehicle collisions, and 19 zones were identified as hotspots for collisions that involve pedestrian violations. It should be noted that 11 zones have been identified as collision-prone zones according to both total collisions and collisions involving pedestrian violations. Figure 5-2 shows the spatial distribution of the identified collision-prone zones.

5.6.3 Identification of hotspots through the SOM model

In this study, R Studio software was utilized to develop the SOM model that aims at identifying collision-prone zones, investigating the features (variables) that can be used to distinguish between collision-prone and non-collision-prone zones, along with ranking these features based on their importance in distinguishing between the two categories.

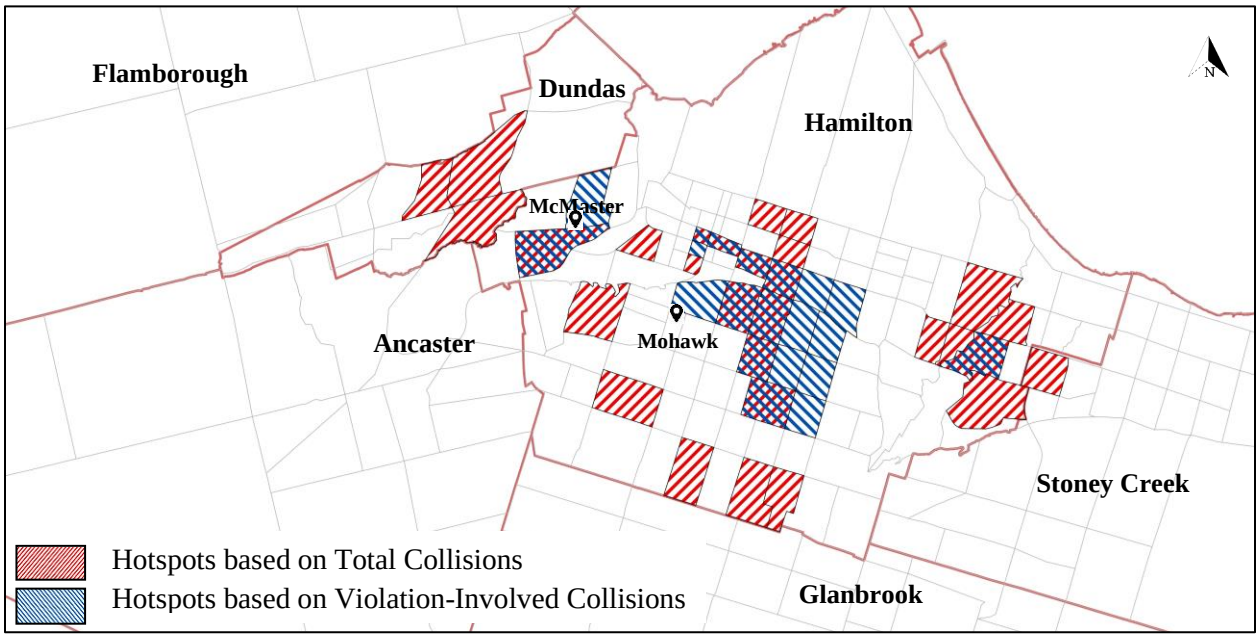


Figure 5-2 Spatial distribution of collision-prone zones

Two datasets with a size of 25 dimensions were projected on a SOM grid in order to reduce their dimension for better visualization. To achieve the highest quality of the SOM map with maintaining the topology of the original dataset on the SOM grid, several SOM grids with various dimensions were investigated and compared based on their topographic error terms. As a result, two hexagonal SOM grids of 5*5 units (25 nodes) were extracted with the lowest value of the topographic error. Figures 5-3 and 5-4 represent the count and neighbourhood distance plots for the total collisions and violation-involved collisions, respectively. The count plot in Figure 5-3 demonstrates the frequency of the total pedestrian-vehicle collision dataset in each node, with the dark red nodes and the beige nodes containing the lowest and the highest frequency of collisions, respectively. On the other hand, the neighbourhood distance plot (the bottom scheme in Figure 5-3 displays the distance between neighboring SOM units of the selected SOM grid.

Similar to the count plots, the color of the nodes gradually changes from beige to dark red with the increase in distance between neighboring SOM units. Similar graphs were developed for the violation-related collisions and are shown in Figure 5-4.

Two SOM models were generated: (SOM model 1) and (SOM model 2) that identify collision-prone zones based on the total collisions and the collisions that involve pedestrian violations, respectively. The first model (SOM model 1) classified the 236 TAZs into two clusters. The first cluster included 212 TAZs, while the second cluster included 24 TAZs.

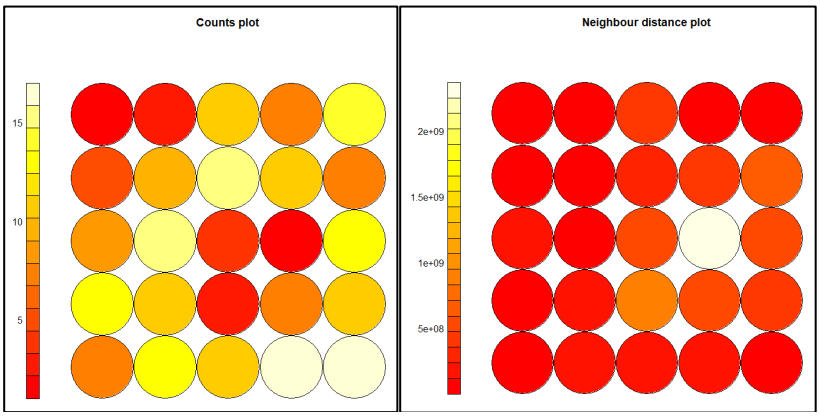


Figure 5-3 Count plot and neighbourhood distance plot for total collisions

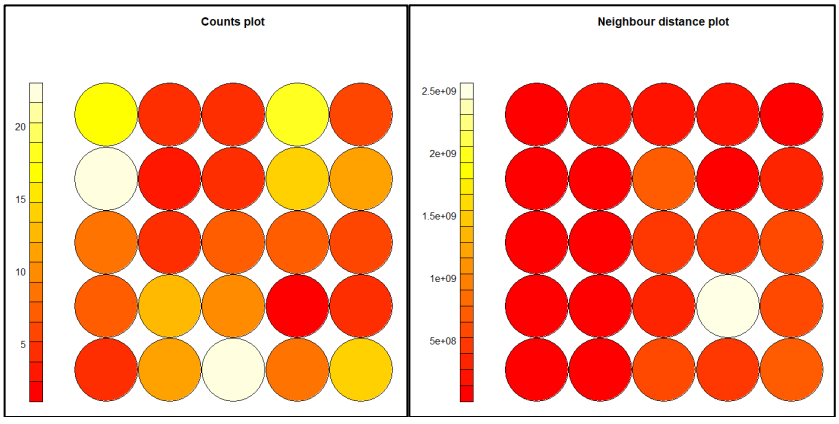


Figure 5-4 Count plot and neighbourhood distance plot for collisions that involve violations

The second model (SOM model 2) defines two clusters as well, with the first cluster containing 221 TAZs and the second cluster containing 15 TAZs. The accuracy of the two models was found to be high, according to the measures presented in Table 5-3.

Table 5-3 Performance of the SOM models

Criteria	SOM model 1	SOM model 2
CCR	88.13%	91.52%
MAE	0.1186	0.0847
RMSE	0.1444	0.1129
Roc Area	0.869	0.892

Afterwards, the two consistency tests were applied to the identified collision-prone zones. According to the results of the within-method consistency test, both SOM and PSI approaches demonstrated a significant high consistency equal to 91% and 87%, respectively, for the collision-prone zones based on total collisions, and 89% and 86% for hotspots including violation-related collisions. Then, a between-method consistency check was generated to check the similarity between the list of the hotspot zones ranked by the PSI method and the list extracted by the SOM model. For the total collisions, the value of the test was equal to 62%. While such value raised to 67% for the violation-involved hotspots. Based on the results, the ranking of collision-prone zones obtained by the SOM model was shown to be more coherent than the ranking obtained by the traditional PSI method.

Following the classification and the consistency check, the two classes that resulted from each of the two SOM models were compared based on four safety-related indices, namely the average collision rate per class, the average collision rate per pedestrian exposure, the average collision

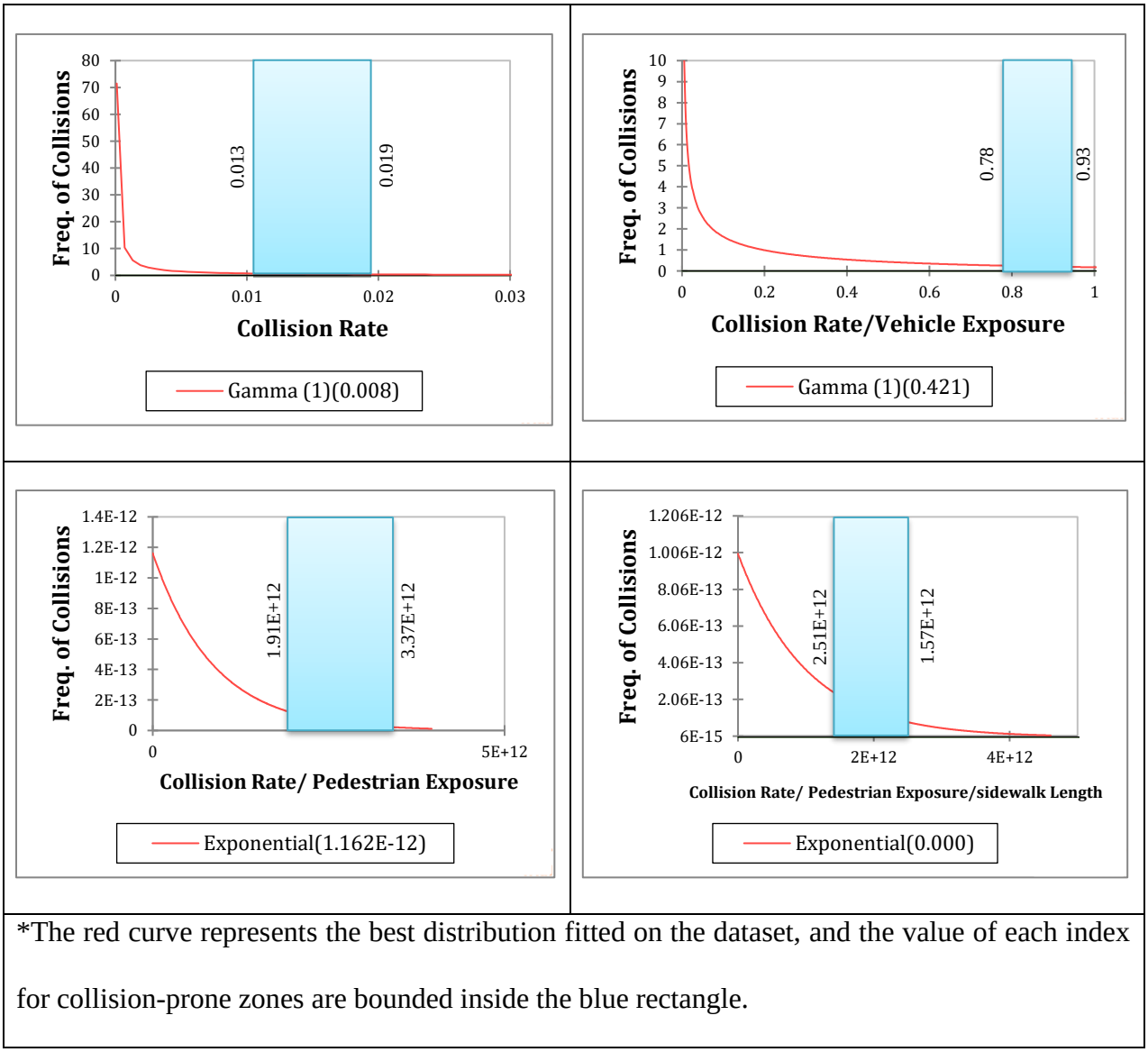
rate per vehicle exposure, and the average collision rate per pedestrian exposure per unit length of the sidewalk network. Table 5-4 shows the values of the four measures for the two SOM models considered in the study. As shown in the table, the second class in both models has significantly higher values in all four presented indices.

Table 5-4 Safety indexes in each class

Indices	SOM model 1 (Total collisions)			SOM model 2 (Collisions involving violations)		
	Class 1	Class 2	P_value	Class 1	Class 2	P_value
Total Number of zones	212	24		221	15	
Collision Rate	0.172	0.729	0.01	0.31	0.652	0.01
Collision Rate/Vehicle Exposure (1000)	0.0158	0.0669	0.01	0.0028	0.0059	0.05
Collision Rate/Pedestrian Exposure (1000)	0.459	1.94	0.05	0.083	0.174	0.002
Collision Rate/Pedestrian Exposure (1000)/ Sidewalk length (km)	0.0002	0.0009	0.02	0.0042	0.0088	0.01

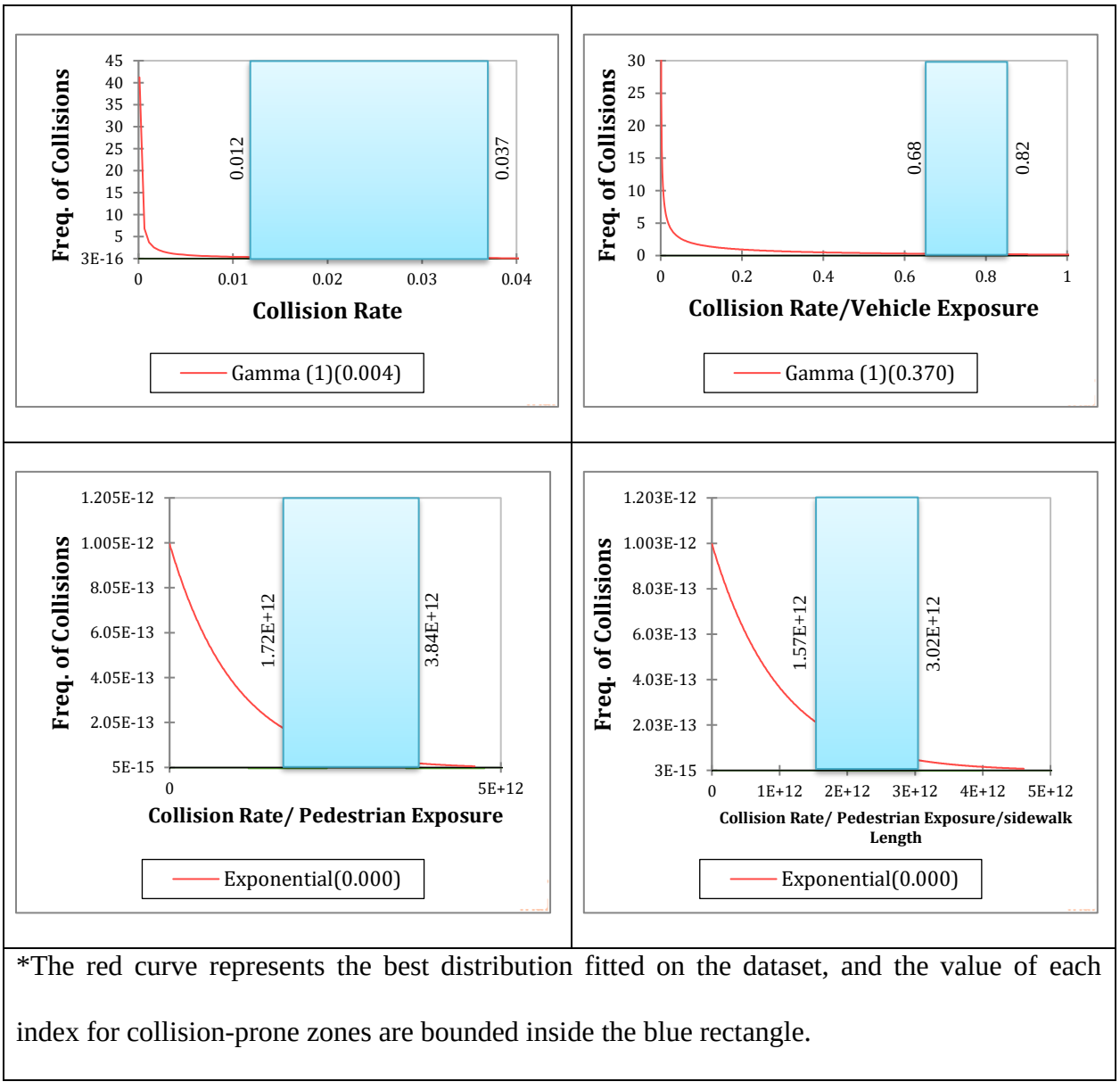
Moreover, Figures 5-5 and 5-6 show the statistical distribution of the four indices in the 236 TAZs for SOM model 1 and model 2, respectively. Several statistical distributions were tested and compared based on the Chi-squared goodness of fit values to select the most accurate distribution of each index. Gamma distribution was found to be the most accurate distribution for the first two indices (i.e., collision rate and collision rate per vehicle exposure), while the exponential distribution demonstrated the best fit for the last two indices. Figure 5-5 shows the range of the four indices for collision-prone zones for the first model (total pedestrian collisions),

while Figure 5-6 shows the range of the corresponding ranges for the second model (collisions that involve pedestrian violations).



*The red curve represents the best distribution fitted on the dataset, and the value of each index for collision-prone zones are bounded inside the blue rectangle.

Figure 5-5 Probability distribution of the four indexes based on total collisions



*The red curve represents the best distribution fitted on the dataset, and the value of each index for collision-prone zones are bounded inside the blue rectangle.

Figure 5-6 Probability distribution of the four indexes based on collisions that involve violations

Meanwhile, the range of each index in the TAZs that were identified as collision-prone zones by the SOM model is displayed. As shown in the two figures, the range of the four indices for the TAZs that were identified as collision-prone zones is clearly located in the tail of the distribution

in all four indices (blue rectangles). For example, Figure 5-5 shows that the pedestrian collision rate in all TAZs follows the Gamma distribution, with shape and scale parameters of 1 and 0.008, respectively (i.e., the mean collision rate = 0.008). The collision rate for the TAZs that are identified as collision-prone locations on this model ranges between 0.013 and 0.019, which is clearly a high collision rate compared to the overall average collision rate in TAZs.

Furthermore, the collision-prone zones identified by the two SOM models were compared to those that were identified using the conventional FB method. Considering the collision-prone zones of the FB method as a reference, the SOM models managed to identify collision-prone locations with an accuracy of 87.3% and 89.4% for the total collisions and collisions that involve pedestrian violations respectively. The F-test was conducted to investigate the similarity between the results of the two approaches (FB and SOM techniques).

According to the F-test, there was not a significant difference between the results of the two methods for both the total collisions and the collisions that involve pedestrian violations. Meanwhile, the performance of the approaches was compared through AIC criteria, which reveals that the SOM approach outperforms the FB method since it has a lower AIC compared to the FB model in both approaches (i.e., 1342.07 vs 1486.11 for the total collisions and 716.52 vs 871.92 for the collisions that involve pedestrian violations).

The P values reported in Table 5-4, the observed high percentile of the four safety-related indices shown in Figures 5-5 and 5-6 for TAZs in the second class, and the results of the F-test reported above evidently show that TAZs in the second class for both SOM models 1 and 2 can be classified as collision-prone zones. These TAZs have a much higher probability of collision

occurrence compared to TAZs in class 1. The SOM model was capable of capturing this trend and identifying collision-prone zones with high accuracy. Moreover, the SOM model enables the calculation of the relative weight of each variable, which shows the relative importance of a factor in distinguishing between the two classes. The SOM model produces several weight vectors (equal to the number of neurons in the map space) and assigns an initial weight to each variable. During the training procedure, the training space will be subdivided into various classes according to the similarity of the weight vectors of the neurons. In the next stage, the classified weight vectors estimate the probability density function of the input variables. Generally, the higher the weight of a factor, the more important this variable is in differentiating between the two classes. Table 5-5 presents the relative weight of the factors along with the average value of each variable for the second class that resulted from each of the two SOM models.

As for the total collisions (SOM model 1), intersection density was found to be the most important factor that distinguishes between collision-prone zones and non-collision prone zones, followed by the pedestrian network directness indicators “average edge length”, the proportional of residential land uses and the two exposure parameters PKT and VKT. According to the table, the collision-prone zones are characterized by higher intersection and signal density, higher density of households, heavier vehicular traffic volume (VKT), and lower average edge length.

Intersection density was also ranked as the first factor that distinguishes between collision-prone zones and non-collision-prone zones based on collisions that involved pedestrian violations, followed by the density of bike-share stations and parking lots, the proportional residential land uses, and the pedestrian network directness indicator “linearity”. Collision-prone zones based on

collisions that involve pedestrian violations are characterized by high intersection density, but low signal density. This means that zones with more unsignalized intersections are more prone to collisions that involve pedestrian violations. Also, these zones are characterized by high bus stop density, high pedestrian network complexity, high residential density, and high network linearity (i.e., high degree of indirectness in the sidewalk network), based on Table 5-5.

Table 5-5 Results of the SOM model in the second class of each model

Factors	**SOM model 1 (Total collisions)- 24 zones**		**SOM model 2 (Collisions involving violations)- 15 zones**	
	Relative weight	**Average value**	**Relative weight**	**Average value**
Intersection Density	67.236	0.96	1.815	0.94
Degree of network coverage	4.037	1.69	0.457	2.17
Complexity	5.067	5.99	0.569	6.12
Average edge length	63.948	71.65	0.0033	76.61
Linearity	0.921	0.88	0.690	0.93
Signal Density	3.915	0.62	0.235	0.16
Bus Stop Density	7.162	0.82	0.628	1.23
Household	8.580	7.16	0.538	6.67
Job	9.191	8.51	0.207	8.96
Residential Density	28.187	38.06	0.724	36.47
Commercial Density	4.967	7.36	0.373	5.81
Institutional/Office Density	4.185	9.35	0.409	8.16
Convenience Store	1.058	0.00081	0.512	0.00067
Park Amenities	3.566	0.0017	0.143	0.0024
Proportion of Parks	0.050	0.0016	0.023	0.0019
Restaurant density	0.604	0.0007	0.292	0.00065
Bike-share stations	0.692	0.0007	1.515	0.0008
Parking lots	0.201	0.0003	0.725	0.0005
Log(VKT)	21.258	5.20	0.431	5.86
Log (PKT)	32.171	22.26	0.196	15.96

The values reported in Table 5-5 suggest that more interest should be given to enhance pedestrian safety in zones with higher intersection density and residential land uses. Increasing the directness of the pedestrian network is key to enhancing the safety of pedestrians on the macro-level. The results also showed that more violation-related collisions would be observed at zones with amenities like bike share stations and parking lots. Thus, it is crucial to investigate the locations where such amenities exist and study pedestrian behaviour there in more detail in order to understand the mechanism of pedestrian violation occurrence and the best strategies to mitigate them.

5.7 Conclusion

In this study, a Full Bayesian model and an unsupervised Deep Learning technique were applied to conduct a macro-level pedestrian safety analysis in the City of Hamilton, Ontario. The study developed FB macro-level collision prediction models for both total pedestrian-vehicle collisions and collisions that involved pedestrian violations. Additionally, the study identified collision-prone zones for both categories of collisions using the developed FB models and the SOM deep learning model. According to the developed FB collision prediction model, several factors, including road user exposure, pedestrian network connectivity indicators (namely, intersection density and network complexity), sidewalk linearity, socio-economic factors (household and job), attractions and amenities in TAZs (specifically bike-share stations, playgrounds, and parks), and residential and commercial land use are directly associated with both categories of collisions. Some factors, however, (namely, degree of network coverage, average edge length, signalized

intersection density, and bus stop density) showed different impacts on total collisions and violation-related collisions.

Moreover, the SOM model was capable of identifying collision-prone zones with high accuracy. The ranked hotspots identified by the SOM model were evaluated with a consistency test and demonstrated higher consistency in comparison with the univariate PSI method. The SOM model identified five zonal variables as the key variables that can differentiate between collision-prone and non-collision-prone zones based on the total pedestrian-vehicle collisions, namely, the intersection density, the pedestrian network directness (represented by the average edge length), the proportion of residential land uses, and road user exposure parameters (VKT and PKT). The model also identified five key variables that distinguish between collision-prone and non-collision-prone zones based on collisions that involved pedestrian violations, namely, intersection density, the density of bike-share stations and parking lots in a TAZ, pedestrian network directness (represented by linearity), and the proportional residential land uses.

Nevertheless, several future research directions can be recommended based on the results and limitations of this study. For example, this study applied the SOM model as an unsupervised deep learning model to identify collision-prone zones. Future studies could investigate the potential use of other Deep Learning models and assess the accuracy of the different models. Additionally, while the study investigated the safety impacts of a wide range of factors, future studies are still needed to investigate the impact of other factors, such as income, car ownership, and household characteristics, among other factors. Moreover, similar analysis can be conducted

in different cities to develop a better understanding of the characteristics of collision-prone zones and account for other undetected factors, such as cultural differences.

Moreover, the work presented in this paper has several practical implications. It is commonly acknowledged that active travelers, such as pedestrians, are considered among the most vulnerable road users, as they are at a higher risk of being killed or severely injured due to road collisions. Also, some pedestrian unsafe behaviours (particularly, pedestrian violations) can contribute to pedestrian-vehicle collisions in urban areas. Thus, conducting a macro-level analysis to identify the collision-prone zones that experience a high frequency of collisions that involve pedestrian violations shall provide a better understanding of the characteristics of such hotspots. Consequently, transportation authorities can direct safety improvement projects effectively to the most hazardous zones and have a better understanding of safety interventions that are most needed in each zone. Furthermore, the results of this study can guide transportation planners engineers in planning future developments so that pedestrian risky behaviours and subsequent collisions are minimized. As well, the deep learning approach proposed in this study shall provide more flexibility in understanding the characteristics of collision-prone zone compared to traditional approaches, since deep learning methods are not dependent on the factors that are pre-defined by the modellers. Accordingly, this approach shall provide an opportunity for transportation engineers to understand the safety impacts of several factors that were not typically studied in macro-level studies, including, for example, the density of bike share stations and parking lots in a traffic analysis zone.

5.8 Reference

Bao, J., Liu, P., & Ukkusuri, S. V. (2019). A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accident Analysis and Prevention, 119, 239-254.

Barker, J., & Baguley, C. (2001). A road safety good practice guide. In Proceedings of the Good Practice Conference. Bristol, UK.

Bhandari, B., & Park, G. (2022). Development of a real-time security management system for restricted access areas using computer vision and deep learning. Journal of Transportation Safety and Security, 14(4), 655-670.

Cai, Q., Abdel-Aty, M., Sun, Y., Lee, J., & Yuan, J. (2019). Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data. Transportation Research Part A, 127, 71-85.

Dabiri, S., & Heaslip, K. (2018). Inferring transportation modes from GPS trajectories using aconvolutional neural network. Transportation Research Part C, 86, 360-371.

Deacon, J. A., Zegeer, C. V., & Deen, R. C. (1975). Identification of hazardous rural highway locations. Transportation Research Record, 543, 16-33.

Ding, C., Chen, P., & Jiao, J. (2018). Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: A machine learning approach. Accident Analysis and Prevention, 112, 116-126.

El-Basyouny, K., & Sayed , T. (2013). Depth-Based Hotspot Identification and Multivariate Ranking using the Full Bayes Approach. Accident Analysis & Prevention, 50, 1082-1089.

El-Basyouny, K., & Sayed, T. (2009). Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention, 41, 820–828.

El-Basyouny, K., & Sayed, T. (2012). Measuring direct and indirect treatment effects using safety performance intervention functions. Safety science, 50(4), 1125-1132.

Ghomi, H., & Hussein, M. (2021). An integrated clustering and copula-based model to assess the impact of intersection characteristics on violation-related collisions. Accident Analysis and Prevention, 159, 106283.

Heckerman, D. (1999). Bayesian learning. In In: Wilson, R., Keil, F. (Eds.), The MIT Encyclopedia of the Cognitive Sciences (pp. 70-72). MIT Press: Cambridge, Massachusetts.

Hollander, J. B., Nikolaishvili, G., Adu-Bredu, A. A., Situ, M., & Bista, S. (2021). Using deep learning to examine the correlation between transportation planning and perceived safety of the built environment. Environment and Planning B: Urban Analytics and City Science, 48(7), 2023-2038.

Hou, Y., & Edara, P. (2018). Network Scale Travel Time Prediction using Deep Learning. Transportation Research Record, 2672(45), 115-123.

Hussein, M., Sayed, T., El-Basyouny, K., & de Leur, P. (2020). Investigating safety effects of wider longitudinal pavement markings. Accident Analysis & Prevention, 142, 105527.

Kim, M., Kho,, S. Y., & Ki, D. K. (2017). Hierarchical Ordered Model for Injury Severity of Pedestrian Crashes in South Korea. Journal of Safety Research, 61, 33–40. doi:doi.org/10.1016/j.jsr.2017.02.011

Kohonen, T. (2013). Essentials of the self-organizing map. Neural Networks, 37, 52-65.

Krzywinski, M., & Altman, N. (2013). Points of significance: Importance of being uncertain. Nature methods, 10(9), 809-811.

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

Library, G. D. (2021, April). McMaster University. Retrieved from https://library.mcmaster.ca/collections/geospatial-data

Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F. Z. (2015). Traffic Flow Prediction with Big Data: A Deep Learning Approach. IEEE Transactions on Intelligent Transportation Systems, 16(2), 865-873.

Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., & Wang, Y. (2017). Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction. Sensors, 17(4), 818.

Ma, X., Gildin, E., & Plaksina, T. (2015). Efficient optimization framework for integrated placement of horizontal wells and hydraulic fracture stages in unconventional gas reservoirs. Journal of Unconventional Oil and Gas Resources, 9, 1-17.

Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C, 54, 187-197.

Miranda-Moreno, L. F., Morency, P., & El-Geneidy, A. M. (2011). The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections. Accident Analysis and Prevention, 43, 1624–1634. doi:doi.org/10.1016/j.aap.2011.02.005

Mohamed, M. G., Saunier, N., Miranda-Moreno, L. F., & Ukkusuri, S. V. (2013). A clustering regression approach: A comprehensive injury severity analysis of pedestrian-vehicle crashes in New York, US and Montreal, Canada. Safety science, 54, 27–37.

Mukherjee, D., & Mitra, S. (2020). A comprehensive study on factors influencing pedestrian signal violation behaviour: Experience from Kolkata City, India. Safety Science, 124, 104610. doi:doi.org/10.1016/j.ssci.2020.104610

Nordback, K., Sellinger, M., & Phillips, T. (2017). Estimating Walking and Bicycling at the State Level. Portland, OR: National Institute for Transportation and Communities (NITC).

Open Hamilton. (2021, April). Retrieved from https://open.hamilton.ca/

OpenStreetMap. (2021, June). Retrieved from https://www.openstreetmap.org/#map=10/43.2607/-79.9352

Osama, A., & Sayed, T. (2016). Evaluating the impact of bike network indicators on cyclist safety using macro-level collision prediction models. Accident Analysis and Prevention, 97, 28-37.

Osama, A., & Sayed, T. (2017). Evaluating the impact of connectivity, continuity, and topography of sidewalk network on pedestrian safety. Accident Analysis and Prevention, 107, 117-125.

Osama, A., Sayed , T., & Sacchi, E. (2018). A Novel Technique to Identify Hot Zones for Active Commuters' Crashes. Transportation Research Record, 2672(38), 266-276.

Peng, C., & Xu, C. (2021). Combined variable speed limit and lane change guidance for secondary crash prevention using distributed deep reinforcement learning. Journal of Transportation Safety & Security, 1-26.

Pour-Rouholamin, M., & Zhou, H. (2016). Investigating the risk factors associated with pedestrian injury severity in Illinois. Journal of Safety Research, 57, 9–17. doi:doi.org/10.1016/j.jsr.2016.03.004

Railway, H. S. (2021, April). The city of Hamilton. Retrieved from https://www.hamilton.ca/hsr-bus-schedules-fares/schedule-routes-maps/pdf-bus-schedules

Sacchi, M., Sayed, T., & El-Basyouny, K. (2015). Multivariate Full Bayesian Hot Spot Identification and Ranking New Technique. Transportation Research Record: Journal of the Transportation Research Board, 2515, 1-9.

Smith, A. M. (1991). Bayesian computational methods. Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences, 337(1647), 369-386.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(4), 583-639.

Statistics Canada. (2021, April). Retrieved from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm

System, T. D. (2021, June). The City of Hamilton. Retrieved from https://hamilton.public.ms2soft.com/tcds/tsearch.asp?loc=Hamilton&mod=

Transport Canada. (2021, September). Retrieved from https://tc.canada.ca/en/road-transportation/statistics-data/canadian-motor-vehicle-traffic-collision-statistics-2019

Walters, C., & Ludwig, D. (1994). Calculation of Bayes posterior probability distributions for key population parameters. Canadian Journal of Fisheries and Aquatic Sciences, 51(3), 713-722.

Wang, K., Bhowmik, T., Yasmin, S., Zhao, S., Eluru, N., & Jackson, E. (2019). Multivariate copula temporal modeling of intersection crash consequence metrics: A joint estimation of injury severity, crash type, vehicle damage and driver error. Accident Analysis and Prevention, 125, 188-197. doi:doi.org/10.1016/j.aap.2019.01.036

Wang, P. (2004). The limitation of Bayesianism. Artificial Intelligence, 158, 97-106.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11, 3571-3594.

Wu, Y., Tan, H., Qin, L., Ran, B., & Jiang, Z. (2018). A hybrid deep learning based traffic flow prediction method and its understanding. Transportation Research Part C: Emerging Technologies, 90, 166-180.

Xie, K., Ozbay, K., Kurkcu, A., & Yang, H. (2017). Analysis of Traffic Crashes Involving Pedestrians Using Big Data: Investigation of Contributing Factors and Identification of Hotspots. Risk Analysis, 37(8), 1459-1476.

Yu, C. Y. (2015). Built Environmental Designs in Promoting Pedestrian Safety. Sustainability, 7(7), 9444–60.

CHAPTER 6

Moving Vision Zero Programs Forward: What Countermeasure combinations work best and where? A Dynamic Copula-based Time-series Approach

Ghomi, H., & Hussein, M. Moving Vision Zero Programs Forward: What Countermeasure combinations work best and where? A Dynamic Copula-based Time-series Approach. Submitted to Accident Analysis & Prevention in January 2023.

Haniyeh Ghomi is the main contributor of this manuscript. The co-author's contributions include guidance, supervision, funding, reviewing the analysis, and editing the manuscript.

6.1 Abstract

Vision Zero stands out as one of the most promising systemic safety action plans. A crucial step to ensure the successful implementation of Vision Zero is to continuously assess the efficiency of the implemented treatments. Traditionally, this is achieved using before-and-after analyses or cross-sectional studies. However, the applicability of these approaches can be limited in assessing Vision Zero initiatives, which usually involve installing multiple treatments at a location, leading to a significant interdependency between treatments. This study proposes a dynamic R-vine copula-based time series model to evaluate the efficiency of treatments implemented as a part of Vision Zero. The proposed approach enables the accurate assessment of the treatments, understanding of their long-term impacts, and identifying the most effective combination of treatments at a location. The study also investigated the association between location characteristics and the performance of treatments. The proposed framework was applied to the City of Toronto at the macro-level (neighbourhood level) and focused on pedestrian-related treatments. Collision data and the implemented countermeasures were obtained from Toronto's Vision Zero Mapping Tool. The results show that the combination of speed limit reduction, leading pedestrian intervals (LPI), and community safety zones was the most frequent combination in terms of efficiency. Enforcement and speed limit reduction were the most effective combination in neighbourhoods with high school density, while LPI was effective in neighbourhoods with high densities of subway stations, and office density, especially when integrated with speed limit reduction and community zones. Driver feedback signs were effective in neighbourhoods with a high density of intersections, but only when combined with automated

enforcement, community safety zones, and speed limit reduction. The results of the study would assist decision-makers in selecting the most effective treatment in a neighbourhood based on the neighbourhood characteristics and the countermeasures that are already installed.

6.2 Introduction

Driven by the massive social and economic burdens of road collisions, governments and transportation authorities are enacting various legislation and policies to reduce the frequency of collisions and their severe consequences. Among a multitude of relevant policies, Vision Zero stands out as one of the most promising safety action plans in many cities around the world. The concept of Vision Zero involves comprehensive strategies and engineering interventions to mitigate traffic fatalities and serious injuries while enhancing the well-being and equitable mobility of all road users (Boodlal et al., 2021). Vision Zero recognizes traffic fatalities and serious injuries as preventable events that can be mitigated through a systemic safe system approach. The idea was first introduced in Sweden in the 1990s and has been adopted by many municipalities in Canada since 2015.

A crucial step to ensure the successful implementation of a Vision Zero strategy is to continuously assess the impact of the adopted plans on enhancing road user safety, which is important to guide future safety improvement plans and revise existing ones. Typically, evaluating the performance of a safety intervention is achieved by conducting a before-and-after analysis or cross-sectional studies. These approaches, particularly, before-and-after analyses, are well-established in the safety literature and have shown enormous success in evaluating the benefits of safety interventions (Fayish and Gross, 2010; Heydari et al., 2014; Wu et al., 2018;

Hussein et al., 2020). Nevertheless, the applicability of before-and-after and cross-sectional studies can be limited in conducting a system-wide evaluation to assess large safety initiatives such as Vision Zero. In Vision Zero, various safety programs are implemented simultaneously, which means a location (e.g., an intersection) or an area (e.g., a neighbourhood) could receive more than one safety treatment at the same time. In this situation, evaluating the performance of each treatment in a separate manner is not accurate as it is not possible to attribute the change in collisions to one intervention. Moreover, safety treatments can show different effects on safety over time. Some treatments can show an immediate impact on collision severity, but over time, the impact can be reduced or even vanished. Other safety treatments may not show an immediate impact on safety, but in the long run, they can be very effective. As such, before-and-after and cross-sectional studies can yield incomplete conceptualization of the impacts of safety treatments, as they cannot easily explain the temporal trend of the treatments' effects on road safety.

Time series models are popular techniques for examining changes in variables over time. A time-series model typically updates the current status of time-dependent variables by using knowledge of their previous values. The results of time series models provide insights regarding the underlying causes of fluctuating trends and patterns of variables in the long run, as well as enable the prediction of future patterns. In the context of evaluating the efficiency of Vision Zero programs, time series analysis can provide transportation planners and policymakers with long-term insights into the effectiveness of the implemented safety treatments. Nevertheless, two main points should be taken into consideration. First, typically, more than one type of treatment is

implemented at a location. Thus, the dependent variable (i.e., the collision frequency or severity) will be affected by more than one common error term due to the presence of interdependency between the impact of countermeasures. Second, the Vision Zero action plans involve introducing several countermeasures as new plans, as well as increasing the frequency of existing treatments. Therefore, newly applied safety measures are always correlated with preinstalled countermeasures and their performance. Consequently, two types of interdependencies between the countermeasures are typically present: cross-sectional (the interdependency between the new countermeasure and existing countermeasures at a location or an area) and serial (the impact of the performance of existing countermeasure overtime on the new countermeasure). Although traditional time series models could investigate the temporal patterns of each countermeasure in a separate manner, they are not capable of handling either of two dependencies due to their linear functionality.

Copula-based multivariate time series models can effectively address this issue as they are capable of capturing both kinds of dependencies at once (Nagler et al., 2022). Specifically, the Vine copula is a type of copula-based model that has proven to be an effective tool for predicting a time-series model with multidimensional dependencies between variables (Patton, 2012). In this model, the performance of countermeasures depends on their individual impact on collisions and on the dependency between each of the installed countermeasures, which is captured by a function called a copula. Additionally, understanding the interdependency among the countermeasures enables the definition of the combination of safety treatments with a high degree of co-movement. This can be very helpful in identifying the combination of treatments

that can work best together under specific location characteristics and suggesting alternative countermeasures with similar temporal patterns to a desired countermeasure that is not applicable (or not economically viable) at a location.

Therefore, this study proposes to apply the copula-based multivariate time series modeling framework to evaluate the efficiency of safety treatments that are conducted as a part of the Vision Zero program in a city. The proposed approach enables the accurate assessment of the impact of safety treatments and understanding of their long-term impact on collision frequency, as well as identifying the most suitable combination of safety measures at a location in order to maximize their effectiveness. The proposed framework was applied to the City of Toronto, Canada as a case study. The City of Toronto adopted Vision Zero in July 2016. According to the city, Toronto's Vision Zero prioritizes the safety of the most vulnerable road users, such as pedestrians, through a range of extensive, proactive, and data-driven initiatives (City of Toronto, 2022). The methodology proposed in this study was implemented on the macro-level (neighbourhood level) and focused on evaluating pedestrian-related treatments.

To that end, Toronto's Vision Zero mapping tool (Mapping tools, 2022) was utilized to obtain pedestrian-vehicle collisions that occurred in the City of Toronto between 2017 and 2021, along with the implemented pedestrian-focused safety measures during this period. The City of Toronto was divided into 158 neighbourhoods. The frequency of severe pedestrian-vehicle collisions (fatal and serious injury collisions) was aggregated to the neighbourhood level. Also, the frequency of locations that receive each of the treatments under investigation in each neighbourhood was obtained. The safety treatments considered in this study included: 1)

engineering improvements (e.g., traffic calming and geometric improvements); 2) automated enforcement (e.g., automated speed enforcement cameras and red light cameras); 3) speed limit reductions; 4) lead pedestrian intervals; 5) accessible pedestrian signals; 6) traffic control measures (such as converting uncontrolled intersection to a signalized one, installing pedestrian crossover signals, and introducing flashing beacons to mid-block crossings); 7) driver feedback signs (such as and LED displays that show drivers' speed); and 8) creating community safety zones (zones in which fines with speeding are doubled).

A dynamic R-vine copula model was implemented in each neighbourhood to understand the temporal trends of the installed countermeasures and the joint interdependency between the treatments in different neighbourhoods. The results of the developed models were used to determine the best combination of countermeasures in each neighbourhood in terms of their safety effectiveness. The prediction power of the proposed models was validated using nine-month collision dataset of 2022. Afterwards, a logistic regression model was developed to investigate the association between neighbourhood characteristics and the performance of the safety treatments. The goal was to understand the neighbourhood characteristics that maximize the benefits of specific countermeasure combinations, aiming to guide the allocation of safety treatments in the future. The rest of the paper is organized as follows: The following section provides a summary of the literature review. Afterward, the research methodology is presented, followed by an overview of the data collection and processing. In the following section, the results of the study are presented and discussed. Finally, the last section of the paper presents the conclusions and recommendations of the study.

6.3 Literature Review

The literature review focused on reviewing previous studies that evaluated the safety benefits of safety improvements that are implemented on a large scale to enhance safety in a city or an urban center. Earlier studies relied on developing simple statistical tests to compare the collision frequency before and after implementing the interventions. For example, Rogers et al., (2016) evaluated the impact of automated enforcement (i.e., red light cameras and automated speed enforcement) that were implemented at 166 locations in the District of Columbia. The study relied on the Pareto analysis to assess the impact of the treatment on both the frequency and the severity of collisions. The study found a significant reduction in the frequency of total and fatal collisions after implementing automated enforcement. In another study, Phares et al., (2021) utilized the t-test to investigate the impact of speed limit reduction on pedestrian and cyclist safety in New York City (NYC). The study analyzed historical collision records in the city from 2012 to 2020 and found a significant reduction in fatalities for pedestrian and cyclist-related collisions after reducing the speed limit from 40 km/h to 30 km/h.

Meanwhile, several studies investigated the safety performance of the countermeasures using statistical regression models. For example, Mammen et al., (2020) developed a difference-in-differences regression model to investigate the impact of speed limit reduction from 30 to 25 mph in New York City streets on both the frequency and the severity of collisions. Based on the results, the study found a 35.8% and 38.7% reduction in total and fatal collisions, respectively. Another study reported a reduction of 63% in fatal collisions for the same treatment (Zhai et al., 2022). In another research, Fridman et al., (2020) conducted a pre-post analysis using a quasi-

experimental model to investigate the impact of reducing the speed limit from 40 km/h to 30 km/h on local roads in the City of Toronto, Ontario. According to the results, the speed limit reduction reduced the frequency of pedestrian collisions by 28% and fatal collisions by up to 60%. Rothman et al., (2022) developed a pilot study to examine the impact of reducing the speed limit and installing speed cameras on school zone safety in the City of Toronto. The study analyzed 34 school zones using a beta regression model. The study reported that the rate of aggressive driving and speeding was reduced near school zones by 4%.

Moreover, several studies used Bayesian techniques to calculate the Collision Modification Factor (CMF) of particular large-scale treatments. For example, Goughnour et al., (2018) investigated the effect of leading pedestrian intervals in three cities in the US (Charlotte, Chicago, and New York City) and the City of Toronto in Canada. The study conducted an Empirical Bayesian before-and-after study to investigate the impact of the treatment on the frequency of pedestrian-vehicle collisions. The study reported a CMF of 0.87 in Charlotte, Chicago, and Toronto, but no significant impact of the treatment was observed in New York City.

As seen in the literature, previous studies evaluated some large-scale treatments using a variety of techniques. Speed reduction, enforcement, and lead pedestrian intervals were among the most evaluated countermeasures. Other treatments, such as driver feedback signs and community safety zones, have been rarely investigated on a large scale in the literature. Systemic evaluations of multiple safety treatments, the association between neighbourhood characteristics and the

benefits of treatments and assessing the long-term impact of interventions are scarce in the literature.

6.4 Methodology

Vector Autoregressive (VAR) model (Sims, 1980) has become a dominant method for analyzing time series data, especially in economics, business, and natural science. Based on the model, a time-dependent variable improves itself over time by capturing the relationship between its past values (i.e., lag or order) and current condition. Such temporal interdependency allows the model to predict future values of a variable using the knowledge of the past. With the same concept, in a multivariate VAR model, the present value of each variable is influenced by its previous values as well as the lagged values of other variables under investigation. In the context of this study, let's suppose that Y_{it} denotes the total number of collisions that are observed in the ith neighbourhood ($i = \{1, \dots, 158\}$) at the t^{th} period. The 5-year of collision data (2017-2021) considered in the study was further classified by season to capture the seasonal variation among the data. Accordingly, the study considered 20 time periods ($t = \{1, \dots, 20\}$). The ith neighbourhood received N types of countermeasures as part of the vision zero program ($N = \{1, \dots, 8\}$ countermeasures). In a multivariate VAR model, Y_{it} will be dependent on $Y_{i(t-1)}$ for each type of countermeasure N. The expanded formula for predicting the total number of collisions at neighbourhood (i) and time period (t) for each type of countermeasure is represented in Equation (6-1):

$$\begin{pmatrix} Y_{i1(t)} \\ \vdots \\ Y_{iN(t)} \end{pmatrix} = \begin{pmatrix} x_{i1(t)}\beta_{11} & \cdots & x_{iN(t)}\beta_{1N} \\ \vdots & \ddots & \vdots \\ x_{iN(t)}\beta_{N1} & \cdots & x_{iN(t)}\beta_{NN} \end{pmatrix} \begin{pmatrix} Y_{i1(t-1)} \\ \vdots \\ Y_{iN(t-1)} \end{pmatrix} \quad (6\text{-}1)$$

The standard form of an N-dimensional multivariate VAR model of an order of 1 can be expressed as Equation (6-2):

$$Y_{it} = X_{it}\beta Y_{i(t-1)} + \varepsilon_t \quad (6\text{-}2)$$

Where Y_{it} is an N-dimensional vector of response variable for the ith neighbourhood at the tth time period, X_{it} is an N-dimensional vector of independent variables (i.e., the total number of countermeasure (n) exists at neighbourhood (i) at time (t)), β is an N*N matrix of coefficients to be estimated, $Y_{i(t-1)}$ is an N-dimensional matrix represents the value of the response variable at the previous time step (delay or lag of 1), and ε_t is an N-dimensional one-step random error term that follows a Gaussian distribution with zero mean and $exp(\varepsilon_t)$ variance (Ma et al., 2021). Nevertheless, conventional multivariate VAR models underestimate the correlation between several time series that are simultaneously entered into the model. To overcome such a limitation, copula models could be used to construct a framework that includes a marginal distribution of each univariate VAR model along with a joint copula function. In fact, the concept of the copula model is based on the idea that a d-dimensional function can be decomposed into d marginal distribution and a copula function which describes the joint dependence structure. In other words, copula-based multivariate models evaluate the marginal distribution of each explanatory variable separately from the joint distribution that is formed by these marginal distributions due to their dependence structure. The standard form of a copula model with d time series models is demonstrated in Equation (6-3) as follows:

$$F_t(Y_t|F_{t-1}) = f_t\left(Y_{t,1}\middle|f_{t-1}\right), \dots, f_t\left(Y_{t,d}\middle|f_{t-1}\right), C_t[f_t\left(Y_{t,1}\middle|f_{t-1}\right), \dots, f_t\left(Y_{t,d}\middle|f_{t-1}\right)] \quad (6\text{-}3)$$

Where $F_t(.)$ denotes the cumulative distribution function at time t, $f_t(.)$ is the marginal distribution of explanatory variables at time t, and C(.) is the joint distribution function. For continuous random variables, C(.) is a unique distribution. However, the estimation of C(.) is more complex for discrete random variables (such as the frequency of collisions in this study). The detailed calculation of copula modelling for discrete random variables is discussed in (Geenens, 2020).

Most copula models assume that the dependence among a pair of variables does not change over time; however, such an assumption is not always accurate. Although several multivariate copula models, such as DCC-GARCH (Engle, 2002), account for the time-varying correlations, they are developed based on the assumption that the error term is normally distributed, which is not an appropriate assumption in the case of collision data. Among several approaches that can overcome these shortcomings of such models, the Vine copula framework stands out as one of the best.

Vine copula has various versions, including the drawable vine (D-Vine) copula, canonical vine (C-Vine) copula, and regular vine (R-Vine) copula, among other versions. The margins have the same structure in all versions, with the only difference being how the different versions make a joint distribution to establish the connection between variables. The R-Vine copula model was selected in this study since it provides more flexibility while predicting the structure of dependency among the countermeasures. In this model, arbitrary R-vines are generated in a cross-sectional tree-based structure. This means that the models could consider any combination

of the variables and investigate the dependency among them by developing sequential trees. Figure 6-1 illustrates the graphical structure of a copula model with six dimensions, assuming that the copulas in trees higher than level 3 are completely independent. To construct the trees in Figure 6-1, a few steps are executed, as follows:

1) $Tree_1$ is a tree with $Nodes_1 = \{1, \dots, 6\}$ and edge set E_1
2) For the subsequent trees, $Tree_j$ is a tree with $Nodes_j = E_{j-1}$ and edge set E_j, where $j = \{1, \dots, 5\}$
3) For the subsequent trees and $\{a, b\} \in E_j$, the proximity condition should be satisfied, which means the edges corresponding to a and b in $Tree_{j-1}$ share a common node.

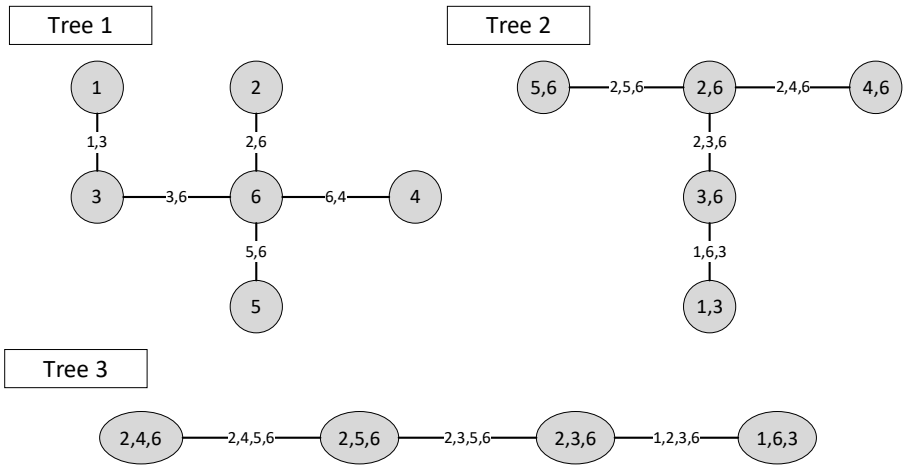


Figure 6-1 Tree structure of R-vine copula model

The model takes into account the uncertainty of parameters on the tree's chronological order. Moreover, copula models could investigate more complicated dependency structures among the variables beyond linear relationships, unlike Pearson and Spearman correlation methods that

capture the linear dependency between a pair of observations and are strongly dependent on the marginal distributions. The most common dependence measure in copula models is Kendall's τ which is only correlated with joint copula structure (Kendall, 1938).

The first step to develop the proposed R-Vine copula model is to define the marginals' distribution. The study considered that the marginals follow skew Student t stochastic volatility models, based on (Kreuzer, 2020). It should be noted that the distribution of the residuals will not be modeled as a Student t distribution. Rather, the Student t copula model only indicates that the dependency between the countermeasures is elliptical with heavy tails. In a stochastic volatility model, the log variance (s_t) of a conditionally distributed vector (Y_t) is modeled with a latent autoregressive model of order 1, as per Equations (6-4 and 6-5):

$$Y_t = exp\left(\frac{s_t}{2}\right)\epsilon_t \qquad (6\text{-}4)$$

$$s_t = \mu + \phi(s_{t-1} - \mu) + \sigma\eta_{ts} \qquad (6\text{-}5)$$

In a Bayesian version of the stochastic volatility models, the posterior distribution of the parameters will be estimated by the Markov Chain Monte Carlo (MCMC) simulation approach. Based on (Kastner, 2016), the prior densities for μ, ϕ, and σ are considered as $\mu \sim N(0, 100^2)$, $\frac{\phi+1}{2} \sim Beta(5,1.5)$, and $\sigma^2 \sim \chi_1^2$, respectively. Therefore, in a skew Student t stochastic volatility model, the data matrix with N=8 countermeasures are as Equations (6-6 and 6-7):

$$Y_{Nt} = exp\left(\frac{s_{Nt}^{st}}{2}\right)\epsilon_{Nt}^{st} \qquad (6\text{-}6)$$

$$s_{Nt}^{st} = \mu_N^{st} + \phi_N^{st}\left(s_{N(t-1)}^{st} - \mu_N^{st}\right) + \sigma_N^{st}\eta_{N(t-1)}^{st} \qquad (6\text{-}7)$$

Where: $\eta_{Nt}^{st} \sim N(0,1)$, $\mu_N^{st} \in |\ \mathbb{R}$, $\phi_N^{st} \in (-1,1)$, $\sigma_N^{st} \in (0,\infty)$, $s_{N(t=0)}^{st} | \mu_N^{st}, \phi_N^{st}, \sigma_N^{st} \sim N\left(\mu_N^{st}, \frac{(\sigma_N^{st})^2}{1-(\phi_N^{st})^2}\right)$, $\epsilon_{Nt}^{st} | \alpha_N^{st}, df_N^{st} \sim sst(\epsilon_{Nt}^{st} | \alpha_N^{st}, df_N^{st})$, $\alpha_N^{st} \in |\ \mathbb{R}$, $df_N^{st} \in (2,\infty)$ for $(t = \{1, \dots, 20\})$. Meanwhile, the prior distributions for α and df are considered as: $\alpha \sim N(0,100)$ and $df \sim N_{>2}(5,25)$.

The second step is to model the joint distribution (i.e., copula family) and define the structure of R-vine decomposition for each pair of copulas located at the edge of the tree simultaneously. In a dynamic R-vine model, a variational Bayesian model will be applied to select the structure of the trees based on the posterior distribution of the parameters. The Gibbs sampler was run for 50000 (with simulation parameters R=iteration parameter and K=thinning parameters set to 2000 and 25, respectively, and 1000 burn-in iterations were utilized). Therefore, the error of the joint distribution is structured as Equation (6-8):

$$C_{t,k}^{r} = ssT\left(y_{t,k} exp\left(-\frac{(s_{t,k}^{st})^{r}}{2}\right) \middle| (\alpha_k^{st})^r, (df_k^{st})^r\right) \qquad (6\text{-}8)$$

For $t = \{1, \dots, 20\}$ t=1,…,m, k=1,…,8, r=1,…,2000, and ssT is the standardized skew Student t distribution function.

Meanwhile, seven families were considered to estimate the copula family on each edge of the tree, including Independence, Gaussian, eGumbel, eClayton, Student t (df=2), Student t (df=4), Student t (df=8). For each copula family, there is a specific Kendall's τ which is described in detail in (Kreuzer, 2020). In most cases, the strategy of vine structure selection is based on maximizing the value of empirical Kendall's τ at each tree level, which is considered in this study as well. The value of Kendall's τ ranges from -1 (perfect inversion) to +1 (completely

positive association). A value of Kendall's τ equals to zero means the absence of association between two parameters.

A d-dimensional R-vine model will have $\frac{d!}{2}2^{\binom{d-2}{2}}$various trees. The detailed procedure of tree construction is described in (Kreuzer, 2020). A summary of the procedure is described in Algorithm (1). It should be mentioned that in constructing the first tree, the empirical Kendall's τ will be calculated for all feasible pairs and the ordering of the nodes in the first level of the tree will be determined based on the top (d-1) pairs. Once the first tree is fixed, the subsequent trees will be generated utilizing the common nodes.

Algorithm (1) Algorithm for generating samples from a dynamic R-vine copula model
Given a fitted dynamic R-vine copula model, repeat the following steps: **Step 1.** To construct the first tree (T_1): calculate the empirical Kendall's τ for all feasible pairs (x,y) with $1 \leq x \leq y \leq d$. The mechanism of T_1 is to maximize the spanning tree on the edges. **Step 2.** For each variable located on the edge of T_1, the Gibbs sampler runs for 50000 iterations. **Step 3.** Therefore, 50000 samples are generated for each parameter, the model will simulate 50000 times and the family will be selected. **Step 4.** All edges of T_1 will be given a unique node (with the highest dependency on its adjacent nodes) **Step 5.** set T_2, and go to step 1.

Utilizing Equations (2 and 3), the frequency of the collisions in each neighbourhood at time period (*t+1*) could be predicted using the information obtained at time (*t*), as shown is Equation (6-9):

$$F_{(t+1)}(Y_{(t+1)}|F_t) = f_{(t+1)}(Y_{(t+1),1}|f_t), \dots, f_{(t+1)}(Y_{(t+1),d}|f_t), C_{(t+1)}[f_{(t+1)}(Y_{(t+1),1}|f_t), \dots, f_{(t+1)}(Y_{(t+1),d}|f_t)] \quad (6\text{-}9)$$

The abovementioned procedure (i.e., defining the marginals' distributions and modelling the joint distribution) was developed to apply the proposed dynamic R-vine copula model. A skew Student t stochastic volatility model and introduced seven families (i.e., Independence, Gaussian, eGumbel, eClayton, Student t (df=2), Student t (df=4), Student t (df=8)) were considered to define the marginal distributions and model the copula function.

6.5 Data

6.5.1 Collision Records and Safety Treatments

The safety countermeasures that are approved or implemented as part of Toronto's Vision Zero action plan are visualized in a dashboard named Vision Zero Mapping Tool, administrated by the City of Toronto (Open Data Toronto, 2022). The dashboard also provides numerous information regarding the historical collision records in the city, including the collision location and the involved road users, among other information. In this study, the Vision Zero Mapping dashboard was used to extract all pedestrian-vehicle collisions that occurred in the City of Toronto between 2017 to 2021, as well as the implemented safety improvements that are related to pedestrian safety. In total, 1874 pedestrian-vehicle collisions were reported in Toronto between 2017 and 2021, resulting in 158 pedestrian fatalities and 603 serious injuries. The spatial distribution of severe pedestrian-vehicle collisions in the City of Toronto is shown in Figure 6-2. The extracted collisions were aggregated to the neighbourhood. The frequency of severe collisions (fatal and serious injury collisions) in each neighbourhood per season was estimated and used as the dependent variable in the developed models. Since 5-years of collision data were used to build

this model, a total of 20 data points were considered in the analysis (5 years of data * 4 seasons per year).

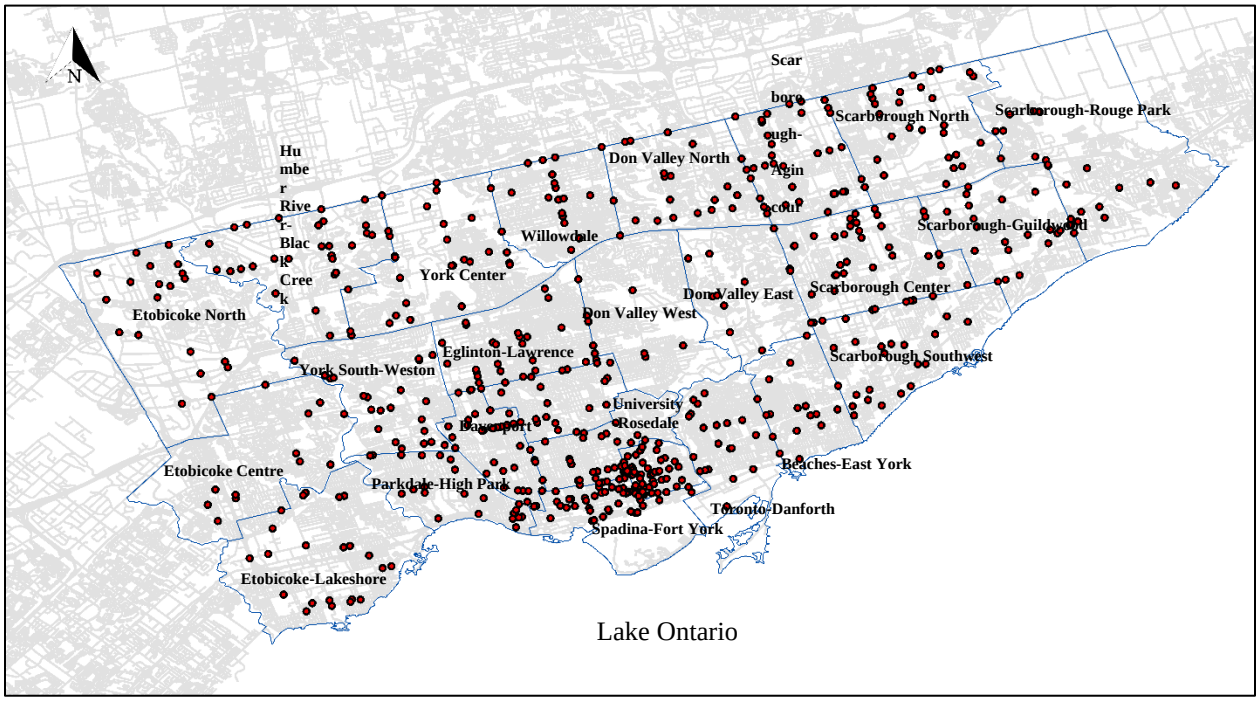


Figure 6-2 Spatial Distribution of fatal and severe injury-related collisions

As for the implemented safety treatments, they were classified into eight categories, as discussed earlier. In total, more than 5000 safety improvement projects were implemented between 2017 and 2021. Table 6-1 provides a breakdown of the safety improvement project by treatment and the number of neighbourhoods in which each treatment is implemented.

Meanwhile, the spatial distribution of the implemented safety treatments is shown in Figures 6-3 and 6-4. Specifically, Figure 6-3 shows the frequency of the implemented treatments in each neighbourhood. Figure 6-4 represents the diversity of the safety treatments in each neighbourhood. According to the figure, all neighbourhoods have received at least four safety

treatments, except a few neighbourhoods. As shown in Figure 6-3, and Figure 6-4, the treatments are mostly installed in the downtown area, the northwest, and the east side of the city, which are aligned with the distribution of pedestrian collisions.

Table 6-1 Safety Treatment Statistics

Safety Measure	Abbreviation	Number of locations receiving treatment	Number of neighbourhoods receiving treatment
Engineering Improvements	EI	120	71
Automated Enforcement	AE	167	148
Speed Limit Reductions	SLR	1186	152
Accessible Pedestrian Signal	APS	309	128
Leading Pedestrian Interval	LPI	372	149
Traffic control measures	TC	281	136
Driver Feedback Sign	DFS	655	132
Community Safety Zones	CSZ	1089	155

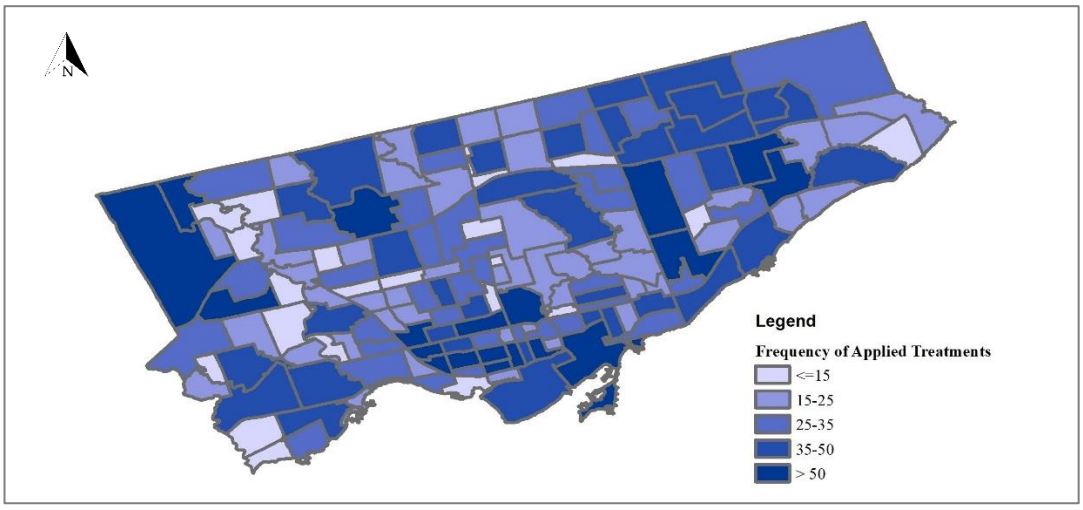


Figure 6-3 The distribution of implemented safety treatments

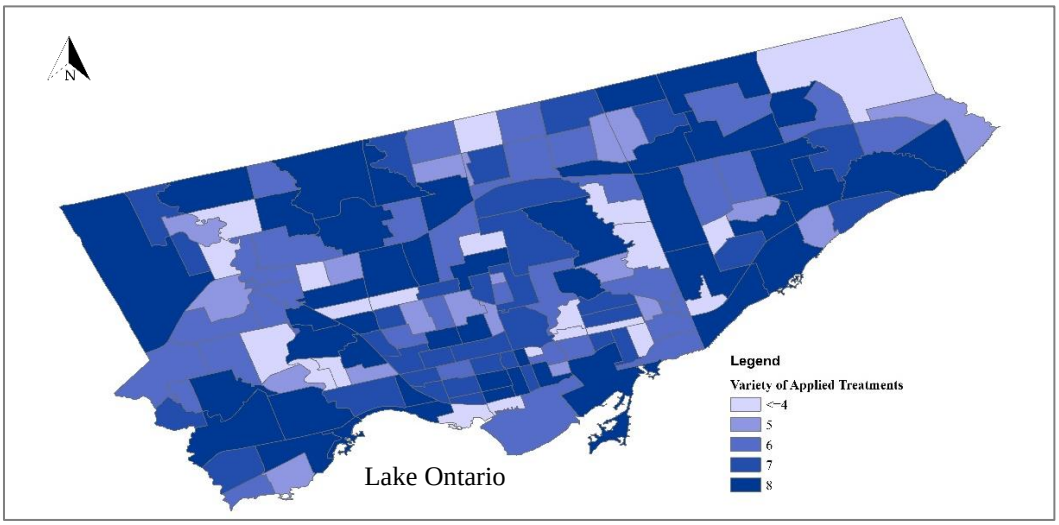


Figure 6-4 Diversity of safety treatments

6.5.2 Neighbourhood Characteristics

This study is focused on pedestrian-vehicle collisions on the neighbourhood level. The city of Toronto is divided into 158 neighbourhoods. The Open Data portal of the City of Toronto (Open Data Toronto, 2022) was utilized as the primary source of extracting the required characteristics of the neighbourhoods. The extracted characteristics can be categorized into four broad categories: exposure-related variables, built-environment factors, land use, and road network factors. All factors were normalized using the min-max normalization method so that each variable is expressed as a value between 0 and 1 before developing the model. This was done mainly to deal with variables with different scales. A descriptive summary of the extracted variables is presented in Table 6-2.

Table 6-2 Descriptive summary of the variables

Category	Variable	Min.	Max.	Mean	Std. Dev.
Exposure	Log (population density) (person/km^2)	0.86	26.74	15.9	5.71
	Household density (households/km^2)	0.12	38.81	21.565	1
Built Environment	Subway Station Density (station/km^2)	0	10.37	7.285	1.76
	School density (School/ km^2)	1.12	12.14	8.73	1.80
Land Use	Residential (% of the neighbourhood area)	0.34	13.11	8.825	0.26
	Commercial (% of the neighbourhood area)	0	4.86	4.53	0.28
	Institutional/Office (% of the neighbourhood area)	0	5.77	4.985	0.51
Road Network	Major Roads Density (km roads/km^2)	1.19	38.81	22.1	6.83
	Intersection density (Intersection/km^2)	0.18	8.84	6.61	1.16
	Sidewalk density (km sidewalk/km^2)	0.01	14.11	9.16	16.82

6.6 Results and Discussions

The dynamic R-vine copula model was developed based on the pedestrian-vehicle collisions that occurred between 2017 and 2021, while the predictive performance of the model was validated based on nine-months 2022 collision data. First, it was crucial to determine the copula family and the structure of vine decomposition for each pair of copulas located at the edge of the tree. In most cases, the strategy of vine structure selection is based on maximizing the value of empirical Kendall's τ at each tree level. The copula families were selected from the seven popular families: Independence, Gaussian, eGumbel, eClayton, Student t (df=2), Student t (df=4), and Student t (df=8). The final copula family for each pair was selected based on the AIC criteria. For example, Table 6-3 shows the selected copula families along with the corresponding parameter estimates for the Milliken neighbourhood (neighbourhood # 30).

Table 6-3 R-vine copula families along with the corresponding parameters

Tree	**Pair**	**Family**	**α**	***df***
1	(LPI,APS)	Student t	0.681	2
1	(SLR,LPI)	Student t	0.354	2
1	(CSZ,TC)	Student t	0.442	2
1	(CSZ,EI)	eGumble	0.011	-
1	(SLR,DFS)	Student t	-0.435	2
1	(SLR,CSZ)	Student t	0.272	2
1	(SLR,AE)	Student t	0.215	2
2	(EI,SLR|CSZ)	Gaussian	0.049	-
2	(TC,SLR|CSZ)	Student t	0.033	4
2	(AE,DFS|SLR)	Student t	-0.128	4
2	(AE,LPI|SLR)	Gaussian	0.117	-
2	(APS,SLR|LPI)	eGumble	0.022	-
3	(EI,AE|CSZ,SLR)	Student t	0.017	8
3	(TC,AE|CSZ,SLR)	Gaussian	0.265	-
3	(EI,LPI|CSZ,SLR)	Student t	0.019	8

In addition, Figure 6-5 represents the graphical representation of the tree of the dynamic R-vine copula model obtained based on the 8-dimensional safety treatment dataset, in the same neighbourhood. It should be noted that the vine copula tree reached level 3 while investigating the dependencies among the combination of countermeasures. This means that all potential dependencies among the countermeasures were completely covered until Tree 4 and the dependency index is equal to zero above Tree 3. Since Milliken neighbourhood is equipped with all 8 types of safety measures, 28 (8×7/2) pair copulas were estimated in total in the first tree, of which 15 were set as independent copulas.

As the R-vine structure is selected based on the maximum spanning tree, the countermeasures that are connected by an edge are highly dependent. For example, driver feedback signs (DFS)

and automated enforcement (AE) are connected to the speed limit reduction (SLR) in the first tree, which represents high dependence among these countermeasures in this neighbourhood.

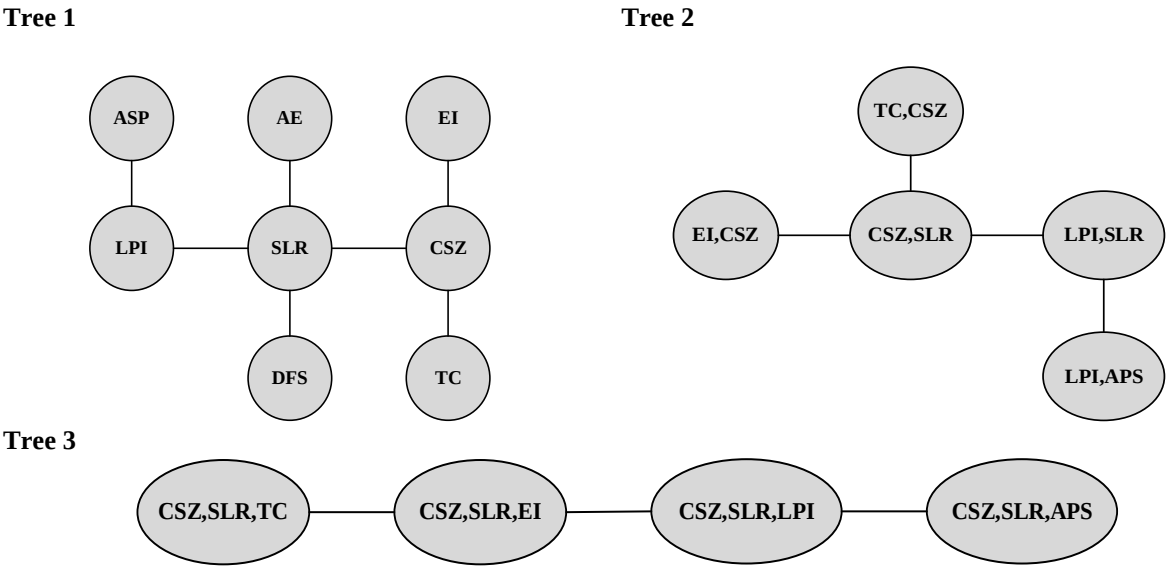
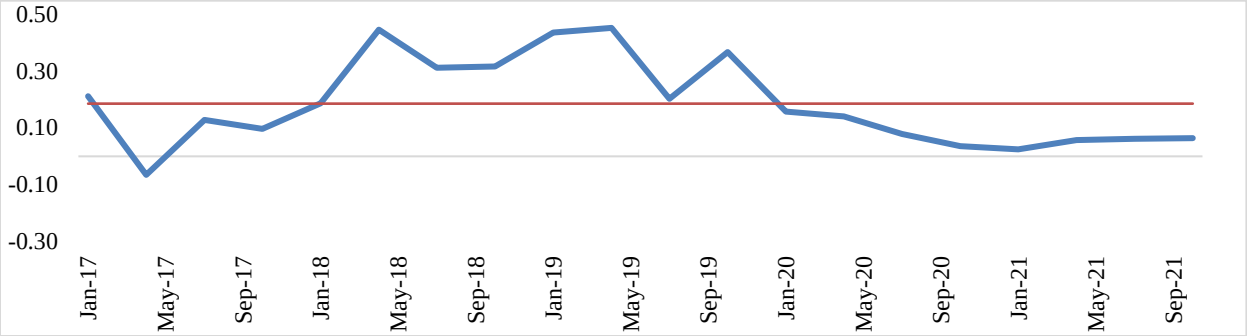


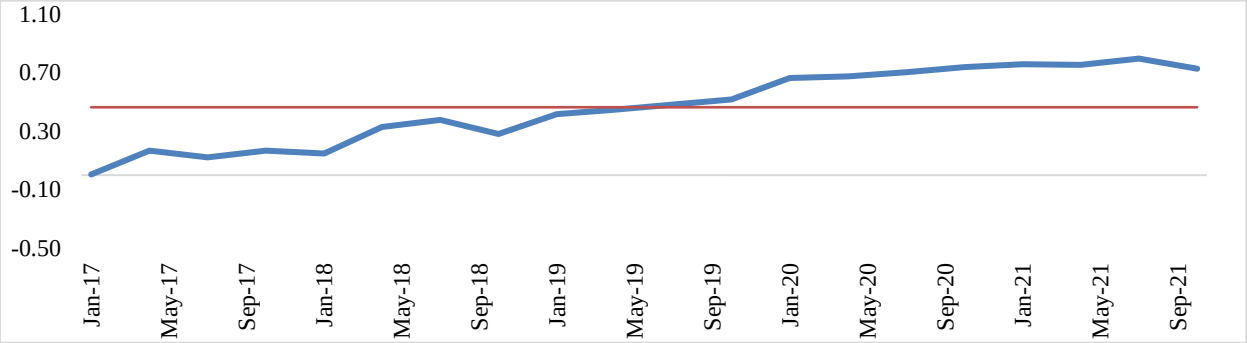
Figure 6-5 The generated trees of the R-Vine copula model in Milliken neighbourhood

Afterwards, the Kendall's τ of each countermeasure combination is estimated over the analysis period to investigate the dependency between two countermeasures and how it changes over time in each neighbourhood. As an example, the estimated Kendall's τ for a pair of countermeasures in each tree level in the Milliken neighbourhood is represented in Figure 6-6. The blue line represents the estimation of Kendall's τ at time t, while the orange line demonstrates the average value of Kendall's τ estimation. It should be noted that Kendall's τ value over 0.3 is considered to represent a strong positive relationship.

a- Level 1: LPI/APS



b- Level 2: EI,SLR|CSZ



c- Level 3: EI,AE|CSZ,SLR

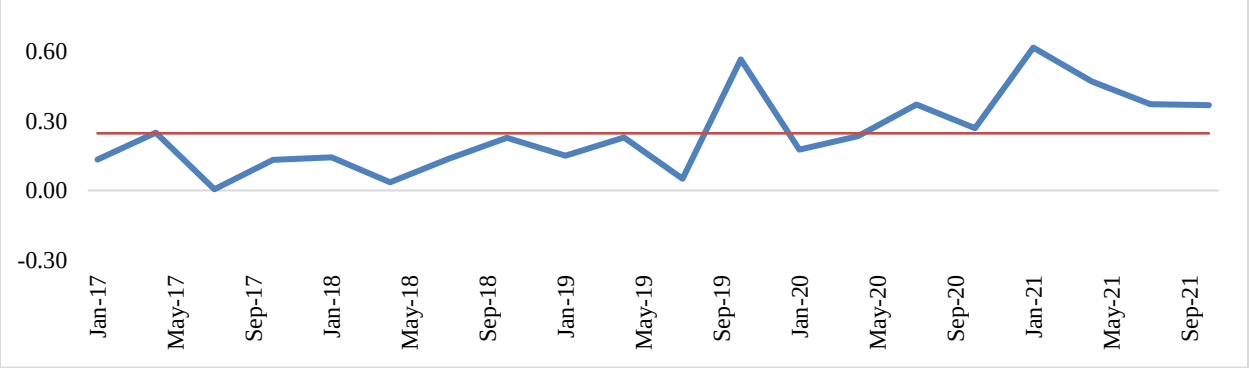


Figure 6-6 A sample of time-varying dependence (in Tree 1) and conditional time-varying dependence for a pair of copula (in Tree 2 and Tree 3)

According to Figure 6-6-a, the dependency between leading pedestrian intervals and accessible pedestrian signals in this neighbourhood was strong in 2018 and most of 2019. However, the dependency between these two countermeasures dropped in 2020 and 2021, which indicates that in the long term, the two countermeasures may not be effective together in reducing the frequency of severe collisions. On the other hand, the Kendall's τ for engineering improvements and speed limit reduction, conditional on the implementation of community safety zone (i.e., implementing engineering improvements and speed limit reduction at locations where community safety zones are already implemented) showed a significant growing trend, which indicates the long-term effectiveness of such combination (Figure 6-6-b). Moreover, Figure 6-6-c indicates that engineering improvements and automated enforcement conditional on speed limit reduction and community safety zone did not start to show a strong positive relationship until after about 3 years of the implementation.

In practice, the proposed model can be used to determine the most effective countermeasure combination that achieve the highest reduction in pedestrian collisions in each neighbourhood. To showcase the applicability of the proposed model, pedestrian-vehicle collisions that occurred in the first three seasons of 2022 in the City of Toronto were collected. In total 245 pedestrian-vehicle collisions are reported in during the nine-month period, leading to 21 fatalities and 69 severe injuries. In addition, the new safety treatments that are installed in 2022 were also obtained. In total, 368 new safety treatments were introduced to 87 neighbourhoods in 2022, as shown in Table 6-4.

Table 6-4 Countermeasures installed in the City of Toronto in 2022

Countermeasure	CSZ	TC	LPI	SLR	AE	APS	DFS	EI
Total number of installations	10	23	197	56	47	21	4	10

First, the model was used to predict the frequency of collisions for the first three seasons of 2022 based on the marginal information and the developed copula structure of the past five years (2017-2021). This was done to validate the model prediction power on a city-wide level. The model showed high accuracy in predicting collision frequency in the city's neighbourhoods, as confirmed by the average value of RMSE, MAE, and MAPE (0.3097, 0.3011, and 4.4960, respectively). Afterwards, the model was used to test the impact of the different pairs of countermeasures on the frequency of serious pedestrian collisions in the future. As an example, Table 6-5 shows the impact of the different combinations in five neighbourhoods in the city, in terms of collision reduction over the nine-month period. As shown in the table, the impact of the different combinations of countermeasures varies significantly among neighbourhoods, based on the characteristics and the countermeasures that are already installed in the neighbourhoods. For example, integrating leading pedestrian interval and accessible pedestrian signals was shown to be the most effective combination to be installed in Wellington Place neighbourhood, with an expected reduction of collision by 7.5 collisions for the last season considered in this study (July-September) in 2022. Nevertheless, the same combination was the least effective combination in Downsview neighbourhood.

Table 6-5 Collision prediction reduction based on each combination of countermeasures

Countermeasures	N 154 West Humber-Clairville	N 13 Fenside-Parkwoods	N 42 Wellington Place	N 8 Downsview	N 21 Woburn North
(TC,APS)	-4.325	-3.271	-5.155	-1.419	-2.611
(LPI,APS)	-5.062	-2.220**	**-7.511***	-1.223**	-4.019
(DFS, SLR)	-3.527	-4.274	-2.835	-5.474	-4.011
(DFS,CSZ)	-1.645	-3.328	-2.691**	-2.712	-2.782
(AE,SLR)	-3.789	-2.822	-4.672	-3.672	-2.317
(EI,LPI)	-1.022**	-4.516	-4.415	-3.367	-5.402
(TC,LPI)	-6.022	-4.751	-3.192	-5.412	-2.172
(AE,CSZ)	-4.341	-5.749	-2.912	-4.335	-1.363
(SLR,LPI,CSZ)	-1.423	-5.105	-3.136	-2.457	-1.482
(EI,SLR,AE)	-3.281	-3.769	-4.891	-2.168	-3.453
(AE,DS,CSZ)	-5.528	-3.567	-4.257	-4.123	-6.098
(TC,LPI,CSZ)	-3.13	**-7.338***	-4.502	-5.563	**-6.213***
(EI,AE,CSZ,SLR)	**-6.764***	-5.244	-4.621	-5.236	-3.562
(TC,LPI,CSZ,SLR)	-3.562	-5.152	-5.542	**-7.56***	-1.216**
(AE,DFS,CSZ,SLR)	-2.564	-3.562	-3.087	-4.911	-4.567

*** Most effective combination in neighbourhood**
**** Least effective combination in neighbourhood**

Figure 6-7 illustrates the combination of countermeasures with the highest impact in each neighbourhood. According to the figure, the most frequent combination of countermeasures was speed limit reduction, leading pedestrian intervals, and community safety zones. Implementing this combination is expected to reduce the frequency of severe pedestrian collisions significantly in the downtown area and in the east of the city, where pedestrian traffic is higher. The integration of engineering improvements, automated enforcement, speed limit reduction, and community safety zones was found to be the most effective combination mainly in neighbourhoods close to major highways (Gardiner Express and 401).

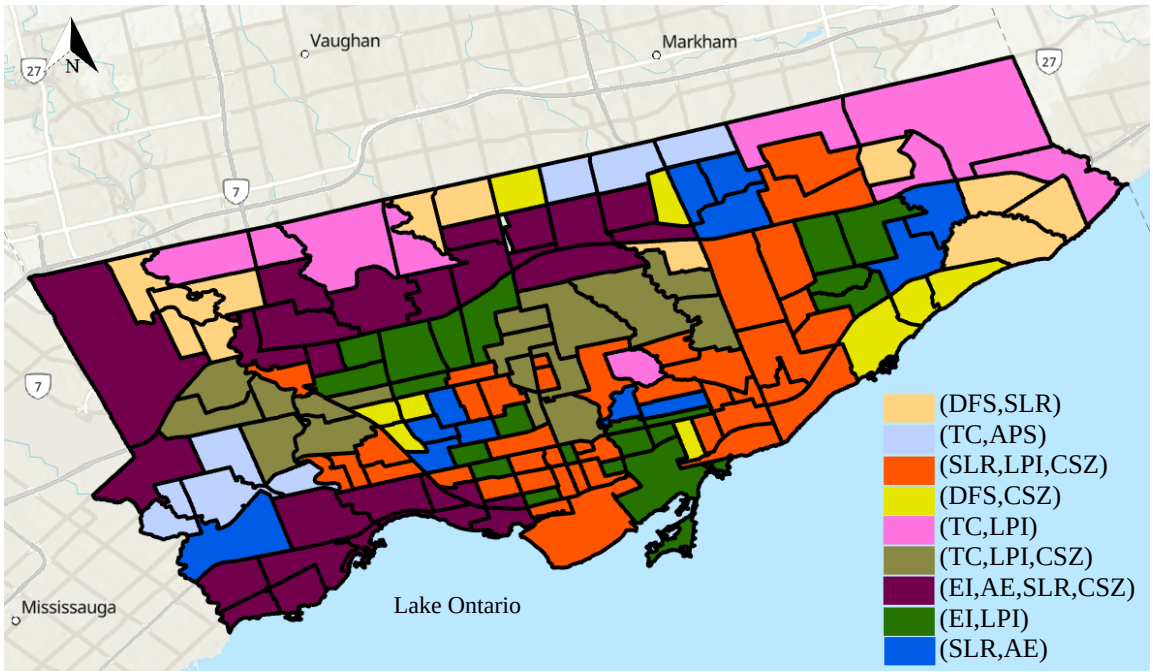


Figure 6-7 Distribution of the most effective combination of countermeasures

Moreover, integrating traffic calming and other engineering-related countermeasures with leading pedestrian intervals was found to be impactful in neighbourhoods near highway 401 and

the part of downtown is filled up with subway stations. Driver feedback signs were not included in any combination of countermeasures in neighbourhoods located near the downtown area. The impact of such treatment was notable when integrated with countermeasures such as speed limit reduction and community safety zones in the southeast part of the city, which is characterized by lower population and higher industrial densities. Finally, traffic control and leading pedestrian interval were found to be the most influential countermeasures in a few neighborhoods (pink color in the figure). However, integrating these countermeasures with community safety zones was shown to be effective in a larger number of neighbourhoods.

Finally, a logistic regression model was utilized to investigate the association between neighbourhood characteristics and the performance of the safety treatments. To develop the model, the following procedures were applied in each neighbourhood: 1) all feasible combinations of countermeasures that were identified by the R-Vine model in each neighbourhood were extracted; 2) the time-series model (i.e., Equation 2) was applied to the extracted combinations to estimate the reduction in severe pedestrian collisions in each neighbourhood; 3) A countermeasure combination is recognized to be effective in the neighbourhoods with the highest expected reduction in collisions (top 15% of the neighbourhoods) based on the results of the previous step, and 4) a binary logistic regression model was applied on each countermeasure combination, with the dependent variable being a binary variable (1 if the countermeasure combination exists is recognized as effective in neighbourhood (i), and 0 otherwise. The neighbourhood characteristics were considered the

independent variables of the models. Therefore, the probability of the outcome was estimated using Equation (6-10), as follows:

$$Y = logit(P) = ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{m=1}^{M} \beta_m X_m + \varepsilon \qquad (6\text{-}10)$$

In this equation, the logit is the natural algorithm of the odds or the likelihood ratio that the dependent variable is 1 as opposed to 0. X_m is the value of mth independent variable (i.e., neighbourhood features), β_m is the corresponding coefficient for mth independent variable, ε is the error term. The details of the implementation of this model can be found in (Sarkar et al., 2011). Table 6-6 presents the results of the developed models. The highlighted cells in Table 6-6 indicate the neighbourhood characteristics that were deemed significant contributors to a countermeasure combination being effective.

The following section provides some key takeaways from the table. To start, neighbourhoods with a high density of schools benefited the most from combining automated enforcement with either speed limit reduction or community safety zones. These treatments help to mitigate some risky behaviours of drivers, such as speeding and red-light running, which can be effective in reducing the frequency of serious pedestrian-vehicle collisions in these areas. Adding other treatments, such as engineering improvements, to those three treatments showed some safety benefits, but they did not lead to significant enhancement of pedestrian safety around school zones.

Table 6-6 Results of binary logistic regression model

Sidewalk density (km/km^2)	**P_Value**	0.016	0.036	0.014	0.005	0.004	0.007	0.001
	Std. Error	1.985	2.456	1.311	0.846	0.346	1.653	1.875
	Coeff.	2.219	2.01	0.384	0.814	0.582	1.081	0.117
Intersection density (Intersection/km^2)	**P_Value**	0.011	0.027	0.005	0.003	0.002	0.001	0.003
	Std. Error	1.136	2.853	2.067	1.007	1.128	1.766	1.769
	Coeff.	0.786	1.553	0.015	0.502	0.244	1.314	0.101
Major roads density (km/km^2)	**P_Value**	0.017	0.004	0.003	0.005	0.008	0.012	0.003
	Std. Error	1.893	2.317	2.187	1.186	1.357	0.984	1.645
	Coeff.	0.612	0.067	0.183	0.172	3.481	1.383	0.593
(% of the neighbourhood	**P_Value**	0.004	0.002	0.002	0.041	0.033	0.011	0.038
	Std. Error	1.222	1.327	1.484	1.576	1.095	1.342	0.286
	Coeff.	1.328	2.058	1.926	0.312	0.776	0.066	0.118
Commercial (% of the neighbourhood area)	**P_Value**	0.002	0.023	0.001	0.046	0.027	0.041	0.032
	Std. Error	1.342	2.513	0.583	2.451	1.732	1.160	1.233
	Coeff.	0.31	1.411	0.032	0.103	0.099	1.732	0.104
Residential (% of the neighbourhood area)	**P_Value**	0.006	0.001	0.004	0.043	0.005	0.015	0.021
	Std. Error	1.876	2.045	2.631	1.675	1.123	2.065	1.385
	Coeff.	1.584	2.742	1.937	0.589	0.773	0.086	0.174
School density (School/ km^2)	**P_Value**	0.011	0.003	0.022	0.001	0.008	0.074	0.018
	Std. Error	1.234	1.168	1.348	2.139	2.094	1.165	2.817
	Coeff.	0.316	0.155	0.083	0.074	2.744	0.109	0.137
Subway station density (station/km^2)	**P_Value**	0.007	0.045	0.006	0.001	0.035	0.016	0.001
	Std. Error	1.924	2.138	1.932	2.810	0.075	0.168	1.445
	Coeff.	1.717	2.141	0.401	0.313	0.911	0.054	0.11
Household density (household/ km2)	**P_Value**	0.003	0.024	0.001	0.037	0.002	0.001	0.002
	Std. Error	0.135	0.946	0.817	0.419	0.312	0.367	0.213
	Coeff.	1.321	0.645	2.111	0.651	0.602	0.094	0.511
Log (population density) (person/km^2)	**P_Value**	0.001	0.050	0.004	0.002	0.002	0.003	0.001
	Std. Error	0.221	1.103	0.316	0.392	0.116	1.463	0.550
	Coeff.	2.353	3.589	0.539	0.416	0.712	0.053	0.194
Countermeasures		**TC,APS**	**LPI, APS**	**DFS, SLR**	**DFS, CSZ**	**AE,SLR**	**EI, LPI**	**EI, CSZ**

0.027	0.032	0.045	0.001	0.002	0.006	0.001	0.002
1.034	1.034	1.034	1.842	2.084	2.683	0.596	1.862
0.133	0.058	2.165	2.484	0.411	0.682	0.072	0.112
0.004	0.008	0.004	0.028	0.004	0.002	0.048	0.039
1.065	1.564	1.098	1.066	2.764	1.654	1.965	0.362
0.009	2.411	0.553	2.102	1.113	0.357	0.05	2.569
0.007	0.035	0.002	0.004	0.002	0.005	0.048	0.032
1.704	1.563	2.054	2.091	1.566	1.746	1.975	2.942
2.081	1.011	0.647	3.117	0.618	1.356	0.081	2.216
0.050	0.006	0.022	0.001	0.001	0.028	0.015	0.017
1.764	1.046	2.583	1.752	3.011	1.854	1.543	1.124
0.164	1.985	1.722	0.333	1.812	0.811	0.138	0.13
0.014	0.005	0.009	0.012	0.032	0.007	0.011	0.034
1.116	1.067	1.547	1.327	1.763	2.063	2.512	1.432
0.134	2.019	1.874	0.114	0.142	0.055	2.109	1.33
0.048	0.003	0.002	0.007	0.006	0.021	0.005	0.003
1.044	1.354	2.178	1.238	2.064	1.983	1.658	1.054
0.213	0.918	2.861	0.219	1.656	2.054	1.151	0.178
0.029	0.005	0.013	0.011	0.001	0.050	0.001	0.003
2.054	2.715	3.026	1.059	2.231	1.463	2.062	1.374
2.182	0.552	1.783	0.314	0.216	1.028	0.115	0.329
0.001	0.002	0.002	0.001	0.001	0.002	0.007	0.001
2.093	1.395	0.128	0.176	1.053	2.118	1.578	1.873
0.461	0.743	2.912	0.45	0.217	2.067	1.117	0.222
0.001	0.022	0.002	0.002	0.001	0.002	0.002	0.001
0.984	0.583	0.341	0.108	0.174	0.685	1.218	0.185
0.089	0.119	2.645	0.204	0.627	0.029	2.164	1.312
0.001	0.033	0.004	0.041	0.001	0.003	0.003	0.002
0.913	0.565	0.436	0.221	0.173	0.146	0.626	0.111
0.156	1.658	3.025	0.383	2.593	0.102	0.196	0.466
AE, CSZ	**LPI,SLR**	**SLR,LPI, CSZ**	**EI,SLR ,AE**	**AE,DFS, CSZ**	**EI,AE,CSZ ,SLR**	**TC,LPI,CSZ ,SLR**	**AE,DFS,CSZ, SLR**

Moreover, Lead pedestrian intervals (LPI) have gained recent popularity as an effective countermeasure at intersections with a high frequency of pedestrian-vehicle conflicts. The results presented in Table 6-6 showed that treatment combinations that involve LPI demonstrated high safety benefits in neighbourhoods that attract higher pedestrian activities, such as neighbourhoods with higher population density, subway station density, commercial, and institutional/office density. In these areas, frequent interactions between pedestrians and turning vehicles are more common. Thus, LPI can be very effective in enhancing pedestrian safety at intersections. For example, two combinations (LPI with accessible pedestrian signals) and (LPI with speed limit reduction and community safety zones) were found to be the most effective treatments in neighbourhoods with higher population and subway station density. (LPI with accessible pedestrian signals) and (LPI with speed limit reduction) were found to be the most effective treatment in neighbourhoods with higher institutional/office density. In neighbourhoods with a high density of commercial areas, measures related to speed limit reduction and traffic signals were found to be the most effective treatments. Specifically, the combinations of (LPI with speed limit reduction) and (LPI with speed limit reduction, community safety zones, and traffic control) were the most effective in these neighbourhoods.

In addition, driver feedback signs (DFS) were found effective in neighbourhoods with a high density of intersections, but only when combined with automated enforcement, community safety zones, and speed limit reduction. It appears that driver feedback sign aid drivers in identifying potential hazards and controlling their speed at busy intersections, which improve the overall safety level in these areas.

Finally, engineering-based improvements seem to be effective in neighbourhoods with a high density of major roads and sidewalks. Combining engineering-based improvements with automated enforcement and speed limit reduction was found to be among the most effective treatment combinations in these neighbourhoods. Engineering-based improvements, which include geometric improvements to the road network elements and enhancement of the pedestrian network, are especially important in neighbourhood with a high density of sidewalks and major roads (i.e., higher pedestrian exposure at major roads). Collisions that occur in these areas are more likely to be severe. Also, these areas typically experience a high frequency of pedestrian violations (jaywalking and crossing on red). Accordingly, improving pedestrian network connectivity and road geometry would help mitigate such behaviours and improve safety.

6.7 Conclusions

This study integrated a dynamic copula R-vine model and a binary logistic regression model to analyze the safety impacts of multiple safety treatments that were implemented in the City of Toronto as part of the city's Vision Zero action plan. The main goal of the study was to investigate the effectiveness of the different safety treatment combinations and assess the relationship between neighbourhood characteristics and the efficiency of the different treatment combinations. The proposed model addressed the interdependency among the countermeasures and considered the temporal trends of the efficiency of different treatments. The results indicated that the effect of different combinations of countermeasures varies between neighbourhoods based on neighbourhood characteristics. The combination of speed limit reduction, leading

pedestrian intervals, and community safety zones was the most frequent combination with the highest efficiency among different neighbourhoods. The results indicated that enforcement and speed limit reduction are the most effective treatments to be implemented in neighbourhoods with high school density, while LPI was very effective in neighbourhoods with high population density, subway station density, commercial, and institutional/office density, especially when integrated with speed limit reduction and community zones. Driver feedback signs (DFS) were found effective in neighbourhoods with a high density of intersections, but only when combined with automated enforcement, community safety zones, and speed limit reduction. Engineering-based improvements seem to be effective in neighbourhoods with a high density of major roads and sidewalks, especially when combined with automated enforcement and speed limit reduction.

The findings of this study would assist engineers and decision-makers in ranking the safety treatment combination based on their effectiveness, selecting the most effective treatment in a neighbourhood based on the countermeasures that are already installed, and deciding on the treatment combinations that can be installed in an area based on the area characteristics. Nevertheless, several future directions are recommended for future studies. First, while macro-level analysis is useful to guide the implementation of countermeasures, micro-level analysis can be conducted to understand the association between treatment combinations and the unique characteristics of intersections and road segments. Moreover, future studies can develop a cost-benefit analysis based on the long-term impact of the safety treatments to guide future safety investments. Finally, future studies can investigate the potential spatial autocorrelation between

collisions in different neighbourhoods, which may impact the performance of the different treatments.

6.8 Reference

Innovation Solutions for econometric analysis, forcasting & simulation. Retrieved from https://www.eviews.com/home.html

Boodlal, L., Garimella, D., Weissman, D., & Shahum, L. (2021). Lessons Learned from Development of Vision Zero Action Plans. U.S. Department of Transportation Federal Highway Administration.

Damsere-Derry, J., Ebel, B. E., Mock, C. N., Afukaar, F., Donkor, P., & Kalowole, T. O. (2019). Evaluation of the effectiveness of traffic calming measures on vehicle speeds and pedestrian injury severity in Ghana. Traffic injury prevention, 20(3), 336-342.

Fayish, A. C., & Gross, F. (2010). Safety effectiveness of leading pedestrian intervals evaluated by a before–after study with comparison groups. Transportation Research Record, 2198(1), 15-22.

Fridman, L., Ling, R., Rothman, L., Cloutier, M. S., Macarthur, C., Hagel, B., & Howard, A. (2020). Effect of reducing the posted speed limit to 30 km per hour on pedestrian motor vehicle collisions in Toronto, Canada-a quasi experimental, pre-post study. BMC public health, 20(1), 1-8.

Geenens, G. (2020). Copula modeling for discrete random vectors. Dependence Modeling, 8(1), 417-440.

Gitelman, V., Carmel, R., & Pesahov, F. (2020). Evaluating Impacts of a Leading Pedestrian Signal on Pedestrian Crossing Conditions at Signalized Urban Intersections: A Field Study. Frontiers in Sustainable Cities, 2(45).

Goughnour, E., Carter, D. L., Lyon, C., Persaud, B., Lan, B., Chun, P., . . . Signor, K. (2018). Safety Evaluation of Leading Pedestrian Intervals on Pedestrian Safety. No. FHWA-HRT-18-060). United States. Federal Highway Administration. Office of Research, Development, and Technology.

Heydari, S., Miranda-Moreno, L. F., & Fu, L. (2014). Speed limit reduction in urban areas: A before–after study using Bayesian generalized mixed linear models. Accident Analysis and Prevention, 73, 252-261.

Hussein, M., Sayed, T., El-Basyouny, K., & de Leur, P. (2020). Investigating safety effects of wider longitudinal pavement markings. Accident Analysis & Prevention, 142, 105527.

Juselius, K. (2006). The cointegrated VAR model: methodology and applications. Oxford university press.

Kreuzer, A. (2020). Bayesian time series modeling with copula structures. (Doctoral dissertation, Technische Universität München).

Lütkepohl, H. (1993). Introduction to Multiple Time Series Analysis. 2nd ed. Berlin: Springer-Verlag.

Ma, J., Shang, Y., & Zhang, H. (2021). Application of Bayesian Vector Autoregressive Model in Regional Economic Forecast. Complexity, 2021.

Mammen, K., Shim, H. S., & Weber, B. S. (2020). Vision Zero: Speed Limit Reduction and Traffic Injury Prevention in New York City. Eastern Economic Journal, 46, 282-300.

McCulloch, R. E., & Tsay, R. S. (1994). Bayesian analysis of autoregressive time series via the Gibbs sampler. Journal of Time Series Analysis, 15(2), 235-250.

Nagler, T., Krüger, D., & Min, A. (2022). Stationary vine copula models for multivariate time series. Journal of Econometrics, 227, 305-324.

Nashad, T., Yasmin, S., Eluru, N., Lee, J., & Abdel-Aty, M. A. (2016). Joint Modeling of Pedestrian and Bicycle Crashes Copula-Based Approach. Transportation Research Record: Journal of the Transportation Research Board, No. 2601, 119-127.

Patton, A. J. (2012). A review of copula models for economic time series. Journal of Multivariate Analysis, 110, 4-18.

Phares, A. C., Hossen, M. A., & Dey, K. (2021). The Impact of Vision Zero Initiatives on Road User Safety in New York City. Mountaineer Undergraduate Research Review, 6(1), 11.

Rana, T. A., Sikder, S., & Pinjari, A. R. (2010). Copula-Based Method for Addressing Endogeneity in Models of Severity of Traffic Crash Injuries: Application to Two-Vehicle Crashes. Transportation Research Record: Journal of the Transportation Research Board, 2147, 75-87.

Rogers, J. M., Dey, S. S., Retting, R., Jain, R., Liang, X., & Askarzadeh, N. (2016). Using automated enforcement data to achieve vision zero goals: A case study. In 2016 IEEE International Conference on Big Data (Big Data) (pp. 4029-4031). IEEE.

Rothman, L., Ling, R., Hagel, B. E., Macarthur, C., Macpherson, A. K., Buliung, R., . . . Howard, A. W. (2022). Pilot study to evaluate school safety zone built environment interventions. Journal of transport and health.

Sims, C. A. (1980). Macroeconomics and Reality. Econometrica, 48(1), 1-48.

The City of Toronto. (2022, July). Retrieved from https://www.toronto.ca/services-payments/streets-parking-transportation/road-safety/vision-zero/vision-zero-plan-overview/

The City of Toronto. (2022, July). Retrieved from https://www.toronto.ca/services-payments/streets-parking-transportation/road-safety/vision-zero/safety-measures-and-mapping/

Toronto's Open Data Portal. (2022, July). Retrieved from https://open.toronto.ca/

Wang, K., Bhowmik, T., Yasmin, S., Zhao, S., Eluru, N., & Jackson, E. (2019). Multivariate copula temporal modeling of intersection crash consequence metrics: A joint estimation of injury severity, crash type, vehicle damage and driver error. Accident Analysis and Prevention, 125, 188-197.

Wu, L., Geedipally, S. R., & Pike, A. M. (2018). Safety evaluation of alternative audible lane departure warning treatments in reducing traffic crashes: an empirical Bayes observational before–after study. Transportation Research Record, 2672(21), 30-40.

Yeo, J., Lee, J., Cho, J., Kim, D. K., & Jang, K. (2020). Effects of speed humps on vehicle speed and pedestrian crashes in South Korea. Journal of safety research, 75, 78-86.

Zegeer, C. V., Richard Stewart, J., Huang, H., & Lagerwey, P. (2001). Safety effects of marked versus unmarked crosswalks at uncontrolled locations: analysis of pedestrian crashes in 30 cities. Transportation research record, 1773(1), 56-68.

Zhai, G., Xie, K., Yang, D., & Yang, H. (2022). Assessing the safety effectiveness of citywide speed limit reduction: A causal inference approach integrating propensity score matching and spatial difference-in-differences. Transportation Research Part A, 157, 94-106.

CHAPTER 7

Conclusion and Future Research

7.1 Summary

The research presented in this dissertation aims at investigating pedestrian violation behaviours and the impact of such dangerous behaviours on pedestrian safety. First, a comprehensive literature review and a meta-analysis were conducted to identify the contributing factors to pedestrian violations and their impact on pedestrian safety levels. Second, two micro-level studies were undertaken to analyze collisions that involved pedestrian violations at intersections and mid-blocks, respectively. The two studies investigated the impact of different attributes on collisions that happened due to pedestrian violations, including built-environment characteristics, amenities and attractions at collision locations, land uses, and road-related features. Third, a macro-level study was conducted to analyze collisions that involve pedestrian violations on the TAZ level. The study applied deep learning techniques to identify collision-prone zones that experience a high frequency of these collisions and understand their characteristics. Finally, the thesis proposed a novel approach to evaluate the safety benefits of various countermeasures that are implemented as part of vision zero programs to enhance pedestrian safety levels in urban areas. In addition, the study introduced an approach to identify the most effective combination of treatments in an area based on the area characteristics. The thesis presented several advanced statistical models and analytical techniques that address various statical issues related to collision data, which were the primary source of data in this thesis.

7.2 Conclusions and Recommendations

This section summarizes the conclusions of each chapter and the recommendations provided by the different studies of the thesis.

7.2.1 Conclusions and recommendations of Chapter 2

The research presented in chapter 2 provided a holistic review of pedestrian violation behaviour in order to develop a solid understanding of the factors that contribute to violations and how violations impact pedestrian safety (i.e., Objective 1). The study utilized a Text Mining method to identify all related studies and conducted a meta-analysis to assess the impact of the different factors on the frequency of pedestrian violations. Meanwhile, the study identified the locations that experience a high frequency of violations, the dominant research methods used to study pedestrian violations, the relationship between violations and safety, and the different strategies that can be adopted to mitigate this behaviour.

The results of the meta-analysis showed that there is a consensus in the literature regarding the positive association between the frequency of pedestrian violations and many factors, including longer waiting times at signalized crosswalks, longer block sizes, and the presence of schools and bus stops near crossing locations. As well, the majority of previous studies agreed that crowded locations that have high traffic volume, a high percentage of heavy vehicles, and a high number of lanes usually experience a lower frequency of pedestrian violations.

Nevertheless, the meta-analysis highlighted that the literature is inconclusive regarding the impact of many factors on pedestrian violations, particularly, vehicle speed, the presence of

refuge islands, countdown signals, weather conditions, and many pedestrian attributes (e.g., age, and group size).

In terms of mitigation strategies that could be applied to mitigate pedestrian violations and related collisions, the research investigated four categories, including the implementation of engineering countermeasures to reduce the frequency of violations, enhancing enforcement, initiating educational programs and public campaigns to change pedestrian behaviour, and developing innovative technology-based solutions that enable the detection of violating pedestrians and warn drivers and pedestrians of the potential risk. While the literature did not provide, for the most part, a quantitative assessment of the efficiency of those mitigation strategies, some general findings were reported, summarized as follows:

- Several engineering-based treatments were found effective in reducing the frequency of pedestrian violations and related collisions, including, for example, the installation of physical barriers and pedestrian call buttons at mid-blocks, developing proper signal timing that minimizes pedestrian waiting time at signalized intersections, and eliminating on-street parking at locations that experience a high frequency of violations.
- Previous studies highlight the potential benefits of educational programs and public campaigns in increasing public awareness of the serious consequences of reckless crossing practices. The most effective programs were installing posters at unsafe locations and educational programs at schools and campus universities.
- The automated detection of violators and the advanced warning of drivers and/or the violating pedestrians of the potential hazard was promoted as a potential solution to

mitigate pedestrian violations and reduce the severity of the consequences of such behaviour.

This research presented in chapter 2 could assist researchers to conduct more lucrative research in the area of pedestrian violations and put more emphasis on under-developed areas. Also, the results will help transportation engineers, planners, and decision-makers to develop better design concepts to mitigate the frequency and severity of violations and enhance pedestrian safety.

7.2.2 Conclusions and recommendations of Chapter 3

The research presented in chapter 3 proposed an integrated clustering and copula-based model to evaluate the impact of intersection characteristics on both frequency and severity of collisions that happened due to pedestrian violations. Specifically, A Latent Class Analysis (LCA) method was applied to divide the collision dataset utilized in the study into a set of homogeneous clusters, based on traffic and intersection characteristics. Then, a copula-based multivariate model for each cluster in order to study the impact of the different factors on collisions under the prevailing conditions of each cluster. The study provides valuable insights for a better understanding of the safety consequences of pedestrian violations. The main findings of the study are summarized as follows:

- The frequency of collisions that involve pedestrian violations is strongly correlated with the presence of bus stops and schools near intersections.
- Increasing the frequency of transit buses helps reduce the total and fatal collisions that involve pedestrian violations at intersections.

- Larger intersections with central refuge islands are more likely to experience collisions happened due to pedestrian violations.
- The study showed an inverse relationship between intersection size and the frequency of collisions that involve violations. Smaller intersections were found to experience a higher frequency of road collisions that involve violations, while larger intersections that are well-served by transit buses tend to have a lower rate of fatal collisions that involve violations.

The research conducted in chapter 3 provided several recommendations that aim at mitigating the safety consequences of pedestrian violations, summarized as follows:

- Transit service operational parameters (particularly, the bus frequency) and the location of bus stops at major intersections should be selected not only based on transit-related factors but also to minimize the impact of such variables on pedestrian behaviour and safety.
- The design of the walking infrastructure at intersections that are located near school zones or along major bus routes needs to be properly evaluated in light of its impact on pedestrian behaviour.
- Large and major intersections that have central refuge islands need to be equipped with other countermeasures to mitigate pedestrian violations, which maximizes the safety benefits of refuge islands.

7.2.3 Conclusions and recommendations of Chapter 4

The study presented in chapter 4 investigated the impact of a variety of factors on the frequency and severity of collisions that involve pedestrian spatial violations at mid-blocks. The study utilized the Structural Equation Modeling (SEM) approach to undertake the analysis. The results of the study provide valuable insights for a better understanding of the factors that encourage pedestrians to engage in spatial violations and increase the risk of collisions at mid-blocks, along with the impact of road and pedestrian network characteristics on pedestrian behaviour and safety. Such understanding assists transportation engineers and planners to develop better design concepts to mitigate the frequency and severity of collisions that are caused by pedestrian spatial violations in urban areas. The key findings of the study are summarized as follows:

- Access to services (particularly, bike-share stations, trail access points, restaurants, and parking lots) were found to be among the most influential factors that increase the frequency of collisions that involve spatial violation at mid-blocks.
- The lack of pedestrian network connectivity and large block size were found to be highly correlated with the frequency and severity of pedestrian collisions that involved spatial violations.
- The study confirmed that mid-blocks with bike-share stations that are located near bus stops increase the probability of spatial violations and exacerbate pedestrian safety levels.
- Violation-related collisions were found more likely at locations that have central refuge islands and a low number of lanes.

Moreover, the study provided several recommendations that aim at mitigating collisions that involve pedestrian spatial violations at mid-blocks, summarized as follows:

- The proper selection of bike-share stations and applying appropriate countermeasures at such locations to mitigate pedestrian spatial violations are essential.
- Locations with poor pedestrian network connectivity and large block size require countermeasures that discourage pedestrian spatial violations.
- Reasonable block sizes, proper connectivity of the pedestrian network, and ensuring that pedestrians can access their desired destination in the shortest possible distance are essential measures to consider when planning new areas.
- Locations with central refuge islands should be investigated to select appropriate countermeasures that aim at reducing the frequency of spatial violations.

7.2.4 Conclusions and recommendations of Chapter 5

The study conducted in chapter 5 investigated pedestrian-vehicle collisions (both total collisions and those that occur due to pedestrian violations) on the macro-level (Traffic analysis zone level). The study proposed a deep learning model to identify hotspot zones that experience a high frequency of collisions that involve pedestrian violations and understand the unique characteristics of such zones. The study provides a better understanding of pedestrian safety on the macro-level and aids engineers and planners in developing specific planning recommendations to enhance safety in unsafe areas where a high frequency of pedestrian violations is observed. The study also aids planners in designing pedestrian-friendly networks

and developing specific mitigation strategies to enhance safety in unsafe zones. The conclusions of the study are summarized as follows:

- The proposed deep learning model identified collision-prone zones with a high accuracy that exceeded the traditional Bayesian approach, based on several consistency tests conducted in the study.
- Intersection density was found to be the most important factor in distinguishing between collision-prone and non-collision-prone zones, based on both total collisions and collisions that involve pedestrian violations.
- The unsupervised deep learning technique identified five zonal variables as the key variables that can differentiate between collision-prone and non-collision-prone zones based on the total pedestrian-vehicle collisions, namely, the intersection density, the pedestrian network directness (represented by the average edge length), the proportion of residential land uses, and road user exposure parameters (VKT and PKT).
- Five zonal variables were also identified as the key variables that distinguish between collision-prone and non-collision-prone zones based on collisions that involve pedestrian violations, including intersection density, the density of bike-share stations and parking lots in a TAZ, pedestrian network directness (represented by linearity), and the proportional residential land uses.

The study recommended that deep learning models should be used in the context of macro-level safety analysis as they outperform traditional methods in identifying collision-prone locations. The study also recommended that creating proper connectivity in pedestrian

networks should be a priority in zones that have pedestrian violation problems as it is key in reducing the frequency of violations and mitigating the subsequent collisions. Finally, the study highlighted the importance of the proper selection of the location of bike share stations and parking lots in a TAZ as they were strongly associated with collisions that involve pedestrian violations. The allocation of these facilities should be optimized considering the location of pedestrian attractions and public transit stations across the TAZs.

7.2.5 Conclusions and recommendations of Chapter 6

The research presented in chapter 6 applied a dynamic R-vine copula-based time series model to evaluate the efficiency of safety measures implemented as a part of Vision Zero programs. The proposed model was applied on a macro-level (the neighbourhood level) to evaluate the effectiveness of various safety treatments and identify the most effective combination of treatments in each neighbourhood, based on the neighbourhood characteristics and the previously implemented treatments in the neighbourhood. The main findings of the study were as follows:

- The effect of different combinations of treatments varies between neighbourhoods based on neighbourhood characteristics.
- The combination of speed limit reduction, leading pedestrian intervals (LPI), and community safety zones was found to be the most efficient combination in most neighbourhoods.
- Automated enforcement and speed limit reduction were the most effective treatments in neighbourhoods with high school density.

- Lead Pedestrian Interval (LPI) was effective in neighbourhoods with high densities of subway stations and office density, especially when integrated with speed limit reduction and community zones.
- Driver feedback signs (DFS) were found effective in neighbourhoods with high intersection density, but only when combined with automated enforcement, community safety zones, and speed limit reduction.
- Engineering-based improvements are effective in neighbourhoods with a high density of major roads and sidewalks, especially when combined with automated enforcement and speed limit reduction.

To the best of the author's knowledge, the study presented in chapter 6 is the first study to propose a methodology to evaluate different treatments implemented as part of Vision Zero rather than evaluating the program as a whole. As such, this research could assist decision-makers in selecting safety treatments in a location based on its characteristics, guide future implementation of vision zero programs in other municipalities and conduct accurate cost/benefit analysis of the installed treatments.

7.3 Limitations and Future Works

The research presented in this dissertation utilized statistical techniques to provide a solid understanding of pedestrian violation behaviour and its impact on pedestrian safety. Nevertheless, the research is subject to several limitations, some of which are discussed in this section. First, the lack of an accurate measure of pedestrian exposure is a key limitation in all studies presented in this dissertation. To overcome this issue, the number of walkers at collision

locations (or zone) was used as a surrogate measure of pedestrian exposure. While the number of walkers is commonly used as a surrogate measure for pedestrian exposure in the safety literature, it is not an accurate representation of exposure, which may impact the accuracy of the results. A more accurate representation of pedestrian exposure is needed to increase the reliability of the results presented in this dissertation. Second, the lack of information related to the vehicle operating speed was another limitation that limited the ability to understand the impact of many factors on collisions that involve pedestrian violations. Integrating the operating speed in the developed models could provide valuable insight regarding the impact of many factors, specifically, transit-related factors and time of day, on pedestrian violation behaviour.

In addition, while the dissertation provided a novel approach to evaluate the safety benefits of countermeasures that are implemented as part of vision zero, the proposed models were applied on the macro level. This did not enable the investigation of the effect of many location-specific characteristics on the results. A micro-level analysis is needed to understand the association between the effectiveness of treatment combinations and the unique characteristics of intersections and road segments. Finally, the research relied mainly on historical records of pedestrian-vehicle collisions. Despite the valuable insights attained through the analysis, collision data are known to have several limitations related to data quality and reliability. Relying on other sources of data, such as traffic conflicts collected from video data at the studied locations, would enable a better understanding of the factors that encourage pedestrians to violate and the failure mechanism that led to collisions.

Moreover, in light of the findings of this dissertation, several future research directions can be recommended. For example, future research can investigate the use of other data sources to capture many explanatory variables with higher accuracy. This can include, for example, the use of cell phone data to capture real-time information regarding pedestrian waiting time before crossing and crossing speed. Future studies can also investigate the problem of the optimal allocation of several facilities that were shown to have an impact on pedestrian violations and related collisions. This may involve developing an optimization framework for the allocation of bike share stations, bike racks, parking lots, and bus stops to balance the accessibility of such facilities and pedestrian safety issues. Finally, future research is encouraged to conduct before-and-after analysis to evaluate the impact of the different treatments, geometric features, and planning concepts presented in this dissertation on reducing predation violations and the risk of collisions.