UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TRANSPORTES

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TRANSPORTES

MESTRADO ACADÊMICO EM ENGENHARIA DE TRANSPORTES

FRANCISCO ALTANIZIO BATISTA DE CASTRO JUNIOR

A CAUSAL INFERENCE ANALYSIS OF INJURY SEVERITY IN MOTORCYCLIST CRASHES

FORTALEZA - CE

2023

FRANCISCO ALTANIZIO BATISTA DE CASTRO JUNIOR

A CAUSAL INFERENCE ANALYSIS OF INJURY SEVERITY IN MOTORCYCLIST CRASHES

M. Sc. Dissertation presented to the Programa de Pós-Graduação em Engenharia de Transportes from Centro de Tecnologia from Universidade Federal do Ceará, as a partial requirement to obtain the Master in Engenharia de Transportes. Field of study: Planejamento e Operação de Engenharia de Transportes.

Supervisor: Flávio José Craveiro Cunto, PhD

FORTALEZA – CE

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C351c Castro Junior, Francisco Altanizio Batista de.
A causal inference analysis of injury severity in motorcyclist crashes / Francisco Altanizio Batista de Castro Junior. – 2023.
113 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Transportes, Fortaleza, 2023.
Orientação: Prof. Dr. Flávio José Craveiro Cunto.

1. motorcyclists. 2. causal inference. 3. severity. 4. road safety. 5. structural equations modeling. I. Título.

CDD 388

FRANCISCO ALTANIZIO BATISTA DE CASTRO JUNIOR

A CAUSAL INFERENCE ANALYSIS OF INJURY SEVERITY IN MOTORCYCLIST CRASHES

M. Sc. Dissertation presented to the Programa de Pós-Graduação em Engenharia de Transportes from Centro de Tecnologia from Universidade Federal do Ceará, as a partial requirement to obtain the Master in Engenharia de Transportes. Field of study: Planejamento e Operação de Engenharia de Transportes.

Approved on: 27/09/2023.

EXAMINATION BOARD

_____
Prof. Flávio José Craveiro Cunto, PhD (Supervisor)
Universidade Federal do Ceará (UFC)

_____
Prof. Francisco Moraes de Oliveira Neto, DSc.
Universidade Federal do Ceará (UFC)

_____
Prof. Sara Maria Pinho Ferreira, PhD.
Faculdade de Engenharia da Universidade do Porto (FEUP)

ACKNOWLEDGEMENTS

To God.

To my close family, Francisco Altanizio and Lucia Castro, and my brother, Kalil Castro, for the unconditional love and emotional support.

To Prof. Flávio Cunto, PhD., for guiding my academic life with exemplary mastery.

To the examination board (Prof. Sara Maria Pinho Ferreira, PhD., and Prof. Francisco Morae de Oliveira Neto, DSc.) for all recommendations to this work.

To my amazing friends Andrezza and Edgar, for all their support.

To all my friends that I made in my life and from PETRAN, for the all-nighters and the moments of sharing. In particular to Aldaianny, Beliza, Bruno, Harley, Diego, Kaio, Lucas Moreira, Lira, Nilso e Renata.

To my coworkers, Camila Maia and Raquel Chaves, and to my boss, Carlos Henrique, for allowing me to develop professionally.

To Claudiane Carvalho for all the support in my life and work.

To my research group, “Grupo de Pesquisa em Segurança Viária”. In particular to Gabriela Martins, Caio Torres, Vanessa Xavier, and Paulo Bruno for all their contributions to my work.

This study was financed by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

RESUMO

A proporção de fatalidades envolvendo motociclistas tem aumentado nos países latinos e asiáticos nos últimos anos. A Segunda Década de Ação para a Segurança Viária 2021-2030 preconiza que a visão sistêmica do Sistema Seguro (SS) é o meio para alcançar zero fatalidades. Os estudos tradicionais da severidade dos sinistros geralmente utilizam uma abordagem baseadas em dados com uma única regressão, podendo não representar adequadamente essa visão. O objetivo principal desta dissertação é desenvolver uma análise causal dos fatores que influenciam a severidade dos sinistros a partir da perspectiva do SS. A consolidação do conhecimento das abordagens de inferência causal em estudos de segurança viária da gravidade dos sinistros envolvendo motociclistas é necessária, portanto, um exemplo simulado de Monte Carlo foi elaborado para compreender as características dessa nova abordagem. Posteriormente, uma extensa revisão da literatura foi realizada para a formulação de uma representação conceitual baseada na abordagem do Sistema Seguro para o processo causal da severidade dos sinistros envolvendo motociclistas. Hipóteses causais foram formuladas com base na representação conceitual e nos dados coletados de sinistros envolvendo motociclistas nas rodovias brasileiras. O modelo causal foi estimado para avaliar a relação entre o uso de álcool e a severidade dos sinistros viários nas rodovias federais do estado do Ceará utilizando a metodologia SEM. Por fim, os resultados apontaram uma relação significativa ao nível de 90% para o álcool em vias urbanas. Outros resultados indicaram que fins de semana e horas noturnas estão associadas ao uso de álcool, enquanto veículos pesados, colisões frontais e horas fora do pico estão diretamente associadas a maiores severidades devido à maior energia de impacto. Além disso, foram encontradas diferenças nas relações em áreas rurais e urbanas. O uso de uma abordagem causal possibilitou obter resultados mais confiáveis, ao controlar variáveis de confusão e utilizar-se de um arcabouço teórico incorporando o Sistema Seguro. Ademais, essa abordagem permitiu obter, em um único modelo, as interrelações entre as variáveis do estudo, o que propicia um maior entendimento dos fatores associados com a severidade. Por fim, os resultados obtidos podem auxiliar os tomadores de decisão a elaborarem um plano de ação capaz de alcançar um sistema seguro para os motociclistas.

Palavras-chave: motociclistas; inferência causal; severidade; segurança viária; modelos de equações estruturais.

ABSTRACT

The proportion of motorcyclist fatalities has increased in Latin and Asian countries in recent years. The Second Decade of Action for Road Safety 2021-2030 has shown that the systematic view of the Safe System (SS) is the means to achieve zero fatalities. Traditional studies often use a data-driven approach and a single regression, which may not adequately represent this view. The main aim of this dissertation is to develop a causal analysis for motorcyclists from the Safe System perspective using observational data. To consolidate the knowledge of causal inference approaches in road safety studies of the motorcyclists' severity, Monte Carlo Simulations were elaborated to understand the characteristics of this new approach. Subsequently, an extensive literature review was carried out to formulate a conceptual model based on the Safe System approach for the causal process of motorcyclists' injury. Causal hypotheses were formulated based on the conceptual model and the data collected from motorcyclist crash data on Brazilian highways. The causal model was estimated to evaluate the relationship between alcohol use and the severity of road crashes on federal highways in the state of Ceará, using the SEM methodology. Finally, the results indicated a significant relationship between alcohol use and severity in urban roads. Other findings suggested that weekends and nighttime hours are associated with alcohol use, while heavy vehicles, head-on collisions, and off-peak hours are directly associated with higher severities due to increased crash energy. Furthermore, differences were found in the relationships in rural and urban areas. The use of a causal approach allowed for obtaining more reliable results by controlling confounding variables and incorporating a theoretical framework that includes the Safe System approach. Moreover, it enabled obtaining, in a single model, the interrelationships among the study variables, leading to a better understanding of road safety. Finally, the obtained results can assist decision-makers in developing an action plan capable of achieving a safe system for motorcyclists.

Keywords: motorcyclists; causal inference; severity; road safety; structural equations modeling.

LIST OF FIGURES

Figure 1 – Dissertation outline 15
Figure 2 – Swiss Cheese model 19
Figure 3 – Speeding Behavior Framework 20
Figure 4 – Wedagama's framework 20
Figure 5 – Swedish Transport Agency's Safe System Framework 22
Figure 6 – Queensland's Safe System Framework 22
Figure 7 – Factors that affect the severity of motorcycle crashes 30
Figure 8 – Chains 34
Figure 9 – Direct and indirect effects (mediation analysis) 35
Figure 10 – Forks 36
Figure 11 – Collider 38
Figure 12 – Effect modification 39
Figure 13 – SEM 42
Figure 14 – Method 47
Figure 15 – Formulation of the causal model 49
Figure 16 – The relationship of interest of the theoretical example 53
Figure 17 – The simplified conceptual model 54
Figure 18 – DAG of the theoretical example 55
Figure 19 – Model 1 56
Figure 20 – Model 2 57
Figure 21 – Model 3 57
Figure 22 – Model 4 58
Figure 23 – Model 5 58
Figure 24 – Model 6 59
Figure 25 – Model 7 60
Figure 26 – Model 8 60
Figure 27 – Model 9 61
Figure 28 – Results of Monte Carlo Simulation on SCM with logit links 61
Figure 29 – DAG and d-sep 62
Figure 30 – Example of SEM, observed (a), estimated (b), and residues (c) matrix 64
Figure 31 – Conceptual model of motorcyclist severity 66
Figure 32 – The causal hypothesis 68
Figure 33 – Motorcycle age and Severity 71

Figure 34 – Motorcycle gender and Severity 71
Figure 35 – Safe Users causal process with observed data 73
Figure 36 – Safe Vehicles causal process with observed data 74
Figure 37 – Safe Roads causal process with observed data 75
Figure 38 – Safe Speeds causal process with observed data 75
Figure 39 – Environmental causal process with observed data 76
Figure 40 – Specific factors causal process with observed data 77
Figure 41 – Tetrachoric correlation among observed variables 79
Figure 42 – *Subjective Norms* relationships 80
Figure 43 – Causal Model 84

LIST OF TABLES

Table 1 – Logit studies in motorcycle crashes 18
Table 2 – Simpson's Paradox 32
Table 3 – ODDs and logit models (Y - fatal (1) and non-fatal (0)) 32
Table 4 – Chains and Monte Carlo Simulation (Y - severity) 34
Table 5 – Direct and indirect effects (Y - severity) 36
Table 6 – Fork (Y - Speed) 37
Table 7 – Collider (Y - Speed) 38
Table 8 – Effect modification 40
Table 9 – D-separation statements using chi-square tests 63
Table 10 – Variables used in the study 69
Table 11 – Relationships in the Causal Model 81
Table 12 – Spatial and Temporal autocorrelation tests 83

CONTENTS

1 INTRODUCTION 12

1.1 Problem Statement 13

1.2 Research Questions 14

1.3 Research Objectives 14

1.4 Dissertation Outline 15

2 THE CAUSAL PROCESS OF INJURY SEVERITY IN MOTORCYCLIST CRASHES 16

2.1 The traditional approaches to finding associations in motorcycle crash severity 16

2.2 Review of efforts of conceptual models of motorcycle crashes 19

2.2.1 The Safe System approach and factors associated with severity in motorcyclist crashes 21

2.2.1.1 Safe Speeds 23

2.2.1.2 Safe Users factors 23

2.2.1.3 Safe Vehicles factors 25

2.2.1.4 Safe Roads factors 26

2.2.1.5 Environmental factors 27

2.2.1.6 Crash specific factors 28

2.2.2 Summary 29

2.3 The paradigm of Causal Inference from the perspective of Road Safety 30

2.3.1 Pearl's theory and Directed Acyclic Graphs (DAGs) 31

2.3.1.1 Chains and Mediation: direct and indirect effects 33

2.3.1.2 Forks: confounder factors and backdoor criterion 36

2.3.1.3 Colliders and selection bias 37

2.3.1.4 Effect modification 39

2.3.1.5 DAGs in summary 40

2.3.1.6 Models based on DAGs 40

2.3.2 Rubin's theory and Propensity Score (PS) approach 44

2.3.3 Causal Inference on crash database issues 45

3 METHOD 47

3.1 Theoretical example with simulated data 47

3.2 Database of motorcycle crashes 48

3.3 Formulation of the causal model 49

3.4 Evaluate and estimate the causal model 50

3.4.1 Spatial and Temporal dependence 51

4 RESULTS 53

4.1 A theoretical example of the causal inference theory on Road Safety 53

4.1.1 Simple logit models 55

4.1.2 Mixed logit models 59

4.1.3 Graphical models 62

4.1.4 Summary 64

4.2 A conceptual model of motorcyclist severity based on the Safe Systems 65

4.3 A practical example of causal inference on Road Safety 67

4.3.1 Relationship of interest and causal hypotheses 67

4.3.2 Motorcyclist database 68

4.3.2.1 Safe users 70

4.3.2.2 Safe vehicles 73

4.3.2.3 Safe Roads 74

4.3.2.4 Safe Speeds 75

4.3.2.5 Environmental 76

4.3.2.6 Specific factors 77

4.3.2.7 Database issues 77

4.3.3 Formulation and estimate of the causal model 78

4.3.3.1 Measurement model 78

4.3.3.2 Structural model 80

4.3.3.3 Interpreting the results of the model 85

5 CONCLUSION AND FUTURE STUDIES 89

REFERENCES 92

APPENDIX A. EXPLORATORY ANALYSES 102

1 INTRODUCTION

Each year, 1.35 million people die in traffic crashes, and half of the victims are vulnerable roadway users (VRUs), such as motorcyclists (WHO, 2018). Furthermore, motorcyclist fatalities occurred nearly 29 times more than passenger vehicles when controlled for miles traveled (NHTSA, 2019).

In the United States, the National Highway Traffic Safety Administration (NHTSA) (2016) stated that motorcycle fatalities increased by 48% between 2002 and 2015. The situation is worse in East Asian and Latin American countries. In China, the average of fatal crashes involving motorcyclists rose by 64% between 2010 and 2018 (FERNÁNDEZ *et al.*, 2020). In Brazilian capitals, the proportion of motorcyclist fatalities nearly increased from 17% to 42% between 2010 and 2019 (DATASUS, 2021; WHO, 2018).

The United Nations (2020) decreed the period between 2021 and 2030 as the Second Decade of Action for Road Safety (UNITED NATIONS, 2020). The Safe System (SS) approach points to a systematic and sustainable long-term road safety strategy that is deemed to be appropriate to reach the 50% reduction in road fatalities goal by 2030 (UNITED NATIONS, 2020). The countries that adopted the Safe System approach achieved the lowest fatality rates per 100,000 inhabitants (WELLE *et al.*, 2018).

Under the SS paradigm, road fatalities and serious injuries are not acceptable. The SS principle recognizes that humans are vulnerable to crash forces in road crashes. However, this new paradigm requires an understanding of the causal relationships between the severity of the crashes and the factors associated with the SS dimensions: Roads, Speeds, Vehicles, and Road Users (ETIKA, 2018; OPAS, 2018; WELLE *et al.*, 2018). The collaboration of all components of these dimensions works towards reducing the likelihood of harm. Therefore, if any of the components fail, the remaining parts must ensure safety to prevent a serious or fatal crash (BAMBACH; MITCHELL, 2015; ETIKA, 2018; ITF, 2016).

Understanding the magnitude of the effects of the main causes associated with road users, speeds, roads, and vehicles can help decision-makers find the best way to avoid serious crashes (CUMMINGS, 2006; DUFOURNET *et al.*, 2016). Despite the benefits of using regression models such as logistic regression to understand associations between factors related to severity, several studies interpret the coefficients as a total effect (AZIMI *et al.*, 2020; CHANG *et al.*, 2016; CUNTO; FERREIRA, 2017; MORRISON *et al.*, 2019). However, it's important to note that the estimated coefficients in single regressions may have an interpretation

of association and partial effect, rather than a causal effect interpretation (WOOLDRIDGE, 2013).

Randomized studies can estimate the causal effect but are rarely used in road safety for ethical and practical reasons. The studies with observational data require a suitable approach for estimating causal effects. Therefore, Pearl (2009) proposes a method to estimate causal effects using observational studies which are common in road crash severity analysis.

Pearl's Causal Inference paradigm emphasizes the importance of identifying and controlling for confounding variables, which can bias the relationship between a cause and its effect. Background knowledge plays a crucial role in identifying potential confounders, and conceptual models (or representation) can aid in visualizing the interrelationships among factors associated with severity (PEARL, 2009; SIQUEIRA, 2020).

The Safe System (SS) approach could be instrumental in elaborating on a conceptual model, as it serves as a cornerstone for understanding the main relationships between crash severity and associated factors. Nevertheless, an extensive literature review is still needed to identify all confounders to evaluate the causal effects of an observational study.

Meanwhile, there is a lack of research that has conducted a comprehensive causal inference analysis based on a thorough literature review to examine the relationships between various factors associated with the severity of motorcyclist crashes (LAUBACH *et al.*, 2021). Additionally, many studies on road safety severity often neglect to address the issue of confounding and interpret coefficients such as total effects, which could be biased. Furthermore, these conventional studies often rely solely on available data, adopting a data-driven approach, and do not incorporate a theoretical framework to guide their analyses and interpret findings.

The causal inference paradigm provides an approach to establishing relationships among factors associated with crash severity. These relationships can be valuable in proposing action strategies to effectively reduce crash severity, empowering decision-makers to develop a comprehensive action plan toward achieving a Safe System for motorcyclists.

1.1 Problem Statement

The research problem is that road safety studies that utilize observational data and single regression models may yield skewed results due to the presence of confounding variables. Additionally, traditional models may not adequately encompass the perspective of

the Safe Systems approach, which takes into account multiple relationships among factors associated with severity.

1.2 Research Questions

The introduction of this research highlights road safety gaps, which leads to the following central question: how can the Causal Inference paradigm be effectively incorporated to assess the causal relationships among factors associated with motorcycle crash severity, taking into account the Safe Systems approach and utilizing observational data from Brazilian highways as the basis of analysis?

Despite there being studies that show how causal inference works with practical examples (LÜBKE *et al.*, 2020), the use in studies of crash severity is not fully understood. Therefore, the first specific question is: how to apply the causal inference approaches in road safety studies of motorcyclists' severity?

The causal relationships between factors and the injury severity of motorcyclists have not been satisfactorily evaluated. Nevertheless, a conceptual model of the factors associated with motorcyclists is necessary to discover the confounders. Therefore, the second specific question is: what is a conceptual model of the interrelationships among factors in motorcycle crashes incorporating the Safe Systems approach and Causal Inference?

The conceptual model is based on causal hypotheses formulated in advance by the researcher. The appropriate statistical methods are necessary to evaluate these hypotheses. It also needs a proper approach to deal with intrinsic issues of crash databases. Thus, the third specific question is: how valid are the causal hypotheses using observational studies of crash severity involving motorcyclists on Brazilian highways based on the Causal Inference paradigm?

1.3 Research Objectives

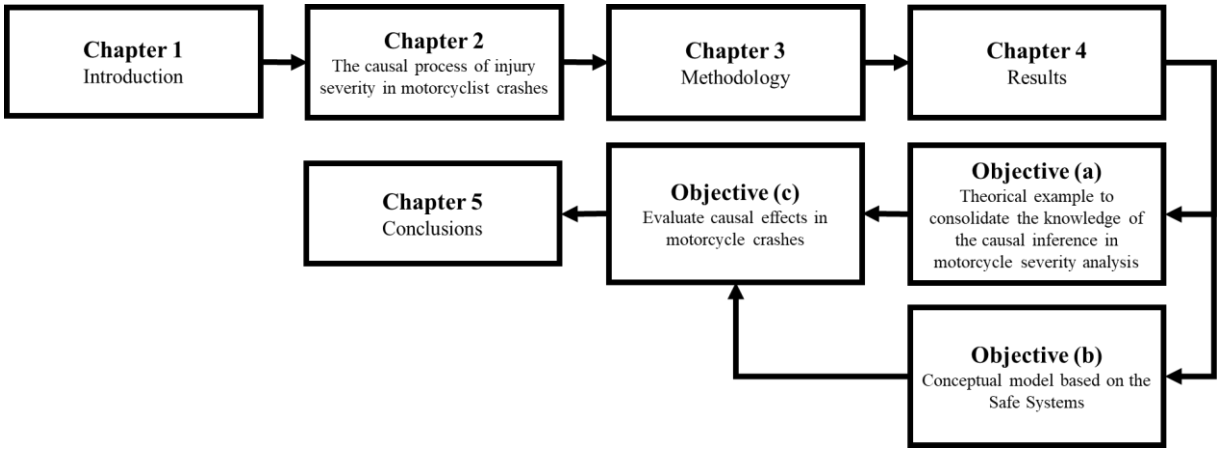
To answer the four research questions, this research is separated into one principal and three specific objectives. The main objective is to develop an analysis of cause-effect relationships between severity and factors on motorcycle crashes from the Safe Systems perspective using observational data and incorporating the Causal Inference paradigm. The specific objectives are as follows:

a) To consolidate the knowledge of the causal inference approach in road safety studies of motorcyclists' severity;
b) To propose a conceptual model based on the Safe System approach identifying the causal hypotheses of the factors impacting the severity of motorcycle crashes; and
c) To verify the causality hypotheses using motorcycle crash data from Brazilian highways.

1.4 Dissertation Outline

This dissertation is structured as follows (Figure 1). Chapter 2 presents an explanation of causal inference and the causal process of injury severity in motorcyclist crashes. Chapter 3 proposes our methodology to find the causal effects of factors using observational data. Chapter 4 presents the analysis results. Finally, Chapter 5 presents the final considerations.

Figure 1 – Dissertation outline



Source: the author.

2 THE CAUSAL PROCESS OF INJURY SEVERITY IN MOTORCYCLIST CRASHES

This chapter is structured into three sections. The first section outlines traditional approaches for identifying factors associated with motorcycle crash severity. In the second section, conceptual models of motorcycle crash severity are reviewed and the primary factors affecting it are highlighted. The third and final section examines the application of the theory of causal inference, providing examples from road safety.

2.1 The traditional approaches to finding associations in motorcycle crash severity

The traditional approaches to finding associations between the factors and severity in motorcyclist crashes are the use of categorical models such as logit and probit models. Some classifications of these models are as follows (ALNAWMASI; MANNERING, 2019; RAHMAN *et al.*, 2021):

- Binary logit/probit: This type of logistic regression is used to model the relationship between a binary outcome variable (*e.g.,* fatal and no-fatal) and one or more predictor variables;
- Ordered logit/probit: This type of logistic regression is used to model the relationship between an ordinal outcome variable (*e.g.,* uninjured < serious < fatal) and one or more predictor variables;
- Multinomial logit: the outcome has more than two categories and the categories do have not an ordered nature. Generally used when parallel lines (same slope) assumption in the ordered model has not been satisfied, or when the interest is to know how each category of a risk factor affects each severity category;
- Nested logit: when have a hierarchical dependency in outcome categories. Used when multinomial logit has independence of irrelevant alternatives (IIA) specification errors;
- Multi-collinearity logistic regression: This type of logistic regression is used when there is multi-collinearity between predictor variables.
- Regularized logistic regression: This type of logistic regression is used when there is a large number of predictor variables or when some predictor variables are highly

correlated. Regularization methods such as L1 or L2 penalization are used to address this issue.

- Logit with random parameters (mixed): it is a type of logistic regression that allows for the modeling of both fixed and random effects. Random parameters allow some predictor variables to be considered random and their effects could be estimated based on both the data and a probability distribution. This allows for the incorporation of within-group variation and can lead to more accurate predictions. It is also known as mixed-effects logistic regression.

Mixed logit models are advanced categorical models that attempt to capture the effect of unobserved heterogeneity through the inclusion of random parameters. Therefore, mixed logit models are robust as they allow for the use of variables with both fixed and random effects, following some probability distribution.

For example, making the age variable random parameters means that within an age range, the effects can change due to factors such as motorcycle driving experience, which was not observed, *i.e.*, it is not in the model. However, from a Causal Inference perspective, these unmeasured variables can be confounding variables, which can lead to biased results. It should be noted that mixed logit models are capable of capturing these effects, but they do not completely remove the bias (GUNASEKARA; CARTER; BLAKELY, 2008). Nevertheless, the studies that used mixed models frequently do not systematize what causes unobserved heterogeneity and do not explain if the unobserved factors could cause confounding effects.

Available studies generally use a single model to evaluate the effect of all factors, and they interpret the coefficients such as total effect. These coefficients are only partial effects and may be biased because of confounding or other reasons, which will be explained in the next sections. Moreover, the misinterpretation of the results is mentioned as “The Table 2 Fallacy” (WESTREICH; GREENLAND, 2013). Specifically, this fallacy refers to the practice of selecting variables or analyses based on their statistical significance, rather than on a priori hypotheses or theoretical considerations, which can lead to false conclusions and overinflated effect sizes.

To prevent biases, it is important to utilize pre-existing knowledge, which could be usually represented by a conceptual model, as a basis for the analysis. This approach can facilitate the development of causal hypotheses and the identification of confounding variables. Table 1 shows some studies that used logit models in motorcycle crash severity analyses.

Table 1 – Logit studies in motorcycle crashes

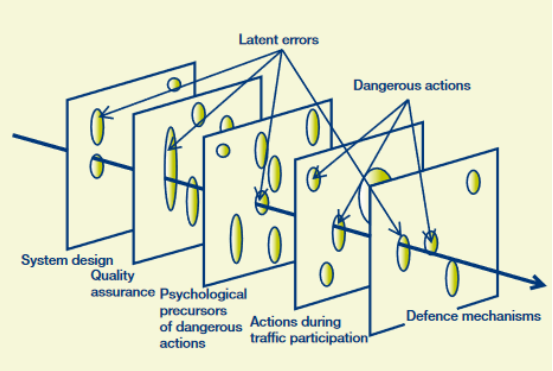
Authors	Study Location	Study Period	Sample Size	Model Used	Main Factors Found to Influence Motorcycle Crash Severity
Rahman *et al.* (2021)	Dhaka	2006–2015	316	Binary Logistic Regression	Weekends, Rainy Season, Dawn and Night Periods, Non-Intersections, Straight and Flat Roads, Highways, Hit Pedestrian-type crashes, No Defect Motorcycles, Heavier Vehicles, No Helmets, Alcohol
Li *et al.* (2021)	California	2011-2016	322	Latent Class-Ordered Probit	Single-, Two-, and Multi-Vehicle Crashes
Abdul Manan (2018)	Malaysia	2010-2012	9,176	Multinomial and mixed logit models	curve road sections, no road markings, smooth, ruts and corrugation of road surface, and wee hours
Geedipally, Turner, and Patil (2011a)	Texas	2003-2008	48,871	Multinomial Logit	Alcohol, Female Riders, Helmet Use, Old Riders
Jones, Gurupackiam, and Walsh (2013)	Alabama	2006-2010	Not Available	Multinomial Logit	Behaviors, Opponent Vehicles, Roadway Geometry
Abrari Vajari *et al.* (2020)	Australia	2006-2018	7,714	Multinomial Logit	Old Motorcyclists, Weekends, Midnight/Early Morning, Rush Hours, Give-Way, Roundabouts, Uncontrolled Intersections
Salum *et al.* (2019)	Tanzania	2013-2016	784	Multinomial Logit	Speeding, Alcohol, Horizontal Curves, Reckless Riding, Off-Peak, Violation, Riding Without a Helmet
Eustace, Indupuru, and Hovey (2011)	Ohio	2003-2007	21,914	Multinomial Probit	Alcohol/Drugs, Speeding, Single-Vehicle Crashes, Segment Roadways
Rifaat, Tay, and De Barros (2012)	Calgary	2003-2005	466	Ordered Logit	Frequent Curves, Alcohol, Speed
Cunto and Ferreira (2017)	Brazil	2004-2011	3,232	Ordered Logit	Helmet Use, Old Riders
Chung, Song, and Yoon (2014)	Korea	2007-2009	792	Ordered Probit	Heavy Vehicles, Violations, Nighttime, Speed
Ijaz *et al.* (2021)	Pakistan	2017-2019	8,770	Random Parameter Logit model	Weekdays, old riders, and heavy vehicle shock
Pervez, Lee, and Huang (2021)	Pakistan	2014-2015	28,894	Random parameter logit model	summer season, weekends, nighttime, elderly riders, heavy vehicles, and single-vehicle collisions
Islam (2022)	Florida	2012-2016	747	Random Parameter Multinomial Logit	Roadway Characteristics, Work-Zone Geometry, Urban Interstate, Large Shoulder Width, Work-Zone Types
Salum et al. and Mannering (2019)	Florida	2012-2016	1,058	Random Parameters Multinomial Logit	Temporal Instability in Risk Factors
Se *et al.* (2021)	Thailand	2016-2019	13,794	Random Parameters Ordered Probit models with heterogeneity in means	Male Riders, Improper Overtaking, Drowsiness, Four-Lane or Wider Highway, Flush and Depressed Median, Road on a Slope, Weekend, Nighttime with Light, Hitting a Van/Minibus, Rear-Ending, Side-Swiping (Rural); Barrier-Median, Crashes between 18:00 and 23:59, Hitting a Passenger Car (Urban)

Source: developed by the author using multiple studies cited in Table 1.

2.2 Review of efforts of conceptual models of motorcycle crashes

One of the first representations of road crash generation is the Swiss Cheese (Figure 2). This represents a defense model that illustrates the chronological sequence of the crash and the possible gaps ("holes or latent error") in defense layers. Swiss Cheese representation can also demonstrate the rise in of severity injuries. The defense layers ("slices of the cheese") are represented by factors related to users, roads, vehicles, and speeds, and when one of these fails there are holes in this slide (REASON, 1997; WEGMAN; AARTS; BAX, 2008). For example, a failure in the design of a curve could lead a motorcycle to get off the road and collide with a tree. If the curve had been well-dimensioned, the crash could have been avoided. Alternatively, if the motorcyclist would be speeding less, he could avoid going off the road.

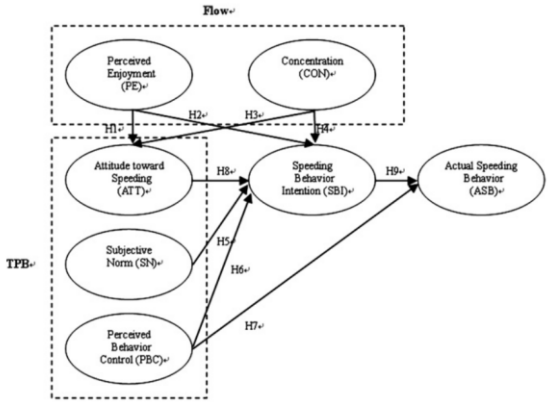
Figure 2 – Swiss Cheese model



Source: Wegman, Aarts and Bax (2008) adapted from Reason (1997).

Chen *et al.* (2011) investigated speeding behavior and other factors using a theory of planned behavior (TBP) and the SEM approach. Figure 3 shows the proposed SEM structure to investigate psychological factors that affect speeding behavior. These factors were measured using indicators of a questionnaire applied to 277 riders of heavy motorcycles. Trinh and Linh (2018) used the same framework but related the intention of helmet use to speeding behavior.

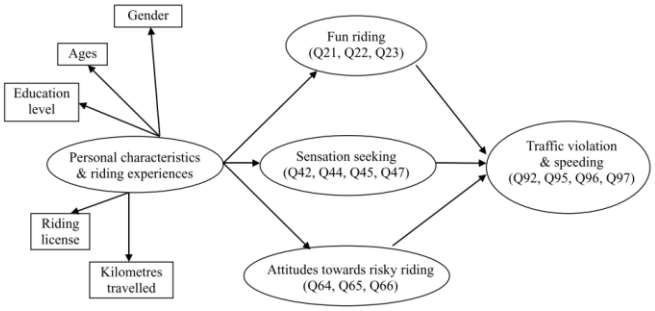
Figure 3 – Speeding Behavior Framework



Source: Chen *et al.* (2011).

Wedagama (2015) studied the intentions of traffic violations and speeding of motorcyclists. The conceptual model (Figure 4) shows that personal characteristics influence perceptions and attitudes toward riding a motorcycle, which influences future traffic violations and speeding. The author used 300 questionnaires to test the representation. The results show that sensation-seeking and attitudes are significant, and male motorcyclists are more likely to be involved in sensation-seeking situations.

Figure 4 – Wedagama's framework



Source: Wedagama (2015).

Previous theoretical representations have primarily focused on the process of road crash occurrence. Studies that solely encompass the representation of the severity of the crash are rare, meaning that the crash has already occurred and what will be conceptually represented are the factors that decrease or increase the probability of the injury. Nevertheless, the process of creating this representation may utilize the results and methods proposed in the previous studies. For example, the characteristics of motorcyclists influence their risky behavior, such as adopting higher speeds, shorter braking distances, and the use of safety equipment.

Additionally, there are aspects of subjective norms, perceived attitudes, and others that are challenging to collect, and an SEM modeling approach may be utilized, which allows for the use of latent variables, better representing the causal process of crash severity involving motorcyclists.

Nowadays, the Safe Systems approach is used in many countries as a framework to reduce the fatalities and serious injuries of crashes to zero, this approach is also related to Vision Zero. In the next section, it is presented the Safe System approach, on which the conceptual framework model of this dissertation is based.

2.2.1 The Safe System approach and factors associated with severity in motorcyclist crashes

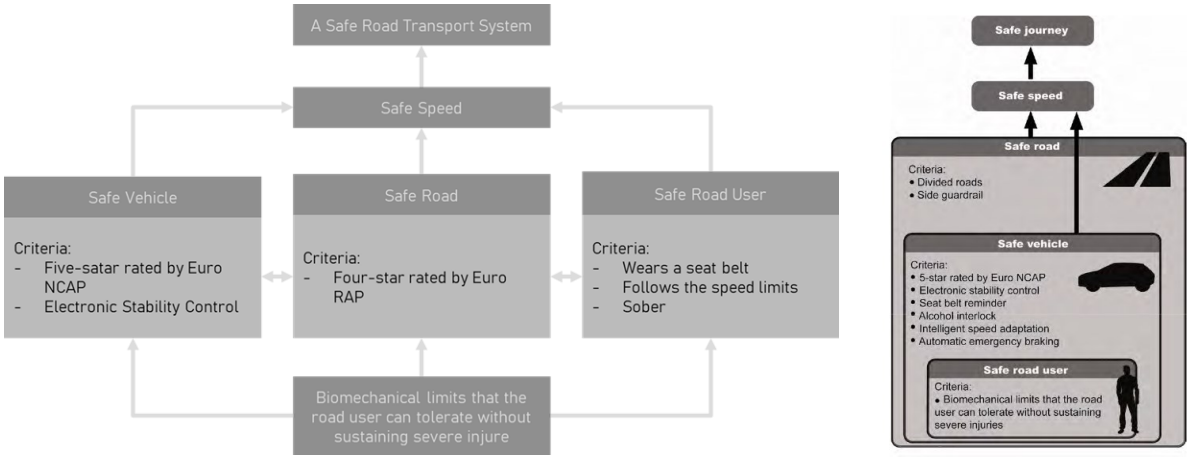
The Vision Zero strategy was developed in 1997 in Sweden to minimize deaths and serious injuries in road crashes. The Safe System approach is based on the best practices of Swedish "Vision Zero" and prior knowledge of Reason's "Swiss Cheese" model (ETIKA, 2018; ITF, 2016; STIGSON, 2009).

The Safe System approach is based on the following principles: i) road users are vulnerable to the energy of impact; ii) people make mistakes; iii) the responsibility is shared with designers, build management, and road users, and; iv) all parts of the system must be strengthened to multiply their effects (proactive approach) (ETIKA, 2018; ITF, 2016).

The Safe System approach provides a holistic view of road safety and has five main cornerstones: Post-crash care, Roads, Speeds, Vehicles, and Users. The union of each part of the system works to reduce the injury risk. Therefore, if one part of the system fails, there should be other parts that provide protection (BAMBACH; MITCHELL, 2015; ETIKA, 2018; ITF, 2016).

There are efforts to develop conceptual models of Safe Systems. The Swedish Transport Agency represented the Safe System in another way (STIGSON; KRAFFT; TINGVALL, 2008). The model (Figure 5) describes the interactions of the three components (roads, vehicles, and road users) under Safe Speed to lead to safe road traffic. The model focuses on safe speed as the most important factor affecting the safety of road users. Furthermore, the representation is based on the tolerance of the impact on road users.

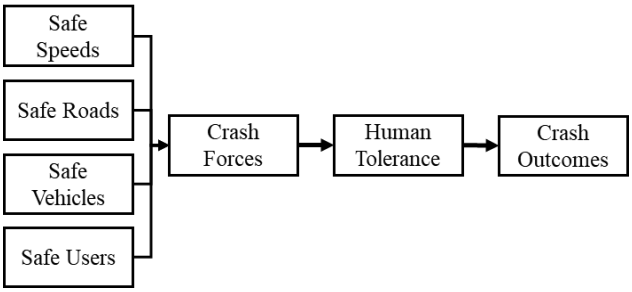
Figure 5 – Swedish Transport Agency's Safe System Framework



Source: Swedish Transport Agency (STIGSON; KRAFFT; TINGVALL, 2008).

In Queensland representation, all elements of the Safe System work together to reduce crash severity. In other words, these elements determine the energy during the crash and the severity outcome (DEPARTMENT OF TRANSPORT AND MAIN ROADS, 2015).

Figure 6 – Queensland's Safe System Framework



Source: Adapted from Department of Transport & Main Roads (2015).

In conclusion, many factors affect the severity of motorcyclists in a crash. These factors are related to speeds, users, vehicles, roads, and the environment. The energy of impact and human tolerance are the two main factors that lead to fatal and severe injuries. The integrated approach implies that road, vehicle, and user factors act simultaneously on different levels to reduce the severity of the crash. The next section presents the factors that contribute to reducing the severity of motorcycle crashes.

2.2.1.1 Safe Speeds

The fatal crash risk increases exponentially with speed. Furthermore, high speeds favor production mistakes by riders and drivers, creating scenarios that the rider does not have control over. The high speed contributes to the severity due to the kinetic energy (CAMERON; ELVIK, 2010; ELVIK; VAA, 2009; OECD, 2015; RIFAAT; TAY; DE BARROS, 2012). According to Jones, Gurupackiam, Walsh (2013) and Salum *et al.* (2019), speeding is associated with 1.3 to 3.0 times more likely to be fatal.

Motorcyclists tend to adopt excessive and inappropriate speeds more than other vehicles. Motorcycles are smaller and have more acceleration capacity, allowing them to overtake high speeds and move better into traffic (OECD, 2015). Other factors that affect speed selection are i) Type of road: the speeds are higher and there are more speed violations on rural roads; ii) age: younger riders may be more likely to engage in riskier riding behavior, while older riders may have decreased physical abilities that can impact their ability to control a motorcycle; and, iii) Network element: motorcyclists are three times more likely to speed than other vehicles (LARDELLI-CLARET *et al.*, 2005; OECD, 2015; SE *et al.*, 2021; WALTON; BUCHANAN, 2012).

2.2.1.2 Safe Users factors

Safe user factors frequently correlate with unsafe behaviors, including alcohol consumption and neglecting proper safety equipment. Moreover, the attributes of the motorcyclist contribute significantly to this dynamic. Acknowledging that speeding constitutes a perilous behavior is vital. Nevertheless, Safe System models have approached speed as an independent dimension of Safe Users (CREASER *et al.*, 2009; OECD, 2015).

The consumption of alcohol by motorcyclists is associated with an increased risk of fatal crashes (GEEDIPALLY; TURNER; PATIL, 2011b; LUNA *et al.*, 1984; OECD, 2015; RAHMAN *et al.*, 2021). Moreover, the effect on motorcyclists is greater than on other vehicles due to the complexity of riding (CREASER *et al.*, 2009; OECD, 2015). Drink-driving is associated with risky behavior such as speeding, and not wearing a helmet (ALNAWMASI; MANNERING, 2019; OECD, 2015; PEEK-ASA; KRAUS, 1996; SODERSTROM *et al.*, 1993).

Studies demonstrate that alcohol was present in 29% to 75% of motorcycle fatal crashes (ALNAWMASI; MANNERING, 2019; DRUMMER *et al.*, 2004; HOLUBOWYCZ;

KLOEDEN; MCLEAN, 1994; OECD, 2015). Rahman *et al.* (2021) studied factors associated with motorcycle severity in Dhaka and found that alcohol use is associated with an increase of two times the odds of a fatal crash. Another study presents evidence linking the absence of motorcycle helmets in fatal motorcycle crashes to the use of alcohol, marijuana, and other drugs (ROSSHEIM *et al.*, 2014).

Crashes involving young men, crashes at nighttime, at weekends, and high speeds are associated with alcohol consumption (GEEDIPALLY; TURNER; PATIL, 2011a; HOLUBOWYCZ; KLOEDEN; MCLEAN, 1994; OECD, 2015; TOPOLŠEK; DRAGAN, 2018). Furthermore, motorcyclists who are aware of the danger of alcohol are more likely to avoid speeding violations (TOPOLŠEK; DRAGAN, 2018).

The consumption of drugs by motorcyclists is another factor that should not be ignored. Motorcyclists are affected in a stronger way than other vehicles, and the consumption is also higher. Moreover, there are associations between the drugs in crashes with other factors, such as age (young), gender (men), time of day (nighttime), and day of the week (weekends) (EUSTACE; INDUPURU; HOVEY, 2011; OECD, 2015).

Younger riders often engage in more hazardous behaviors such as speeding and alcohol consumption (ALNAWMASI; MANNERING, 2019; OECD, 2015; PERVEZ; LEE; HUANG, 2021). In general, young motorcyclists have a higher crash risk because of their lack of experience and a propensity to adopt risky behaviors (CHESHAM; RUTTER; QUINE, 1993; JONES; GURUPACKIAM; WALSH, 2013; OECD, 2015; PERVEZ; LEE; HUANG, 2021).

On the other hand, the increased vulnerability of older riders, often stemming from physical fragility, results in a 22% increase in the likelihood of severe crashes (CUNTO; FERREIRA, 2017; GEEDIPALLY; TURNER; PATIL, 2011a; IJAZ *et al.*, 2021; OECD, 2015). Therefore, the association between age and severity is not linear, varying across age ranges (ALNAWMASI; MANNERING, 2019).

Another factor associated with age is experience. Experienced riders tend to have few crash risks. With more distance traveled, the motorcyclist has a lower risk per kilometer (MULLIN, 2000; OECD, 2015). Professional and experienced riders have a better hazard perception than novice and young riders (BELLET; BANET, 2012; OECD, 2015; WALI; KHATTAK; AHMAD, 2019). Nevertheless, more experienced motorcyclists do not fully imply that are safer because they could have greater self-confidence (TOPOLŠEK; DRAGAN, 2018). Furthermore, there is an increase in crash risk related to riders who do not hold a valid license (ISLAM, 2022; LARDELLI-CLARET *et al.*, 2005; LIN *et al.*, 2003; MAGAZZÙ; COMELLI; MARINONI, 2006; OECD, 2015; SMC, 2014).

In general, male riders are more likely to be involved in fatal crashes as compared to female riders. This has been attributed to the higher propensity observed for male riders to engage in risky behavior, such as speeding, racing, wheel spin, and “wheelies” while riding a motorcycle (ABRARI VAJARI *et al.*, 2020; JONES; GURUPACKIAM; WALSH, 2013; PRIYANTHA WEDAGAMA; WISHART, 2019; SALUM *et al.*, 2019; SE *et al.*, 2021; THEOFILATOS; YANNIS, 2015).

The use of protective equipment is the best way to prevent serious injury. The use of a quality helmet protects against head injuries significantly, reducing about 42% to 69% of fatal crashes (ELVIK; VAA, 2009; LIN; HWANG; KUO, 2001; OECD, 2015; SALUM *et al.*, 2019). Rahman *et al.* (2021) and Salum *et al.* (2019) found that helmet use reduces 0.5 to 0.6 times the odds of a fatal crash. Other protective clothes are gloves, boots, jackets, airbags jackets, and pants. Motorcyclists who were wearing jackets, pants, or gloves were 20% to 60% less likely to be hospitalized (DE ROME *et al.*, 2011; OECD, 2015). Other factors are fluorescent or bright clothing to improve visibility (OECD, 2015; WALI; KHATTAK; AHMAD, 2019).

2.2.1.3 Safe Vehicles factors

Mechanical deficiencies in motorcycles can increase the chances of severe crashes, with tire and brake issues manifesting in 12% of cases (RECHNITZER; HAWORTH; KOWADLO, 2000). Though safer motorcycle selection is a common practice among cautious riders, it is vital to acknowledge that motorcycles cannot alone confer complete protection during collisions (OECD, 2015).

Despite motorcycle defects potentially leading to an increased likelihood of crashes, Rahman *et al.* (2021) found that crashes involving motorcycles without defects are three times more likely to be fatal (RAHMAN *et al.*, 2021). On the other hand, another study found that the age of the vehicle is not associated with severity (GEEDIPALLY; TURNER; PATIL, 2011b).

The characteristics of the motorcycle can affect control and encourage risky behaviors. Larger engine sizes can cause likely more fatal and severe crashes (PAI, 2009; WASEEM; AHMED; SAEED, 2019). Nevertheless, other studies found no relation between severity and engine size (MÖLLER *et al.*, 2020; NGUYEN-PHUOC *et al.*, 2019). The type (sport, tourism, trail, etc.) can lead to a risk of crashes. Moreover, sport bikers are associated

with non-incapacitating injuries (ALNAWMASI; MANNERING, 2019; SAVOLAINEN *et al.*, 2011). The characteristics of the vehicle are related to the other motorcyclists' factors, such as their behavior (BJØRNSKAU; NÆVESTAD; AKHTAR, 2012; OECD, 2015; TEOH; CAMPBELL, 2010).

Advanced Braking Systems (ABS) in motorcycles is a technology that improves the stability when braking for motorcyclists. ABS makes a positive contribution to safety, reducing possibly the risk of a crash or the severity. Studies show that this system could avoid about 25% of fatal crashes (OECD, 2015; RIZZI; STRANDROTH; TINGVALL, 2009; SMC, 2014; TEOH, 2011). Other technologies are not available in third countries, such as motorcycle stability control, speed alert, curve and frontal collision warning, tire pressure monitoring, and e-Call (OECD, 2015). In general, newer and more expensive motorcycles have more safety-enhancing technologies.

2.2.1.4 Safe Roads factors

Road factors are responsible for about 8% of crashes involving motorcyclists (OECD, 2015). Approximately 30% of crashes involving motorcyclists occur in or after a curve. Curves poorly dimensioned, with small radii, are more prone to crash risk and more severity (ACEM, 2009; EUSTACE; INDUPURU; HOVEY, 2011; GEEDIPALLY; TURNER; PATIL, 2011a; RIFAAT; TAY; DE BARROS, 2012; SAVOLAINEN; MANNERING, 2007). Jones, Gurupackiam, and Walsh (2013) and Salum *et al*. (2019) found a twofold increase in fatality risk on curves as compared to straight-road segments because curves are associated with run-off-road crashes. However, Rahman et al. (2021) contrarily reported an increase of 4 times in the chances of fatality crashes for straight road segments allegedly due to more opportunities for speeding as well as less riding focus in long straight stretches.

About 1/3 of fatal motorcyclist crashes happen at junctions. The severity is higher in intersections for motorcyclists than for other road users (HÉRAN, 2017). Objects near intersections could reduce significantly visibility, making it more difficult to notice road users (OECD, 2015).

Other studies show that intersections could decrease the probability of fatal crashes (GEEDIPALLY; TURNER; PATIL, 2011a; JONES; GURUPACKIAM; WALSH, 2013; SALUM *et al.*, 2019; SAVOLAINEN; MANNERING, 2007). Rahman *et al.* (2021) and Li *et al.* (2021) found that intersections reduce 0.4 times the probability of fatal crashes than non-

intersections. The authors argued that in intersections motorcyclists are more cautious and take lower speeds (LI *et al.*, 2021; RAHMAN *et al.*, 2021; ZAFRI *et al.*, 2022).

Other characteristics, such as lane width and number of lanes, are associated with injuries, varying their effect could be positive or negative depending on the model (FLASK; SCHNEIDER; LORD, 2014; ISLAM, 2022; LI *et al.*, 2021; SE *et al.*, 2021). Li *et al.* (2021) found that an increasing number of lanes is associated with a decrease in injury severity in single-crash vehicles and an increase in injury in two-vehicle crashes.

The quality of road surface is another factor that could increase the risk of fatal crashes. The irregularities in the road can lead to a loss of stability and grip (ACEM, 2009; IHIE, 2010; OECD, 2015). However, studies show that good pavement-surface conditions increase the probability of high severity (about 3 times) (ABDUL MANAN *et al.*, 2018; ABRARI VAJARI *et al.*, 2020; GEEDIPALLY; TURNER; PATIL, 2011a; XIN *et al.*, 2017). The authors argue that a good surface is associated with higher speeds.

On the roadside, obstacles, such as vegetation and constructions, could increase the fatal crash risk because of compromised visibility (ACEM, 2009; XIN *et al.*, 2017). Studies show that road barriers, like guard rails, increase the severity of motorcyclists. These contribute to 2% to 4% of motorcyclists fatalities (2-BE-SAFE, 2010; OECD, 2015).

2.2.1.5 Environmental factors

Environmental factors affect more the motorcyclist than other users because of the cognitive load required to control the motorcycle (ABRARI VAJARI *et al.*, 2020; BLACKMAN; HAWORTH, 2013; CUNTO; FERREIRA, 2017; RAHMAN *et al.*, 2021). Weekends are associated with an increase of fatality odds by 1.7 times, attributed to reduced traffic leading to higher chances for speeding. Furthermore, weekends correlate with increased alcohol and drug consumption (ABRARI VAJARI *et al.*, 2020; JONES; GURUPACKIAM; WALSH, 2013; OECD, 2015; RAHMAN *et al.*, 2021; SALUM *et al.*, 2019).

The role of lighting conditions is also discernible. Rahman et al. (2021) found an increase in the chances of fatal crashes (OR=5.3) in the morning (dawn) and in segments insufficiently illuminated during nighttime (OR=12.3) as compared to daylight. Li et al. (2021) also found a decline in critical injuries during daylight. Nighttime periods can be associated with factors such as low vehicular flow, speeding, and alcohol consumption.

Water on the road reduces the skid resistance, causing more risky situations for the road user. Furthermore, rainy seasons are associated with problems such as fatigue, poor visibility, and malfunctioning vehicles (OECD, 2015; RAHMAN *et al.*, 2021; SE *et al.*, 2021). Rainy seasons are also associated with twice more probability of fatal crashes than summer seasons (RAHMAN *et al.*, 2021). On another hand, other studies showed that clear weather increases the probability of fatal crashes (ABRARI VAJARI *et al.*, 2020; JONES; GURUPACKIAM; WALSH, 2013; SAVOLAINEN; MANNERING, 2007; WASEEM; AHMED; SAEED, 2019). The authors argued that clear weather may lead to high-speed and risky behaviors.

According to a study by Se *et al.* (2021), rural areas pose a higher risk for motorcyclists, with male riders, pillion riders, speeding, improper overtaking, and fatigue being significant determinants of severe and fatal injuries. The findings highlight the importance of safety education for motorcyclists, particularly in rural areas, to increase awareness of the risks of severe injuries. Additionally, increased enforcement efforts, such as reducing the number of unlicensed riders and providing more riding training for rural riders, may be effective strategies to reduce motorcycle-related injuries (SE *et al.*, 2021).

2.2.1.6 Crash specific factors

Crash-specific factors are associated with another characteristic of the crash (RAHMAN *et al.*, 2021), such as the type of collision (*e.g.*, head-on collisions) and vehicles that the motorcycle collided with (*e.g.*, heavy vehicles). The nature of collisions can have a significant impact on the resulting impact energy and severity. In instances where a motorcycle collides with another vehicle, the energy transferred during the collision can be substantial, leading to serious injuries or fatalities. When a heavy truck and a lighter motorcycle collide, the energy of the impact is greater than if both vehicles had been of similar mass. Additionally, if the collision occurs at a high speed, the energy of impact will be even greater. This can result in significant damage to both vehicles and potentially lead to severe or even fatal injuries (PERVEZ; LEE; HUANG, 2021; SE *et al.*, 2021).

Collisions involving heavy vehicles lead to a 34% rise in incapacitating injuries and a 14% elevation in fatality likelihood (CHUNG; SONG; YOON, 2014; PERVEZ; LEE; HUANG, 2021; SE *et al.*, 2021). This result is also supported by Jones et al. (2013) who found a 5-fold fatality increase with heavy vehicles. Given that the vehicular masses are directly

proportional to the kinetic energy to be dissipated during a given crash, this variable is crucial to be considered in crash severity models. It is also important to highlight that for crashes involving at least two vehicles, vulnerable road users such as motorcyclists, might have higher chances of sustaining more severe injuries when large vehicles are involved due to vehicular height and width, which can significantly influence the first point of impact on the rider.

More than one vehicle involved in the crash increases the probability of fatal injury (SE *et al.*, 2021). Multi-vehicle crashes are almost 1.5 times and 2.0 times more likely to be fatal compared with minor and serious injuries respectively (ABRARI VAJARI *et al.*, 2020).

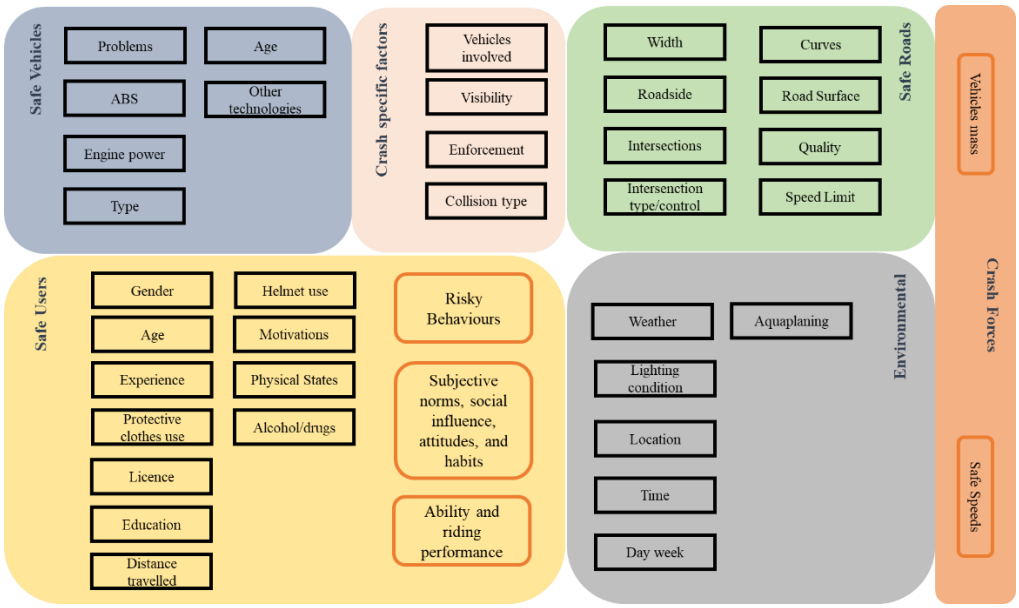
The angle of collision is another important factor that can greatly impact the severity of a collision. Head-on collisions will typically result in a more severe outcome compared to an angled collision (SALUM *et al.*, 2019). Salum *et al.* (2019) showed that head-on crashes increased 1.1, 2.0, and 2.4 times the chances of a fatal crash relative to severe injuries, minor injuries, and no injuries respectively (SALUM *et al.*, 2019).

2.2.2 Summary

Figure 7 shows a summary of the main factors affecting the severity of motorcycle crashes. The previous research findings revealed a pronounced awareness of the intricate interrelationships among the elements contributing to the severity of road crashes. Moreover, a multitude of factors exhibited diverse effects, occasionally even contradictory, on the realm of road safety investigations.

Nevertheless, the results obtained from these investigations may not consistently exhibit precision. This inconsistency can be attributed to the lack of a Causal Inference Approach. This specific approach holds significance due to its ability to mitigate the influence of confounding variables or unaccounted factors that could potentially distort the outcomes. A thorough understanding of these underlying principles, combined with the integration of causal inference methodologies in road safety studies, plays a pivotal role in bolstering the credibility and dependability of the acquired results.

Figure 7 – Factors that affect the severity of motorcycle crashes



Source: The author.

2.3 The paradigm of Causal Inference from the perspective of Road Safety

Identifying the causes of crashes is crucial for developing action plans to reduce them (CUMMINGS, 2006; DUFOURNET *et al.*, 2016). Randomized Controlled Trials (RCTs) are the traditional method for determining causal effects. Nevertheless, RCTs are rarely used in road safety, which typically relies on observational data (DAVIS, 2021). The traditional methods for modeling the severity or frequency of crashes can result in biased outcomes due to factors such as endogeneity and a data-driven approach[1] (HAUER, 2010). To overcome these challenges, new approaches have emerged.

Endogeneity refers to a situation in which an explanatory variable is correlated with the error term in a statistical model. This can happen due to various reasons, such as omitted variable bias, simultaneous causality (X causes Y but Y also causes X), or measurement error, as described by Wooldridge. However, sometimes endogeneity is used specifically to refer to simultaneous causality and measurement error is considered separately, particularly if it is assumed to be random, such as when individuals are equally likely to underestimate or overestimate their health status (GUNASEKARA; CARTER; BLAKELY, 2008).

[1]A data-driven approach refers to a method of problem solving where decisions and insights are based primarily on the analysis of data and information. The main idea is to make decisions and predictions based on patterns and relationships found in the data, rather than relying solely on intuition or prior knowledge.

There are two main theories of Causal Inference, Rubin (1974) and Pearl (2009). Rubin's theory (1974), also known as the potential outcomes framework (PO), is based on the concept of counterfactuals. It defines a causal effect as the difference between what would have happened if an intervention had occurred, and what happened. Rubin's theory uses a model-based approach and relies on the assumption of stable unit treatment value (SUTVA) and the ignorability of the treatment assignment mechanism.

Pearl's theory (2009) is based on the concept of causal diagrams (also called Directed Acyclic Graphs or DAGs). It uses a graphical representation of variables and their relationships to identify and control for confounding factors, and to estimate causal effects. Pearl's theory (2009) also emphasizes the importance of understanding the underlying mechanisms that generate the data, and how these mechanisms affect the estimation of causal effects.

In summary, Pearl's theory focuses on understanding the underlying causal mechanisms and uses the graphical representation (*i.e.*, DAGs) to identify and control confounding, while Rubin's theory focuses on counterfactuals and uses a model-based approach to estimate causal effects. Both theories have their own set of assumptions and limitations and can be complementary in certain situations.

For pedagogical purposes, this dissertation will initially elucidate Pearl's theory, which offers a more illustrative understanding of the Causal Inference approach. To gain a better understanding of Pearl's theory, the next section will provide examples of a confounder (with Simpson's Paradox), and the theory of DAGs. These concepts are essential to comprehending Pearl's causal inference framework and will help illustrate how confounding variables can impact study results, how seemingly contradictory conclusions can arise from the same data, and how DAGs can be used to represent causal relationships between variables.

2.3.1 Pearl's theory and Directed Acyclic Graphs (DAGs)

To exemplify Simpson’s Paradox (Simpson, 1951) look at this fictional example (Table 2). This example is based on the original Simpson study (1951), using the same numeric values. Nevertheless, this example was modified to represent a Road Safety study. It consists of roads with similar characteristics, differentiating from treatment (*i.e.* speed bump) and type (*i.e.* intersections and road segments). The number of fatal crashes was counted on sites and separated by type.

Table 2 – Simpson's Paradox

Group	Treatment	No treatment
Intersections	81 out of 87 non-fatal (93%)	234 out of 270 non-fatal (87%)
Road segments	192 out of 263 non-fatal (73%)	55 out of 80 non-fatal (69%)
Total	273 out of 350 non-fatal (78%)	289 out of 350 non-fatal (83%)

Source: Adapted from Pearl (2009).

Intersections with treatment have more non-fatal crashes (93%) than no treatment (87%). In the same way, road segments have more chances to be safe as well with treatment (73% vs. 69%). Nevertheless, using the total (union of the groups) the chance to be safe is less with the treatment (78%) than without (83%). Therefore, if the type of entity is known, the treatment is effective (reduces fatalities). Nevertheless, if the type is not known, the treatment is ineffective. This illogical affirmation is called Sympson's Paradox.

Sympson's Paradox occurs because of the causal mechanism. The road segments are more propensity to receive the treatment and the treatment is less effective in this group. Because of this, there is a "confounding" when the groups are joined.

It is possible to calculate the total odds ratio[2] of Table 2 using this equation: Odds ratio = (289/273) / (61/77) = 1.34. In other words, the sites with treatment have 1.34 more chances to have a fatal crash. This result is skewed because of the causal mechanism that causes endogeneity in the relationship between treatment and outcome. Therefore, it is necessary to control the confounder factor: "Type". One way to control a confounder is to put it in a model, for example, a logit regression[3]. The result of this model, with treatment and type, is an odds ratio of 0.7 for the treatment (Table 3). This means that the treatment is efficient when the Type is controlled/made constant/adjusting. Therefore, this result is without bias and represents the causal effect in this example.

Table 3 – ODDs and logit models (Y - fatal (1) and non-fatal (0))

Variables	Logit Model 01 - OR	Logit Model 02 - OR
Intercept	0.21	0.14
Type (Road segments (1) and Intersections (0))	x	3.53
Treatment (1), no treatment (0)	**1.34**	**0.70**

x - variable not included in the model; OR - odds ratio
Source: the author.

[2] Odds Ratio (OR) is a statistic that quantifies an association between variables. OR represents the odds that an outcome occurs in a group compared to another group..
[3] In statistical software, use the fatal crashes as the outcome and the type and treatment as explicative variables: "P(fatal) = logit(Treatment + Type)". The exponential coefficient of the treatment is the causal odds ratio.

A DAG has several properties to help find causal effects. The four main configurations in a DAG are chains, forks, colliders, and effect modification. The following sections will provide examples of each configuration using practical examples (with simulated data) from road safety. For the sake of clarity and understanding, 100 observations were used and the simulation was run 1000 times (Monte Carlo simulation) using linear regressions with normally distributed data. The simulations were developed based on the research by Siqueria (2020) and Lübke (2020).

2.3.1.1 Chains and Mediation: direct and indirect effects

To explain the chain structure (denoted as **A →…→ B**), a hypothetical example of crash severity was made, relating age, helmet use, and severity variables (Figure 8). Figure 8 also shows a series of functions with hypothetical terms that describe these relationships. These functions and the DAG are part of Structural Causal Modeling (SCM).

SCM is a framework for understanding and inferring causality from observational data. It uses mathematical models, known as structural equations, to represent the relationships between variables in a system. These equations define the causal structure of the system and specify how variables are affected by one another. SCM allows for the estimation of causal effects by controlling for confounding factors and making assumptions about the underlying mechanisms of the system (PEARL; GLYMOUR; JEWELL, 2016).

Structural equations are functions of any type with non-parametric variables and random terms to determine statistically the value of variables. Random terms refer to variables or factors that are included in a statistical model but are not part of the primary causal relationship being studied. They are often included to account for sources of variation or noise in the data and are typically treated as random variables with a probability distribution (PEARL; GLYMOUR; JEWELL, 2016).

The properties of this chain are that Helmet use and Severity are dependent. If the value of the Helmet use is known (*e.g.* 1), then it is possible to infer the value of Severity (close to -5). In the same way, Age and Helmet use are dependent. Therefore, Age and Severity are likely dependent (PEARL; GLYMOUR; JEWELL, 2016).

Figure 8 – Chains

Structural Causal Modeling (1)

Severity = -5 * Helmet use + N(0,1)*

Helmet use = 2 * Age + N(0,1)*

Age = N(5,2)*

*N(m, sd) - normal distribution with mean equal to m and standard deviation equal to sd

Source: the author.

When the Helmet use is controlled, the Severity and Age are conditionally independent. There are three ways to control a variable, *i.e.*, keep the variable constant. First, placing it in a model, such as regressions, in other words, systematizing the variable. Second, by statistically restricting its value, for example, limiting the Age only to a range (*e.g.*, 50-64). Finally, restricting the value throughout a study, for example, the study was designed only to obtain data on older motorcyclists (SHIPLEY, 2000).

It is possible to demonstrate how a chain works with simulation. When Helmet use is in the model (Model 02 - Table 4), the value of the Age is insignificant because the value of the Helmet use is constant. In other words, the chains are like an energy system, where the edges (links between nodes [variables]) are the wires and the energy is the causal effect. When the Helmet has been controlled, the connection between Age and Severity is closed, like a switch in the off condition (LÜBKE *et al.*, 2020; PEARL; GLYMOUR; JEWELL, 2016; PEARL; MACKENZIE, 2018; SHIPLEY, 2000).

Table 4 – Chains and Monte Carlo Simulation (Y - severity)

Variables	Linear Model 01	Linear Model 02
Intercept	0.10*	0.04
Age	-10.02	-0.01*
Helmet use	x	-5.00

x - variable not included in the model; *not significant

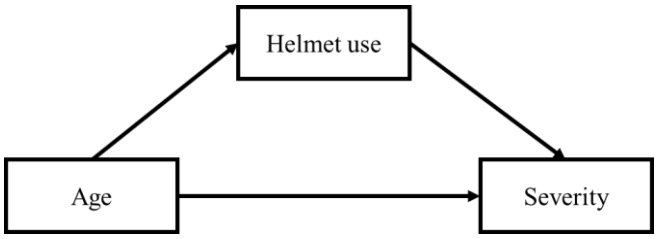
Source: the author.

The conditional independence is also called the d-separation statement. D-separation gives sufficient conditions for two variables (nodes) in a DAG to be independent upon conditioning on other variables (nodes). For example, Age is independent of Severity given Helmet, in mathematical language: Age ⊥ Severity | Helmet. Furthermore, d-separation

is the translation between the causal language and the statistic language (PEARL; GLYMOUR; JEWELL, 2016; SHIPLEY, 2000).

In Figure 9 and SCM 2, Age indirectly affects the severity in the chain **Age → Helmet use → Severity,** and directly in **Age → Severity**. The total effect of Age in Severity is the sum of the two paths (ROBINS M. JAMES, 2020). The analysis of these paths is also known as mediation analysis (PEARL, 2019). The direct and indirect paths are also called front-door paths (PEARL; GLYMOUR; JEWELL, 2016; SHIPLEY, 2000).

Figure 9 – Direct and indirect effects (mediation analysis)



Structural Causal Modeling (2)

Severity = -5 * Helmet use + 3 * Age + N(0,1)

Helmet use = 2 * Age + N(0,1)

Age = N(5,2)

Source: the author.

Using linear regression, it is possible to obtain the direct and total effect. When Helmet use is not in the model (Model 1 - Table 5), the β of the Age is the total effect. Therefore, in SCM 2 (Figure 9), the Severity = -5 * Helmet use + 3 * Age. When substituting the value of Helmet use = 2 * Age, the value of Severity = -10 * Age + 3 * Age, giving -7 * Age. When Helmet use is in the model (Model 2 - Table 5) the value is the direct effect of Age in Severity (= 3). The indirect effect is the total effect minus the direct effect (-7 - 3 = -10).

The value of the direct effect is positive, and the total effect is negative. Therefore, it is important to know what is measured in each model. Generally, in road safety studies the modelers used all variables available in the databases in a single model (SONG; KOU; WANG, 2021). In these cases, the βs or odds ratios obtained have the meaning of partial effect or association when controlled for the other variables (WOOLDRIDGE, 2013), however, researchers in most cases interpreted it as a total effect. This misinterpretation is called “the table 2 fallacy” (WESTREICH; GREENLAND, 2013).

Table 5 – Direct and indirect effects (Y - severity)

Variables	Linear Model 01	Linear Model 02
Intercept	0.09*	0.00*
Age	-6.98	3.00
Helmet	x	-5.00

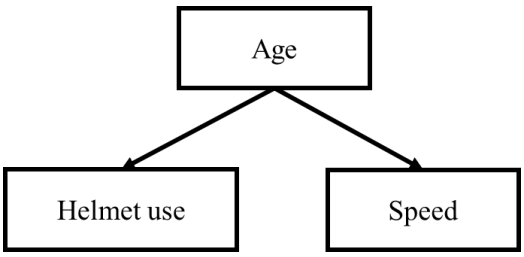
x - variable not included in the model; *not significant
Source: The author.

2.3.1.2 Forks: confounder factors and backdoor criterion

A fork (**A ← … ← C → … → B**) transmits an association between **A** and **B**, but it is not causal (ROHRER, 2018). To demonstrate this structure a hypothetical example was made, where age influences both helmet use and Speed, denoted as **Helmet use ← Age → Speed** (Figure 10 and SCM 3). When the value of the Age increases both Speed and Helmet use increase too, creating a spurious correlation (PEARL; GLYMOUR; JEWELL, 2016).

According to this DAG, **Age and Helmet use** and **Age and Speed** are dependent. Furthermore, Helmet use and Speed are likely dependent, because of the spurious correlation. Finally, Helmet use and Speed are independent, conditioning on Age (PEARL; GLYMOUR; JEWELL, 2016).

Figure 10 – Forks



Structural Causal Modeling (3)

Speed = 5*Age + N(0,1)

Helmet use = 2 * Age + N(0,1)

Age = N(5,2)

Source: The author.

When Age is constant (*e.g.* 20), both equations (Speed and Helmet use) are equal to N(0,1) plus the constant (20). Therefore, the Speed and Helmet use are independent because the random terms are independent (assumption).

When Age is not in the model (Model 01 - Table 6), there is a spurious association (2.00) between Helmet use and Speed (this value is not in SCM 3). Nevertheless, when Age is in the model (Model 02 - Table 6), Helmet use and Speed are independent. The Age in this DAG is called a confounder factor because it confounds the relationship between Helmet use and Speed (LÜBKE *et al.*, 2020; PEARL; GLYMOUR; JEWELL, 2016).

Table 6 – Fork (Y - Speed)

Variables	Linear Model 01	Linear Model 02
Intercept	5.04	0.02*
Age	x	5.00
Helmet use	2.00	0.00*

x - variable not included in the model; *not significant
Source: The author.

A path of variable (F) that is a causal ancestor of both other two variables (X and Y) is called the **backdoor path** (X ←...← F →...→ Y). The foundation of causal analysis is to block all backdoor paths (PEARL; GLYMOUR; JEWELL, 2016; SHIPLEY, 2000). This is also related to the terms confounding, do-calculus, endogeneity, omitted variable bias, ignorability assumption, and exchangeability. For details about these terms consult: (GUNASEKARA; CARTER; BLAKELY, 2008; PEARL, 2009; ROBINS M. JAMES, 2020; RUBIN, 1974).

When all backdoors have been blocked, the relationship between the treatment and the outcome could be related to random assignment. Therefore, when all spurious paths are blocked, all directed paths are left unperturbed, and spurious paths are not created, then the causal effect can be calculated with a statistical test (PEARL; GLYMOUR; JEWELL, 2016; ROBINS M. JAMES, 2020).

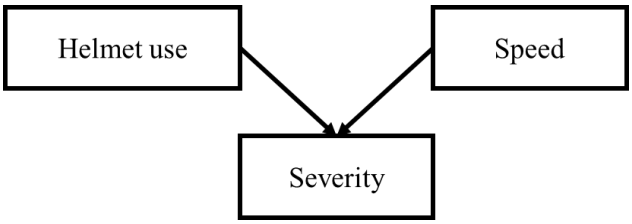
2.3.1.3 Colliders and selection bias

Colliders or inverted forks have the structure **A → … → C ← … ← B,** this does not transmit association, but it can transmit if controlled (ROHRER, 2018). To demonstrate this structure a hypothetical example was made, where helmet use and Speed influence the severity, denoted as **Helmet use → Severity ← Speed** (Figure 11 and SCM 4). When the value of Severity is constant (*e.g.* 15), the values of Helmet use and Speed will be associated. For

example, if Helmet use is 1, Speed must be 4 for the final result to be 15. Therefore, is possible to infer the value of Speed (PEARL; GLYMOUR; JEWELL, 2016).

If for some reason, the severity is controlled, for example, in a study that only collected fatal and severe injury crashes (selection bias). Therefore, the correlation between Helmet use and Speed will be not zero. The control of a collider variable creates a spurious association (ROBINS M. JAMES, 2020).

Figure 11 – Collider



Structural Causal Modeling (4)

Speed = N(7,1)

Helmet use = N(1,1)

Severity = -5 * Helmet use + 5 * Speed + N(0,1)

Source: The author.

In SCM 4, Helmet use is independent of Speed. Hemet use and Severity are dependent because of the direct link. Speed and Severity are dependent, in the same way. Nevertheless, Helmet use and Speed are dependent when Severity is conditioned/controlled (PEARL; GLYMOUR; JEWELL, 2016).

When Severity is not in the model (Model 1 - Table 7), there is not a spurious relation (0) between Helmet use and Speed. Nevertheless, when Severity is in the model (Model 2 - Table 7), Helmet use and Speed are dependent (0.96). The Severity of this DAG is called a collider factor and causes spurious association when controlled (LÜBKE *et al.*, 2020; PEARL; GLYMOUR; JEWELL, 2016).

Table 7 – Collider (Y - Speed)

Variables	Linear Model 01	Linear Model 02
Intercept	7.00	0.27
Helmet use	0.00*	0.96
Severity	x	0.19

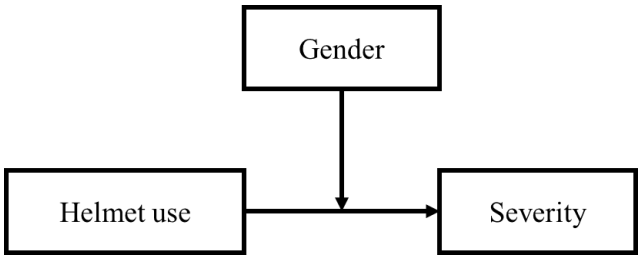
x - variable not included in the model; *not significant

Source: The author.

2.3.1.4 Effect modification

When the effect of Helmet use in Severity varies between values of another variable (*e.g.*, gender), it is called the effect modification or moderation (PEARL; GLYMOUR; JEWELL, 2016). A hypothetical example: helmets are more efficient for female motorcyclists. Gender does not affect (cause) either Severity or Helmet use, changing only the relationship between them (*i.e.*, there is heterogeneity in the effect across levels of gender). It is possible to represent the modification effect using an arrow pointing towards the arrow between the variables (Figure 12) (LAUBACH *et al.*, 2021).

Figure 12 – Effect modification



Structural Causal Modeling (5)

Gender = Binom(1,0.5)*

Helmet use = N(1,1)

Severity = -2 * Helmet use - 1.5 * Helmet use * Gender + N(0,1)

*Binom(size, prob) - binomial distribution with the number of trials equal to size and probability of success on each trial equal to prob. In other words, there are two levels (0 - men, 1 - women), and the probability is 50% for each.

Source: The author.

When the interaction between Helmet use and Gender is not in the model (Model 1 - Table 8), there is not a spurious relation (-2.75 is the total effect) between Helmet use and Severity. Nevertheless, when Helmet use * Gender is in the model (Model 2 - Table 8), the coefficient of Helmet use (-2.0) is the effect in the men group (*i.e.*, gender equals 0). Furthermore, the coefficient of interaction (-1.5) plus Helmet use (-2.0) is the effect (-3.5) in the women group (*i.e.*, gender equals 1). Furthermore, all these values are causal effects.

Table 8 – Effect modification

Variables	Linear Model 01	Linear Model 02
Intercept	0.00*	0.00*
Helmet use	-2.75	-2.00
Helmet use * Gender	x	-1.50

x - variable not included in the model; *not significant
Source: The author.

An effect modifier is different from a confounder. Confounders skew the relationship between the effect and the cause, while effect modifiers show different effects on the cause. Nevertheless, a variable can be both a confounder and an effect modifier if it affects the effect, the cause, and the relationship (LAUBACH *et al.*, 2021; PEARL; GLYMOUR; JEWELL, 2016).

2.3.1.5 DAGs in summary

There are four conditions to have an association between two variables (X and Y). First, X causes Y, the directed link causes dependency (front-door paths). Second, similarly, Y causes X. Third, there is another variable (Z) that causes Y and X, creating a spurious correlation (backdoor paths). Finally, there is another variable (C) that was controlled (the value is constant), and C is caused by X and Y (*i.e.*, C is a collider), then creating a spurious correlation (selection bias). The first and second conditions are causal, and the third and fourth are only associations (LAUBACH *et al.*, 2021; PEARL; GLYMOUR; JEWELL, 2016).

2.3.1.6 Models based on DAGs

The estimation of coefficients of relationships and evaluating DAGs are crucial components of the Causal Inference of Pearl. These tasks can be achieved using various modeling techniques. The purpose of using a DAG in causal inference is to visually represent the relationships between the exposure and the outcome and to identify any confounding variables that may need to be adjusted in the analysis. The advantage of using DAGs is that they clearly show the different types of relationships, including direct and indirect effects, back-door paths, and colliders.

The Structural Causal Model (SCM) uses the graph theory, representing the causal hypotheses in a DAG where relationships of observed and unobserved variables can be

depicted. SCMs are a set of endogenous (**V**) and exogenous (**U**) variables[4] linked by functions (**F**) (**V** = **F**(**U**)), and this causal mechanism is represented by a DAG (KLINE, 2015; PEARL; GLYMOUR; JEWELL, 2016).

When the set of functions (**F**) is linear in its parameters, SCM could be associated with a Path Analysis (PA). PA is a method used to measure associations or causality of a DAG, using a unique estimation of **F**. In other words, the set of linear functions (**F**) is estimated at the same time. The goal of this estimation is to minimize the difference between the observed covariances (data) and the predicted by the causal model (estimated) (SHIPLEY, 2000).

The estimation is usually realized using the maximum likelihood (ML). At this point, the PA differs from a simple multivariate regression, which is estimated using the least square. Furthermore, ML has some assumptions: endogenous variables are a numeric continuum, relationships are linear in the parameters, and the data follow multivariate normal (HOYLE, 2012; SHIPLEY, 2000).

When the endogenous variables are categorical or ordinal, the use of ML as an estimator is not indicated. In this case, there are other estimators based on least squares. The most used are the diagonally weighted least squares (DWLS) and the weighted least squares with mean- and variance adjusted (WLSMV) (HOYLE, 2012). In these cases, the link between ordered/binary variables and others is the logit or probit functions.

A DAG has a set of conditional probabilistic independencies (d-sep) that could be tested using the PA model or other techniques, such as conditional correlations (SHIPLEY, 2000). The estimation of PA gives the relation between observed and estimated correlation matrices. These matrices will be different if the DAG is not consistent with observed data. Therefore, there are two principal metrics to evaluate a given model. The first is the test of hypotheses, in which the null hypothesis is that the two matrices are equal. The second is the Root Mean Square Error (RMSEA) which shows the difference between the two matrices. Acceptable values are p-value > 0.05 and RMSEA < 0.05 (SCHUMACKER; LOMAX, 2010). Nevertheless, the p-value is not recommended when there is a large sample size (400>) because this metric is sensitive to sample size (HAIR *et al.*, 2009; SHIPLEY, 2000). Furthermore, it is possible to see the individual values of the residuals. Standardized residuals are like Z-scores, then values greater than 2.58 (99% confidence) indicate that a particular relationship is not well computed by the model (SCHUMACKER; LOMAX, 2010).

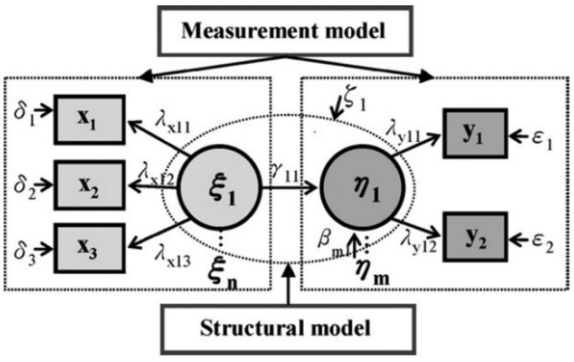
[4] Endogenous variables are those that are influenced by other variables in a system or model. On the other hand, exogenous variables are variables that are not influenced by other variables in the system, but rather affect the endogenous variables.

When there is an unobserved variable problem, latent variables can be used. A latent variable refers to an underlying factor that affects the relationship among observed variables but is not directly observed or measured. For example, a study may observe that helmet use is related to reduced fatalities in motorcycle crashes, but there may be other factors such as driver experience or road conditions that are not directly measured but could be influencing the relationship. In this case, driver experience and road conditions can be considered latent variables. To account for these latent variables, researchers may use statistical techniques such as Structural Equation Modeling (SEM), which uses other available variables to estimate latent variables (BROWN, 2015; HOYLE, 2012; NEWSOM, 2015).

SEM is a collection of statistical techniques formed by two principal parts. First is the confirmatory factor analysis (CFA). CFA consists of measurement models, which are the relationships between the observed variables (indicators) and latent variables (factors). Second is the structural model, *i.e.*, the relationship between the other variables that are not indicators (latent and other observed variables) (BROWN, 2015; HOYLE, 2012; NEWSOM, 2015).

Figure 13 portrays the errors in measuring the variables (δ and ε). Latent variables are commonly denoted by circles or ellipses (ξ and η), while λ and γ represent the coefficients of the models. The error in estimating the relationship between the latent variables of the structural model is shown by ζ. Observed variables are usually represented by rectangles or squares. A statistical dependence is typically indicated by a directional arrow or path, where the variable at the tail of the arrow causes the variable at the point. A correlation between variables is denoted by a double-headed arrow. In the path structure shown in Figure 13, the models assume the structures presented in Equation 1 (AL-MAHAMEED *et al.*, 2019; TORRES; XAVIER; CUNTO, 2020).

Figure 13 – SEM



Source: Leen, Chungm and Son (2008).

$$\begin{bmatrix} y \\ x \end{bmatrix} = \begin{bmatrix} A_y & 0 \\ 0 & A_x \end{bmatrix} \begin{bmatrix} \eta \\ \xi \end{bmatrix} + \begin{bmatrix} \varepsilon \\ \delta \end{bmatrix} \quad (1)$$

The equation involves two column vectors (y and x) representing observed variables. The variables may have measurement errors, which are represented by ε and δ. Additionally, there are two coefficient matrices, A_y and A_x, which correspond to the latent indicators of observed variables. Equation 1 expresses the vector η that encompasses all the variables in the structural model (AL-MAHAMEED *et al.*, 2019; TORRES; XAVIER; CUNTO, 2020).

The vectors β and γ represent the estimated regression coefficients for the dependent and independent variables, respectively, while ζ is a vector of regression errors, and ξ is a vector of collected independent variables. Structural equation models are based on the assumption that the observed variables' variance-covariance matrix is a function of the model parameters. When the model is correctly specified, this variance-covariance matrix is equivalent to the population matrix (AL-MAHAMEED *et al.*, 2019; TORRES; XAVIER; CUNTO, 2020).

Studies that evaluated the injuries in motorcycle crashes using only PA analysis are rare (LEE *et al.*, 2017). Road safety studies of motorcyclists that used SEM usually applied questionnaires to find relationships among motorcyclists' risky behavior, perception, and user characteristics variables (CHOU *et al.*, 2022; GOH; LEONG; CHEAH, 2020; NADIMI *et al.*, 2021; ZIAKOPOULOS; NIKOLAOU; YANNIS, 2021). A study used SEM in a database of motorcycle crashes (KASHANI *et al.*, 2020), and others used SEM with a focus on the severity of motorcyclists (HASANZADEH; ASGHARIJAFARABADI; SADEGHI-BAZARGANI, 2020; LEE *et al.*, 2017).

Lee *et al.* (2017) developed a path analysis using a logit regression link to evaluate the effect of helmet use on the severity of motorcycle crashes. The results showed that helmets affected fatalities through other variables, such as the prevalence of head injuries, craniotomies, and complications. Helmet use reduces the rates of injuries by 34.5%.

Kashanbi *et al.* (2020) studied crashes involving motorcyclists using SEM. They elaborated a latent variable called "accident size", elaborated by Lee *et al.* (LEE *et al.*, 2018). The accident size is measured by the number of injured individuals, the number of fatalities, and the number of vehicles involved and damaged. The data used is from Iran between 2011 and 2018, containing 204,299 rural and urban traffic motorcycle crashes. The study showed

that the factors of motorcyclists and the road are important in mitigating the crash severity (KASHANI *et al.*, 2020).

Hasanzadeh, Asguarujafarabadi, and Sadeghi-bazargani (2018) used a SEM model with an artificial neural network to estimate the severity of motorcycle crashes in Iran. The authors found that marital status, education level, riding for fun, engine volume, dark hour riding, cell phone answering, and driving license are significant to the prediction (HASANZADEH; ASGHARIJAFARABADI; SADEGHI-BAZARGANI, 2020).

None of the motorcycle studies reviewed utilizing SEM have shown any concern for Causal Inference theory. They neglect critical concepts like blocking back doors and remain indifferent to collider bias. Similarly, in the realm of road safety studies, these fundamental aspects are seldom given due consideration. Issues in crash databases, including missing data, balancing, and selection bias, tend to be overlooked in this domain as well.

2.3.2 Rubin's theory and Propensity Score (PS) approach

Another approach employed for conducting causal inference is Rubin's causal model. This model relies on the Propensity Score (PS), a statistical method employed to equalize treatment groups in observational studies, particularly when group assignments are not random. The idea behind the PS approach is to estimate the probability of a subject receiving a certain treatment based on their observed characteristics, such as age, gender, socio-economic status, *etc*. The PS is calculated for each subject and then used to match or attribute weights to the subjects in the treatment and control groups so that they have similar distributions of observed characteristics. This helps to reduce the influence of confounding variables and improves the validity of the causal inferences. The use of PS can help to control for selection bias and increase the comparability of treatment and control groups, making it a useful tool for estimating the treatment effect in observational studies (SASIDHARAN; DONNELL, 2014).

The results of models/approaches based on Pearl's and Rubin's theory are likely to yield similar outcomes if the model is well-specified, meaning that all confounding variables are properly controlled. However, PS models are typically used in situations where treatments are involved, such as cases where the effect being studied can be changed by an individual. In severity studies, variables like helmet use may be suitable as treatment variables, while others such as weather conditions may not be appropriate for the same purpose. Therefore, Pearl's approach, which utilizes SEM or other graphical models, can be more appropriate for estimating effects in severity studies.

Stating that the difference between the two approaches is that PS approaches lack a theoretical model is a false statement, as it is possible to use PS based on a DAG to estimate causal effects. However, PS models may not be as suitable as SEM, for example, when it comes to testing whether the DAG is appropriate for the data. SEM provides more robust tools for testing the adequacy of the DAG to the data, allowing for a more comprehensive assessment of fit and accuracy.

2.3.3 Causal Inference on crash database issues

Crash databases usually contain several issues that jeopardize causal inference. The issues could be selection bias, missing data, unbalanced data, omitted variables, underreporting crashes, and spatial/temporal dependency. The causal inference literature refers to the selection bias when a collider variable is controlled (ELWERT; WINSHIP, 2014; ROBINS M. JAMES, 2020; SHIPLEY, 2000). The selection of individuals could occur because of various factors. First, the collider variable has its value constrained (*e.g.,* if the severity is a collider, and the study uses only fatal crashes). Second, when the treatment is not chosen randomly (known as self-selection), which is quite common in before-after road safety studies. Finally, if the outcome variable has missing data, then restricting the individuals in the analysis may result in bias (LORD; QIN; GEEDIPALLY, 2021; ROBINS M. JAMES, 2020; WOOD, 2016).

The imbalanced data in road crash databases is a result of the low number of cases reported for high-level injury severity, for instance. This imbalance can cause issues with modeling as the model will be heavily weighted towards the majority class. To handle this issue, there are methods such as oversampling and under-sampling. Oversampling replicates samples from minor classes, but this could lead to an overfit model. Under-sampling could eliminate most of the records, losing information (LANE; CLARKE; HENDER, 2012; TOPUZ; DELEN, 2021).

Omitted variable bias occurs when a relevant explanatory variable is not included in the model. This can lead to incorrect estimates of the effects of the included variables and bias in the results. For example, if the level of experience of a driver is omitted in a road crash model, this can lead to biased estimates of the effects of other factors such as helmet use. This is because the relationship between the included variables and the outcome (crash) may be confounded by the omitted variable, and therefore the estimates of their effects will not accurately reflect the true relationship. To avoid omitted variable bias in road safety studies, it is important to consider all relevant variables and include them in the analysis, either through

direct measurement or through the use of proxies/latent variables (YANG; WANG; DING, 2019).

Underreporting crash bias refers to the phenomenon where some road crashes are not recorded or reported in official data sources, leading to an incomplete and inaccurate representation of the true number and nature of crashes. This can have serious consequences for road safety research and decision-making, as it can lead to incorrect conclusions and ineffective interventions. To address this bias, it is important to use multiple data sources and to collect data through various methods, such as police reports, hospital records, and surveys, to obtain a more comprehensive and accurate understanding of road crashes. Underreporting crashes can occur when the crashes involve people who are unaware that road crashes should be reported, crashes without injuries, and crashes involving people who are affected by drugs or alcohol (LI, 2014). These cases could lead to a selection of individuals in the study, causing a selection bias.

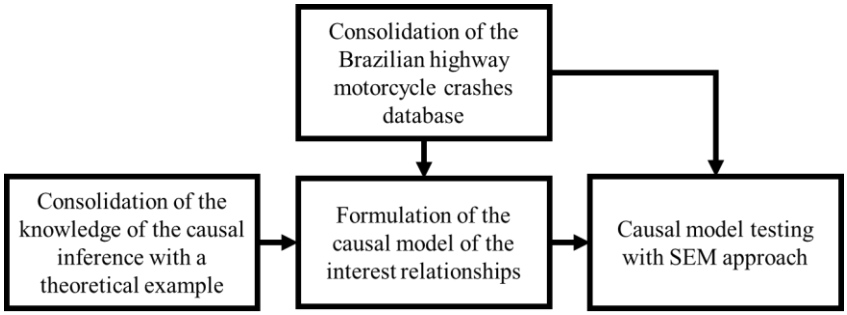
The spatial or temporal autocorrelation could also lead a biased results. Individuals who suffered the same crash or were on the same road could be correlated. To account for spatial autocorrelation, analysts often use spatial statistical models or apply spatial weighting methods to their data. Moran's I test is one method commonly used to assess the presence of spatial autocorrelation. The test can help identify clusters of crash occurrences in a given area and determine whether the crashes are randomly dispersed or spatially related (LI, 2014).

Statistical models typically assume that observations are independent. However, spatial or temporal dependence can violate this assumption, leading to biased coefficients and underestimated standard errors, which in turn can result in an overestimation of the significance of traditional tests. Moreover, as Shipley (2016) points out, temporal dependence can reduce the amount of information in observed data, thereby giving rise to the concept of effective sample size.

3 METHOD

The method of this dissertation is based on the studies of Siqueira (2020), Laubach *et al.* (2021), and Pearl (2021). Figure 14 shows the representation of the process for finding causal effects using observational data of crashes involving motorcyclists.

Figure 14 – Method



Source: The author.

The approach of this dissertation starts with a theoretical example, where the main components of the causal inference theory are presented using simulated data. Secondly, the motorcycle crashes database is consolidated. Thirdly, a causal model is formulated based on prior knowledge. The structure of the causal model depends on the structure of the data as well. Therefore, the second and third steps are dependent. Finally, the testable implications of the causal model are evaluated using SEM and observational data.

3.1 Theoretical example with simulated data

Chapter 2 showed Pearl's causal inference paradigm. To consolidate this knowledge in severity road safety studies, a theoretical example was developed using a DAG, which was used to understand the configurations (chains, forks, colliders, and modification) and their implications. This simplified and hypothetical example illustrates the difference between traditional and causal inference modeling in injury severity studies.

Firstly, the relationships of interest are defined, and only these relationships will be analyzed. Secondly, backdoor paths and confounders are defined. These bias the relationship of interest. Finally, the SCM is defined, *i.e.*, the functions that define each variable are described, then it is possible to simulate the data. The concept of simulations was partially inspired by the research conducted by Siqueira (2020) and Lübke (2020).

The *simstudy* package of the R language was used to simulate categorical data with logit links, with previously established coefficients and distributions. Each sample has 400 observations and was simulated 500 times – Monte Carlo Simulation. Subsequently, the average coefficients were derived from the 500 simulations. For graphical models, where is not possible to obtain an average result, the data were simulated with 200,000 observations (400*500), which is equivalent to the previous simulation.

Logit regressions and Structural Equation Modeling were used to compare traditional and causal approaches. These models are likely to have similar results if the model is well specified and the causal inference theory is used.

3.2 Database of motorcycle crashes

The motorcycle crash data was obtained from the Brazilian Federal Highway Police Department (PRF) from 2017 to 2019. There are databases separated by crashes, victims, and causes. The data were organized using the victims' database, using only riders of motorcycles or other Powered-Two-Wheelers (PTW) users. The use of only the riders is because passengers are likely to be statistically dependent on riders. Therefore, these dependencies could create biased outputs in models because there is an assumption of independence of observations in most models.

Only crashes involving two vehicles were used, since crashes involving only the motorcyclist or more than two vehicles may have different characteristics that need to be analyzed separately. The data underwent spatial filtering, focusing specifically on the state of Ceará. This approach was adopted to avoid potential spatial dependence issues that could arise from including all crashes in Brazil. Furthermore, the residuals of the final model were subjected to a Moran's I test to evaluate spatial dependence, as explained in the subsequent steps.

The database contains information about the characteristics of crashes, vehicles, environments, and users. For example, there is information about the motorcyclists, such as age and gender, and the possible causes, filled out by the police, of the crash, *e.g.*, alcohol use and speeding. The available variables need to be sufficient to close back doors and evaluate the causal hypotheses of interest. If any variable is needed and it is not available in the database, it is necessary to use some proxy or latent variable to represent it.

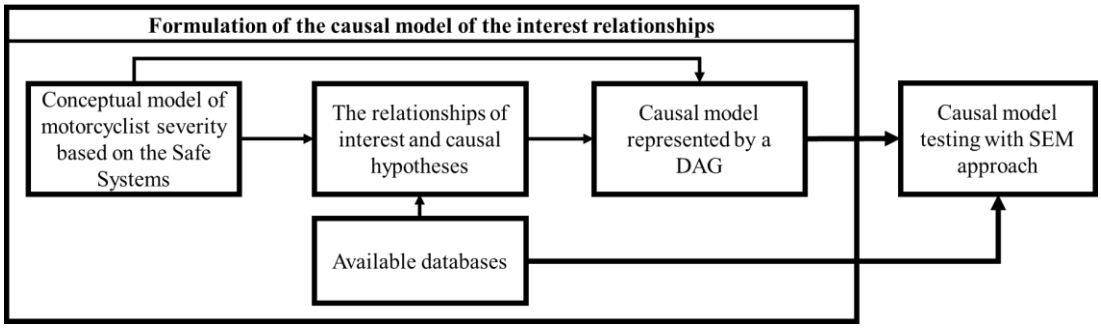
An exploratory analysis of the variables was elaborated. The association between the variables and the severity of the crash was depicted in a graph showing Pearson's independence chi-square test. This test verifies the association between two categorical variables, showing the p-value (H_0: there is no association, $p > 0.05$; H_1: there is an association, $p < 0.05$) and the significant residuals. The significant level (usually 95%) was corrected using the size of the contingency table (dividing the level by the number of rows and the number of columns) because some tables are large and this could lead to misinterpretations (MACDONALD; GARDNER, 2000; SHARPE, 2015).

The chi-square test assumes that the expected values are above five. If this assumption is false, Fisher's exact test was used in addition to the chi-square. Because the database is large, the p-value tends to be significant. Therefore, it was computed Cramer's V statistic which measures the force of the association of categorical variables, varying between zero and one. Values above 0.5 are considered a high association and values below 0.1 are considered a low association, but these values depend on the degrees of freedom.

3.3 Formulation of the causal model

To ascertain the factors contributing to the severity of motorcycle crashes, it is imperative to construct a comprehensive causal model that illustrates the underlying causal hypotheses. As highlighted by Siqueira in 2020, this causal model should be built upon a conceptual framework that synthesizes all existing knowledge about the phenomenon. Figure 15 illustrates the systematic approach adopted in this dissertation for the development of this causal model.

Figure 15 – Formulation of the causal model



Source: the author.

The initial phase involves the development of a conceptual model rooted in the principles of the Safe System approach and using an extensive literature review with a focus on showing the relationship between the factors and the severity of motorcycle crashes. The main goal of this stage is to identify the most important causal elements that could bias the results, such as potential confounders and colliders. Consequently, the conceptual model could aid in identifying potential biased paths.

Due to the complex causal mechanisms of crashes, the second step allows the investigator to propose different causal hypotheses. This entails the definition of confounders and the determination of direct and indirect effects within the proposed causal model. For instance, within the conceptual model, one might explore specific relationships, such as the impact of alcohol consumption on crash severity or its influence on speeding behavior. Moreover, it is important to note that only relationships that can be adequately represented by the available databases are subjected to analysis.

The third step entails the graphical representation of the causal hypotheses using a DAG. The causal model is a composite of the causality hypotheses and backdoor paths of the relationships of interest (associations). Variables within the causal model can be either observable or unobservable, often referred to as latent variables. These latent variables may be used to represent abstract concepts (e.g., safety perception) or variables that are challenging to be directly observed (e.g., impact energy in a traffic collision).

It is crucial to emphasize that the conceptual model plays a pivotal role in constructing the DAG, as it encapsulates vital information about the interconnections among the variables of interest. This information is integral in distinguishing relationships that represent causal hypotheses (structural coefficients) from those intended to close backdoor pathways (regression coefficients).

3.4 Evaluate and estimate the causal model

The *lavaan* package of R was used to estimate the model. Because most variables used are categorical, the Weighted Least Squares with mean- and variance adjusted (WLSMV) estimator was used (BROWN, 2007). In this case, the link between ordered variables and others is the probit function.

Four metrics were used to evaluate the model. The first is the test of hypotheses, in which the null hypothesis is that the observed and estimated correlation matrices are equal. A

p-value greater than 0.05 suggests a good fit, as it indicates that the observed data is likely to be observed under the hypothesized model. When dealing with large samples, even small deviations from the null hypothesis may result in significant p-values. This can lead to the potential overinterpretation of the significance of findings (HAIR *et al.*, 2009).

The second metric is the Root Mean Square Error (RMSEA) which shows the difference between data and estimated matrices. A lower RMSEA value indicates a better fit, with values below 0.05 or 0.08. The third and fourth metrics, the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI) were used as additional goodness-of-fit measures. These indices assess the relative improvement in model fit by comparing the hypothesized model with a baseline model. CFI and TLI values close to 1 indicate a good fit, with values above 0.90 generally considered acceptable (Hair et al., 2009).

Standardized coefficients were utilized since they offer high interpretability and remain unaffected by scaling, allowing for meaningful comparisons between coefficient values. However, even if the causal model is supported by the data, the causal hypotheses are not fully confirmed due to the presence of multiple other DAGs that could also fit the same data. Nevertheless, the a priori formulated causal model retains its reliability compared to data-driven approaches.

3.4.1 Spatial and Temporal dependence

The models underwent a thorough evaluation by analyzing the residuals to detect any spatial and temporal dependencies. In the case of SEM with probit links, the residuals of endogenous variables were not readily available in R language libraries. Consequently, a custom function was developed to calculate the residuals, specifically the deviance residuals. For additional details, please refer to the GitHub repository: https://github.com/altanizio/deviance_resid_probit_ov_binary.

Spatial dependence was assessed using Moran's I test, which involved testing the residuals of the final model. The concept of neighbors was defined as all points within a 1 km radius, considering the diverse locations of the crashes in the study. Various configurations were tested, but the 1 km radius was chosen as it provided the most favorable results by focusing on crashes that occurred on nearby roads. Additionally, alternative configurations produced similar outcomes, further supporting the selection of the 1 km radius (ANSELIN; IBNU SYABRI; YOUNGIHN KHO, 2006; CLIFF; ORD, 1973; MORAN, 1948).

The null and alternative hypotheses for the Moran test are as follows:

- Null Hypothesis (H0): There is no spatial autocorrelation in the variable of interest. In other words, the values of the variable are randomly distributed across the spatial locations.
- Alternative Hypothesis (H1): There is spatial autocorrelation in the variable of interest. This suggests that the values of the variable exhibit some degree of spatial clustering or spatial dependence, meaning that similar values tend to occur near each other.

In addition to spatial dependence, the analysis also examined temporal dependence by including the month and year variables. The goal was to explore potential associations between the residuals and these temporal factors and identify any existing temporal patterns. To assess the relationship between the residuals and the year and month variables, Kruskal-Wallis tests were conducted. The Kruskal-Wallis test is a non-parametric statistical test used to compare the median ranks of two or more independent groups. The null and alternative hypotheses for the Kruskal-Wallis test are as follows:

- Null Hypothesis (H0): The median ranks of the groups are equal. In other words, there is no difference in the distribution of the variable of interest among the groups.
- Alternative Hypothesis (H1): The median ranks of at least one group are different from the others. This suggests that there is a difference in the distribution of the variable of interest among the groups.

Given the large sample size used in the models, which often leads to the rejection of null hypotheses, the analysis took into account the estimative of the Moran Test and the Effect Size of Kruskal-Wallis. These measures were analyzed to gain a better understanding of the significance and magnitude of the observed effects.

4 RESULTS

The results are separated into three sections. The first section presents a theoretical example of the causal inference theory on road safety. The second section presents a conceptual model of motorcyclist severity based on the literature review of the Safe Systems approach. The third section provides a practical example of the causal inference theory based on the previous conceptual model using observational data from Brazilian highways.

4.1 A theoretical example of the causal inference theory on Road Safety

To better illustrate and clarify the causal inference process, this section describes a theoretical example of road safety. Figure 16 illustrates the relationship of interest, the effect of *alcohol use* on crash *severity* through *speed*, that will be applied as the baseline for the theoretical example.

Figure 16 – The relationship of interest of the theoretical example

Source: The author.

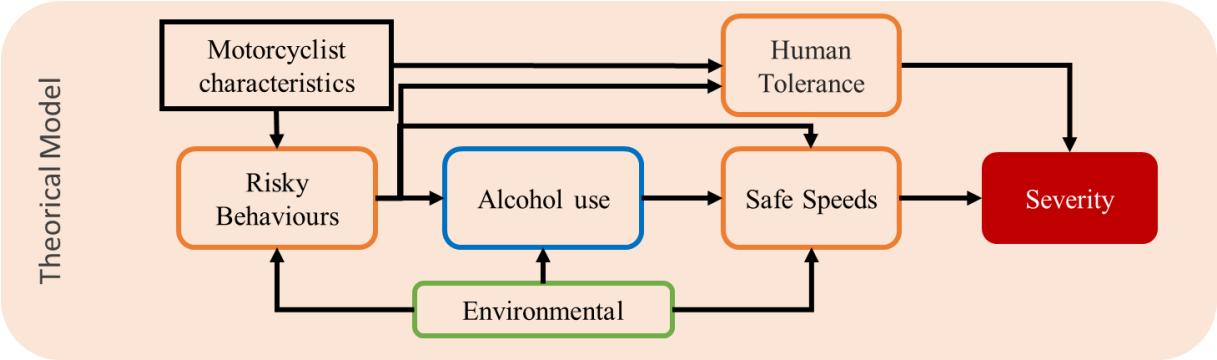
Figure 17 presents a simplified conceptual model of the literature review, showcasing only some of the existing relationships. It should be noted that in subsequent sections, this conceptual model will be further explored.

As seen in the previous chapters, studies found associations between alcohol use and risky behaviors. Risky behavior in the context of motorcycling can be characterized by the adoption of dangerous driving practices, such as operating a motorcycle while under the influence of alcohol, excessive speeds, and non-adherence to safety equipment protocols. This last one has a direct impact on the human body's ability to absorb part of the kinetic energy generated during an impact event, thereby increasing the likelihood of non-serious injuries.

The risky behavior of motorcyclists is influenced by multiple factors, including the environment and individual characteristics of the riders. For example, rural areas and weekends have been shown to increase the likelihood of risky behaviors. Additionally, younger

motorcyclists tend to exhibit a higher frequency of risky behaviors compared to their older counterparts.

Figure 17 – The simplified conceptual model

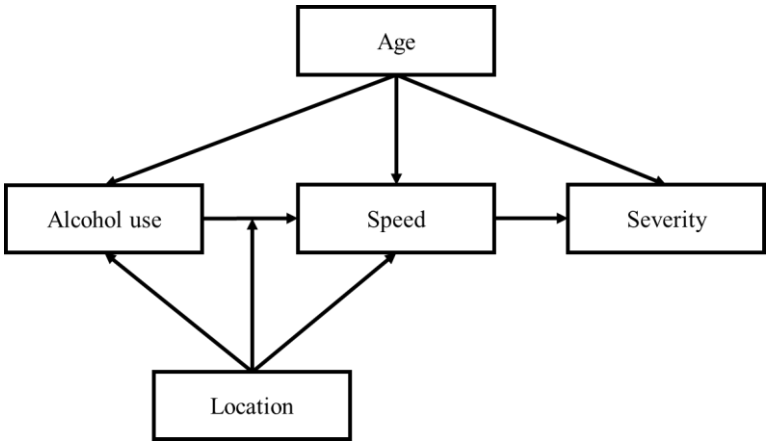


Source: The author.

Figure 18 shows a simplified DAG based on Figure 17. The *Alcohol use* → *Speed* → *Severity* relationship (Figure 16) can be seen as having two backdoor paths, *Location* (Rural/Urban) and motorcyclist *Age*. Additionally, the effect of *Alcohol* on *Speed* varies between urban and rural areas, represented by the modifying effect. Relationships were previously formulated with assigned coefficients and functions in a pre-elaborated SCM. Therefore, this simplified example used only the motorcyclist's *Age* and crash *Location* as confounders to facilitate the understanding. The data were simulated using the *simstudy* package of R language.

Age has three categories: 1, 2, and 3 (30%, 50%, and 20% of data, respectively). Category 1 was renamed to *Age_18_30*, representing motorcyclists between the ages of 18 and 30. Similarly, *Age_30_50* represents between 30 and 50, and *Age_50* is above 50. These categories were chosen because studies found that age has a non-linear effect, given that old motorcyclists are more vulnerable and young are more prone to exhibit riskier behavior.

Figure 18 – DAG of the theoretical example



Structural Causal Modeling (SCM)

Age = categorical(0.3;0.5;0.2)

Age_18_30 = binary(Age=1)

Age_30_50 = binary(Age=2)

Age_50 = binary(Age=3)

Location = binary(0.5)

Alcohol use = logit(Age_18_30 - Age_50 + Location)

Speed = logit(Age_18_30 - Age_50 + 0.5*Location + 0.5*Alcohol + Location*Alcohol)

Severity = logit(- Age_18_30 + Age_50 + Speed)

Source: The author.

Location is another confounder representing whether the crash was on a rural (1) or urban (0) road. This variable also is a modifier of the *Alcohol* and *Speed* relationship. In other words, besides the bias in the relationship of interest, the effect is different on rural and urban roads. Rural traffic crashes with alcohol use are more inclined to have speeding in this theoretical example. Using the SCM, on urban roads, *i.e.*, if *Location* is equal to 0, the effect of *Alcohol* on *Speed* is 0.5. However, on rural roads (1), the effect is 0.5 plus 1 (modification effect). Both effects are causal and show the importance of investigating the effect in different strata of the population. For some strata, the treatment could have a positive effect, and for others a negative effect. This illustration demonstrates the causal inference process, where a conceptual model is transformed into a DAG representing relationships of interest and causal hypotheses.

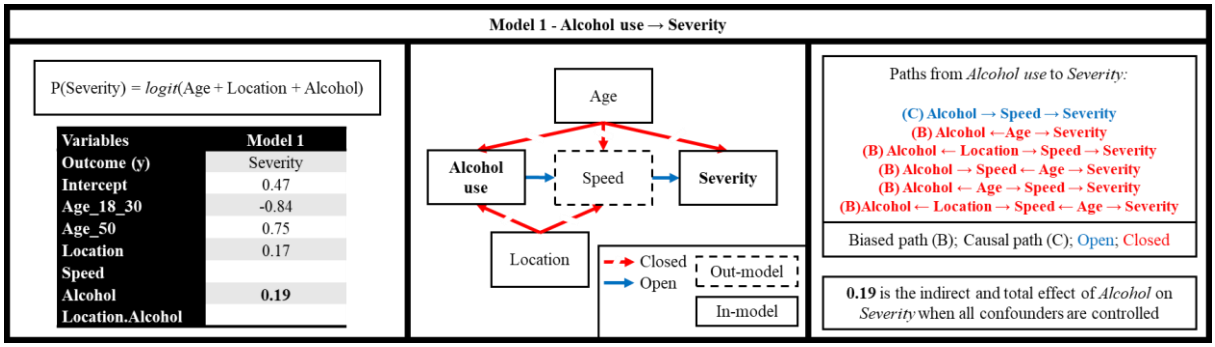
4.1.1 Simple logit models

In this section, simple logit models were used to illustrate the characteristics of the DAG of Figure 18. For this purpose, Monte Carlo simulations were estimated using the

simulated data. Six logit models were estimated, varying the dependent and independent variables.

Model 1 (Figure 19) demonstrates the overall (and indirect) impact of *Alcohol use* on *Severity*. There are six pathways connecting *Alcohol use* to *Severity*, but only one of these pathways is causal. To accurately determine the causal effect, it is necessary to include variables that can control the other five pathways. In this case, *Location* and *Age* must be incorporated into the model. The table within the figure presents the coefficient values from a logistic regression analysis, with the bolded coefficients representing the effects of interest. The other values in the models should not be considered as implying causality, as they are only used to control for confounding factors and are not representative of the main relationship of interest.

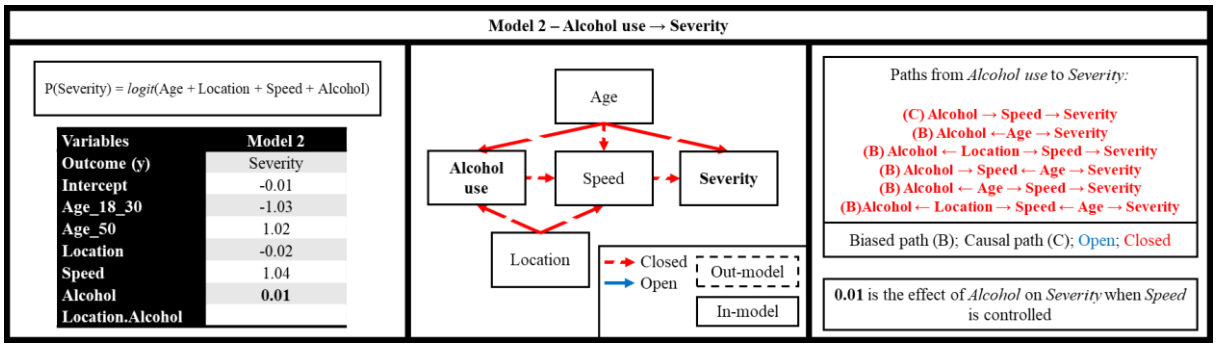
Figure 19 – Model 1



Source: The author.

Incorporating *Speed* into the model can obstruct all pathways, including the causal pathway between *Alcohol use* and *Severity*. As a result, the causal effect may be significantly reduced or even disappear, as depicted in Figure 20. This highlights the importance of selecting variables carefully when determining the causal effect, as the inclusion of certain variables may distort the true relationship being analyzed.

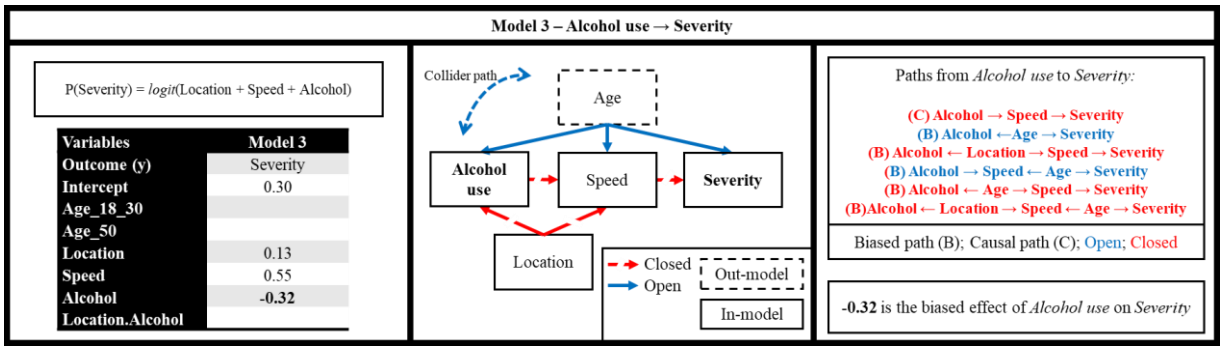
Figure 20 – Model 2



Source: The author.

When a confounder, such as *Age*, is not controlled in the model, it can significantly impact the relationship being analyzed. In Model 3 (Figure 21), the absence of control for *Age* results in a skewed relationship between *Alcohol use* and *Severity*, with the coefficient value being *-0.32*. However, if *Age* were controlled (as in Model 2), the relationship would likely be closer to zero, given that *Speed* is also controlled in the model. Additionally, two biased pathways remain open in Model 3. The first pathway, *Alcohol* ← *Age* → *Severity*, represents a 'backdoor' relationship, while the second pathway, *Alcohol* → *Speed* ← *Age* → *Severity*, involves a 'backdoor' relationship *Speed* ← *Age* → *Severity* and a collider relationship *Alcohol* → *Speed* ← *Age*, which all contribute to a biased effect. It is important to remember that a collider (a variable that is influenced by two or more other variables) can transmit an association when it is controlled in the model.

Figure 21 – Model 3

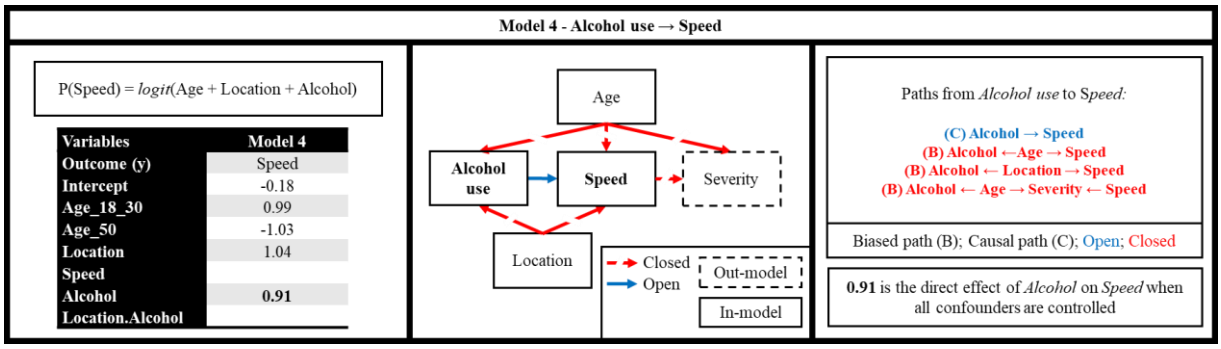


Source: The author.

Model 4 displays the direct effect of Alcohol consumption on Speed. By accounting for all confounding variables and ensuring that only the causal pathways are open, the coefficient

value in this model is in line with established SCMs, as demonstrated in Figure 18, approaching 1.

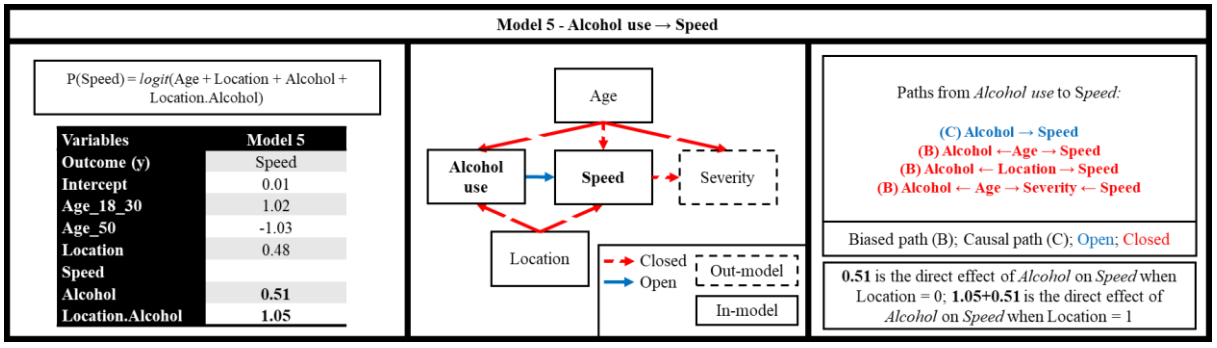
Figure 22 – Model 4



Source: The author.

Model 5 (Figure 23) illustrates the modification effect scenario. The interaction between *Location* and *Alcohol use* is included in the model, leading to a coefficient value of *0.51* for *Alcohol* and *1.05* for the interaction term *Location*Alcohol*. If Location is equal to 0 (rural areas), the effect of *Alcohol use* on *Speed* in these areas would be 0.51. In urban areas, the effect would be *0.51 + 1.05 = 1.56*. The overall effect of *Alcohol use* is calculated as the average of these two values, resulting in *(0.51 + 1.56)/2 = 1.04*, which is close to the established effect of *1* (Figure 18) and the effect found in Model 4 (Figure 22).

Figure 23 – Model 5

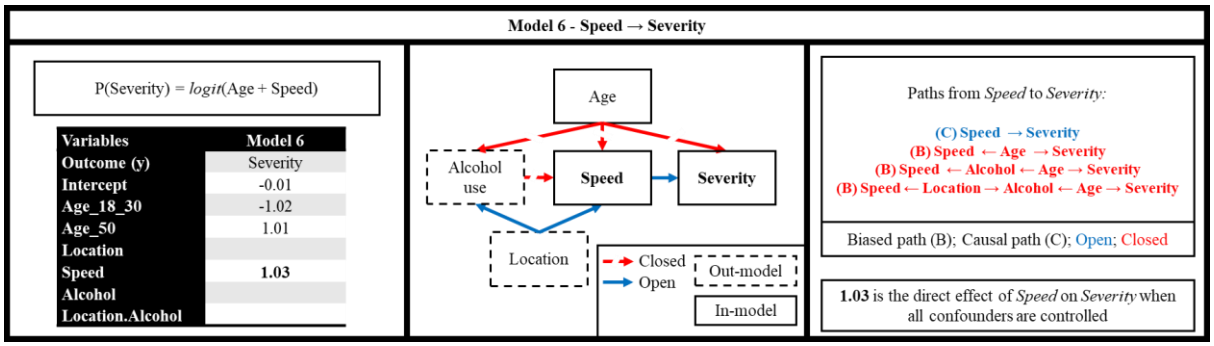


Source: The author.

Model 6 (Figure 24) displays the causal effect of *Speed* on *Severity*. To accurately estimate this effect, it is necessary to control for the confounder *Age*. The variable *Location* does not need to be included in the model to evaluate the causal effect of *Speed* on *Severity*. The model contains four pathways, with one being the causal pathway. By controlling for *Age*

in the model, all biased pathways are blocked, leaving only the causal pathway open for evaluation.

Figure 24 – Model 6



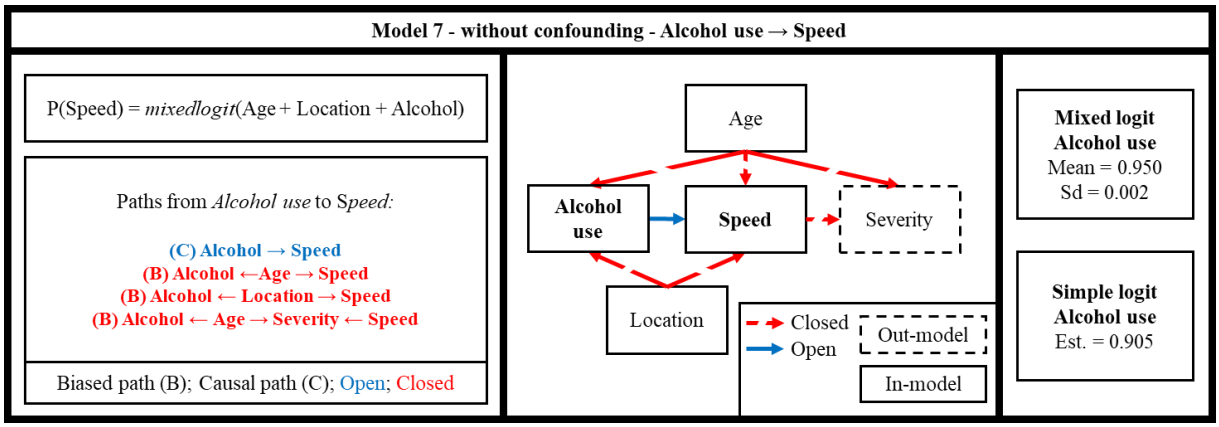
Source: The author.

4.1.2 Mixed logit models

Recent studies used mixed logit and probit (random parameters) to find the effects of factors. These models allow a more generalized structure since it can include unobserved heterogeneity. To demonstrate the use of these models within the framework of causal inference (confounding and modification effect), Monte Carlo simulations were conducted using simulated data, with 500 iterations and 400 observations.

The baseline model used was similar to Model 4 in the previous section. Model 7 (Figure 25) shows a mixed logit model in which the *Alcohol use* coefficients are randomly generated following a Normal distribution. All backdoor paths are closed and the causal path is open, yielding the causal coefficients. The standardized deviation (sd) of the random coefficients is low, implying that most of the coefficients are close to an average of 0.95. A simple logit model was also estimated, producing a result of 0.91, which is similar to the previous one. The slight difference between the two results may be due to differences in the type of estimation and the specifications of the model.

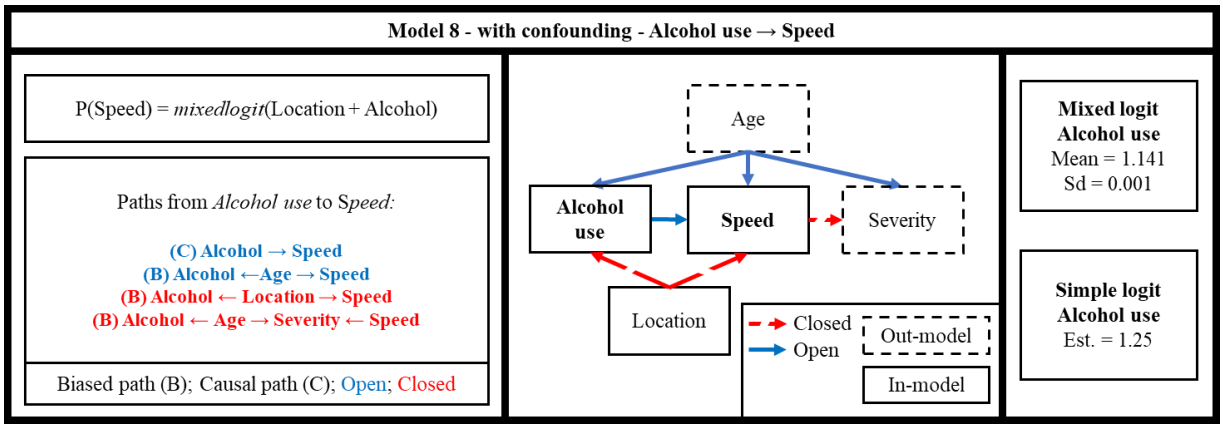
Figure 25 – Model 7



Source: The author.

The next model (Figure 26) shows the impact of opening a backdoor path when using both random and traditional logit estimations. Both estimations exhibit bias. The mixed logit model has a low *sd* once more, and the mixed logit coefficient approaches the causal one. Nevertheless, relying solely on the mixed logit is insufficient to block the backdoor. Thus, it is crucial to incorporate causal theory to accurately evaluate the causal effect.

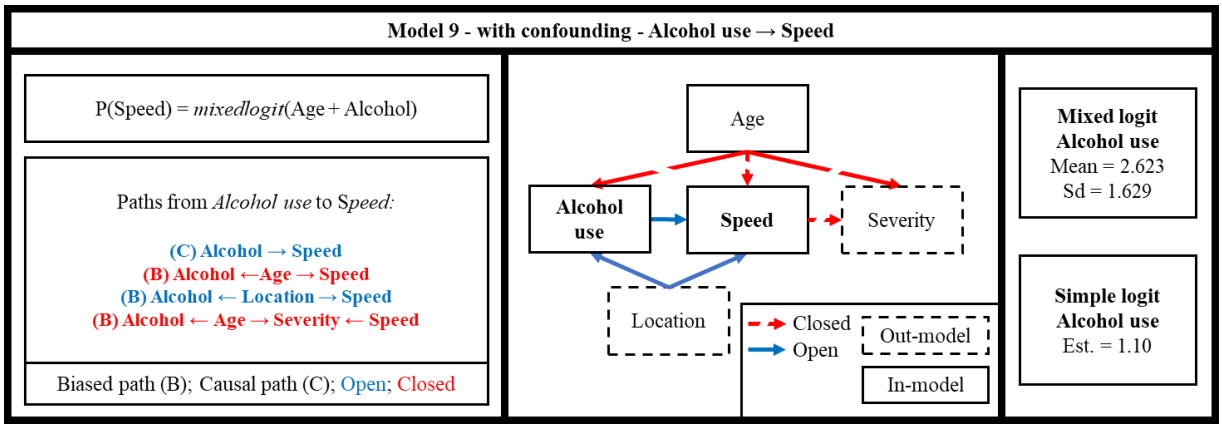
Figure 26 – Model 8



Source: The author.

Model 9 (Figure 27) highlights another case of biased results, this time due to a different confounding variable, *Location*. The variable *Location* also has an effect modification on the relationship between *Alcohol use* and *Speed*, which may explain the higher standard deviation of the mixed model compared to other models. However, all coefficients in the model are biased and diverge from the true causal relationship.

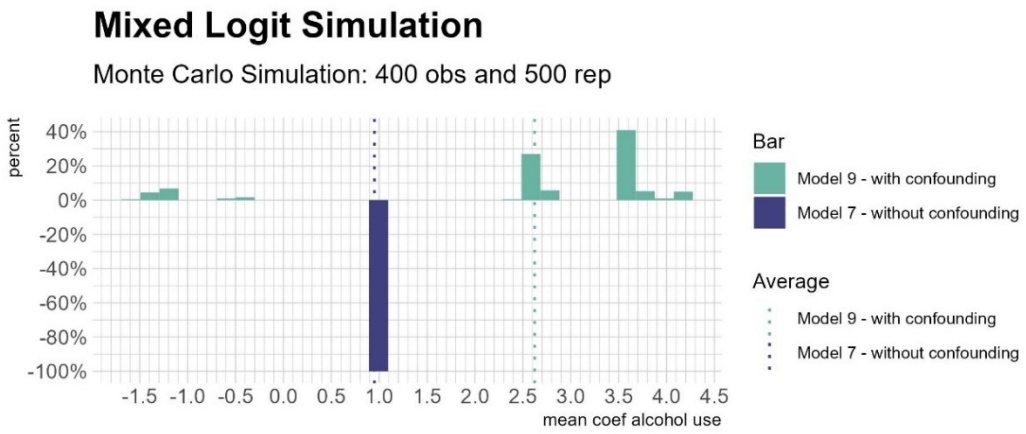
Figure 27 – Model 9



Source: The author.

Figure 28 presents the results of the coefficients (per observation) of *Alcohol use* on *Speed* for the mixed logit models (Model 7 and Model 9). The biased model (Model 9) has an open backdoor, *Alcohol use* ← *Location* → *Speed*. In this scenario, there is an unobserved heterogeneity as the coefficients of *Alcohol use* on *Speed* display high variability. This variability occurs probably because the effect varies between individuals in rural and urban areas (*Location*). When the *Location* variable is included in the model, the *Alcohol* coefficient tends to converge to the overall value. Thus, while mixed models alone may not be sufficient to address endogeneity (*confounding*), they can provide insight into the unobserved heterogeneity caused by the modifying effect of the *Location* variable.

Figure 28 – Results of Monte Carlo Simulation on SCM with logit links



Source: The author.

4.1.3 Graphical models

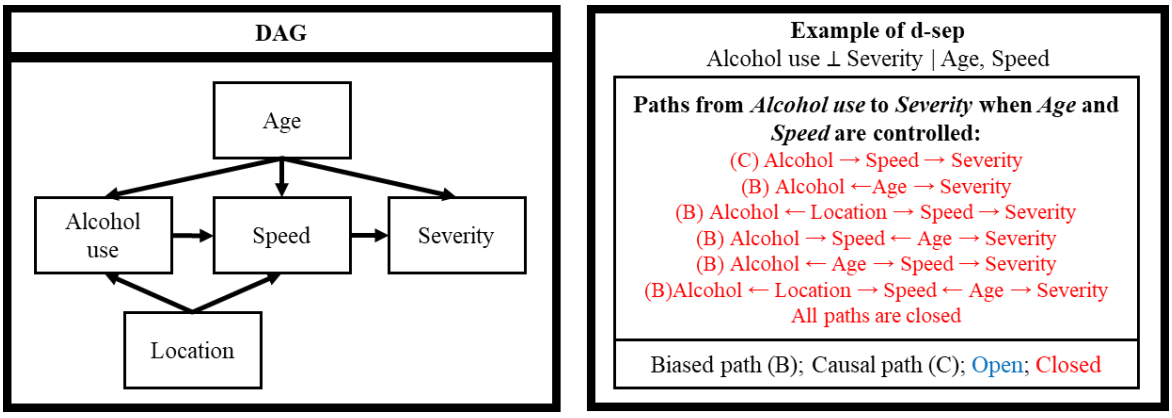
In this section, only one simulation was used because it is not possible to obtain average results from a graphical representation. To compensate for this, 200,000 observations were generated, equivalent to 400 times 500, to ensure reliable data.

The graphical models or DAGs can be generated in two ways. The first is by using a theoretical model, where the relationships between variables are formulated based on previous knowledge and a review of the literature. The second method is by creating a DAG using observed data, for instance, by using Bayesian network methods.

This section provides an example of a DAG that was created based on a review of the literature. The DAG can then be evaluated using observed data, such as through the use of Structural Equation Modeling (SEM). To fully understand how SEM models work, it is important to be familiar with the concept of d-separation. A DAG has a set of conditional probabilistic independencies that can be tested using SEM or other techniques, such as conditional correlations.

D-separation statements describe all conditions under which variables are independent given other variables by a DAG. For instance, in the DAG shown in Figure 29, if Age and Speed are controlled, Alcohol use and Severity are independent (symbol "⊥"), as all paths between them are blocked (Alcohol use ⊥ Severity | Age, Speed).

Figure 29 – DAG and d-sep



Source: The author.

If this conditional independence is not observed in the data, the DAG will not be consistent with the data. The set of d-separations in the previous DAG can be expressed as follows:

1) *Age ⊥ Location*: Age is independent of Location;
2) *Alcohol use ⊥ Severity | Age, Speed*: Alcohol use is independent of Severity when Age and Speed are controlled;
3) *Location ⊥ Severity | Age, Speed*: Location is independent of Severity when Age and Speed are controlled.

One can explore the testable implications of the DAG by employing conditional correlations. In the case of categorical data, Chi-square tests provide a suitable analysis method. Table 9 presents the results obtained from the simulated data, which consisted of 200,000 observations. Despite some of the p-values (H_0: there is no association, $p > 0.05$; H_1: there is an association, $p < 0.05$) being smaller than 0.05, indicating a not true statement and a potential inconsistency between the DAG and the data, the large sample size often leads to lower p-values. Nonetheless, all the Root Mean Square Error of Approximation (RMSEA) values were less than 0.05, indicating a relatively good fit between the DAG and the data. An essential consideration is to acknowledge that the data used for analysis was generated based on the DAG, elucidating the complexities arising from working with large sample sizes.

Table 9 – D-separation statements using chi-square tests

d-sep	rmsea	χ^2	df	p.value	rmsea 2.5%	rmsea 97.5%
Age ⊥ Location	0.001	2.54	2	0.28	0.000	0.006
Alcohol ⊥ Severity | Age, Speed	0.007	17.28	6	0.01	0.000	0.022
Location ⊥ Severity | Age, Speed	0.002	3.70	6	0.72	0.000	0.016

Source: The author.

The estimation of SEM provides the relationship between the observed and estimated correlation matrices of the endogenous[5] variables in the model. If the DAG is not consistent with the observed data, meaning that the d-separation statements are not true, then the matrices will differ (SCHUMACKER; LOMAX, 2010). To evaluate the discrepancy, the RMSEA and p-value metrics can be used.

Figure 30 illustrates an example of the estimated correlation matrices using simulated data. The model was estimated using the WLSMV and with probit link, utilizing the *lavaan* library. Figure 30 (c) displays the difference between the observed (a) and estimated (b) matrices, and it is expected that the values should be as close to zero as possible.

[5] Endogenous variables are those that are influenced by other variables in a system or model. On the other hand, exogenous variables are variables that are not influenced by other variables in the system, but rather affect the endogenous variables.

Figure 30 – Example of SEM, observed (a), estimated (b), and residues (c) matrix

Source: The author.

The model's evaluation metrics, including a p-value of 0.00 and a RMSEA of 0.031 [0.028, 0.033], suggest that the model is in line with the data, given that the RMSEA value falls below the threshold of 0.05. However, it is essential to highlight that this consistency does not necessarily imply that this is the only valid DAG for the data. There may be other DAGs that share the same d-separation statements and are also consistent with the data.

The SEM can estimate both direct and indirect effects in a single estimation. In the example, the indirect effect of Alcohol use on Severity was calculated to be 0.103 [0.099, 0.106]. This value differs by 0.19 from the previous logit model (Model 1 - Figure 19) because the SEM used a probit link instead of a logit link. Currently, the *lavaan* library in R does not support the estimation of models with logit links, only probit links.

Furthermore, it is possible to use SEM with groups to generate two models simultaneously. For example, it is possible to create two models with distinct Locations (group 1 is Urban and group 2 is Rural) and estimate both models at the same time. The effect of Alcohol use on Speed in group 1 was estimated to be 0.548 [0.533, 0.562] and in group 2 was estimated to be 1.179 [0.169, 0.189]. These values are close to the theoretical values of 0.5 for group 1 and 1.5 for group 2. This example demonstrates the usefulness of SEM models for causal analysis.

4.1.4 Summary

In this section, multiple models were utilized to demonstrate the process of making causal inferences based on observed data. The examples highlighted the difficulties that arise when all relevant variables are not included in the model, which is a common issue in road

safety studies that focus on the severity of crashes. The logit models can measure direct and indirect effects, but SEM is more robust. SEM allows for the simultaneous estimation of all relationships, and it is possible to use group analysis to measure the effects across different subpopulations in the data.

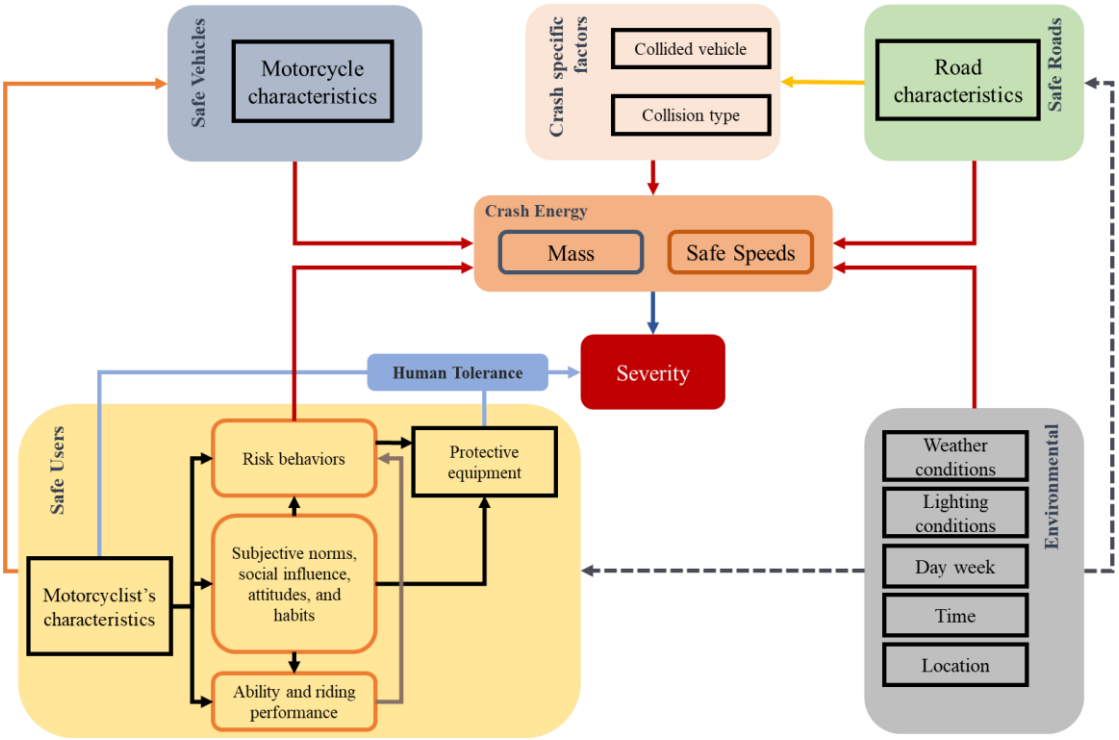
It is important to be aware of database issues, such as temporal and spatial dependence and missing data before interpreting the coefficients as causal. When there is a problem of unobserved variables, latent variables can be employed to address this issue. A latent variable refers to a hidden factor that influences the relationship between observed variables but cannot be directly observed or measured. For instance, a study may observe that helmet use is associated with a decrease in motorcycle crash fatalities, but there may be other underlying factors such as driver skill or road conditions that are not directly measured but could be affecting the relationship. In this scenario, driver skill and road conditions can be considered latent variables. To account for these latent factors, researchers may apply statistical methods such as SEM, which leverages other available variables to reflect latent variables.

4.2 A conceptual model of motorcyclist severity based on the Safe Systems

The conceptual model presented in this dissertation is based on Chapter 2, which focuses on the Safe System approach. This representation considers crash energy and human tolerance as the main components of crash severity. Figure 31 shows the proposed interrelationships between factors and the dimensions of the Safe System. However, it should be noted that the model only represents the severity of a crash that has already occurred and does not take into account post-crash factors.

This conceptual model representation of the impact of factors on the severity of crashes involving motorcyclists is based on a comprehensive review of existing literature. Some of these hypotheses will be tested using observed data in subsequent sections of this dissertation.

Figure 31 – Conceptual model of motorcyclist severity



Source: the author.

Human tolerance in a crash is shaped by the combination of the protective equipment worn and the characteristics of the motorcyclist. The crash energy, as a result of the interaction between mass and speed, is the most crucial factor that determines the fatal outcome of a crash.

Safe motorcyclists typically use safer equipment that increases their tolerance to crash forces and reduces the severity. Certain characteristics of the motorcyclist, such as age, gender, education, experience, and riding license, have a direct impact on their tendency to engage in risky behaviors. In this study, risky behavior is defined as actions that increase the likelihood of crashes or the severity of injury in the event of a crash. Examples of such behaviors include speeding, riding under the influence, lane splitting, and not wearing protective gear. Furthermore, the motorcyclist's characteristics can also have a direct impact on the motorcyclist's tolerance to crash forces, based on factors such as age, gender, and other individual characteristics.

Additionally, there are also unobserved factors that can impact a motorcyclist's behavior, such as *subjective norms, social influence, attitudes, and habits*. These factors may not be directly observable, but they can have a significant impact on a motorcyclist's tendency

to engage in risky behaviors. This behavior is also influenced by the environment, such as weekends, rainy days, time of day, and location, among others.

Safer motorcyclists often choose safer motorcycles. However, despite their efforts, the motorcycle itself is unable to provide complete protection to the rider in the event of a crash, as the rider is often ejected from the vehicle.

Speed is interrelated with motorcyclist, motorcycle, road, and environmental factors. Safe motorcyclists tend to adopt lower speeds and exhibit behaviors that avoid dangerous situations, such as reckless overtaking. Vehicles with larger engines are often associated with higher speeds. The absence of an arrow linking Safe Roads to Safe Users stems from the hypothesis positing that risky behaviors are intrinsic to the users, rather than being instigated by the travel environment.

The vehicles involved in the crash (such as a car, bus, truck, or motorcycle) and the collision type (such as frontal or rear) are related to the kinetic energy involved in the impact. It is worth noting that the environment and crash location (urban or rural) exert considerable influence on all the factors and relationships involved, as evidenced by previous studies mentioned in Chapter 2. To obtain more reliable results, it is important to study the effect of each causal hypothesis in different locations.

4.3 A practical example of causal inference on Road Safety

This section presents a practical example of the causal inference theory on road safety by using observational data from Brazilian highways. The purpose of this section is to establish a relationship of interest and formulate causal hypotheses between the independent variables and the outcome of road crashes. A causal model is crafted to estimate and assess the impact of diverse factors on road safety. Its primary purpose is to offer insights into the most critical contributors to road crashes and recommend potential interventions aimed at enhancing road safety.

4.3.1 Relationship of interest and causal hypotheses

The relationship between alcohol use and road safety was selected as a practical example of causal inference due to its ability to be tested using observational data from Brazilian highways. The first hypothesis is that *ALCOHOL* leads to *SPEEDING* (**H1**).

Subsequently, *SPEEDING* is a major factor affecting *SEVERITY* because of the impact energy (**H2**).

The last hypothesis (**H3)** represents the direct relationship between *ALCOHOL* and *SEVERITY*. This hypothesis considers the various ways in which alcohol use can contribute to road crashes, such as failure to wear protective gear, loss of control while operating a vehicle, violation of traffic laws, etc. By examining the direct impact of *ALCOHOL* on *SEVERITY*, a comprehensive understanding of the role of alcohol use in road safety may be achieved. Furthermore, **H1** and **H2** may not fully represent the effect of *ALCOHOL* on road safety. As a result, **H3** is a means of verifying the other impacts that alcohol use causes on severity, beyond excessive speed.

Unfortunately, the reliability of the *SPEEDING* variable may be compromised as it relies on the subjective determination made by a police officer to assess whether a crash was caused by speeding. Consequently, this variable was not utilized, and instead, the measurement of the total effect of alcohol on Severity was chosen (T1 = H3 + H1*H2) (Figure 32).

Figure 32 – The causal hypothesis

Source: the author.

4.3.2 Motorcyclist database

The data were collected from the Brazilian Federal Highway Police Department (PRF) between 2017 and 2019 (1909 observations). Table 10 shows the variables that were collected and the respective groups to which they belong (motorcyclist characteristics, vehicle characteristics, speed, environmental, and specific factors), a brief description, and the percentage observed for each class.

The data were selected to consider only crashes involving at least one motorcycle and one other vehicle. Additionally, the data only include crashes that occurred on the national highways (BRs) in the state of Ceará. The causal model, which includes the process of *ALCOHOL* → *SEVERITY*, and all the variables that influence this relationship, will be formulated in section 4.3.3. The following sections present only exploratory analyses of the variables and the construction of interrelationships in each group.

Table 10 – Variables used in the study

Group	Variables	Levels	Description	Count (%)
Safe users	*US_AGE*	age_18_30	Riders between 18 and 30 years old	672 (38%)
		age_30_50	Riders between 30 and 50 years old	843 (48%)
		age_50	Rider over 50 years old	232 (13%)
	US_GEN	**female**	Female motorcyclist	115 (6.4%)
		male	Male motorcyclist	1,691 (94%)
	ALCOHOL	Yes	One of the reasons for the crash was the use of alcohol	167 (8.7%)
	LACK_ATT	Yes	One of the reasons for the crash was the lack of attention	1,035 (54%)
	N_SAFE_DIST	Yes	One of the reasons for the crash was not keeping a safe distance	261 (14%)
	OVERTAK	Yes	One of the reasons for the crash was not making a safe overtaking	38 (2.0%)
	TRAF_RU_DIS	Yes	One of the reasons for the crash was disobedience of traffic rules	421 (22%)
Safe vehicles	*ENG_SIZE*	**cc_150**	Engine size below 150cc	1,261 (73%)
		cc_above_150	Engine size above or equal 150cc	477 (27%)
	VCLE_AGE	**vehicle_age_2016**	Other years	1,401 (82%)
		vehicle_age_above_2016	Vehicle made after 2016	313 (18%)
	VCLE_PROB	Yes	One of the reasons for the crash was a mechanical defect in the vehicle	48 (2.5%)
Safe roads	*RD_TYPE*	curve	Road with curve or roundabout	164 (9.4%)
		intersection	Intersections of roads	123 (7.0%)
		straight	Straight roads, viaducts, tunnels, or bridges	1,467 (84%)
	RD_LANES	double	Double lanes	644 (34%)
		multiple	Multiple lanes	375 (20%)
		simple	Single lane	890 (47%)
	RD_PROB	Yes	One of the reasons for the crash was problems on the road, visibility, or signaling	89 (4.7%)
Safe Speeds	*SPEEDING*	Yes	One of the reasons for the crash was incompatible speed	53 (2.8%)
Environmental	*HR_NIGHT*	**day**	Others	1,231 (64%)
		night	Between 06:00 pm and 05:00 am	678 (36%)
	HR_RUSH	**no_rush**	Others	1,199 (63%)
		rush	Between 07:00 am and 09:00 am, and	710 (37%)

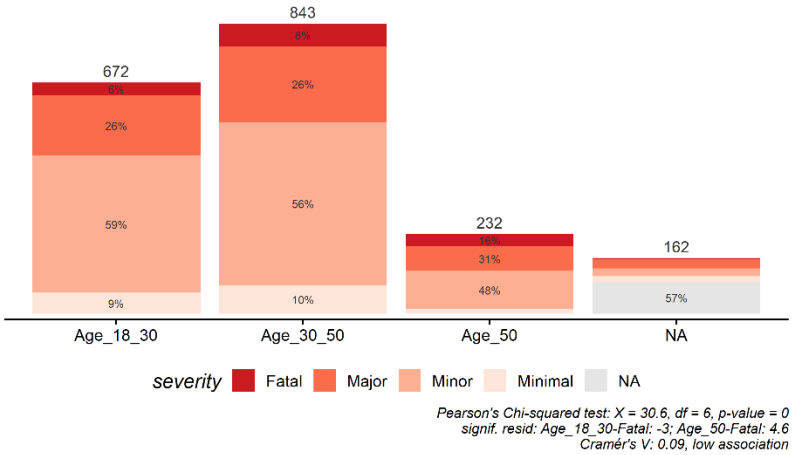
Group	Variables	Levels	Description	Count (%)
			between 05:00 pm and 07:00 pm	
	LAND_USE	rural	Rural area	628 (33%)
		urban	Urban area	1,281 (67%)
	WEATHER	**sunny**	Weather with clear sky or sun	1,539 (82%)
		cloudy	Weather with fog	241 (13%)
		rainy	Weather with rain or drizzle	93 (5.0%)
	WEEKDAY	**workday**	Workday	1,315 (69%)
		weekend	Weekend	594 (31%)
Specific	*VCLE_COLL*	**car**	The vehicle that collided was a car or an SUV	1,009 (54%)
		heavy	The vehicle that collided was a truck or a bus vehicle	335 (18%)
		light	The vehicle/person that collided was a bike or a pedestrian	199 (11%)
		PTW	The vehicle that collided was a PTW	342 (18%)
	COLL_TYPE	frontal	Frontal collision	169 (8.9%)
		others	Rear collision	1740 (91,1%)
Severity	*SEVERITY*	**Minimal and Minor**	Minimal and Minor injury	1180 (64,5%)
		Major and Fatal	Major and Fatal injury	636 (34,8%)

*** reference group**
Source: the author.

4.3.2.1 Safe users

Regarding the age (*US_AGE*) an initial exploratory analysis (Figure 33) showed that young motorcyclists (aged 18-30) had lower odds of being involved in a fatal crash (residual = -3), while older motorcyclists had higher odds (residual = 4.6). The Cramer's V and chi-square test revealed that this variable had a significative association with severity. As explained in the methodology, Cramer's V is a correlation coefficient that measures the strength of association between two categorical variables.

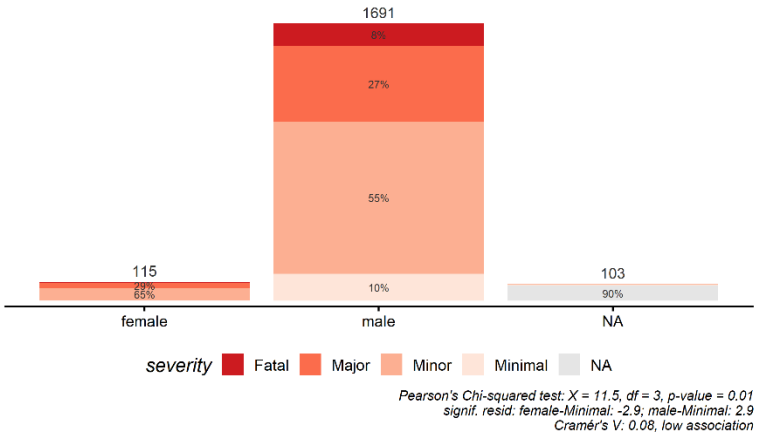
Figure 33 – Motorcycle age and Severity



Source: The author.

An analysis of the gender of motorcyclists (*US_GEN*) (Figure 34) indicated that female riders have lower odds of sustaining a minimal injury compared to male riders (residual = -2.9). However, there were no significant differences between male and female riders in other severity levels. The results of the Cramer's V and chi-square test showed that this variable had a statistically significant association with severity.

Figure 34 – Motorcycle gender and Severity



Source: The author.

The characteristics of motorcyclists, such as *US_GEN* and *US_AGE*, play a crucial role in determining the severity of a crash. This is primarily because these characteristics are associated with two critical factors - *Human tolerance* and *Risky behavior*, as previously

defined. *Human tolerance* varies across different genders and age groups, which can affect the level of injury sustained in a crash. In the same vein, risky behaviors like speeding or driving under the influence are more likely to be prevalent within specific age and gender groups, increasing the probability of severe crashes.

Risky behavior is a factor that is difficult to measure directly, as it is likely a "latent construct". As explained before, a latent construct is a theoretical concept that is not directly measurable but can be measured indirectly by collecting data on related observable indicators. Therefore, in this dissertation, risky behavior was treated as a latent variable that could be measured using other variables in the dataset.

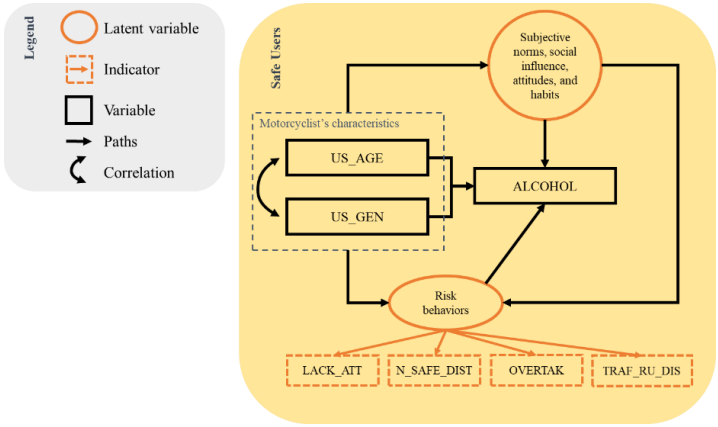
The indicators used to measure the latent construct of *Risky behavior* in this study were lack of attention (*LACK_ATT*), failure to maintain a safe distance (*N_SAFE_DIST*), overtaking (*OVERTAK*), and disobedience of traffic rules (*TRAF_RU_DIS*). These factors are expected to be representative of the *Risky behavior* of motorcyclists.

Another variable that is difficult to measure directly is the combination of *Subjective norms, social influence, attitudes, and habits*. While these can all contribute to motorcycle crash severity, they are more commonly associated with the field of social psychology and therefore the formulation is not within the scope of this dissertation.

Another important point to consider is the reliability of the variable *ALCOHOL* as it is based solely on suspicion by the police officer who collected the crash data. Therefore, the issue of data collection could introduce bias into the results, despite the specification of the model taking all necessary precautions to estimate causal effects.

Figure 35 illustrates all of the processes mentioned in this section. It is a visual representation that has been derived from the conceptual model presented earlier in this dissertation (Figure 31).

Figure 35 – Safe Users causal process with observed data



Source: The author.

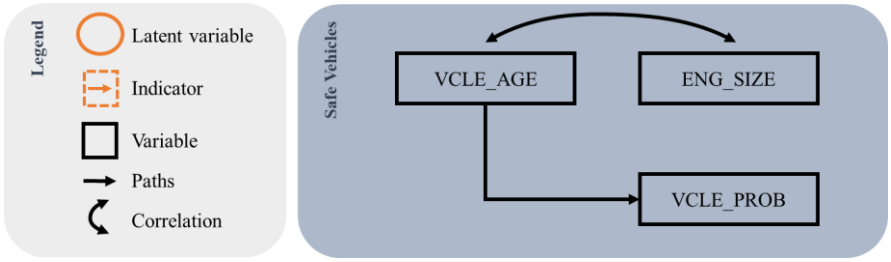
4.3.2.2 Safe vehicles

The age of a vehicle (*VCLE_AGE*) can provide important information about its safety features, including whether it has an Anti-lock Braking System (ABS) and other safety equipment. To capture this information, the *VCLE_AGE* was categorized into two categories: vehicles manufactured in or after 2016, when ABS became mandatory in Brazilian legislation, and vehicles manufactured before 2016. This classification can also reflect both the maintenance conditions (*VCLE_PROB*) and the presence of safety devices. *VCLE_PROB* is a variable recorded by the field agent to determine the motive of the crash was a problem in the vehicle.

The speed adopted by motorcycle riders is likely to be linked to the engine of the motorcycle, and this factor could also be associated with risky behavior. The engine size of the motorcycles (*ENG_SIZE*) in the study is divided into two categories: engines with a size above or equal to 150cc (cubic centimeters) and engines below 150cc. The 150cc threshold was chosen because it is the most common engine size for motorcycles in Brazil. The Cramér's V and Chi-square tests conducted on this variable showed an association with severity. The tests and graphs can be found in Appendix A.

Figure 36 provides a visual representation of the proposed causal process of safe vehicles using the observed data. The two-point arrow (correlation) represents the relationship between variables, but it is not the primary focus of interest in this analysis. Therefore, it is not essential to establish whether the age of the motorcycle causes the engine size, the engine size causes vehicle problems, or whether there is another variable that causes both. However, it is important to note that these relationships could occur.

Figure 36 – Safe Vehicles causal process with observed data



Source: The author.

4.3.2.3 Safe Roads

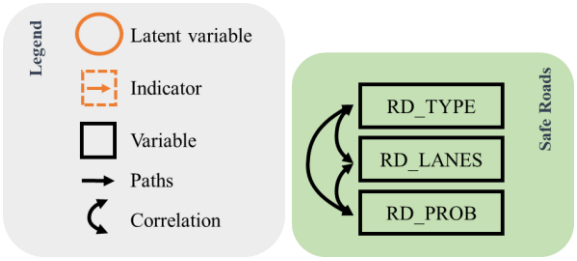
The nature of a road (*RD_TYPE*) can be classified into three types: curved, straight, or intersection. This classification has a significant impact on the behavior of motorcyclists on the road. For instance, on straight roads, motorcyclists are more likely to speed or overtake other vehicles. On curved roads, motorcyclists must exercise extra caution and take additional precautions while navigating to avoid exiting the road and potentially falling off their motorcycles. At intersections, the angle of collision can lead to higher impact energy, resulting in more severe crashes.

The second variable to consider is the number of lanes on the road (*RD_LANES*), which can be classified as simple, double, or multiple lanes. It is important to note that the number of lanes on the road can affect the behavior of road users, with drivers on multiple lanes being more likely to engage in lane changes, overtaking, and speeding.

The final variable is *RD_PROB*, which indicates whether any issues related to the road were identified as contributing factors to the crash by the agent in the field. This variable helps to understand the role that road conditions played in the occurrence of the crash.

Figure 37 depicts the causal process underlying Safe Roads, using observed data. The figure includes variables such as *RD_TYPE* and *RD_LANES*. The variables in the figure are likely correlated, as indicated by the two-point arrows, but for this dissertation, these relationships are not the focus.

Figure 37 – Safe Roads causal process with observed data



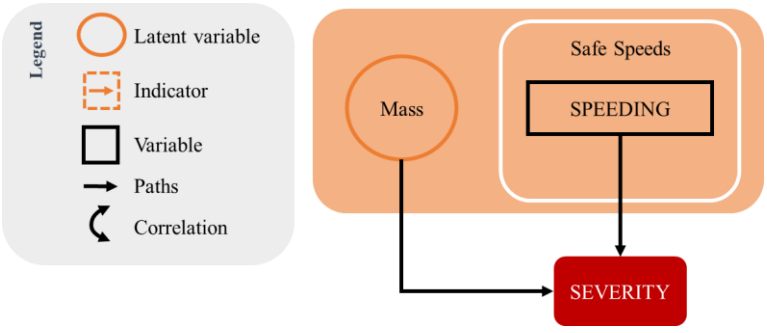
Source: The author.

4.3.2.4 Safe Speeds

The variable that represents speeding is affected by the behavior of the motorcycle and the use of alcohol. It indicates if the motorcycle was traveling at a high speed during the crash, as reported by the police officer. Nevertheless, the reliability of this variable may be compromised due to its dependence on the subjective judgment of the police officer in determining if speeding was the motive of the crash. Consequently, this variable was not considered in the analysis. It is worth noting that other variables could also present similar challenges; however, speed is comparatively less reliable since there is a higher propensity for false statements and limited availability of information data.

On the other hand, the mass variable consists of characteristics related to the vehicles involved in the crash. Both speed and mass contribute to the crash energy, which is a factor in determining the severity of the collision (Figure 38).

Figure 38 – Safe Speeds causal process with observed data



Source: The author.

4.3.2.5 Environmental

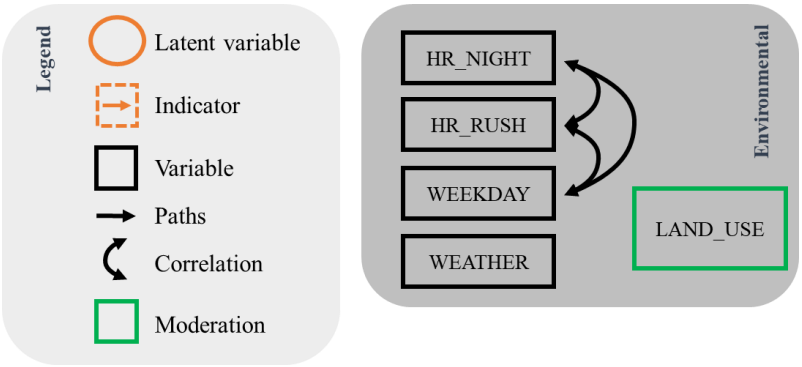
The time-related variables included night and rush hours. *HR_NIGHT* referred to crashes occurring between 6:00 PM and 5:00 AM and was associated with an increased incidence of speeding and alcohol use. Rush hours (*HR_RUSH*), occurring from 7:00 AM to 9:00 AM and 5:00 PM to 7:00 PM, were also related to speed, though to a lesser degree due to increased traffic congestion, particularly in urban areas. The final time-related variable was whether the crash occurred on a weekday (*WEEKDAY*), which was also associated with an increased incidence of alcohol use and speeding.

The *WEATHER* was categorized as sunny for clear or sunny conditions, cloudy for foggy conditions, and rainy for conditions with rain or drizzle. However, the relationship between rainy weather and crash severity is complex and may vary depending on local and road characteristics.

The *LAND_USE* was classified into rural and urban areas and was used as a modifier variable in the analysis. The effects of all other variables were estimated separately for each group (*i.e.*, rural and urban) using group modeling in SEM, as detailed in section 4.1.4.

Figure 39 depicts the causal process using observed data. The environmental factors, including night hour, rush hour, weather, and weekday, are shown to influence the outcome variable. The two-point arrows between the hour night, hour rush, and weekday variables represent their likely correlations, which are not of interest in this dissertation. The land use variable was used as a moderator, allowing the effects of all other variables to be analyzed separately for rural and urban groups.

Figure 39 – Environmental causal process with observed data



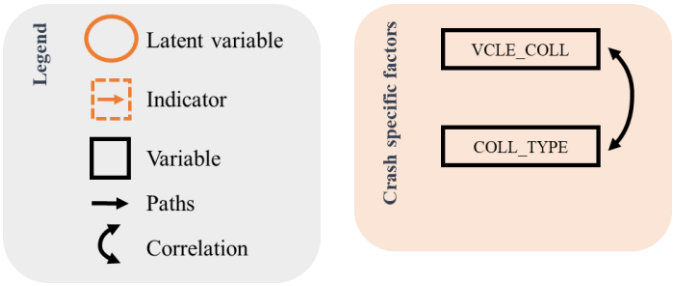
Source: The author.

4.3.2.6 Specific factors

The types of collision (*COLL_TYPE*) analyzed in this study were head-on and others. Head-on collisions are more likely to result in fatal outcomes, while other types of collisions, such as rear-end collisions, typically result in material damage with lower chances of fatalities. The vehicles involved in the collision (*VCLE_COLL*) were categorized as car, heavy, light, and PTW. Heavy vehicles included trucks and buses, light vehicles included bikes and pedestrians, PTW included motorcycles, and cars included all other vehicles. The type of vehicle and collision are closely related to the energy of the impact.

Figure 40 depicts the causal process of specific factors using the observed data. The variables are likely correlated, as represented by the two-point arrows in the diagram. However, as these relationships are not of interest to this dissertation, they are not further analyzed.

Figure 40 – Specific factors causal process with observed data



Source: The author.

4.3.2.7 Database issues

The first type of issue that could lead to skewed results is temporal and spatial dependence. Moran's I test indicated a small spatial dependence with a test value of 0.145, which was significant. Nevertheless, other variables, such as *LAND_USE* (urban and rural), will be included in the model to try to account for this spatial relationship. The residual of the model in the following sections will be tested to verify if there is still spatial and temporal dependence.

The last type of issue in data is missing data. There is a 2.2% rate of missing data, meaning that at least one value in one variable is missing. It was found that variables related to motorcycles have more missing data and appear to be related. Missing data related to *US_AGE*,

US_GEN, and *SEVERITY* may be interrelated. For instance, if *US_AGE* was not recorded, it is possible that *SEVERITY* and vehicle characteristics were not recorded either, indicating that these variables may be related. Although the missing data are few in this database, they could still potentially lead to biased coefficients in the next models. However, the extensive dataset size could enhance the reliability of the results.

4.3.3 Formulation and estimate of the causal model

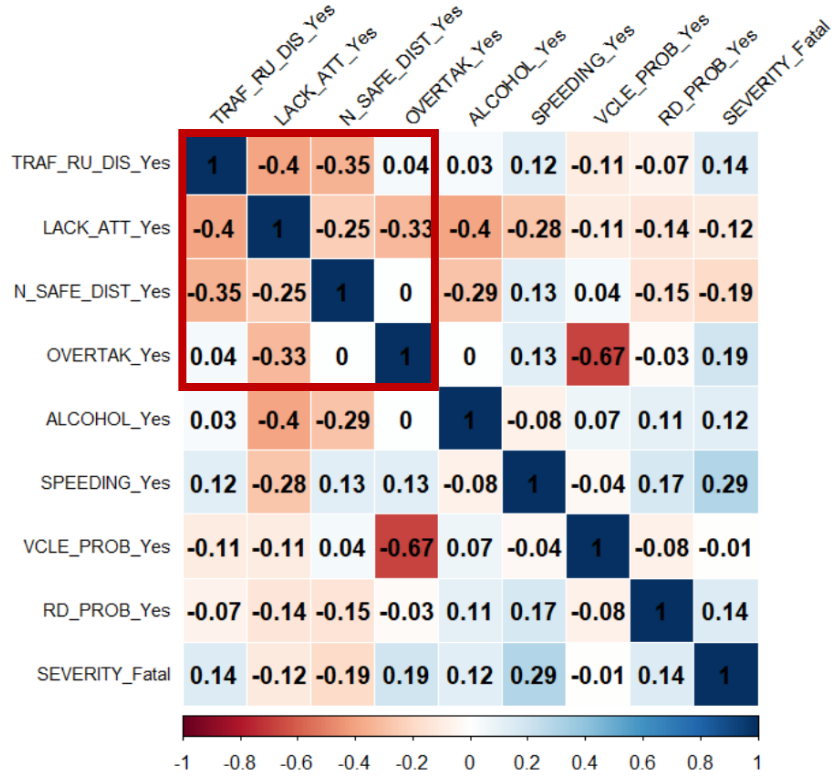
The first part of SEM is the measurement model, which consists of the latent variables. In this study, the latent variables are *Subjective Norms* and *Risk Behaviors*. The second part is the structural model, which illustrates all the relationships between the exogenous and endogenous variables in this study. Estimating these two parts separately is recommended to ensure reliable results (BOLLEN, 1989; BOLLEN; BAULDRY, 2011; GRACE; BOLLEN, 2008; HAIR *et al.*, 2009; HOYLE, 2012; MORRISON; MORRISON; MCCUTCHEON, 2017).

4.3.3.1 Measurement model

Risk Behaviors can be assessed through the following indicators: LACK_ATT, N_SAFE_DIST, OVERTAK, and TRAF_RU_DIS. Initially, a positive correlation between these variables was expected, as they were assumed to adequately represent the latent variable. However, it has been observed that in most cases, the correlation between these variables is negative. This negative correlation suggests a potential issue with the database specification (Figure 41).

One potential explanation for this observation is that police officers frequently select “overtaking” as the reason for a traffic violation without simultaneously indicating “lack of attention”, for example. This inconsistency presents difficulties in accurately assessing risk behaviors. As a result, it was decided to include only the TRAF_RU_DIS variable to represent *Risk Behaviors*. This choice was considered suitable since this variable captures a substantial portion of the user's problematic behavior, displaying a positive correlation with severity and alcohol use.

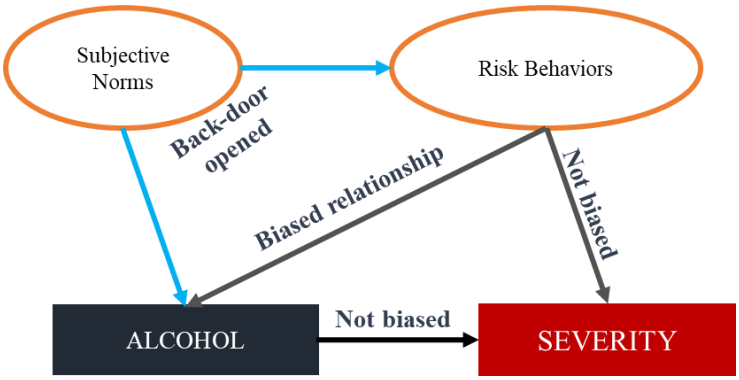
Figure 41 – Tetrachoric correlation among observed variables



Source: The author.

After analyzing the database, it was found that there were no observed variables suitable for representing the latent variable of *Subjective Norm* in the model. Consequently, the decision was made to exclude this factor from the analysis. Importantly, this exclusion does not introduce biased results for causal effects; it only impacts the relationship with *Risk Behaviors*. Figure 42 illustrates that the relationship between alcohol use and severity remains unbiased when *Subjective Norms* are not controlled (omitted from the model). However, the relationship between *Risk Behaviors* and ALCOHOL becomes biased, highlighting the significance of properly categorizing relationships and assessing the effects of including or excluding specific variables in the model. Therefore, in the analysis of the results, this particular relationship will not be considered due to the likelihood of bias.

Figure 42 – *Subjective Norms* relationships



Source: The author.

4.3.3.2 Structural model

Adjustments were necessary to properly fit the model. Firstly, all endogenous variables (SEVERITY, ALCOHOL, ENG_SIZE, VCLE_AGE, TRAF_RU_DIS, and COLL_TYPE) were treated as binary variables (Table 10) with probit links. Secondly, the variable VCLE_PROB was excluded from the analysis due to convergence issues in the model, likely resulting from its high correlation with other variables in the study. However, the removal of this variable does not introduce bias, as the potential back-door effect it could have created can be mitigated by considering VCLE_AGE. Finally, all exogenous variables are freely estimated for correlation among them, adhering to the standard approach of SEM.

Through iterative model fitting, it was identified that adjustments needed to be made to the model to estimate the impact of ALCOHOL on SEVERITY. However, it is crucial to acknowledge that continuously altering the causal model in this manner carries the potential risk of overfitting[6]. Despite this concern, all modifications made to the model were thoroughly justified. Moreover, these adjustments did not result in significant alterations to the fundamental causal structure, indicating a minimal likelihood of any adverse consequences. Table 11 illustrates all the relationships within the causal model, with the operators "~" denoting regression and "~~" indicating correlation.

[6] Overfitting happens when a model becomes too focused on the training data and performs poorly when faced with new, unseen data.

Table 11 – Relationships in the Causal Model

Effect	Relationship	Hypotheses
T1	SEVERITY~ALCOHOL	The total causal effect of alcohol use is greater injuries
A1_AGE	SEVERITY~US_AGE	The age of a motorcyclist is associated with human vulnerability, which can lead to increased severity
A1_GEN	SEVERITY~US_GEN	The gender of a motorcyclist is associated with human vulnerability, which can lead to increased severity
A2_AGE	ALCOHOL~US_AGE	The age of a motorcyclist is associated with alcohol use
A2_GEN	ALCOHOL~US_GEN	The gender of a motorcyclist is associated with alcohol use
A3_AGE	ENG_SIZE~US_AGE	The age of a motorcyclist is associated with motorcycle characteristics
A3_GEN	ENG_SIZE~US_GEN	The gender of a motorcyclist is associated with motorcycle characteristics
A4_AGE	VCLE_AGE~US_AGE	The age of a motorcyclist is associated with motorcycle characteristics
A4_GEN	VCLE_AGE~US_GEN	The gender of a motorcyclist is associated with motorcycle characteristics
A5_AGE	RISK_BEHAV~US_AGE	The age of a motorcyclist is associated with risk behaviors
A5_GEN	RISK_BEHAV~US_GEN	The gender of a motorcyclist is associated with risk behaviors
A6	ALCOHOL~RISK_BEHAV	The risk behaviors of a motorcyclist are associated with alcohol use, and this relationship is often influenced by subjective norms
A7	SEVERITY~RISK_BEHAV	The risk behaviors of a motorcyclist are associated with severity
A8	ALCOHOL~WEEKDAY	Alcohol use is more commonly observed on weekends
A9	ALCOHOL~HR_NIGHT	Alcohol use is more commonly observed during nighttime hours
A10	SEVERITY~WEEKDAY	Weekends are associated with severity due to speeding
A11	SEVERITY~HR_NIGHT	Some drivers may be more prone to speeding during nighttime hours due to reduced traffic
A12	SEVERITY~HR_RUSH	Rush hours are associated with traffic congestion and slower driving speeds
A13	SEVERITY~WEATHER	Poor weather conditions can contribute to unsafe driving behaviors
A14	SEVERITY~ENG_SIZE	Motorcycle engine size is associated with higher speeds
A15	SEVERITY~VCLE_AGE	Motorcycle age size is associated with safe equipment
A16	SEVERITY~RD_TYPE	Road type is associated with the severity
A17	SEVERITY~RD_LANES	The number of lanes is associated with the severity
A18	SEVERITY~RD_PROB	Road problems are associated with the severity
A19	COLL_TYPE~RD_TYPE	The type of road can influence the nature of collisions between vehicles
A20	COLL_TYPE~RD_LANES	The number of lanes can influence the nature of collisions between vehicles
A21	SEVERITY~COLL_TYPE	The type of collision between two vehicles influences the energy of impact, which can result in more severe injuries
A22	SEVERITY~VCLE_COLL	The type of vehicle involved in a collision with a motorcycle influences the energy of the impact, which can result in more severe injuries
C1	ENG_SIZE~~VCLE_AGE	Engine size is also correlated with vehicle age due to factors beyond the rider's characteristics. For example, technological advancements and economic conditions
U and R	LAND_USE	Moderation - The effects of the DAG vary beyond rural and urban areas

Source: the author.

The model was estimated using data on motorcyclist crashes taking into account only valid observations and handling missing data through listwise deletion[7]. To enhance the analysis, a multigroup Structural Equation Modeling approach was employed, simultaneously estimating two models (Rural and Urban). This allowed for comparing estimations and obtaining improved Goodness-of-Fit measures. Additionally, the model utilized the weighted least squares mean and variance adjusted (WLSMV) estimator with a probit link between variables. The model was also able to estimate intercepts, facilitating the estimation of residuals. Moreover, all results were presented as standardized values, enabling easy comparison of the estimates (Figure 43).

The model exhibits favorable performance based on key fit indices. The RMSEA value is found to be less than 0.05, indicating a good fit between the model and the observed data. Additionally, both the CFI and TLI exceed the recommended threshold of 0.90, further confirming the model's strong fit. The significance of the p-value aligns with expectations, considering the considerable sample size of over 400 observations.

It is crucial to acknowledge that while there may exist alternative DAGs that yield a satisfactory fit to the data, the current model was specifically developed based on a well-established theoretical framework. This theoretical foundation enhances its credibility and renders it more reliable compared to other competing DAGs.

Currently, there is a lack of available functions in contemporary R language packages that allow for the computation of residuals for endogenous variables in SEM with probit links. To address this limitation, novel functions were developed, as outlined in the methodology section, to facilitate the calculation of these residuals.

To evaluate the adequacy of the developed functions, rigorous testing was performed on the deviance residuals associated with all endogenous variables. These tests aimed to examine potential spatial and temporal dependence within the residuals, which could impact the reliability and validity of the model. The methodology section provides a detailed explanation of the methods employed in conducting these tests.

The tests revealed the presence of statistically significant values; however, their effect sizes are found to be low, indicating their negligible impact. The significance of these p-values may be attributed to the sample size, while the lower values of the statistics suggest that spatial or temporal dependence within this model is irrelevant.

[7] Listwise deletion is a method in statistics where cases with missing data in any of the variables of interest are completely excluded from the analysis, resulting in a reduced sample size.

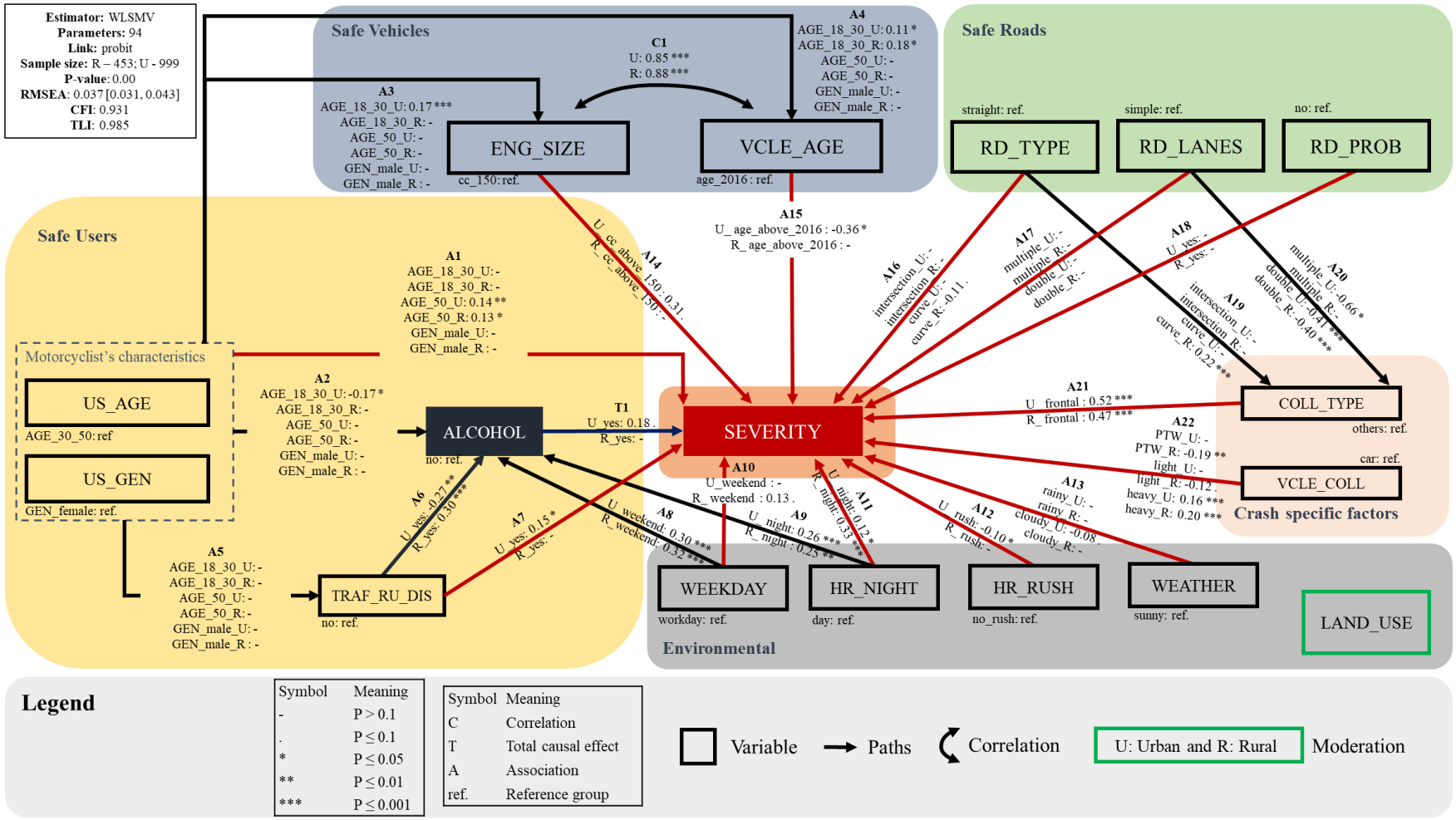
Table 12 – Spatial and Temporal autocorrelation tests

Endogenous Variable	Moran's I test for spatial autocorrelation		Kruskal-Wallis Rank Sum Test for temporal autocorrelation			
			YEAR		MONTH	
	Rural	Urban	Rural	Urban	Rural	Urban
SEVERITY	0.03 (ns)	0.081 (***)	0.002 (ns)	0 (ns)	-0.007 (ns)	-0.001 (ns)
ALCOHOL	-0.043 (ns)	0.023 (.)	-0.002 (ns)	0.003 (ns)	0.016 (.)	-0.001 (ns)
ENG_SIZE	0.046 (ns)	-0.01 (ns)	0.003 (ns)	0.003 (.)	-0.005 (ns)	-0.001 (ns)
VCLE_AGE	0.001 (ns)	0.005 (ns)	0.011 (*)	0.006 (*)	-0.001 (ns)	-0.004 (ns)
TRAF_RU_DIS	-0.001 (ns)	0.064 (***)	0.005 (ns)	0.004 (*)	-0.004 (ns)	0.003 (ns)
COLL_TYPE	0.071 (.)	0.075 (***)	-0.003 (ns)	-0.002 (ns)	0.007 (ns)	0.004 (ns)

Legend: "ns" - Not significant ($P \geq 0.1$); "." – Statistically significant at 90% ($P < 0.1$); "*" - Statistically significant at 95% ($P < 0.05$); "**" - Highly statistically significant ($P < 0.01$); "***" - Extremely statistically significant ($P < 0.001$)
Source: the author.

It is important to recognize that the SEVERITY variable exhibits a Moran's I test value of 0.081, significantly lower than the previously obtained value without the model (0.145). This compelling result demonstrates the model's effectiveness in mitigating spatial dependence by incorporating other influential factors.

Figure 43 – Causal Model



Source: the author.

4.3.3.3 Interpreting the results of the model

It is important to highlight that the *a priori* formulated causal hypotheses of the model indicate that alcohol use increases the probability of speeding and not using safety equipment, thereby resulting in greater impact energy and reduced human tolerance.

The relationship denoted as T1 represents the only causal effect in which all potential back doors have been closed. The DAG depicted in Figure 45 encompasses additional variables, notably road characteristics, specific factors, rush hours, and weather, none of which are designated to close back doors. Nevertheless, these variables serve the purpose of enhancing estimations and fostering comprehension of the analyzed phenomenon concerning road crashes involving motorcyclists.

The study's findings highlight a substantial influence of alcohol (T1) on the severity of motorcycle crashes, especially in urban areas (0.18), where the effect was statistically significant at a 90% confidence level. These results suggest that alcohol has a more pronounced influence on the increase in crash severity in urban settings. A study on motorcycle rider severity yielded a similar result, where the driver of the striking vehicle with alcohol suspicion had a 0.79 probability (p-value = 0.08) of a fatal crash (RAHMAN *et al.*, 2021).

In rural areas, the effect was found to be insignificant in the model, contradicting findings from other studies (CZECH *et al.*, 2010; LOWENSTEIN; KOZIOL-MCLAIN, 2001; TSUI *et al.*, 2010; WUNDERSITZ; RAFTERY, 2017). However, a study revealed that alcohol intoxication does not exhibit a correlation with a higher occurrence of severe injury or mortality in road crashes. However, it does emerge as a prominent predictor for post-injury morbidity (SHIH *et al.*, 2003).

Some hypotheses could explain these results. Urban areas typically have higher population densities and more congested traffic conditions, which may amplify the consequences of impaired judgment and reduced reaction times caused by alcohol consumption. Additionally, the presence of numerous vehicles in urban environments could increase the likelihood and severity of collisions involving motorcyclists under the influence.

Moreover, urban areas often feature complex road networks, including intersections, roundabouts, and multi-lane highways, which can pose greater challenges for intoxicated motorcyclists. Navigating through these intricate traffic patterns while impaired increases the risk of misjudgments and potentially severe crashes.

Another important point to consider is the reliability of the variable ALCOHOL as it is based solely on suspicion by the police officer who collected the crash data. Therefore, the

issue of data collection could introduce bias into the results, despite the specification of the model taking all necessary precautions to estimate causal effects.

The model uncovers noteworthy associations between the age of motorcyclists and their engagement in alcohol-related road crashes in urban areas (A2). Riders between the ages of 30 and 50 exhibit a higher propensity for alcohol consumption as compared to those aged between 18 and 30 years old. The findings also indicate that younger individuals are more likely to be associated with new motorcycles (A4) and those equipped with larger engines (A3). Conversely, the gender of motorcyclists does not display any significant direct associations with other variables in the study.

Motorcyclists above 50 years old are directly associated with the severity of crashes (A1). It is essential to clarify that this coefficient should not be interpreted as the total effect. The total effect is a culmination of all pathways through which the age variable influences severity. In this particular analysis, the direct effect measurement conveys the influence of age in addition to the effect of alcohol use (A2) and the inclination to choose safer vehicles (A3 and A4). Consequently, the direct effect uncovered may be indicative of human vulnerability.

The analysis reveals a significant association between *Risk Behavior* (represented by TRAF_RU_DIS) and alcohol consumption (A6), with rural areas displaying a positive relationship, while urban areas exhibit a negative one. However, this association is influenced by the *Subjective Norms* variable as indicated by the measurement model discussed earlier (Figure 42). As a result, the interpretation of this association may lack coherence or meaningfulness.

Crashes occurring during weekends presented a positive association with alcohol consumption in both rural and urban areas (A8). In rural areas, weekends have a limited but noticeable impact on the severity (A10). During nighttime hours, both rural and urban areas witness an increase in alcohol use (A9) and severity (A11). Several factors could be imagined to influence these results, including more access to alcohol consumption opportunities (open bars), especially during weekends, and reduced traffic flow, which can lead to higher instances of speeding.

Recent advancements in motorcycle technology may have led to a notable trend: newer motorcycles are typically linked to lower severity (A15) in rural crashes, whereas motorcycles with larger engines tend to exhibit higher severity in urban areas (A14). The integration of safety-enhancing features like ABS (Anti-lock Braking System) in modern models required by recent improvements in vehicle regulation in Brazil is likely to play an important part in this positive safety aspect. On the other hand, motorcycles with larger engine

sizes often have a greater potential for achieving higher speeds, thereby increasing the severity of crashes. It is important to note that, the correlation found between the two variables highlights that newer motorcycles frequently feature larger engines (C1). This specific trend in the Brazilian market may reduce the real positive gains brought by the technological advances of motorcycle safety systems.

Curves show a slight but limited association with a decrease in severity in rural areas (A16), potentially attributed to the reduced speed typically observed in these areas. However, curves are also associated with a higher risk of frontal collisions in rural areas (A19), which can amplify the impact energy and result in fatal crashes. When calculating the overall impact of curves on severity (A16 + A19 * A21), the effect is found to be statistically insignificant in both rural and urban areas (the analysis was performed using the R programming language).

The results also confirmed the findings from several studies in terms of the type of crash and type of vehicle involved (JONES; GURUPACKIAM; WALSH, 2013; PERVEZ; LEE; HUANG, 2021; SE *et al.*, 2021). Frontal crashes and crashes with heavy vehicles tended to result in higher severity, due to the increased energy of impact impact both from the combination of speeds and greater mass (A21 and A22).

These effects are a subset of the analyzed variables, and other relationships could be examined using a similar approach. It is crucial to highlight the divergent results between urban and rural areas, as certain variables, such as vehicle age, are only significant in rural areas. This emphasizes the importance of conducting severity analyses considering the specific characteristics of each land-use type. Furthermore, most of the findings align with the literature presented in chapter two. However, it is important to note that the causal effect (T1) was analyzed more diligently, whereas the other relationships are associations and should be interpreted as such.

It is crucial to acknowledge that while there may exist alternative DAGs that yield a satisfactory fit to the data, the current model was developed based on prior knowledge structured from the theoretical framework. This theoretical foundation enhances its credibility and reliability compared to other potential DAGs.

The findings of this study exhibit notable distinctions from conventional approaches due to several compelling factors. Firstly, the outcomes obtained are characterized by reduced biases, attributed to the identification and control of confounding variables. This approach enabled the discernment of potentially biased effects, as the model inherently

recognizes a priori instances of effects influenced by latent factors (e.g., the bias of effect A6 by subjective norms).

Secondly, the use of SEM facilitated the exploration and estimation of relationships across multiple endogenous variables (Ys), thereby enhancing the comprehension of relationships within the Safe System approach. Thirdly, this modeling approach enabled a more conceptual differentiation between direct effects and total effects of variables. Thirdly, the causal model not only allows the examination of the underlying hypotheses through a DAG but also permits the capacity to validate the compatibility of observed data with a given conceptual model. Lastly, a model capable of simultaneously estimating two models, one for rural and another for urban areas, allows for direct comparisons between these estimations. This approach enhances the statistical power due to the larger sample size.

5 CONCLUSION AND FUTURE STUDIES

The specific objectives of this dissertation were threefold. The first objective was to consolidate the knowledge of causal inference approaches in road safety studies concerning the severity of motorcyclist crashes. This was achieved by utilizing a theoretical example and simulated data, illustrating the relationship between alcohol use, speed, and severity. A DAG was established, considering various relationships, including back-door paths. Logit regression and graphical models were employed to determine the causal effects and demonstrate the practical application of these methods to observed crash data. The example illustrates the distinction between causal and traditional modeling methodologies, highlighting how causal modeling can effectively address the presence of open backdoors. In contrast, traditional modeling fails to distinguish between direct, indirect, and total effects, potentially leading to biased interpretations of coefficients. Nevertheless, traditional models remain valuable for identifying associations and uncovering potential relationships.

The second objective of this dissertation was to propose a conceptual model based on the Safe System approach, targeting specifically the identification of causal hypotheses concerning the factors impacting the severity of motorcycle crashes. To accomplish this, an extensive literature review was conducted, with specific emphasis on the Safe System approach for motorcyclists. By conducting an in-depth analysis of numerous studies, this research explores potential relationships among factors influencing crash severity. The findings from this comprehensive review were then synthesized to develop a conceptual model, which played a pivotal role in constructing the causal model. The resulting conceptual model illustrates the intricate interconnections among variables associated with the severity of motorcycle crashes, grouped into six main categories: Safe Users, Safe Vehicles, Safe Roads, Crash Energy, Environmental, and Crash-specific factors. Notably, the model highlights that crash severity is primarily determined by two key factors: human tolerance and crash energy.

The last objective of this study focused on examining causal hypotheses using motorcycle crash data obtained from federal highways in the state of Ceará-Brazil. The collected data underwent thorough processing and analysis, employing graphs and statistical tests. Despite encountering certain challenges, such as missing data and spatial dependence, it is important to note that these issues do not significantly impact the results. This is mainly due to the low effects found when the specific tests were performed in residuals of the final model.

In contrast to traditional models, the causal model utilized in this context is founded on a theoretical framework, derived from comprehensive studies across diverse areas, to

propose relationships. By incorporating these aspects, the employed model enhances reliability compared to relying solely on a data-driven approach to uncover associations. Consequently, the causal model offers a more precise analysis. Additionally, the model effectively identifies relationships among variables, which is not a common focal point in traditional severity studies that only analyze the severity variable with others.

The estimated causal model shows compatibility with the data, resulting in favorable metrics and confirming the consistency between the formulated DAG and the pre-established theoretical model. The findings of the study underscore a significant relationship between alcohol consumption and the severity of motorcycle crashes, particularly in urban areas (with an effect size of 0.18). The results suggest that alcohol has a more pronounced impact on increasing crash severity in urban settings. Given the higher population densities and congested traffic conditions typical of urban areas, impaired judgment and reduced reaction times due to alcohol consumption may exacerbate the consequences of motorcycle crashes.

In contrast to traditional models, the estimated model not only highlights relationships with severity but also uncovers other important factors. Particularly, it reveals a significant link between weekends and increased alcohol consumption in both rural and urban areas. While weekends have a limited but noticeable impact on alcohol-related issues in rural areas, nighttime hours witness a surge in alcohol use and severity in both settings. This increase can be attributed to factors such as easier access to open bars and reduced traffic congestion, which may lead to higher instances of speeding.

Furthermore, the analysis unveiled notable disparities between rural and urban areas, as certain variables demonstrated significance exclusively within each setting. For instance, factors such as hour rush and vehicle age exhibited significance solely within urban areas, whereas road type emerged as a significant factor specifically in rural areas. These findings emphasize the importance of considering contextual factors when developing targeted interventions and policies to address road safety in different geographical areas.

The initial causal hypothesis proposed that alcohol use might be linked to an increase in severity. The causal model was constructed on a theoretical framework, and all potential confounders were controlled for. Nevertheless, the findings only demonstrate a rise in severity in urban areas.

It is crucial to recognize the limitations of this study, particularly in its representation of some of the variables driven by constraints posed by the available dataset. The variable representing speeding was coded in the database based on an on-site judgment made by the Police Officer and, therefore, was not used in this study. Risky behaviors were

represented by a surrogate variable named traffic violations, also reported by the officer attending the scene. This surrogate variable yielded a counterintuitive result mainly because the dataset was not able to adequately describe subjunctive norms variables. Furthermore, a potential bias coming from missing data was not tested. While this has to be acknowledged, this issue is likely to have a limited impact on the results due to the large sample size. Despite these limitations, employing causal approaches can lead to more reliable results compared to traditional ones.

For future studies, it is advisable to investigate data-related challenges by employing spatial models and methodologies to address missing data. This would enhance the overall robustness and accuracy of the analysis. Furthermore, the causal approach presented in this study should be extended to other scenarios, encompassing other vulnerable road users such as cyclists and pedestrians, to validate its applicability and utility in diverse contexts.

REFERENCES

2-BE-SAFE. **Rider / Driver behaviours and road safety for PTW**. Brussels: European Commission, 2010.

ABDUL MANAN, M. M. *et al.* Road characteristics and environment factors associated with motorcycle fatal crashes in Malaysia. **IATSS Research**, [*s. l.*], v. 42, n. 4, p. 207–220, 2018.

ABRARI VAJARI, M. *et al.* A multinomial logit model of motorcycle crash severity at Australian intersections. **Journal of Safety Research**, [*s. l.*], v. 73, p. 17–24, 2020.

ACEM. **In-depth investigations of accidents involving powered two wheelers**. Brussels: Association des Constructeurs Européens de Motocycles, 2009. Disponível em: https://www.maids-study.eu/pdf/MAIDS2.pdf. Acesso em: 9 dez. 2023.

AL-MAHAMEED, F. J. *et al.* Analyzing Pedestrian and Bicyclist Crashes at the Corridor Level: Structural Equation Modeling Approach. **Transportation Research Record: Journal of the Transportation Research Board**, [*s. l.*], v. 2673, n. 7, p. 308–318, 2019. Disponível em: http://journals.sagepub.com/doi/10.1177/0361198119845353. Acesso em: 8 dez. 2023.

ALNAWMASI, N.; MANNERING, F. A statistical assessment of temporal instability in the factors determining motorcyclist injury severities. **Analytic Methods in Accident Research**, [*s. l.*], v. 22, 2019. Disponível em: https://www.sciencedirect.com/science/article/pii/S2213665719300168. Acesso em: 8 dez. 2023.

AZIMI, G. *et al.* Severity analysis for large truck rollover crashes using a random parameter ordered logit model. **Accident Analysis and Prevention**, [*s. l.*], v. 135, 2020. Disponível em: https://www.sciencedirect.com/science/article/pii/S0001457519305731. Acesso em: 8 dez. 2023.

BAMBACH, M. R.; MITCHELL, R. J. Safe system approach to reducing serious injury risk in motorcyclist collisions with fixed hazards. **Accident Analysis and Prevention**, [*s. l.*], v. 74, p. 290–296, 2015.

BELLET, T.; BANET, A. Towards a conceptual model of motorcyclists' Risk Awareness: A comparative study of riding experience effect on hazard detection and situational criticality assessment. **Accident Analysis and Prevention**, [*s. l.*], v. 49, p. 154–164, 2012.

BJØRNSKAU, T.; NÆVESTAD, T.-O.; AKHTAR, J. Traffic safety among motorcyclists in Norway: A study of subgroups and risk factors. **Accident Analysis & Prevention**, [*s. l.*], v. 49, p. 50–57, 2012.

BLACKMAN, R. A.; HAWORTH, N. L. Comparison of moped, scooter and motorcycle crash risk and crash severity. **Accident Analysis & Prevention**, [*s. l.*], v. 57, p. 1–9, 2013.

BOLLEN, K. A. **Structural Equations with Latent Variables**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 1989. *E-book*. Disponível em: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118619179. Acesso em: 8 dez. 2023.

BOLLEN, K. A.; BAULDRY, S. Three Cs in Measurement Models: Causal Indicators, Composite Indicators, and Covariates. **Psychological Methods**, [*s. l.*], v. 16, n. 3, p. 265–284, 2011.

BROWN, T. A. **Confirmatory factor analysis for applied research**. 2. ed. New York: Guilford: The Guilford Press, 2015.

CAMERON, M. H.; ELVIK, R. Nilsson's Power Model connecting speed and road trauma: Applicability by road type and alternative models for urban roads. **Accident Analysis & Prevention**, [*s. l.*], v. 42, n. 6, p. 1908–1915, 2010.

CHANG, F. *et al.* Injury severity of motorcycle riders involved in traffic crashes in Hunan, China: A mixed ordered logit approach. **International Journal of Environmental Research and Public Health**, [*s. l.*], v. 13, n. 7, p. 1–15, 2016.

CHEN, C. F.; CHEN, C. W. Speeding for fun? Exploring the speeding behavior of riders of heavy motorcycles using the theory of planned behavior and psychological flow theory. **Accident Analysis and Prevention**, [*s. l.*], v. 43, n. 3, p. 983–990, 2011. Disponível em: http://dx.doi.org/10.1016/j.aap.2010.11.025. Acesso em: 8 dez. 2023.

CHESHAM, D. J.; RUTTER, D. R.; QUINE, L. Motorcycling safety research: A review of the social and behavioural literature. **Social Science & Medicine**, [*s. l.*], v. 37, n. 3, p. 419–429, 1993.

CHOU, C.-C. *et al.* Effectiveness evaluation on cross-sector collaborative education programs for traffic safety toward sustainable motorcycle culture in Vietnam. **IATSS Research**, [*s. l.*], v. 46, n. 2, p. 258–268, 2022. Disponível em: https://doi.org/10.1016/j.iatssr.2022.01.001. Acesso em: 8 dez. 2023.

CHUNG, Y.; SONG, T. J.; YOON, B. J. Injury severity in delivery-motorcycle to vehicle crashes in the Seoul metropolitan area. **Accident Analysis and Prevention**, [*s. l.*], v. 62, p. 79–86, 2014. Disponível em: http://dx.doi.org/10.1016/j.aap.2013.08.024. Acesso em: 8 dez. 2023.

CREASER, J. I. *et al.* Effects of alcohol impairment on motorcycle riding skills. **Accident Analysis & Prevention**, [*s. l.*], v. 41, n. 5, p. 906–913, 2009.

CUMMINGS, P. Changes in traffic crash mortality rates attributed to use of alcohol, or lack of a seat belt, air bag, motorcycle helmet, or bicycle helmet, United States, 1982-2001. **Injury Prevention**, [*s. l.*], v. 12, n. 3, p. 148–154, 2006.

CUNTO, F. J. C.; FERREIRA, S. An analysis of the injury severity of motorcycle crashes in Brazil using mixed ordered response models. **Journal of Transportation Safety and Security**, [*s. l.*], v. 9, n. May, p. 33–46, 2017.

CZECH, S. *et al.* Comparing the cost of alcohol-related traffic crashes in rural and urban environments. **Accident Analysis & Prevention**, [*s. l.*], v. 42, n. 4, p. 1195–1198, 2010.

DATASUS. **Sistema de Informações sobre Mortalidade**. [*S. l.*], 2021. Disponível em: http://tabnet.datasus.gov.br. Acesso em: 8 dez. 2023.

DAVIS, G. A. Mechanisms, mediators, and surrogate estimation of crash modification factors. **Accident Analysis and Prevention**, [*s. l.*], v. 151, 2021. Disponível em: https://www.sciencedirect.com/science/article/pii/S0001457521000099. Acesso em: 8 dez. 2023.

DE ROME, L. *et al.* Motorcycle protective clothing: Protection from injury or just the weather?. **Accident Analysis and Prevention**, [*s. l.*], v. 43, n. 6, p. 1893–1900, 2011.

DEPARTMENT OF TRANSPORT AND MAIN ROADS. **Safer Roads, Safer Queensland**. Queensland: [*s. n.*], 2015. Disponível em: https://www.tmr.qld.gov.au/-/media/Safety/roadsafety/Strategy-and-action-plans/roadsafetystrategy201521.pdf?la=en. Acesso em: 8 dez. 2023.

DRUMMER, O. H. *et al.* The involvement of drugs in drivers of motor vehicles killed in Australian road traffic crashes. **Accident Analysis and Prevention**, [*s. l.*], v. 36, n. 2, p. 239–248, 2004.

DUFOURNET, M. *et al.* Causal inference to detect selection bias in road safety epidemiology. **arXiv**, [*s. l.*], 2016. Disponível em: http://arxiv.org/abs/1607.03775. Acesso em: 8 dez. 2023.

ELVIK, R.; VAA, T. **The handbook of road safety measures**. 2. ed. Bingley, UK: Emerald Publishing Limited, 2009.

ELWERT, F.; WINSHIP, C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. **Annual Review of Sociology**, [*s. l.*], v. 40, n. 1, p. 31–53, 2014.

ETIKA, A. **Developing safe system road safety indicators for the UK**. [*S. l.: s. n.*], 2018. Disponível em: https://www.pacts.org.uk/wp-content/uploads/PactsReport_-Developing-Safe-System-Road-Safety-Indicators-for-the-UK_Oct18-FINAL.pdf. Acesso em: 8 dez. 2023.

EUSTACE, D.; INDUPURU, V. K.; HOVEY, P. Identification of risk factors associated with motorcycle-related fatalities in ohio. **Journal of Transportation Engineering**, [*s. l.*], v. 137, n. 7, p. 474–480, 2011.

FERNÁNDEZ, O. *et al.* **Road Safety Annual Report 2020**. [*S. l.: s. n.*], 2020. Disponível em: https://www.itf-oecd.org/sites/default/files/docs/irtad-road-safety-annual-report-2020_0.pdf. Acesso em: 8 dez. 2023.

FLASK, T.; SCHNEIDER, W. H.; LORD, D. A segment level analysis of multi-vehicle motorcycle crashes in Ohio using Bayesian multi-level mixed effects models. **Safety Science**, [*s. l.*], v. 66, p. 47–53, 2014.

GEEDIPALLY, S. R.; TURNER, P. A.; PATIL, S. Analysis of motorcycle crashes in Texas with multinomial logit model. **Transportation Research Record**, [*s. l.*], n. 2265, p. 62–69, 2011a.

GEEDIPALLY, S. R.; TURNER, P. A.; PATIL, S. Analysis of Motorcycle Crashes in Texas with Multinomial Logit Model. **Transportation Research Record: Journal of the Transportation Research Board**, [*s. l.*], v. 2265, n. 1, p. 62–69, 2011b.

GOH, W. C.; LEONG, L. V.; CHEAH, R. J. X. Assessing significant factors affecting risky riding behaviors of motorcyclists. **Applied Sciences (Switzerland)**, [*s. l.*], v. 10, n. 18, 2020. Disponível em: https://doi.org/10.3390/app10186608. Acesso em: 8 dez. 2023.

GRACE, J. B.; BOLLEN, K. A. Representing general theoretical concepts in structural equation models: The role of composite variables. **Environmental and Ecological Statistics**, [*s. l.*], v. 15, n. 2, p. 191–213, 2008.

GUNASEKARA, F. I.; CARTER, K.; BLAKELY, T. Glossary for econometrics and epidemiology. **Journal of Epidemiology and Community Health**, [*s. l.*], v. 62, n. 10, p. 858–861, 2008.

HAIR, J. F. J. *et al.* **Análise Multivariada de Dados**. 6. ed. Porto Alegre: Bookman Companhia Editora Ltda, 2009.

HASANZADEH, S.; ASGHARIJAFARABADI, M.; SADEGHI-BAZARGANI, H. A hybrid of structural equation modeling and artificial neural networks to predict motorcyclists' injuries: A conceptual model in a case-control study. **Iranian Journal of Public Health**, [*s. l.*], v. 49, n. 11, p. 2194–2204, 2020.

HAUER, E. Cause, effect and regression in road safety: A case study. **Accident Analysis and Prevention**, [*s. l.*], v. 42, n. 4, p. 1128–1135, 2010.

HÉRAN, F. Les deux-roues motorisés en milieu urbain solution ou problème ?. **Transports urbains**, [*s. l.*], v. N° 131, n. 2, p. 14–19, 2017.

HOLUBOWYCZ, O. T.; KLOEDEN, C. N.; MCLEAN, A. J. Age, sex, and blood alcohol concentration of killed and injured drivers, riders, and passengers. **Accident Analysis & Prevention**, [*s. l.*], v. 26, n. 4, p. 483–492, 1994.

HOYLE, R. H. **Handbook of Structural Equation Modeling**. [*S. l.: s. n.*], 2012.

IHIE. **Guidelines for Motorcycling - Improving safety through engineering and integration**. London: Institute of Highway Engineers, 2010.

IJAZ, M. *et al.* Investigation of factors influencing motorcyclist injury severity using random parameters logit model with heterogeneity in means and variances. **International Journal of Crashworthiness**, [*s. l.*], v. 0, n. 0, p. 1–11, 2021.

ISLAM, M. An analysis of motorcyclists' injury severities in work-zone crashes with unobserved heterogeneity. **IATSS Research**, [*s. l.*], v. 46, n. 2, p. 281–289, 2022. Disponível em: https://doi.org/10.1016/j.iatssr.2022.01.003. Acesso em: 8 dez. 2023.

ITF. **Zero Road Deaths and Serious Injuries**. Paris: OECD, 2016. Disponível em: https://www.oecd.org/publications/zero-road-deaths-and-serious-injuries-9789282108055-en.htm. Acesso em: 8 dez. 2023.

JONES, S.; GURUPACKIAM, S.; WALSH, J. Factors influencing the severity of crashes caused by motorcyclists: Analysis of data from Alabama. **Journal of Transportation Engineering**, [*s. l.*], v. 139, n. 9, p. 949–956, 2013.

KASHANI, A. T. *et al.* Factors affecting the accident size of motorcycle-involved crashes: a structural equation modeling approach. **International Journal of Injury Control and Safety Promotion**, [*s. l.*], v. 0, n. 0, p. 1–6, 2020.

KLINE, R. B. **Principles and Practice of Structural Equation Modeling**. [*S. l.: s. n.*], 2015.

LANE, P. C. R.; CLARKE, D.; HENDER, P. On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. **Decision Support Systems**, [*s. l.*], v. 53, n. 4, p. 712–718, 2012.

LARDELLI-CLARET, P. *et al.* Driver dependent factors and the risk of causing a collision for two wheeled motor vehicles. **Injury Prevention**, [*s. l.*], v. 11, n. 4, p. 225–231, 2005.

LAUBACH, Z. M. *et al.* A biologist's guide to model selection and causal inference. **Proceedings of the Royal Society B: Biological Sciences**, [*s. l.*], v. 288, n. 1943, 2021. Disponível em: https://doi.org/10.1098/rspb.2020.2815. Acesso em: 8 dez. 2023.

LEE, J. *et al.* Analysis of crash proportion by vehicle type at traffic analysis zone level: A mixed fractional split multinomial logit modeling approach with spatial effects. **Accident Analysis and Prevention**, [*s. l.*], v. 111, n. September 2017, p. 12–22, 2018.

LEE, J. *et al.* How motorcycle helmets affect trauma mortality: Clinical and policy implications. **Traffic Injury Prevention**, [*s. l.*], v. 18, n. 6, p. 666–671, 2017. Disponível em: https://www.tandfonline.com/doi/full/10.1080/15389588.2016.1204650. Acesso em: 9 dez. 2023.

LEE, J. Y.; CHUNG, J. H.; SON, B. Analysis of traffic accident size for Korean highway using structural equation models. **Accident Analysis and Prevention**, [*s. l.*], v. 40, n. 6, p. 1955–1963, 2008.

LI, J. *et al.* A Motorcyclist-Injury Severity Analysis: A Comparison of Single-, Two-, and Multi-Vehicle Crashes Using Latent Class Ordered Probit Model. **Accident Analysis and Prevention**, [*s. l.*], v. 151, n. December 2020, 2021. Disponível em: https://doi.org/10.1016/j.aap.2020.105953. Acesso em: 8 dez. 2023.

LI, H. **Impacts of Traffic Interventions on Road Safety**. 2014. 1–228 f. Thesis (PhD degree) - Imperial College London, London, 2014. Disponível em: https://spiral.imperial.ac.uk/bitstream/10044/1/18068/1/Li-H-2013-PhD-Thesis.pdf. Acesso em: 8 dez. 2023.

LIN, M.-R. *et al.* Factors associated with severity of motorcycle injuries among young adult riders. **Annals of Emergency Medicine**, [*s. l.*], v. 41, n. 6, p. 783–791, 2003.

LIN, M.-R.; HWANG, H.-F.; KUO, N.-W. Crash Severity, Injury Patterns, and Helmet Use in Adolescent Motorcycle Riders. **The Journal of Trauma: Injury, Infection, and Critical Care**, [*s. l.*], v. 50, n. 1, p. 24–30, 2001.

LORD, D.; QIN, X.; GEEDIPALLY, S. R. **Highway Safety Analytics and Modeling**. [*S. l.*]: Elsevier Science, 2021.

LOWENSTEIN, S. R.; KOZIOL-MCLAIN, J. Drugs and Traffic Crash Responsibility: A Study of Injured Motorists in Colorado. **The Journal of Trauma: Injury, Infection, and Critical Care**, [*s. l.*], v. 50, n. 2, p. 313–320, 2001.

LÜBKE, K. *et al.* Why We Should Teach Causal Inference: Examples in Linear Regression With Simulated Data. **Journal of Statistics Education**, [*s. l.*], v. 28, n. 2, p. 133–139, 2020.

LUNA, G. K. *et al.* The Influence of Ethanol Intoxication on Outcome of Injured Motorcyclists. **The Journal of Trauma: Injury, Infection, and Critical Care**, [*s. l.*], v. 24, n. 8, p. 695–700, 1984.

MACDONALD, P. L.; GARDNER, R. C. Type I Error Rate Comparisons of Post Hoc Procedures for I j Chi-Square Tables. **Educational and Psychological Measurement**, [*s. l.*], v. 60, n. 5, p. 735–754, 2000.

MAGAZZÙ, D.; COMELLI, M.; MARINONI, A. Are car drivers holding a motorcycle licence less responsible for motorcycle—Car crash occurrence?. **Accident Analysis & Prevention**, [*s. l.*], v. 38, n. 2, p. 365–370, 2006.

MÖLLER, H. *et al.* Crash risk factors for novice motorcycle riders. **Journal of Safety Research**, [*s. l.*], v. 73, p. 93–101, 2020.

MORRISON, C. N. *et al.* On-road bicycle lane types, roadway characteristics, and risks for bicycle crashes. **Accident Analysis and Prevention**, [*s. l.*], v. 123, n. August 2018, p. 123–131, 2019.

MORRISON, T. G.; MORRISON, M. A.; MCCUTCHEON, J. M. Best Practice Recommendations for Using Structural Equation Modelling in Psychological Research. **Psychology**, [*s. l.*], v. 08, n. 09, p. 1326–1341, 2017.

MULLIN, B. Increasing age and experience: are both protective against motorcycle injury? A case-control study. **Injury Prevention**, [*s. l.*], v. 6, n. 1, p. 32–35, 2000.

NADIMI, N. *et al.* Analyzing traffic violations among motorcyclists using structural equation modeling. **International Journal of Injury Control and Safety Promotion**, [*s. l.*], v. 28, n. 4, p. 454–467, 2021. Disponível em: https://doi.org/10.1080/17457300.2021.1942922. Acesso em: 8 dez. 2023.

NEWSOM, J. T. **Longitudinal Structural Equation Modeling**. [*S. l.: s. n.*], 2015.

NGUYEN-PHUOC, D. Q. *et al.* Exploring the prevalence and factors associated with self-reported traffic crashes among app-based motorcycle taxis in Vietnam. **Transport Policy**, [*s. l.*], v. 81, n. June, p. 68–74, 2019.

OECD. **Improving Safety for Motorcycle, Scooter and Moped Riders**. [*S. l.: s. n.*], 2015. *E-book*. Disponível em: https://www.oecd.org/publications/improving-safety-for-motorcycle-scooter-and-moped-riders-9789282107942-en.htm. Acesso em: 8 dez. 2023.

OPAS. **Salvar VIDAS - Pacote de medidas técnicas para a segurança no trânsito**. Brasília, DF: [*s. n.*], 2018. Disponível em: https://iris.paho.org/handle/10665.2/34980. Acesso em: 8 dez. 2023.

PAI, C.-W. Motorcyclist injury severity in angle crashes at T-junctions: Identifying significant factors and analysing what made motorists fail to yield to motorcycles. **Safety Science**, [*s. l.*], v. 47, n. 8, p. 1097–1106, 2009.

PEARL, J. **Causality**. 2. ed. University of California, Los Angeles: Cambridge, 2009.

PEARL, J. **The causal foundations of structural equation modeling**. New York: Guilford Press, 2021. Disponível em: https://ftp.cs.ucla.edu/pub/stat_ser/r370.pdf. Acesso em: 9 dez. 2023.

PEARL, J. The seven tools of causal inference, with reflections on machine learning. **Communications of the ACM**, [*s. l.*], v. 62, n. 3, p. 54–60, 2019.

PEARL, J.; GLYMOUR, M.; JEWELL, N. P. **Causal Inference in Statistics: A Primer**. [*S. l.*]: Wiley, 2016.

PEARL, J.; MACKENZIE, D. **The Book of Why**. [*S. l.: s. n.*], 2018. v. 1

PEEK-ASA, C.; KRAUS, J. F. Alcohol Use, Driver, and Crash Characteristics among Injured Motorcycle Drivers. **The Journal of Trauma: Injury, Infection, and Critical Care**, [*s. l.*], v. 41, n. 6, p. 989–993, 1996.

PERVEZ, A.; LEE, J.; HUANG, H. Identifying Factors Contributing to the Motorcycle Crash Severity in Pakistan. **Journal of Advanced Transportation**, [*s. l.*], v. 2021, p. 1–10, 2021. Disponível em: https://doi.org/10.1155/2021/6636130. Acesso em: 8 dez. 2023.

PRIYANTHA WEDAGAMA, D. M.; WISHART, D. Analysing local motorcyclists' perception towards road safety. **MATEC Web of Conferences**, [*s. l.*], v. 276, 2019. Disponível em: https://www.matec-conferences.org/10.1051/matecconf/201927603002. Acesso em: 8 dez. 2023.

RAHMAN, M. H. *et al.* Identification of factors influencing severity of motorcycle crashes in Dhaka, Bangladesh using binary logistic regression model. **International Journal of Injury Control and Safety Promotion**, [*s. l.*], v. 28, n. 2, p. 141–152, 2021. Disponível em: https://doi.org/10.1080/17457300.2021.1878230. Acesso em: 8 dez. 2023.

REASON, J. **Managing the risks of organizational accidents**. [*S. l.: s. n.*], 1997.

RECHNITZER, G.; HAWORTH, N.; KOWADLO, N. The Effect of Vehicle Roadworthiness on crash incidence and severity. **Monash University Accident research centre**, [*s. l.*], v. 164, n. 164, p. 74, 2000.

RIFAAT, S. M.; TAY, R.; DE BARROS, A. Severity of motorcycle crashes in Calgary. **Accident Analysis and Prevention**, [*s. l.*], v. 49, p. 44–49, 2012.

RIZZI, M.; STRANDROTH, J.; TINGVALL, C. The effectiveness of antilock brake systems on motorcycles in reducing real-life crashes and injuries. **Traffic injury prevention**, [*s. l.*], v. 10, n. 5, p. 479–487, 2009.

ROBINS M. JAMES, M. A. H. Causal Inference - what if. **Foundations of Agnostic Statistics**, [*s. l.*], p. 235–281, 2020.

ROHRER, J. M. Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. **Advances in Methods and Practices in Psychological Science**, [*s. l.*], v. 1, n. 1, p. 27–42, 2018.

ROSSHEIM, M. E. *et al.* Associations Between Drug Use and Motorcycle Helmet Use in Fatal Crashes. **Traffic Injury Prevention**, [*s. l.*], v. 15, n. 7, p. 678–684, 2014.

RUBIN, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. **Journal of Educational Psychology**, [*s. l.*], v. 66, n. 5, p. 688–701, 1974.

SALUM, J. H. *et al.* Severity of motorcycle crashes in Dar es Salaam, Tanzania. **Traffic Injury Prevention**, [*s. l.*], v. 20, n. 2, p. 189–195, 2019.

SASIDHARAN, L.; DONNELL, E. T. Propensity scores-potential outcomes framework to incorporate severity probabilities in the Highway Safety Manual crash prediction algorithm. **Accident Analysis and Prevention**, [*s. l.*], v. 71, p. 183–193, 2014.

SAVOLAINEN, P. T. *et al.* The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. **Accident Analysis & Prevention**, [*s. l.*], v. 43, n. 5, p. 1666–1676, 2011.

SAVOLAINEN, P.; MANNERING, F. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. **Accident Analysis & Prevention**, [*s. l.*], v. 39, n. 5, p. 955–963, 2007.

SCHUMACKER, R. E.; LOMAX, R. G. **A beginner's guide to structural equation modeling**. [*S. l.: s. n.*], 2010.

SE, C. *et al.* Empirical comparison of the effects of urban and rural crashes on motorcyclist injury severities: A correlated random parameters ordered probit approach with heterogeneity in means. **Accident Analysis and Prevention**, [*s. l.*], v. 161, n. July, p. 106352, 2021. Disponível em: https://doi.org/10.1016/j.aap.2021.106352. Acesso em: 8 dez. 2023.

SHARPE, D. Chi-Square Test is Statistically Significant: Now What?. **Practical Assessment, Research, and Evaluation**, [*s. l.*], 2015. Disponível em: https://scholarworks.umass.edu/pare/vol20/iss1/8/. Acesso em: 8 dez. 2023.

SHIH, H.-C. *et al.* Alcohol intoxication increases morbidity in drivers involved in motor vehicle accidents. **The American Journal of Emergency Medicine**, [*s. l.*], v. 21, n. 2, p. 91–94, 2003.

SHIPLEY, B. **Cause and Correlation in Biology**. [*S. l.: s. n.*], 2000.

SIQUEIRA, M. F. **Metodologia de análise dos determinantes da demanda por transportes no paradigma da inferência causal**. 2020. Dissertação (Mestre em Engenharia de Transportes) - Universidade Federal do Ceará, [*s. l.*], 2020.

SMC. **Motorcycle Vision version 2.0**. [*S. l.: s. n.*], 2014. Disponível em: https://www.svmc.se/smc_filer/SMC%20centralt/Rapporter/2014/The%20Motorcycle%20Vision%202.0%20English%20print.pdf. Acesso em: 8 dez. 2023.

SODERSTROM, C. A. *et al.* Alcohol use, driving records, and crash culpability among injured motorcycle drivers. **Accident Analysis & Prevention**, [*s. l.*], v. 25, n. 6, p. 711–716, 1993.

SONG, Y.; KOU, S.; WANG, C. Modeling crash severity by considering risk indicators of driver and roadway: A Bayesian network approach. **Journal of Safety Research**, [*s. l.*], v. 76, p. 64–72, 2021.

STIGSON, H. **A safe road transport system – factors influencing injury outcome for car occupants**. Departmented. Stockholm: Karolinska Institutet, 2009.

STIGSON, H.; KRAFFT, M.; TINGVALL, C. Use of Fatal Real-Life Crashes to Analyze a Safe Road Transport System Model, Including the Road User, the Vehicle, and the Road. **Traffic Injury Prevention**, [*s. l.*], v. 9, n. 5, p. 463–471, 2008.

TEOH, E. R. Effectiveness of Antilock Braking Systems in Reducing Motorcycle Fatal Crash Rates. **Traffic Injury Prevention**, [*s. l.*], v. 12, n. 2, p. 169–173, 2011.

TEOH, E. R.; CAMPBELL, M. Role of motorcycle type in fatal motorcycle crashes. **Journal of Safety Research**, [*s. l.*], v. 41, n. 6, p. 507–512, 2010.

THEOFILATOS, A.; YANNIS, G. A review of powered-two-wheeler behaviour and safety. **International Journal of Injury Control and Safety Promotion**, [*s. l.*], v. 22, n. 4, p. 284–307, 2015.

TOPOLŠEK, D.; DRAGAN, D. Relationships between the motorcyclists' behavioural perception and their actual behaviour. **Transport**, [*s. l.*], v. 33, n. 1, p. 151–164, 2018.

TOPUZ, K.; DELEN, D. A probabilistic Bayesian inference model to investigate injury severity in automobile crashes. **Decision Support Systems**, [*s. l.*], v. 150, 2021. Disponível em: https://doi.org/10.1016/j.dss.2021.113557. Acesso em: 8 dez. 2023.

TORRES, C. A.; XAVIER, V. J. M.; CUNTO, F. J. C. Analyzing the Relationship between Road Safety Pillars and the World Health Organization Member States' Mortality Rate using Structural Equation Modeling Approach. **Transportation Research Record**, [*s. l.*], v. 2674, n. 4, p. 1–10, 2020.

TRINH, T. A.; LINH LE, T. P. The Association between Risk-taking Behavior and Helmet Use among Motorcyclist. **IOP Conference Series: Earth and Environmental Science**, [*s. l.*], v. 143, n. 1, 2018.

TSUI, K. *et al.* Association between Drink Driving and Severity of Crash Injuries to Road Users. **Hong Kong Journal of Emergency Medicine**, [*s. l.*], v. 17, n. 1, p. 34–39, 2010.

UNITED NATIONS. **Improving global road safety**. New York: [*s. n.*], 2020.

WALI, B.; KHATTAK, A. J.; AHMAD, N. Examining correlations between motorcyclist's conspicuity, apparel related factors and injury severity score: Evidence from new motorcycle crash causation study. **Accident Analysis and Prevention**, [*s. l.*], v. 131, n. April, p. 45–62, 2019.

WALTON, D.; BUCHANAN, J. Motorcycle and scooter speeds approaching urban intersections. **Accident Analysis & Prevention**, [*s. l.*], v. 48, p. 335–340, 2012.

WASEEM, M.; AHMED, A.; SAEED, T. U. Factors affecting motorcyclists' injury severities: An empirical assessment using random parameters logit model with heterogeneity in means and variances. **Accident Analysis & Prevention**, [*s. l.*], v. 123, p. 12–19, 2019.

WEDAGAMA, D. M. P. Local Motorcyclists' Intentions towards Traffic Violations and Speeding. **Journal of the Eastern Asia Society for Transportation Studies**, [*s. l.*], v. 12, p. 1871–1883, 2015.

WEGMAN, F.; AARTS, L.; BAX, C. Advancing sustainable safety. **Safety Science**, [*s. l.*], v. 46, n. 2, p. 323–343, 2008. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S092575350700094X. Acesso em: 8 dez. 2023.

WELLE, B. *et al.* **Sustentável e Seguro**. Washington: [*s. n.*], 2018. Disponível em: https://www.wribrasil.org.br/sites/default/files/Sustentavel_Seguro.pdf. Acesso em: 8 dez. 2023.

WESTREICH, D.; GREENLAND, S. The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. **American Journal of Epidemiology**, [*s. l.*], v. 177, n. 4, p. 292–298, 2013.

WHO. **The Global status report on road safety 2018**. [*S. l.: s. n.*], 2018. Disponível em: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/. Acesso em: 8 dez. 2023.

WOOD, J. S. **Causal Inference in Traffic Safety Research**. 2016. 211 f. Dissertation (Degree of Doctor of Philosophy) - The Pennsylvania State University, State College, 2016. Disponível em: https://etda.libraries.psu.edu/catalog/28774. Acesso em: 8 dez. 2023.

WOOLDRIDGE, J. M. **Introductory econometrics**. 5. ed. Mason : South-Western Cengage Learning, 2013.

WUNDERSITZ, L.; RAFTERY, S. Understanding the context of alcohol impaired driving for fatal crash–involved drivers: A descriptive case analysis. **Traffic Injury Prevention**, [*s. l.*], v. 18, n. 8, p. 781–787, 2017.

XIN, C. *et al.* Modeling Safety Effects of Horizontal Curve Design on Injury Severity of Single-Motorcycle Crashes with Mixed-Effects Logistic Model. **Transportation Research Record: Journal of the Transportation Research Board**, [*s. l.*], v. 2637, n. 1, p. 38–46, 2017.

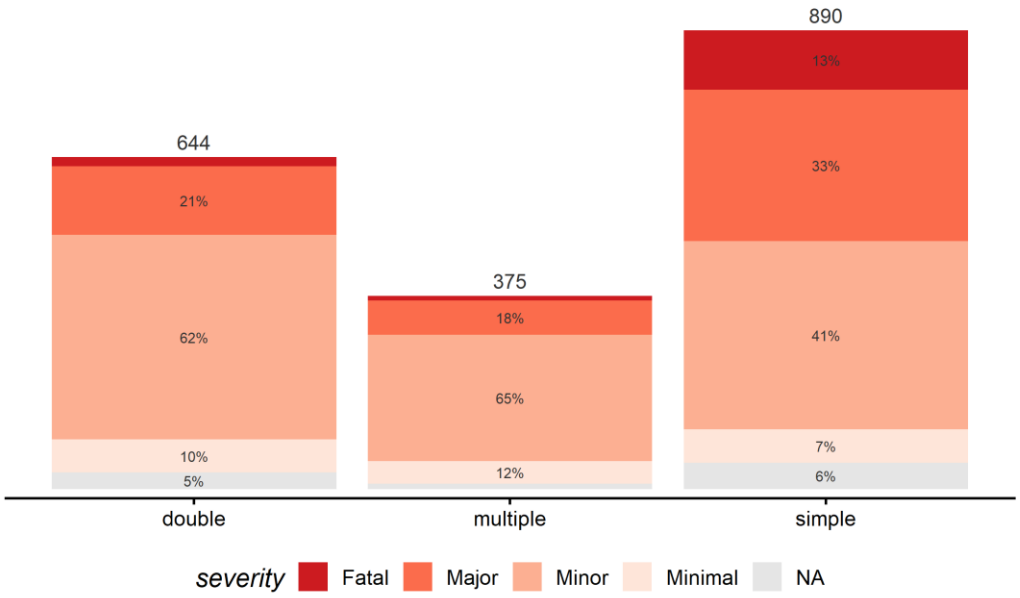
YANG, S.; WANG, L.; DING, P. Causal inference with confounders missing not at random. **Biometrika**, [*s. l.*], v. 106, n. 4, p. 875–888, 2019.

ZAFRI, N. M. *et al.* Comparative risk assessment of pedestrian groups and their road-crossing behaviours at intersections in Dhaka, Bangladesh. **International Journal of Crashworthiness**, [*s. l.*], v. 27, n. 2, p. 581–590, 2022.

ZIAKOPOULOS, A.; NIKOLAOU, D.; YANNIS, G. Correlations of multiple rider behaviors with self-reported attitudes, perspectives on traffic rule strictness and social desirability. **Transportation Research Part F: Traffic Psychology and Behaviour**, [*s. l.*], v. 80, p. 313–327, 2021.

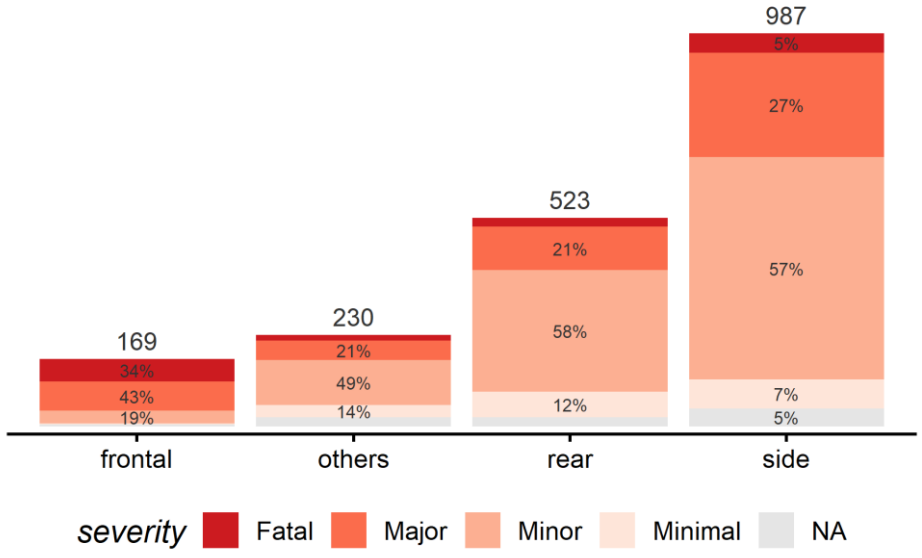
APPENDIX A. EXPLORATORY ANALYSES

RD_LANES



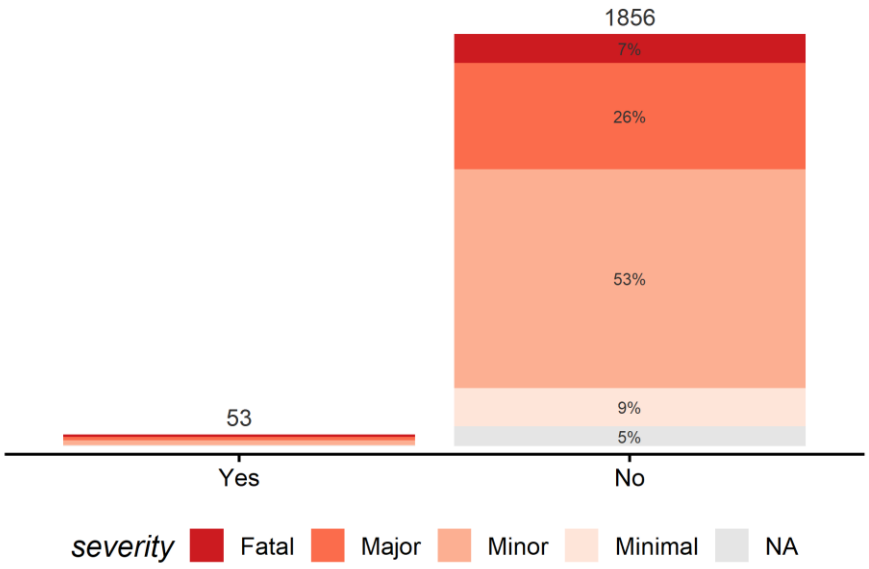
Pearson's Chi-squared test: X = 151.5, df = 6, p-value = 0
signif. resid: double-Minor: 5.8;multiple-Minor: 5;simple-Minor: -9.5;
double-Major: -3.7;multiple-Major: -4.2;simple-Major: 7;
double-Fatal: -5.5;multiple-Fatal: -4.3;simple-Fatal: 8.7
Cramér's V: 0.2, moderate association

RD_TYPE



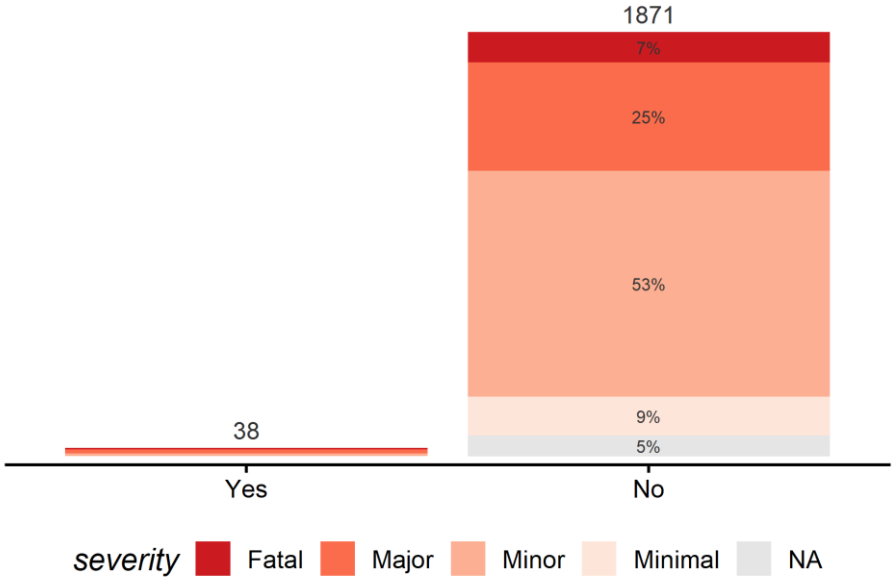
Pearson's Chi-squared test: X = 262.4, df = 9, p-value = 0
signif. resid: frontal-Minimal: -3.5;others-Minimal: 3.1;rear-Minimal: 3.1;
frontal-Minor: -9.8;side-Minor: 3.4;frontal-Major: 5.2;
rear-Major: -3.1;frontal-Fatal: 13.4;rear-Fatal: -3.6;
side-Fatal: -4.3
Cramér's V: 0.22, moderate association

SPEEDING



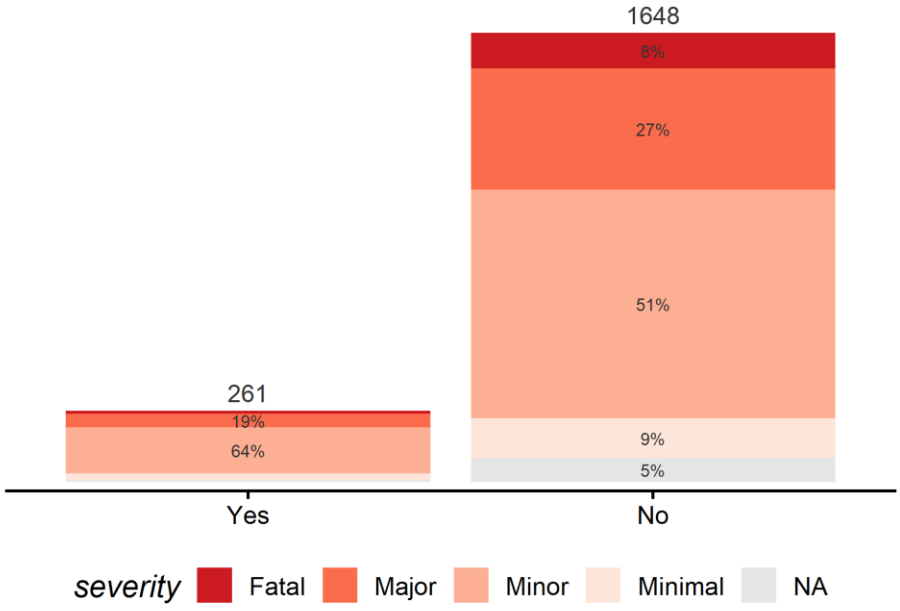
Pearson's Chi-squared test: X = 17.9, df = 3, p-value = 0
signif. resid: Yes-Fatal: 3.8; No-Fatal: -3.8
Fisher's Exact Test for Count Data: p-value = 0
Cramér's V: 0.1, low association

OVERTAK



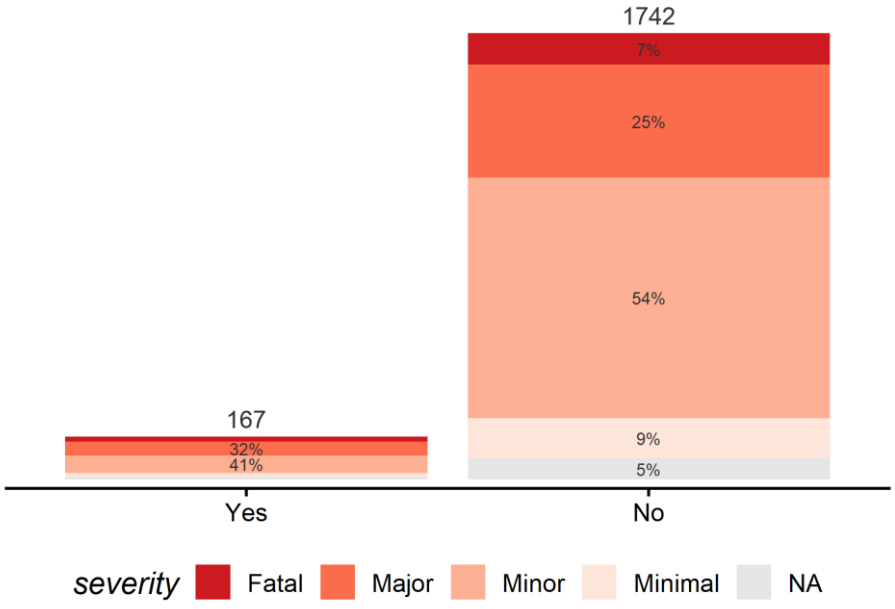
Pearson's Chi-squared test: X = 16.3, df = 3, p-value = 0
signif. resid: Yes-Minor: -2.8; No-Minor: 2.8
Fisher's Exact Test for Count Data: p-value = 0
Cramér's V: 0.09, low association

N_SAFE_DIST



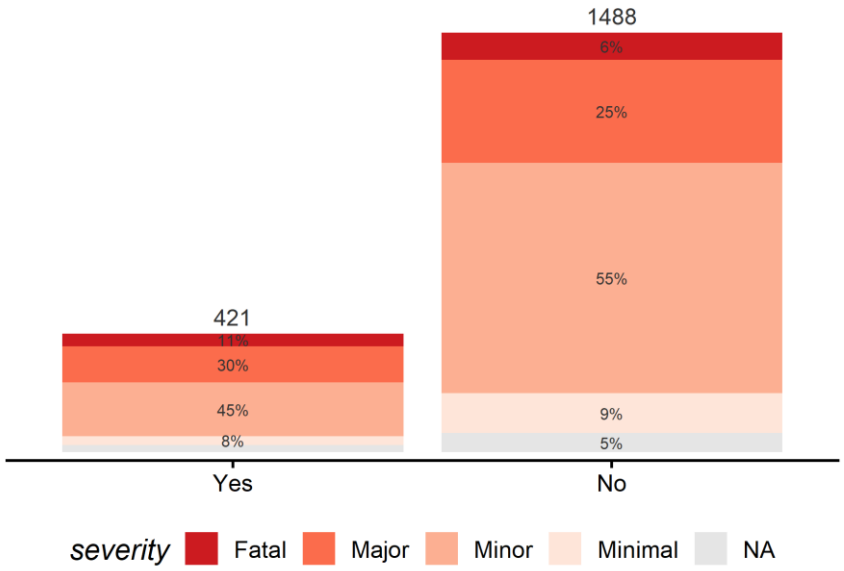
Pearson's Chi-squared test: X = 16.9, df = 3, p-value = 0
signif. resid: Yes-Minor: 3.6; No-Minor: -3.6; Yes-Major: -2.9; No-Major: 2.9
Cramér's V: 0.1, low association

ALCOHOL



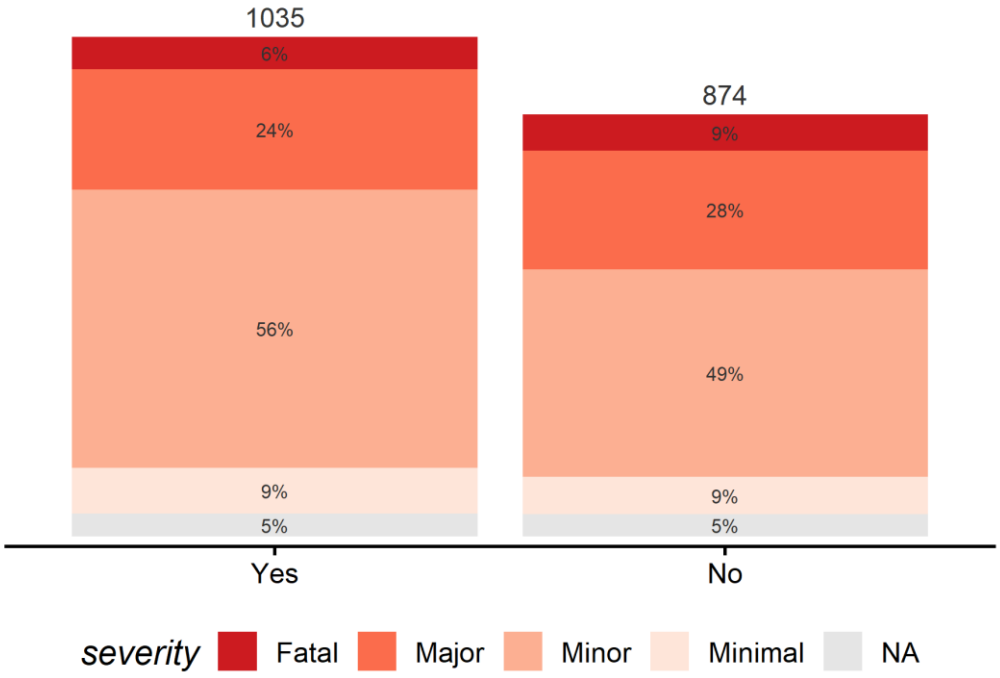
Pearson's Chi-squared test: X = 11.9, df = 3, p-value = 0.01
signif. resid: Yes-Minor: -3.2; No-Minor: 3.2
Cramér's V: 0.08, low association

TRAF_RU_DIS



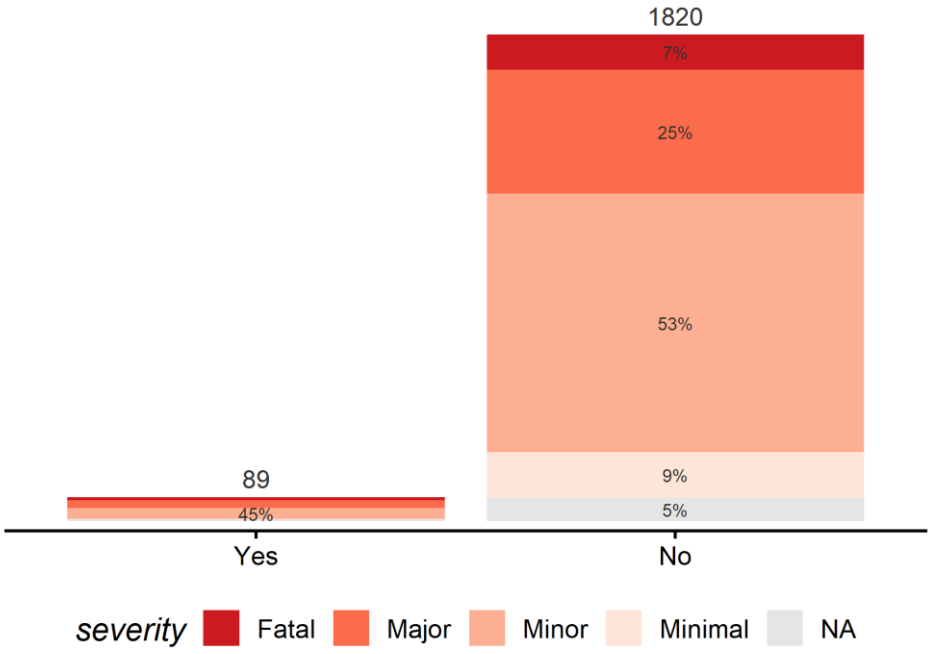
Pearson's Chi-squared test: X = 20.4, df = 3, p-value = 0
signif. resid: Yes-Minor: -3.4; No-Minor: 3.4; Yes-Fatal: 3.2; No-Fatal: -3.2
Cramér's V: 0.11, low association

LACK_ATT



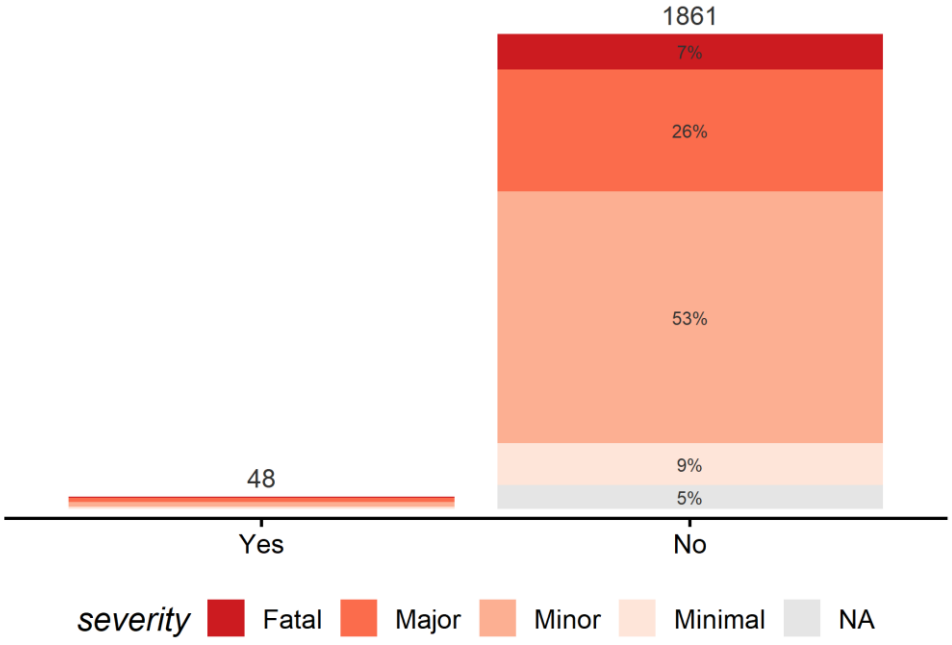
Pearson's Chi-squared test: X = 9.6, df = 3, p-value = 0.02
signif. resid: Yes-Minor: 2.8; No-Minor: -2.8
Cramér's V: 0.07, low association

RD_PROB



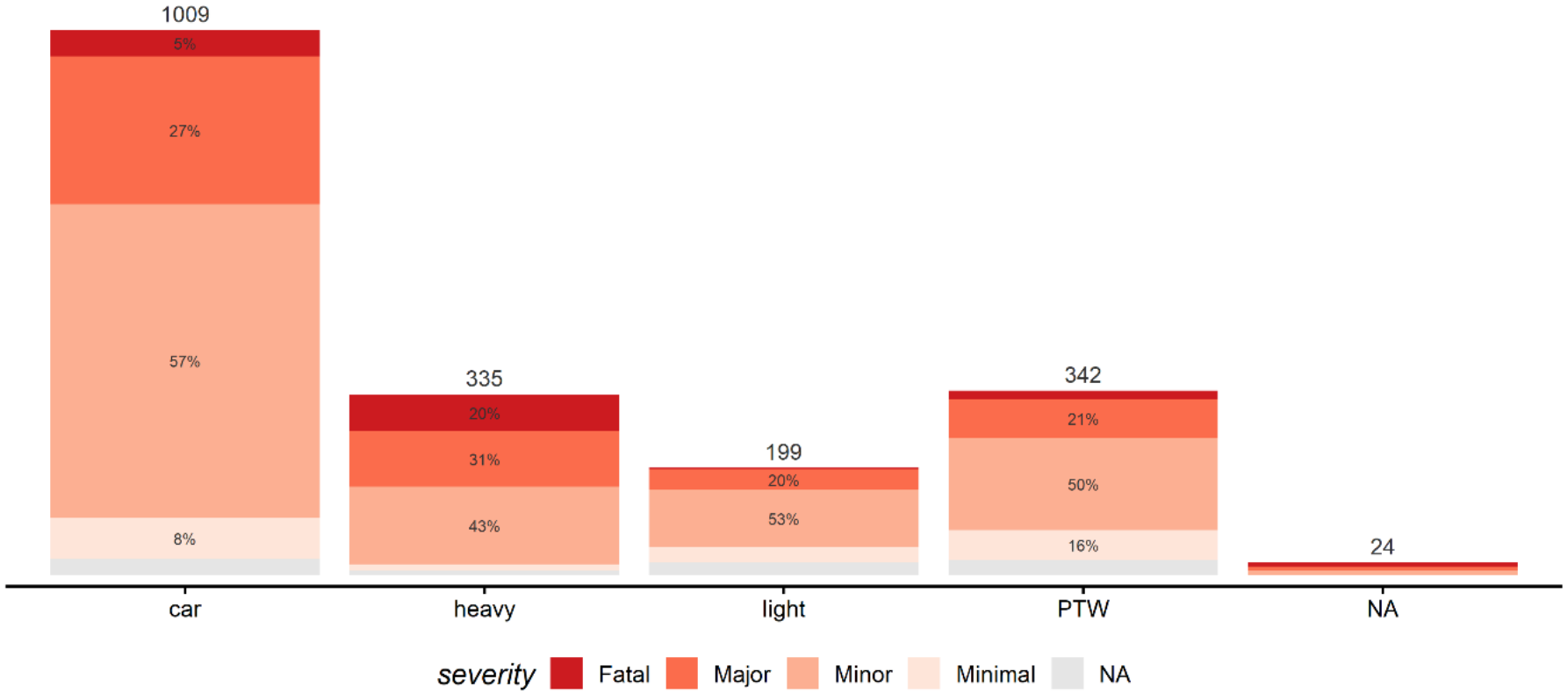
Pearson's Chi-squared test: X = 11.3, df = 3, p-value = 0.01
Cramér's V: 0.08, low association

VCLE_PROB



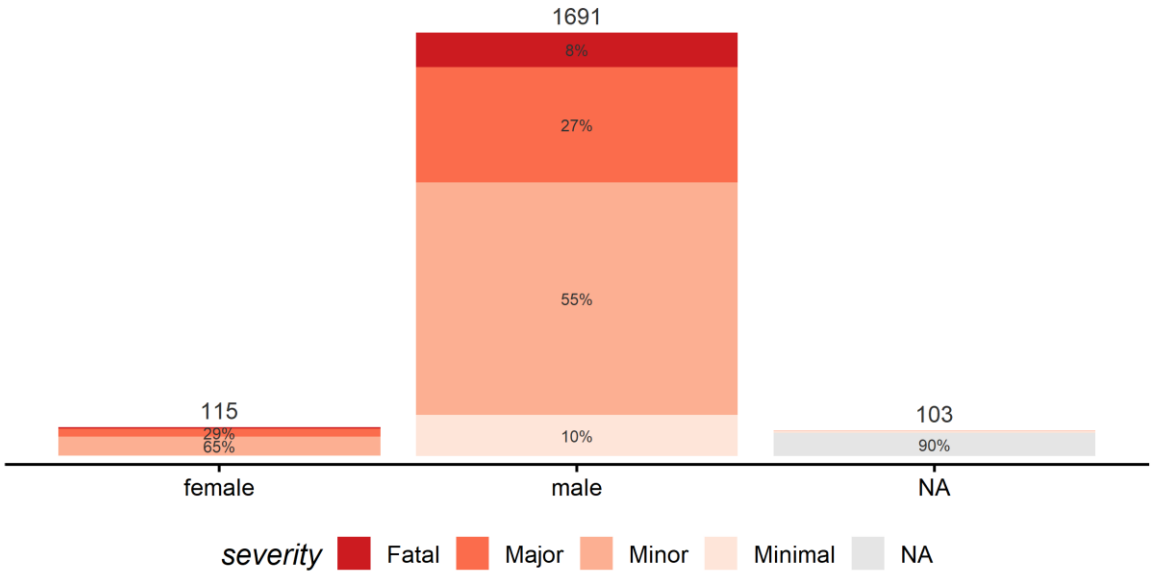
Pearson's Chi-squared test: X = 7.3, df = 3, p-value = 0.06
Fisher's Exact Test for Count Data: p-value = 0.06
Cramér's V: 0.06, low association

VCLE_COLL



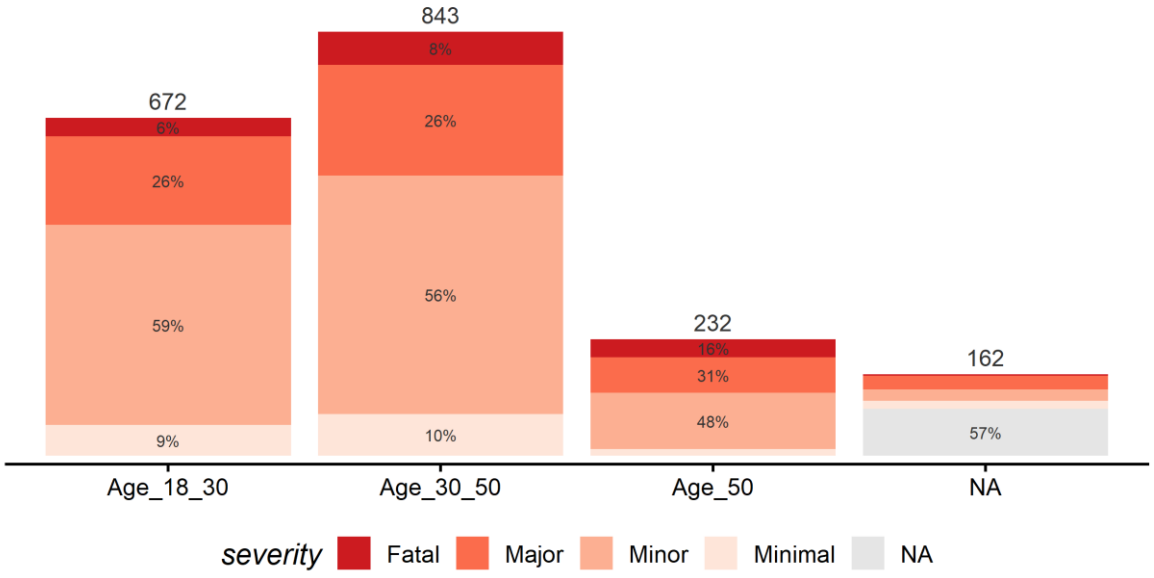
Pearson's Chi-squared test: X = 160.6, df = 9, p-value = 0
signif. resid: heavy-Minimal: -4.2; light-Minimal: 3.3; PTW-Minimal: 5.3; car-Minor: 3.3; heavy-Minor: -4.7; car-Fatal: -4.5; heavy-Fatal: 10.1; light-Fatal: -3.4
Cramér's V: 0.17, moderate association

US_GEN



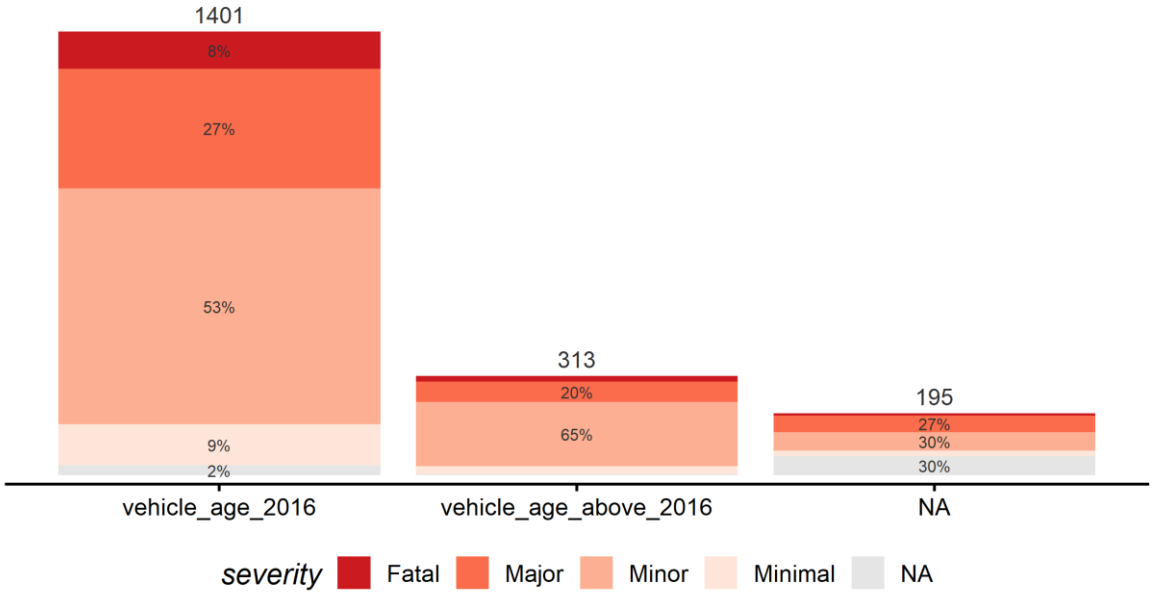
Pearson's Chi-squared test: X = 11.5, df = 3, p-value = 0.01
signif. resid: female-Minimal: -2.9; male-Minimal: 2.9
Cramér's V: 0.08, low association

US_AGE



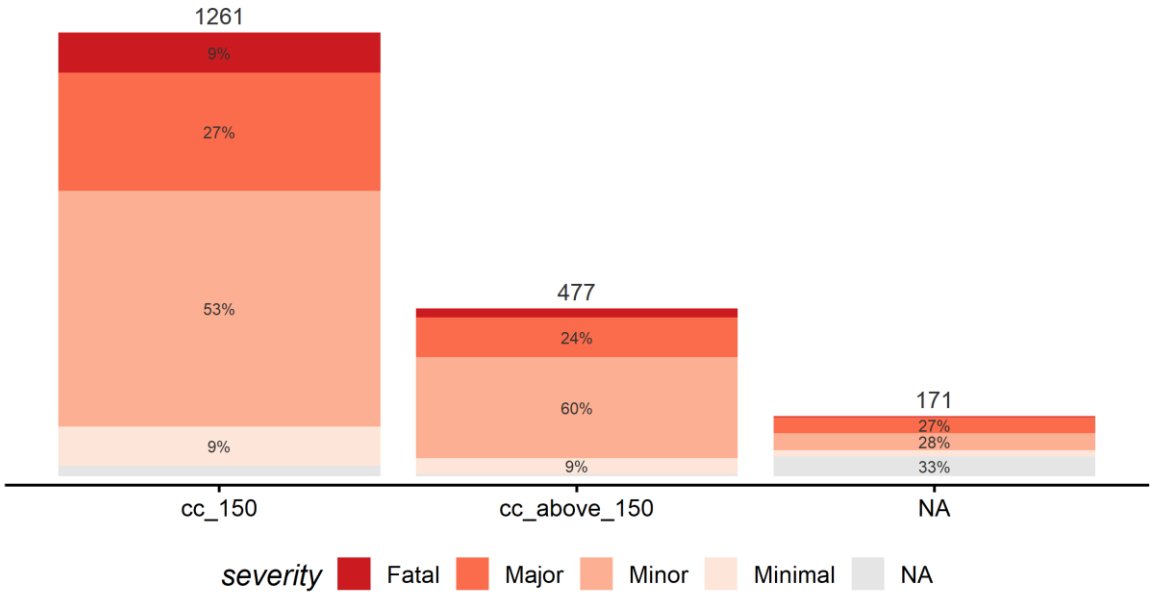
Pearson's Chi-squared test: X = 30.6, df = 6, p-value = 0
signif. resid: Age_18_30-Fatal: -3; Age_50-Fatal: 4.6
Cramér's V: 0.09, low association

VCLE_AGE



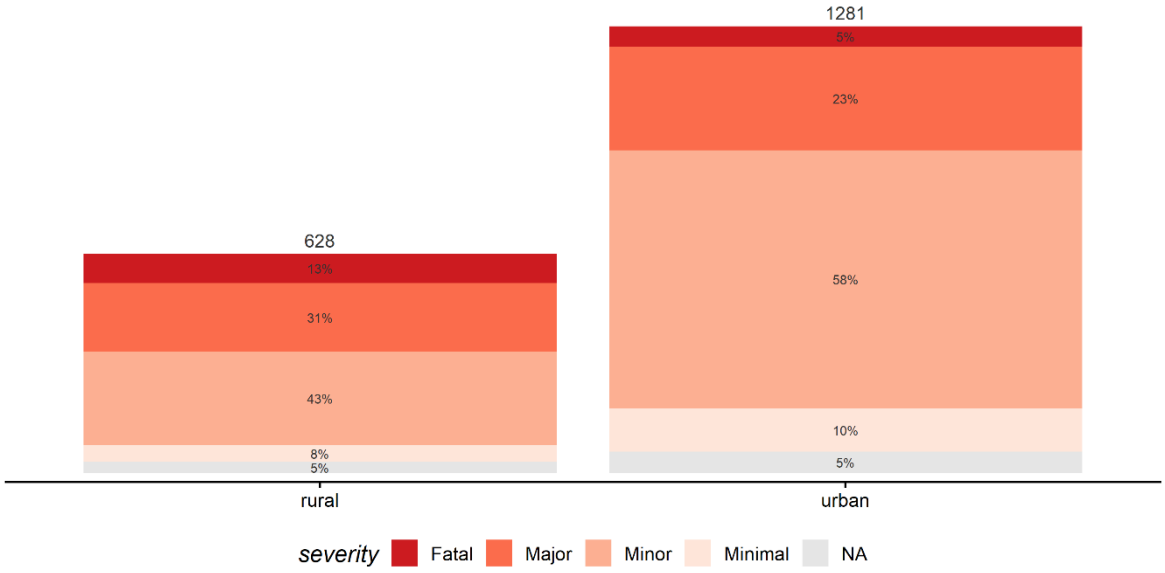
Pearson's Chi-squared test: X = 10.6, df = 3, p-value = 0.01
Cramér's V: 0.08, low association

ENG_SIZE



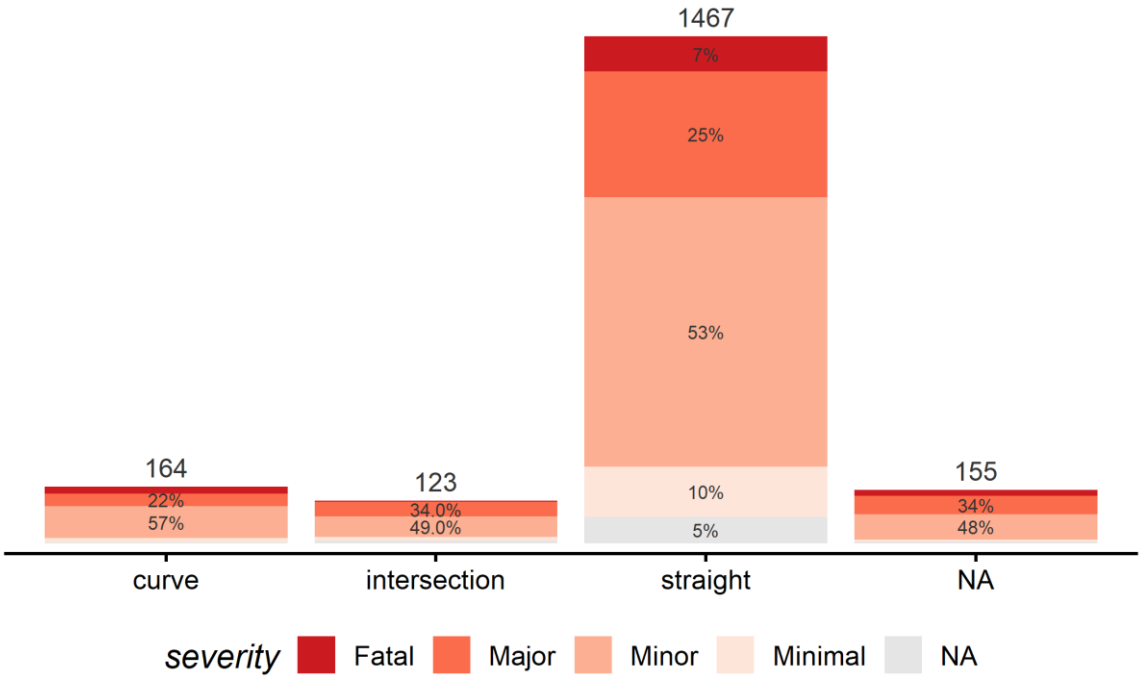
Pearson's Chi-squared test: X = 10.6, df = 3, p-value = 0.01
Cramér's V: 0.08, low association

LAND_USE



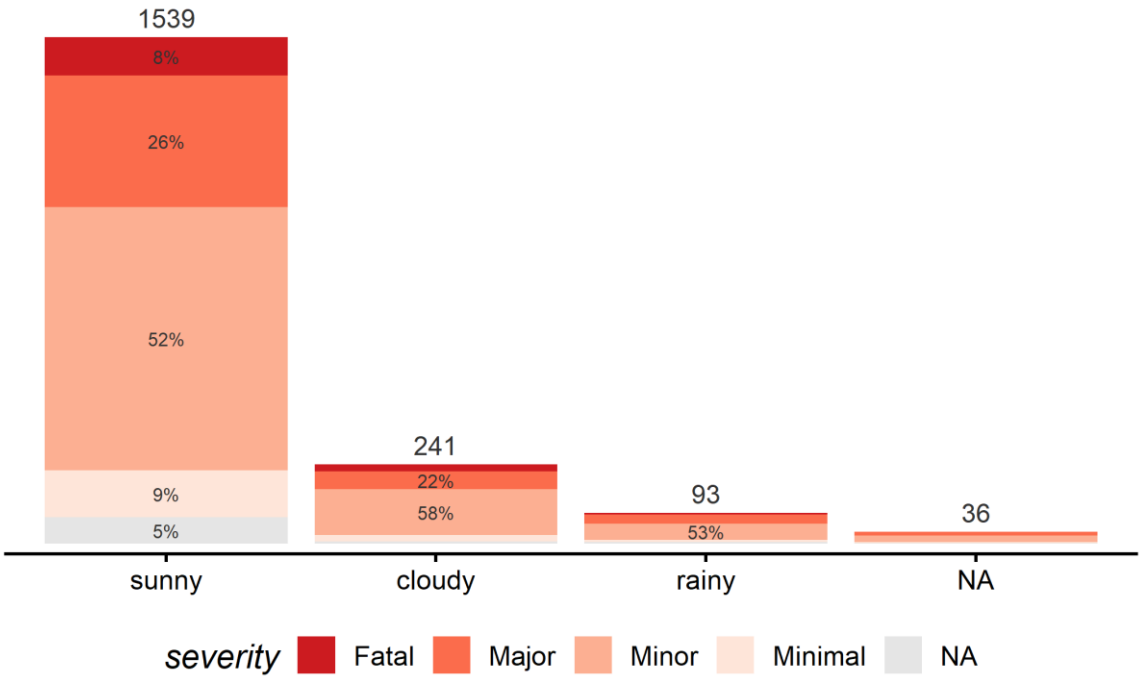
Pearson's Chi-squared test: X = 73.1, df = 3, p-value = 0
signif. resid: rural-Minor: -6.3; urban-Minor: 6.3; rural-Major: 3.9; urban-Major: -3.9; rural-Fatal: 6.8; urban-Fatal: -6.8
Cramér's V: 0.2, moderate association

RD_TYPE



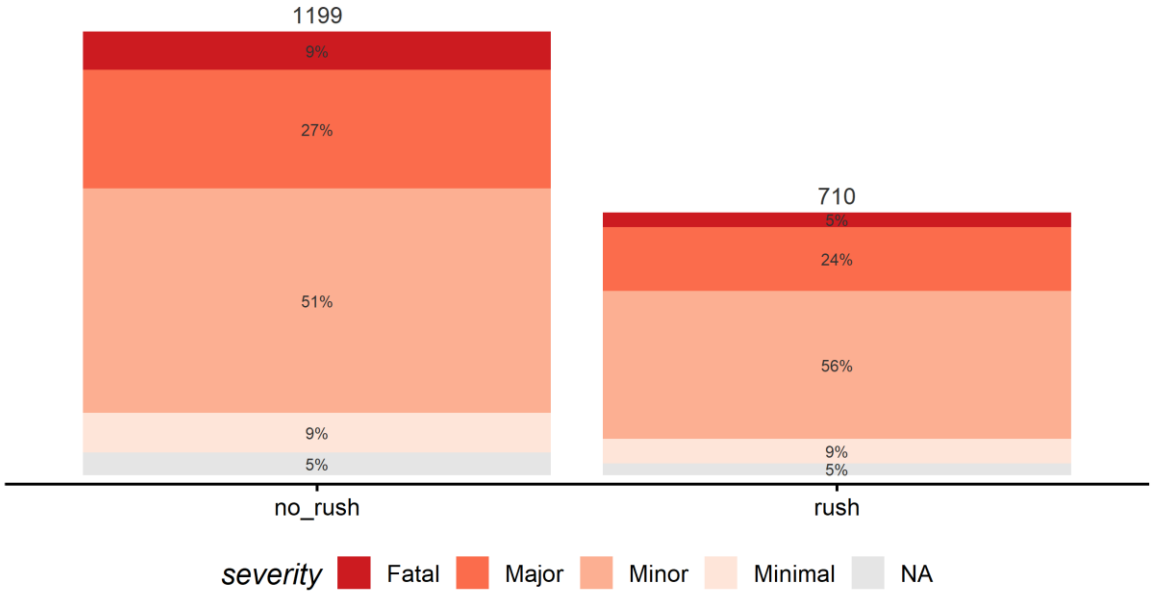
Pearson's Chi-squared test: X = 16.4, df = 6, p-value = 0.01
Cramér's V: 0.07, low association

WEATHER



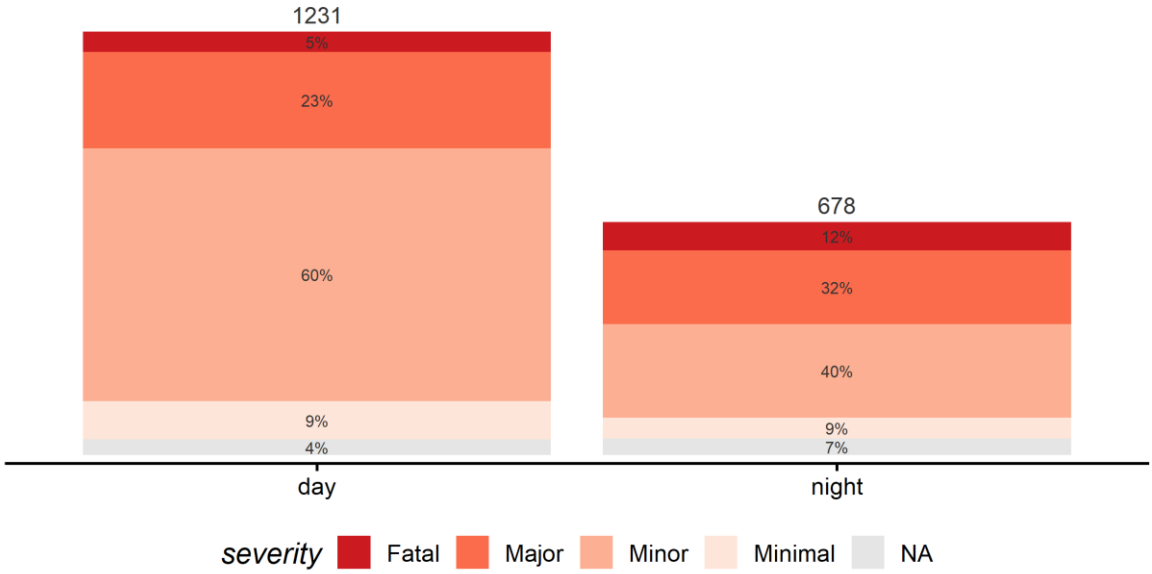
Pearson's Chi-squared test: X = 5.4, df = 6, p-value = 0.49
Cramér's V: 0.04, little association

HR_RUSH



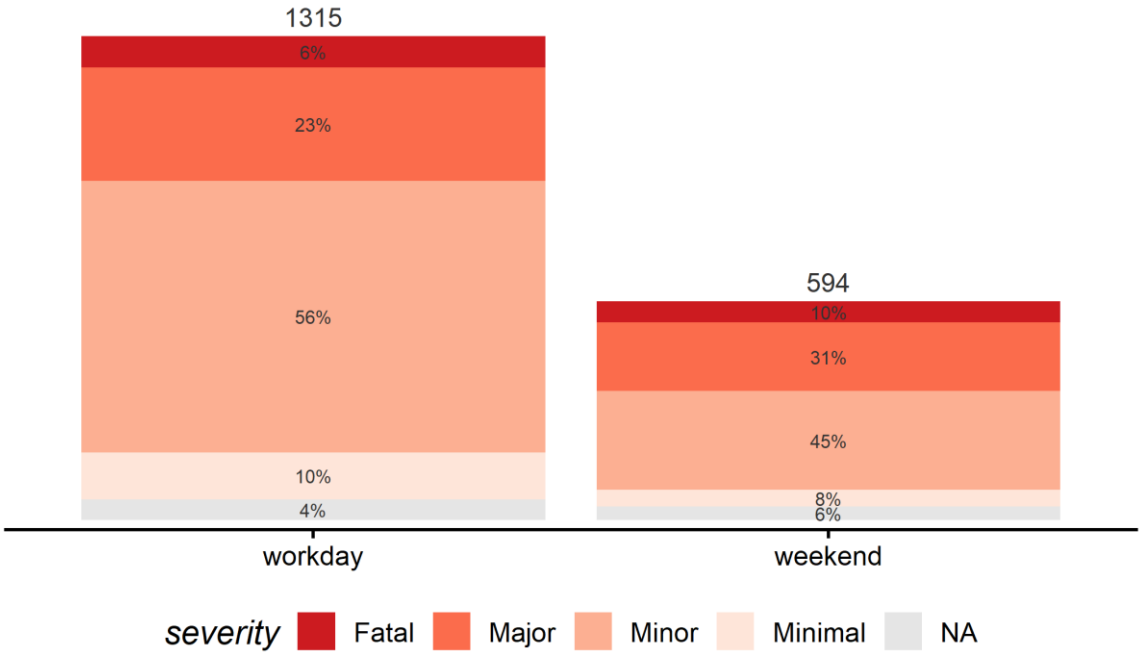
Pearson's Chi-squared test: X = 9.5, df = 3, p-value = 0.02
Cramér's V: 0.07, low association

HR_NIGHT



Pearson's Chi-squared test: X = 77.9, df = 3, p-value = 0
signif. resid: day-Minor: 7.7; night-Minor: -7.7; day-Major: -4.7; night-Major: 4.7; day-Fatal: -6.2; night-Fatal: 6.2
Cramér's V: 0.21, moderate association

WEEKDAY



Pearson's Chi-squared test: X = 26.5, df = 3, p-value = 0
signif. resid: workday-Minor: 4.1; weekend-Minor: -4.1; workday-Major: -3.9; weekend-Major: 3.9
Cramér's V: 0.12, low association