

Traffic Crash Patterns and Causations based on Sequence of Events:
Preparing for a Transition into Automated Transportation

By

Yu Song

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Civil and Environmental Engineering)

at the

UNIVERSITY OF WISCONSIN-MADISON

2021

Date of final oral examination: 8/24/2021

The dissertation is approved by the following members of the Final Oral Committee:

David A. Noyce, Professor, Civil and Environmental Engineering

Soyoung Ahn, Professor, Civil and Environmental Engineering

Bin Ran, Professor, Manufacturing Systems Engineering

Madhav V. Chitturi, Assistant Research Scientist, Civil and Environmental Engineering

Samuel Younkin, Assistant Research Scientist, Global Health Institute

Table of Contents

<i>Abstract</i>	<i>iv</i>
<i>Acknowledgements</i>	<i>viii</i>
<i>List of Figures</i>	<i>ix</i>
<i>List of Tables</i>	<i>xi</i>
Chapter 1 Introduction	1
1.1 Background and Motivation.....	1
1.2 Conceptual Framework and Research Focus.....	5
1.3 Objective, Research Questions, and Study Design	6
1.4 Dissertation Outline.....	8
Chapter 2 Background	11
2.1 Future of Transportation and Safe Automation	11
2.1.1 Future of Transportation	11
2.1.2 Safe Automation.....	17
2.2 Crash Sequences and Scenarios	21
2.2.1 Crash Sequences	21
2.2.2 Scenarios for Automated Vehicle Testing.....	30
Chapter 3 A Methodology for Traffic Crash Sequence Analysis: Impact of Event Encoding and Dissimilarity Measures	39
3.1 Introduction.....	39
3.2 Literature Review	41
3.3 Crash Sequence Analysis Methodology	43
3.4 Data Processing	45
3.5 Crash Sequence Encoding.....	48
3.6 Crash Sequence Dissimilarity Measures	50
3.6.1 Distances between element distributions	52
3.6.2 Measures based on counts of common attributes.....	53
3.6.3 Edit distances	54
3.6.4 Summary of dissimilarity measures.....	59
3.7 Case Study: Single-Vehicle Crashes on Interstate Highways	64
3.7.1 Encoding schemes	66
3.7.2 Comparison of dissimilarity measures	68
3.7.3 Sequence clustering results	77

3.8 Conclusions	84
<i>Chapter 4 Automated Vehicle Crash Sequences: Patterns and Potential Uses in Safety Testing.....</i>	87
4.1 Introduction.....	87
4.2 Literature Review	89
4.2.1 Patterns in AV Crashes	89
4.2.2 Patterns in AV Disengagements	91
4.2.3 Relationship Among AV Crashes, Disengagements, and Contributing Factors	92
4.2.4 Safety Performance of AVs Compared with Conventional Vehicles	94
4.2.5 Sequence of Events in Traffic Crashes.....	95
4.3 Data.....	96
4.3.1 AV Crashes.....	96
4.3.2 AV Disengagements in Crashes	100
4.3.3 AV Crash Sequences	101
4.4 Methodology	103
4.4.1 Transition Matrix.....	104
4.4.2 Measuring Sequence Dissimilarity and Optimal Matching.....	105
4.4.3 Cluster Analysis.....	108
4.5 Results	111
4.5.1 Most Frequent Subsequences	111
4.5.2 Transitions to and from Disengagement.....	112
4.5.3 Sequence Characterization.....	114
4.5.4 Cross-tabulation between Sequence Group and Other Variables	118
4.6 Discussion.....	122
4.7 Conclusions	126
<i>Chapter 5 Intersection Two-Vehicle Crash Scenario Specification for Automated Vehicle Safety Evaluation Using Sequence Analysis and Bayesian Networks.....</i>	129
5.1 Introduction.....	129
5.2 Literature Review	130
5.2.1 Scenario Generation Using Historical Crash Data	130
5.2.2 Crash Sequence Analysis	131
5.2.3 Bayesian Networks for Crash Analysis	132
5.3 Data.....	132
5.3.1 Subsetting	133
5.3.2 Numbering Crash Participants	134
5.3.3 Encoding Sequences.....	136
5.3.4 Other Crash Attributes	138
5.4 Methodology	140
5.4.1 Crash Sequence Comparison and Clustering.....	141
5.4.2 Bayesian Network Modeling.....	143

5.5 Results	145
5.5.1 Sequence Types.....	145
5.5.2 Variable Dependencies.....	152
5.5.3 Scenario Specification.....	157
5.6 Discussion and Conclusions	161
<i>Chapter 6 Conclusions.....</i>	<i>163</i>
6.1 Summary of Dissertation.....	163
6.2 Contribution	165
6.3 Limitations	166
6.4 Future Directions	167
<i>References</i>	<i>170</i>
<i>Appendices</i>	<i>183</i>
Appendix for Chapter 3	183
Appendix for Chapter 5.....	188

Abstract

The development of automated vehicle (AV) technology is suggesting a promising future of safer and more efficient transportation. However, there are still many challenges in ensuring the operational safety of AVs before their deployment. Scenario-based testing of AVs is an essential part of the safety verification of this technology, and generating challenging scenarios is critical for scenario-based testing of AVs. Research presented in this dissertation focuses on developing a methodology for crash sequence analysis which is used to generate scenarios for AV safety testing. AVs with SAE Driving Automation Levels 3 and 4 are expected to share the roads and handle conflicts with human drivers. Building a scenario library based on comprehensive samples of historical crash data would be the most efficient way to set up the foundation of a scenario-based AV verification system. Crash scenarios are temporally ordered scenes that consist of 1) participants' actions and interactions, and 2) the relatively static surrounding environment. To incorporate both elements, this dissertation's scenario-generating procedure included two steps – 1) characterization of crashes based on sequences of events, and 2) specification of interrelationships between crash sequences and other crash attributes that depict the surrounding environment. Research tasks developed and demonstrated the crash-sequence-based scenario-generating procedure with three studies.

In the first study, a first-of-its-kind crash sequence analysis methodology was developed to serve as the foundation of this dissertation research. The methodology was designed to select the most appropriate sequence encoding schemes and dissimilarity

measures for crash sequence analysis, and be applicable to various use cases of crash analysis.

In the second study, crash sequence analysis methods were applied to California AV collision data to query and identify representative crash sequence types. Seven sequence types were found to be significantly associated with crash outcomes and environmental conditions. Based on the findings, the research proposed a scenario-based AV testing framework with crash sequences as the core component.

In the third study, the scenario-generating procedure incorporates a sequence analysis and a Bayesian network analysis. The procedure was demonstrated by specifying intersection two-vehicle crash scenarios. Fifty-five crash sequence types were identified. The interrelationships among sequence types, crash outcomes, and operational design domain (ODD) variables were depicted by a Bayesian network. Based on the network, scenarios could be specified as combinations of crash sequences and ODDs.

This dissertation contributes to the understanding of traffic crashes and efficient testing of AVs by developing a first-of-its-kind crash sequence analysis methodology and a novel test scenario generating procedure. This dissertation laid the foundation for traffic crash sequence analysis and the use of crash data for AV test scenario generation. As future crashes happen, new data can be added to the database to add greater depth and further understanding to the critically important topic of scenario-based AV safety evaluation. Findings from this dissertation will have further influences in improving transportation safety and supporting the transition into automated transportation. Knowledge of crash

sequences will help future research in analyzing crash causations. A comprehensive test scenario library will speed up large-scale AV safety testing with the help of simulation.

To my wife C. J. Zhang

Acknowledgements

I would like to extend my deepest gratitude to my advisor and dissertation committee chair, Dr. David Noyce, for his guidance and support through my Ph.D. training, as well as his profound belief in my abilities and work. I'm also extremely grateful to my dissertation committee members, Dr. Sue Ahn, Dr. Bin Ran, Dr. Madhav Chitturi, and Dr. Sam Younkin for all their help to me in finishing this dissertation. Especially, I would like to thank Madhav for sharing his knowledge and giving invaluable suggestions on my research. Thanks also go to the SAFER-SIM University Transportation Center for sponsoring part of my dissertation research. Throughout my four and a half years at UW-Madison, I have enjoyed working with TOPS Lab colleagues and have frequently received their help. Specially, I would like to thank Hiba, Beau, Lingqiao, Boris, Kelvin, Andi, Steve, Bill, Adam, and Jon. Finally, I would like to express my appreciation and gratitude to my wife, my parents, and my in-laws for their endless support.

List of Figures

Figure 1 Conceptual framework.....	6
Figure 2 Dissertation outline	8
Figure 3 Trend in AAA survey results (22)	13
Figure 4 Cartoon by Saint Louis Star (November 6, 1923, p. 14) (23).....	14
Figure 5 Trends in annual U.S. motor vehicle fatalities (1921–2017) (31).....	16
Figure 6 Generalized function/event sequence (87).....	26
Figure 7 Illustration of causal pattern “dart-out first half” (87).....	27
Figure 8 Conceptual model of the crash generation process (88).....	28
Figure 9 Illustration of a type of motor vehicle-bicycle crash, “bicycle rideout – intersection controlled by sign” (88).....	29
Figure 10 Levels of scenarios (92).....	32
Figure 11 Example of three levels of scenarios (92)	32
Figure 12 Layered model of variables describing an AV test scenario (93).....	33
Figure 13 Procedure of crash sequence analysis	44
Figure 14 Linkage structure of the CRSS data files (144).....	46
Figure 15 Tree-structured subsetting of crash data.....	48
Figure 16 Sequence lengths.....	65
Figure 17 Process of developing encoding schemes	67
Figure 18 Clustering quality of the OMlev measure	72
Figure 19 ARI sensitivity to the LOM parameter e	75
Figure 20 Alluvial diagram of clustering results with the OMlev measure	78
Figure 21 Form of transition matrix P (64)	105
Figure 22 Silhouette widths	110
Figure 23 Transition rates from preceding events to disengagement and from disengagement to succeeding events	113
Figure 24 Graph illustrations of AV crash sequence patterns.....	117
Figure 25 Crash severity distribution by sequence group.....	119
Figure 26 Manner of collision distribution by sequence group.....	120
Figure 27 Sequence group distribution by facility type.....	121

Figure 28 Sequence group distribution by time of day	121
Figure 29 Sequence group distribution by year	122
Figure 30 AV safety testing framework with sequence of events embedded.....	124
Figure 31 CRSS two-vehicle crash types (144)	135
Figure 32 Sequence structure.....	137
Figure 33 Clustering quality indices for CC D (rear end) sequences.....	146
Figure 34 Bayesian network generated from hill climbing learning.....	153
Figure 35 Alternative Bayesian network.....	155
Figure 36 Bayesian network of sequence types and crash outcomes.....	156
Figure 37 Bayesian network of sequence types, human factors, and environmental conditions	156
Figure 38 Distribution of sequence types resulting in fatalities	158

List of Tables

Table 1	SAE levels of driving automation (2)	2
Table 2	U.S. government automated vehicle technology principles (18)	12
Table 3	Example of an NHTSA ADS scenario descriptor (95)	34
Table 4	Sequence alignment costs	55
Table 5	Dissimilarity measures	61
Table 6	Sensitivity of dissimilarity measures to sequence attributes	62
Table 7	Conditions for obtaining data for case study from CRSS	64
Table 8	Example of encoding schemes	68
Table 9	Mantel test results	70
Table 10	Comparison of ARIs	76
Table 11	Sequence clustering results with the OMlev measure	80
Table 12	Detailed interpretation of sequence clustering results with OE	83
Table 13	Summary of data from California AV collision reports	97
Table 14	2015-2019 AV test mileages and crash rates by organization	99
Table 15	Causes of disengagements involved in AV crashes	100
Table 16	Event encodings	102
Table 17	Example of crash event sequences	103
Table 18	Example of ways to align two sequences	107
Table 19	Cluster size and cluster average silhouette width	111
Table 20	Top 15 most frequent subsequences	112
Table 21	Clusters of AV crash sequences	115
Table 22	Subsetting criteria	133
Table 23	Vehicle renumbering	136
Table 24	Sequence lengths	138
Table 25	Crash outcomes	139
Table 26	Human factors and environmental conditions	140
Table 27	Sequence alignment	142
Table 28	Distribution of CRSS intersection two-vehicle crash configurations	143
Table 29	Sequence clustering results	148

Table 30 Interpretation of representative sequences	150
Table 31 Arc strengths.....	154
Table 32 Distribution of intersection type and TCD in k3 crashes.....	159
Table 33 Distribution of speeding behavior and time of day in k3 crashes	159
Table 34 Distribution of sequence types at signal-controlled intersections	160

Chapter 1 Introduction

1.1 Background and Motivation

Transportation is moving toward a connected and automated future. Automated vehicle¹ (AV) technology is one of the many advanced transportation technologies acting as both a promoter and a disrupter in future transportation (1). The Society of Automotive Engineers (SAE) defines 6 levels of driving automation, as shown in Table 1 (2). From Driving Automation Level 0 to Level 5, the involvement of driving automation technology increases, and the involvement of human driver decreases. Levels 0 to 2 require a human driver to be actively engaged in driving tasks. Starting from Level 3, driving automation technology takes care of most driving tasks. Level 3 automation requires a human driver to be present and prepared to take over control of the vehicle at any time and especially when in an emergency, and vehicle automation can only operate under limited conditions (i.e., operational design domains, or ODDs). Level 4 automation does not need a human driver to take over control but only enables the vehicle to operate in limited ODDs. Vehicles with Level 5 automation do not need human drivers to take over control and can operate in unlimited ODDs. The AVs discussed in this dissertation are assumed to have or are aimed to have SAE Driving Automation Levels 3 and 4.

¹ “Automated vehicle”, “autonomous vehicle”, “self-driving vehicle”, and “driverless vehicle” are terms commonly seen used to describe a vehicle with a high level of driving automation. There is not a universally agreed term to describe this new technology. In the “AV 4.0” document, USDOT used the term “automated vehicle”, which is also used throughout this research.

Table 1 SAE levels of driving automation (2)

Level	Narrative Definition	DDT		DDT Fallback	ODD	
		Sustained Lateral and Longitudinal Vehicle Motion Control	OEDR			
Driver Performs Part or All of the DDT						
Driver Support	0 – No Driving Automation	The performance by the driver of the entire DDT, even when enhanced by active safety systems.	Driver	Driver	Driver	n/a
	1 – Driver Assistance	The sustained and ODD-specific execution by a driving automation system of either the lateral or the longitudinal vehicle motion control subtask of the DDT (but not both simultaneously) with the expectation that the driver performs the remainder of the DDT.	Driver and System	Driver	Driver	Limited
	2 – Partial Driving Automation	The sustained and ODD-specific execution by a driving automation system of both the lateral and longitudinal vehicle motion control subtasks of the DDT with the expectation that the driver completes the OEDR subtask and supervises the driving automation system.	System	Driver	Driver	Limited
ADS (“System”) Performs the Entire DDT (While Engaged)						
Automated Driving	3 – Conditional Driving Automation	The sustained and ODD-specific performance by an ADS of the entire DDT with the expectation that the DDT fallback-ready user is receptive to ADS-issued requests to intervene, as well as to DDT performance-relevant system failures in other vehicle systems, and will respond appropriately.	System	System	Fallback-ready user (becomes the driver during fallback)	Limited
	4 – High Driving Automation	The sustained and ODD-specific performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will need to intervene.	System	System	System	Limited
	5 – Full Driving Automation	The sustained and unconditional (i.e., not ODD-specific) performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will need to intervene.	System	System	System	Unlimited

Note: DDT: Dynamic Driving Task; ODD: Operational Design Domain; OEDR: Object and Event Detection and Response; ADS: Automated Driving System. The DDT does not include strategic aspects of the driving task, such as determining destination(s) and deciding when to travel.

Organizations developing AVs are promoting the idea of automated vehicles saving lives as they are expected to partially and eventually fully replace human drivers, who are blamed (i.e., driver error) for causing 94% of the motor vehicle crashes in the United States (3–5). At the same time, AV developing organizations are testing the performance of AVs intensively, on public roads, closed-course tracks, and in simulations (6). The goal is to ensure an efficient and safe deployment of this technology.

One of the challenges with simulation-based testing of AVs is developing scenarios. Scenarios need to reflect real-world challenges existing in a variety of ODDs (7, 8). AVs with Level 3 and Level 4 automation are expected to be operating in a mix of AVs and human-driven vehicles (HDVs), thus developing scenarios incorporating human driver behavior is essential for validating Level 3 and Level 4 AVs (9). National-level historical crash data offers a comprehensive sample of crashes covering different ODDs, thus is an excellent data source for developing challenging scenarios for AV safety evaluation.

Scenario is defined as a sequence of scenes that consist of actions, events, goals, and values (10). Prior efforts developed AV test scenarios by characterizing crashes based on multiple dimensions of crash contributing factors and attributes (11–13). Without incorporating the sequential connections among events and actions in crashes, the generated crash scenarios lack specificity (14). There is a gap between the crash scenarios developed by prior efforts and crash scenarios ready to be implemented in simulation-based tests. Characterizing crashes based on crash sequences can generate more detailed, comprehensive, and realistic scenarios to narrow this gap.

Crash sequences are sets of chronologically ordered pre-crash and crash events. In a prior study by Wu et al., crash sequences were found to directly affect injury severity, and crash sequence analysis to be useful for understanding the dynamics and causations of crashes (15). Sequence analysis is widely applied in biological and social sciences but has not been explored extensively in traffic safety research. The purpose of sequence analysis in traffic safety research is similar with that in biological and social sciences, focusing on identifying representative components of sequences, analyzing sequence similarities and differences, and evaluating the relationships between sequences and potential outcomes (15). However, as there are numerous theories and techniques involved in sequence analysis, a methodology needs to be developed to appropriately apply the theories and techniques to crash sequences. Questions to be addressed concern specific aspects of crash sequence analysis including sequence data processing, sequence encoding, comparison, clustering, and interpretation. Further questions to be addressed include: How are crash sequence types associated with other attributes of crashes such as human factors, environmental conditions, and crash outcomes? Also, how to identify statistically significant scenarios for specific ODDs for AV testing? Association and causal analysis would help answer those questions.

Another concern in crash-sequence-based test scenario generation is what data to use. As this research focuses on Level 3 and Level 4 AVs, which are expected to share the roads with HDVs, historical crash data can be used to generate test scenarios. An assumption is that HDV crash cases are challenging for AVs, as they reflect the dynamics of human driver behaviors and environmental conditions that would be encountered by AVs. Some vehicle interaction challenges will not be covered by scenarios generated from

historical HDV crash data. Fortunately, as open-road tests are being carried out, new data from those tests can provide us insights in unique scenarios of AV crashes (6, 16). However, unusual scenarios or edge cases need to be addressed from the AV system safety design perspective (17).

In this research, some HDV scenarios were developed using data from a national crash database – the National Highway Traffic Safety Administration (NHTSA) Crash Report Sampling System (CRSS), and unique AV crash scenarios were developed using data from the California Department of Motor Vehicles (DMV) AV collision reports. The CRSS databases provide a comprehensive sample of crashes covering various ODDs, and chronologically ordered crash sequences of events. The California AV collision reports offer detailed narratives for extraction of crash sequence data.

1.2 Conceptual Framework and Research Focus

A conceptual framework of scenario generation for AV safety evaluation was developed for this research, as illustrated in Figure 1. The framework consists of four components, and this research focuses on three of them. **Crash sequence analysis and clustering** generate an array of representative sequence types. **Analysis of associations and causations** specify the relationships between crash sequence types and other crash attributes including human factors, environmental conditions, and crash outcomes. **Test scenarios** that are statistically significant for specific ODDs are identified through that procedure. Included in the framework for completeness but not a focus of this dissertation, **rubrics** (including metrics and benchmarks) can be developed for AV evaluation, as a

further study of crash causal analysis. Test scenarios and rubrics are expected to affect and strengthen each other in the process of scenario-based testing of AVs.

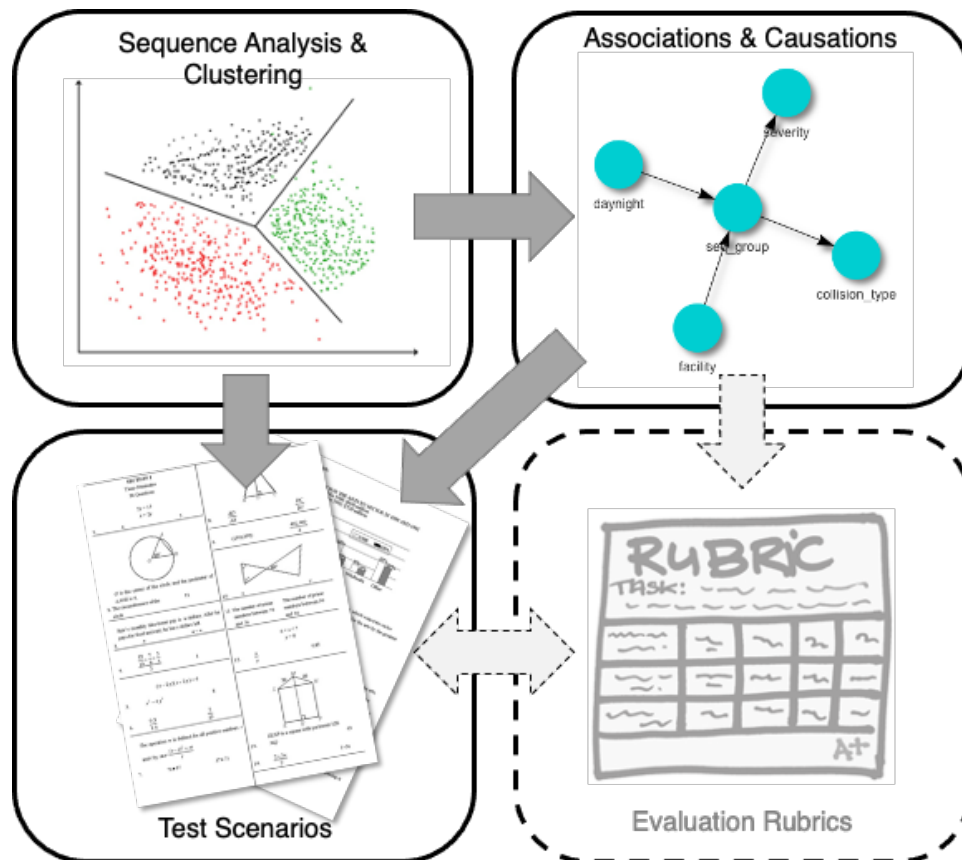


Image sources: lucidworks.com; management.ind.in; sc.edu

Figure 1 Conceptual framework

1.3 Objective, Research Questions, and Study Design

The objective of this dissertation is to develop a methodology of crash sequence analysis and apply a sequence-analysis-based procedure to generate test scenarios for AV safety evaluation. Specifically, the following four research questions are addressed:

- How to analyze HDV and AV (Level 3 and 4) crash sequences?
- What can we know from analyzing crash sequences of events?

- What unique scenarios can be developed using available AV crash data?
- What scenarios can be developed using HDV crash data?

To address the four research questions, this research focuses on two tasks: 1) Developing and demonstrating a methodology of crash sequence analysis, focusing on selecting the most appropriate encodings and dissimilarity measures for specific use cases of crash analysis. 2) Developing and demonstrating a procedure to generate crash scenarios for AV testing, which include two steps: crash characterization based on sequence analysis, and specifying relationships between crash dynamics (i.e., sequences) and ODD depictions (e.g., environmental conditions, human factors).

The two tasks were performed in three separate studies. The first study developed a crash sequence analysis methodology, and used the NHTSA CRSS data to demonstrate its effectiveness and usefulness. The second task was performed with two studies, one applying the scenario generating procedure on data from California AV collision reports, and the other applying the procedure on NHTSA CRSS crashes.

The analytical methods applied in this dissertation research can be classified into two categories – analysis of sequence patterns, and analysis of variable relationships. Sequences were studied at the levels of element, subsequence, and whole sequence, using methods such as transition matrix analysis, optimal matching, and clustering. The relationships between sequence types and ODD variables were studied using methods of cross-tabulation and graphical modeling. Detailed descriptions of methods are provided in each of Chapters 3, 4, and 5.

To summarize, this dissertation research:

- Developed a novel methodology of crash sequence analysis (data processing, encoding, comparison, and clustering).
- Innovatively applied sequence analysis to California AV crash data, developed unique AV crash scenarios, and proposed a framework of scenario-based AV safety evaluation.
- Designed and applied the test scenario generating procedure based on crash sequence analysis using historical HDV crash data.
- This research showed that crash sequence analysis is effective in characterizing HDV and AV crashes, and useful in generating statistically significant scenarios for AV safety evaluation.

1.4 Dissertation Outline

The dissertation outline is illustrated in Figure 2.

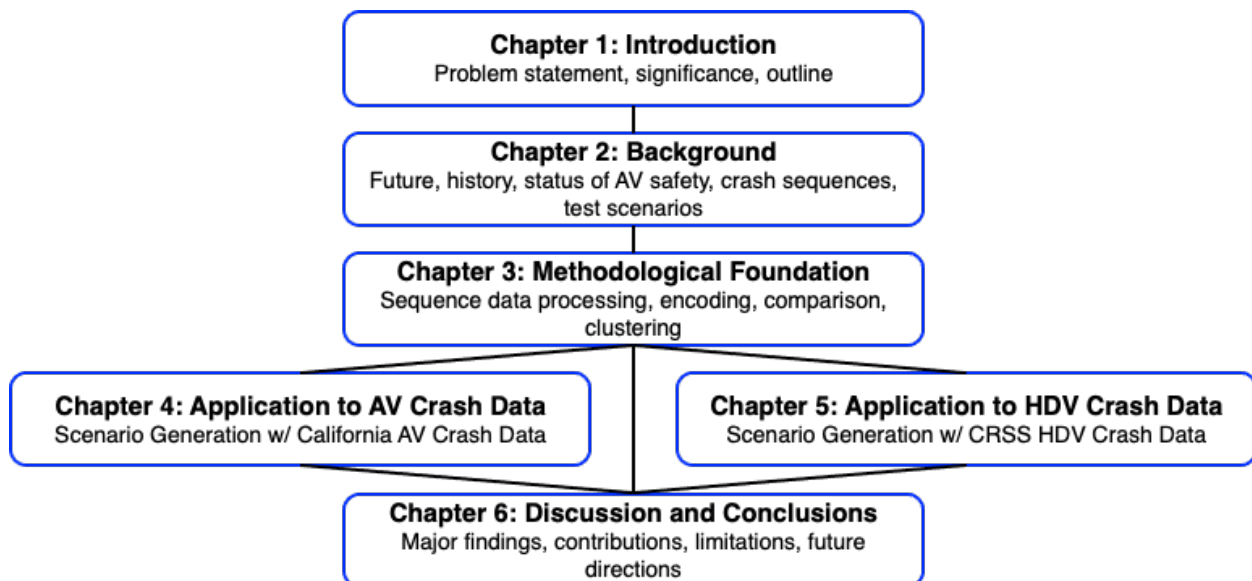


Figure 2 Dissertation outline

This dissertation is organized as follows:

Chapter 1 introduces the background, motivation, conceptual framework, focus, objective, and contribution of this dissertation.

Chapter 2 gives a detailed background of automated transportation and AV safety, and reviews literature related to crash sequences and AV test scenarios.

Chapter 3 develops a methodology of crash sequence analysis. A procedure of crash sequence data processing, sequence encoding, dissimilarity measuring, and clustering was introduced. Using data from the NHTSA CRSS database, different sequence encoding schemes and dissimilarity measures were compared. The optimal dissimilarity measures were identified for each encoding schemes based on the agreements with a benchmark crash typology. Apart from developing AV test scenarios, the crash sequence analysis methodology is applicable to the modeling of crash injury severity and analysis of crash causations.

Chapter 4 applies sequence analysis to California AV crash data. The sequence analysis characterized AV crashes into seven distinctive types based on patterns in sequences of events. The seven types showed unique scenarios of AV crashes, including AVs' hesitant driving style and the involvement of disengagements. Based on an analysis of associations between AV crash sequence types and other crash attributes, a preliminary framework of scenario-based AV testing was proposed.

Chapter 5 applies sequence analysis to the CRSS data and generated scenarios for intersection two-vehicle crashes. Sequence encodings were specifically designed for the purpose of test scenario developing, and best performing dissimilarity measure was used

to compare and cluster crash sequences. Fifty-five sequence types were developed as representative scenarios. A Bayesian network (BN) was developed to analyze the causal relationships among sequence types, human factors, environmental conditions, and crash outcomes. BN analysis can identify statistically significant scenarios for specific ODDs.

Chapter 6 concludes the dissertation with a discussion of the dissertation's major findings, contributions, limitations, and several directions for future work.

Chapters 3 to 5 are organized in the format of manuscript. Chapter 3 is a slightly modified version of a submitted paper (under review). Chapter 4 is a slightly modified version of a published paper. Chapter 5 is a working paper to be submitted. List of references for the original papers:

- Chapter 3: Song, Y., M.V. Chitturi, & D.A. Noyce. A Methodology for Traffic Crash Sequence Analysis: Impact of Event Encoding and Dissimilarity Measures. Submitted to Accident Analysis & Prevention (Under Review).
- Chapter 4: Song, Y., M.V. Chitturi, & D.A. Noyce. Automated Vehicle Crash Sequences: Patterns and Potential Uses in Safety Testing. Accident Analysis & Prevention, 153, 2021.
- Chapter 5: Song, Y., M.V. Chitturi, & D.A. Noyce. "Intersection Two-Vehicle Crash Scenario Specification for Automated Vehicle Safety Evaluation Using Sequence Analysis and Bayesian Networks". (To Be Submitted to the Journal of Safety Research).

Chapter 2 Background

2.1 Future of Transportation and Safe Automation

2.1.1 Future of Transportation

There is not a clear understanding of what the transportation system will look like in terms of operations and safety when most or all vehicles are automated. Some potential benefits of AVs, as frequently mentioned in government documents and media coverages, may include but not limited to: improved safety, improved mobility, improved accessibility, improved supply chain management, improved land use, and improved energy efficiency (1, 18–20). Similarly, some potential negative effects of AVs include but not limited to: increased vehicle-miles-traveled (VMT), induced congestion (due to “empty” AV miles or new patterns for land use and development), increased urban sprawl, and increased cybersecurity risks for transportation systems (19, 20). Although a significant research effort could be initiated on each of the potential improvements and impacts of the implementation of AVs, this research effort is focused on safety. Specifically, how AVs will affect road safety and how to quantifiably determine if and how AVs achieve their proposed goal of maximizing safety benefits and minimizing safety risks. Improving traffic safety is one of the primary motivations for developing AVs, and is also prioritized as one of the U.S. Government Automated Vehicle Technology Principles, as listed in Table 2 (18).

Table 2 U.S. government automated vehicle technology principles (18)

Core Interest	Sub-Area
I. Protect Users and Communities	1. Prioritize Safety
	2. Emphasize Security and Cybersecurity
	3. Ensure Privacy and Data Security
	4. Enhance Mobility and Accessibility
II. Promote Efficient Markets	5. Remain Technology Neutral
	6. Protect American Innovation and Creativity
	7. Modernize Regulations
III. Facilitate Coordinated Efforts	8. Promote Consistent Standards and Policies
	9. Ensure a Consistent Federal Approach
	10. Improve Transportation System-Level Effects

Currently, public attitudes toward AVs are mixed, with safety being the most significant topic of concern. Based on the 2020 American Automobile Association (AAA) surveys of consumer sentiment on AVs, only 12% of the surveyed drivers trust a fully automated vehicle to drive itself, 28% do not know how they feel about the technology, and the remaining 60% would be afraid to ride a fully automated vehicle. The results also showed that 72% of surveyed drivers would feel safer riding in an AV permitting drivers to take over control, and 47% would feel safer knowing the AV has passed rigorous testing and inspections (21). Before 2020, AAA had conducted 5 automated vehicle surveys in 4 years, with results showing a trend in percentage of drivers afraid to ride in a fully automated vehicle changing as illustrated in Figure 3. Anecdotal evidence based on University of Wisconsin-Madison AV experiences provided to the public suggest that experience with automated vehicles and transport in general tends to reduce the anxiety of AV transportation.

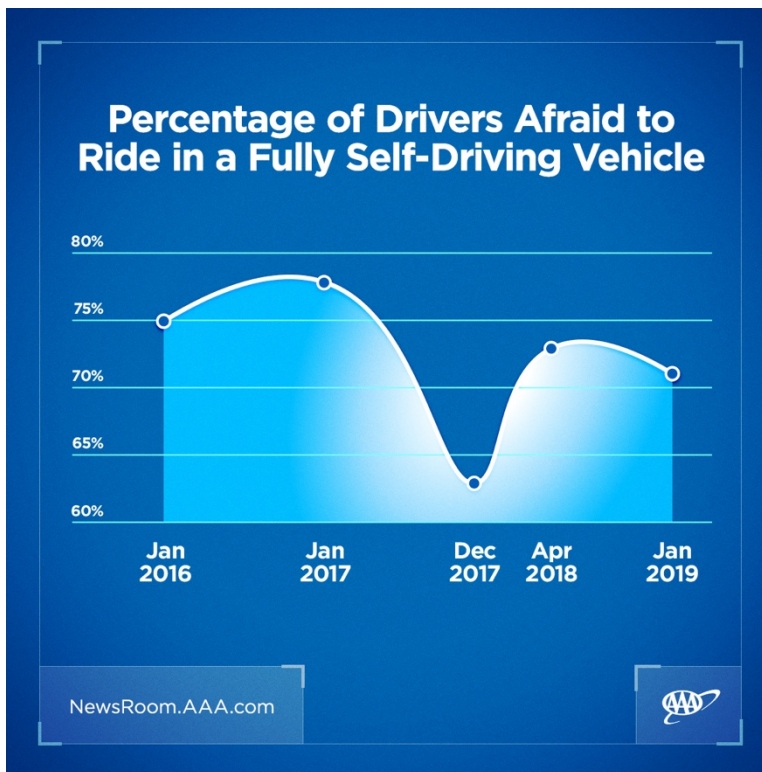


Figure 3 Trend in AAA survey results (22)

Will people ever build trust in AVs? One might assume that the answer is “yes” when AV technology becomes well established and ubiquitous. Nevertheless, a short-term answer may be found from the history of the occurrence and reception of disruptive transportation technologies. When bicycles were first appearing on streets in the U.S. in the 1880s, people had mixed opinions, with some people viewing it as a dangerous machine (23). Not long into the 1900s, the introduction of motor vehicles destabilized the U.S. urban street norm again (23). Motor vehicles were viewed as “juggernauts”, “needless and inherently dangerous machines”, and were portrayed by the media then (see Figure 4) as a “machine age Moloch to which motorists sacrificed generous offerings of child victims” (23). More than a century has passed since the first chaos brought by motor vehicles to the

streets, people seem to have well-accepted the co-existence of people walking and bicycling with motor vehicles, varying in type, size, and the fuel burned. Once a disruptive technology, motor vehicles are a necessary part of society and the primary variable in new land development. In 2018, there were 269,424,328 registered motor vehicles in the U.S., in which 111,242,132 were automobiles (private or commercial), with about 338 automobiles per 1,000 people (24). According to National Household Travel Survey, the average number of motor vehicles owned per U.S. household has grown significantly, from 1.16 (1969), to 1.77 (1990), to 1.89 (2001), and to 2.27 (2017) (25, 26).

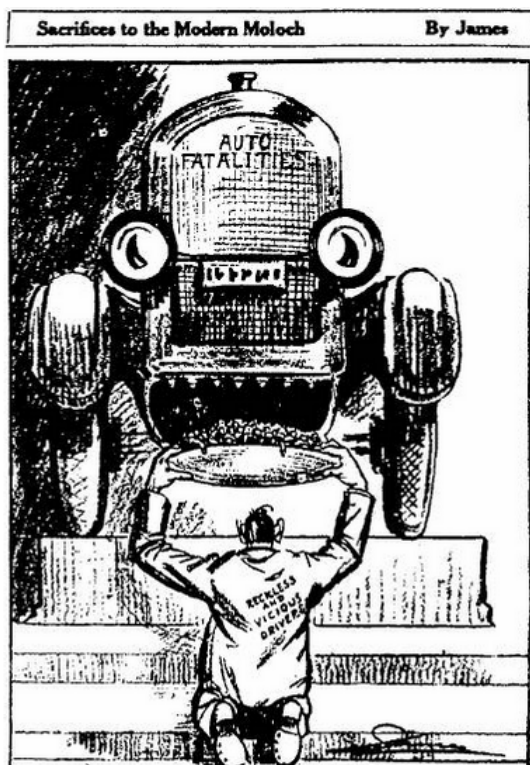


Figure 4 Cartoon by Saint Louis Star (November 6, 1923, p. 14) (23)

Data showing trends in vehicle crash fatalities, as plotted in Figure 5, offers insights into the safety record of motor vehicles. The fatalities per capita (million people) measure

shows the significant impact and cost to public safety and health brought by motor vehicles, and the fatalities per billion vehicle-miles-traveled (VMT) shows the safety effects of motor vehicles on a usage basis. Fatalities per capita trended higher in the 1920s, the dawn of motor vehicle transportation in the U.S., and peaked in the 1930s. World War II brought the number down, which picked up an immediate increase after the war and reached another peak in the 1960s and 1970s. Since the 1970s energy crisis, fatalities per capita has generally trended down. Fatalities per VMT, however, since the beginning of the motor vehicle age, has decreased steadily. Automobiles did not start with a well-accepted safety record, which raised great social concerns. However, with the advancement of vehicle and traffic safety technologies, as well as enhanced understanding and education of traffic safety, motor vehicles started receiving more social acceptance. That does not mean that motor vehicles have reached a satisfactory safety record. There were still 36,560 people killed from motor vehicle crashes on U.S. roads in 2018 (27). According to the Centers for Disease Control and Prevention (CDC) data, in the same 2018, 655,381 deaths in the U.S. were attributed to heart disease, 599,274 deaths to cancer, 159,486 deaths to chronic lower respiratory diseases, 147,810 deaths to stroke, and 122,019 deaths to Alzheimer's disease (28, 29). Motor vehicle traveling is one of the leading causes of death in the U.S., especially for younger age groups (27, 30). Yet few people think about concern for their safety when getting into their vehicles each day.

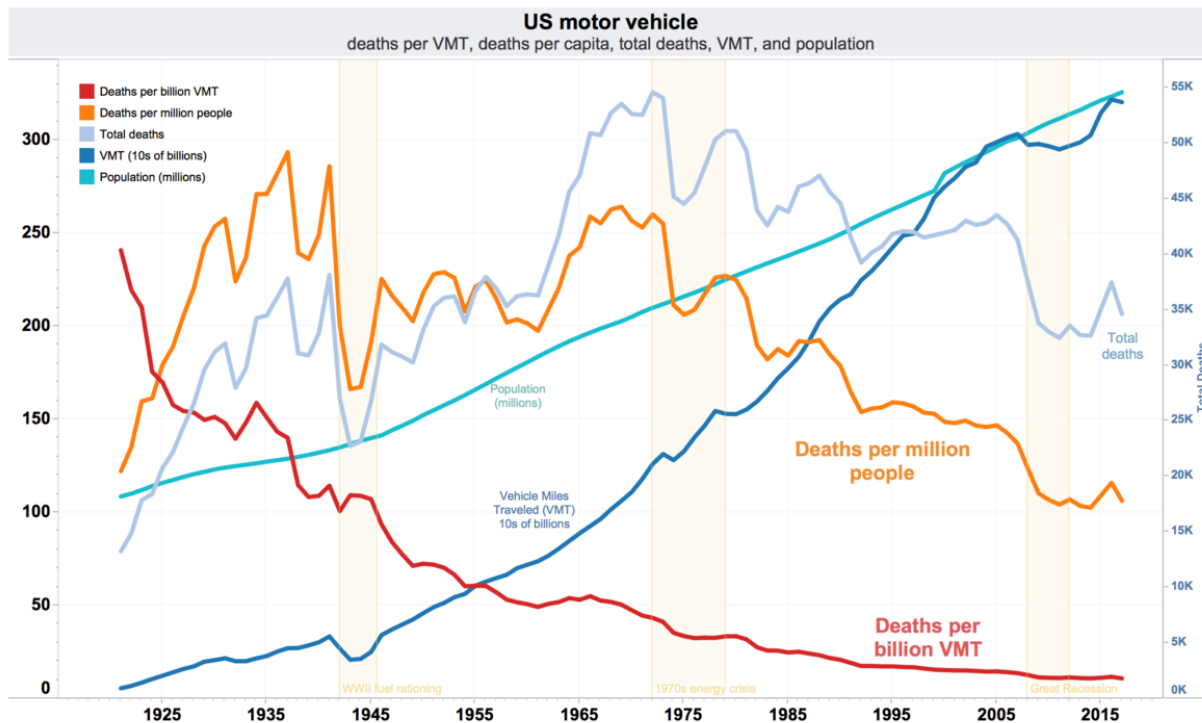


Figure 5 Trends in annual U.S. motor vehicle fatalities (1921–2017) (31)

It is fully expected that AV technology will lead to a significant reduction in the number of transportation-related fatalities each year. However, this belief is not universally accepted. Potentially, AVs will be part of a solution to bring the traffic fatality number down, but it is necessary to make sure these new technologies operate the way as we expected, including but not limited to following traffic rules, cooperating with surrounding road users, and properly reacting to emergency situations. Automobiles took a long time and too many lives to be finally accepted, but still pose safety risks to the public today. AVs should not go on the old path of the 1920s' automobiles, they should be proved safe before being deployed to accomplish their goal of making transportation safer and more efficient.

2.1.2 Safe Automation

For a considerably long period of time, AVs will share the existing infrastructure and follow the same traffic rules with human-driven vehicles (HDVs) (1, 32, 33). It is impossible to predict how long such a period will be, but data may give us some idea. According to data from the Bureau of Transportation Statistics, the average age of cars and light trucks on U.S. roads was 11.8 years in 2019 (34). Typically, people in the U.S. do not replace their vehicles frequently (35). Not to say that there is yet any certified fully automated vehicle rolling out of an assembly line for customers to purchase. Also, there will be people that would like to stick with conventional vehicles. Passenger vehicles with automatic transmissions were firstly being sold in the 1930s, but till now vehicles with manual transmissions are still being sold (35, 36). The presumably long co-existence of AVs and HDVs is also implied by the fact that AVs currently being tested (e.g., Waymo deployments) were designed to drive on existing roadways, recognize existing traffic control devices, and even mimic human driving behaviors (37, 38). Although AVs are expected to reduce some errors and overcome some difficulties in driving tasks by minimizing the involvement of human factors, with a mixture of AVs and HDVs, AVs will still face most safety challenges that human drivers are facing now on our roadways. In addition to that, AVs may encounter even more various challenges which have not been encountered by human drivers, due to uncertainties in AVs' automated driving systems (ADS) and AVs' unique driving styles. Uncertainties in complex automated systems come with possible hardware failures, software bugs, algorithmic errors, or cyber-safety issues. Efforts could be put into minimizing each uncertainty, but no system is 100% bullet-proof. AVs' unique driving styles include overly paranoid reactions (e.g., hesitation when making turns or merging into traffic, sudden

acceleration or braking) and incapability of socially communicate with other road users (33, 39–42). AVs' unique driving styles may not be a significant issue when all vehicles are automated and all road users are connected in our transportation system, but before such a system is in place, AVs' overly paranoid and uncooperative driving styles may lead to violations of other road users' expectancy, which potentially increase safety risks.

What does AVs' safety record look like now? Although there is not a wide deployment of any AVs on public roads, field tests are being carried out. In California, crashes happened to AVs being tested on public roads have been reported to the state Department of Motor Vehicles (DMV) since 2014 (43). AVs tested in California are or aimed at SAE driving automation Levels 3 and 4. As of November 3, 2020, 274 AV crashes have been reported to California DMV (44). Earlier comparisons between crash records of AVs from California and those of HDVs from national crash databases showed that it is still difficult to tell if AVs performed better than conventional vehicles (45–47). One study by Banerjee et al. did claim that current AVs are 15 to 4,000 times worse than human drivers in safety performance measured by crashes per cumulative mile driven (48). All California AV crashes during testing have caused property damage only or minor injuries with no fatal crashes reported. Nationwide, there have been 4 reported fatal crashes in the U.S. led by motor vehicles with different levels of driving automation in or before November 2020. Three fatal crashes occurred with Tesla vehicles on "Autopilot" (a Level 2 automation feature) during customer use, with drivers killed. One fatal crash occurred with an Uber vehicle (with Level 3 automation) in automatic driving mode during testing, killing a pedestrian. The three Tesla crashes and one Uber crash were investigated by the National Transportation Safety Board (NTSB) (49–52). Investigations concluded that the Tesla and

Uber crashes were all primarily caused by some vehicle control system functioning or design issues. In other words, some part of the vehicle's operating system failed. More needs to be done to prove the safety of AVs as they are being developed.

What does "safe" mean for AVs? Carnegie Mellon University AV safety expert, Koopman, gives the following definition (9):

"Safe" means at least correctly implementing vehicle-level behaviors such as obeying traffic laws (which can vary depending upon location) and dealing with non-routine road hazards such as downed power lines and flooding. But it also means things such as fail-over mission planning, finding a way to validate inductive-based learning strategies, providing resilience in the face of likely gaps in early-deployed system requirements, and having an appropriate safety certification strategy to demonstrate that a sufficient level of safety has actually been achieved.

Testing is an essential part of the interdisciplinary efforts to ensure AV safety (9). Private and public organizations are testing their AVs extensively on both closed courses and public roads (53, 54). In 2019, more than 1,400 automated vehicles were tested by more than 80 companies across 36 U.S. states and Washington, D.C. (55).

Field testing of AVs is a time-consuming process. Currently, there is no government requirements on how many miles AVs should be tested on road to prove their safety. The National Highway Traffic Safety Administration (NHTSA) issues Federal Motor Vehicle Safety Standards (FMVSS) for conventional vehicle safety regulation, but the standards do not cover much about advanced driver-assistance systems or automated driving systems, nor do the standards mention requirements about miles for conventional or automated

vehicle safety testing (56, 57). A RAND Corporation study estimated that with only field tests, AVs need to drive hundreds of millions of miles to ensure desirable safety performance, and extra millions of miles to verify changes in performance as the vehicles are improved over time (58). A significant amount of time is needed to design and prepare AV field tests. Also, AVs need to be tested under a variety of scenarios to ensure their capability in handling different situations. To speed up the AV evaluation process, simulation-based testing is also performed, and is regarded as essential as field testing. Waymo has run their AVs for 10 billion miles in simulation to train their vehicles, in complement to field tests, which, on the other hand, help improve their simulation scenarios (59). In Europe, projects such as PEGASUS (Germany) and HumanDrive (UK) are developing simulation-based AV safety evaluation frameworks (60, 61).

For simulation-based AV safety evaluation, defining test scenarios is important and challenging. Simulation scenarios need to provide a reliable estimate of real-life performance (12). Test scenario design needs to be supported by a good understanding of traffic crashes. Crashes are complex, with certain unique features in each case, but it is not possible to use each individual case as a scenario for AV testing. Similar with developing a set of standardized test problems to examine students' academic capability, a set of representative test scenarios must be designed for AV safety evaluation. Previous studies have attempted to develop test scenarios for simulation-based AV safety evaluation, through clustering historical human-driven vehicle crash data and finding representative cases (11–13, 62, 63). For example, Nitsche et al. developed pre-crash scenarios for testing ADS in intersection environments, by clustering historical crash data from a U.K. database, based on crash attributes (11). Sander and Lubbe also examined the potential of using

clustering methods in defining intersection test scenarios for automated emergency braking (AEB) system, using historical crash data from a German database (12). Sui et al. developed pre-crash scenarios for vehicle AEB testing, applying the same methods used by Sander and Lubbe, on a Chinese data set of car-to-two-wheeler crashes (13). Clustering was proved to be a suitable method for developing representative pre-crash scenarios, but scenarios developed from previous studies are far from comprehensive and detailed. Crash dynamic information, as pointed out by Sander and Lubbe, can potentially improve representative scenarios developed using historical crash data (12).

2.2 Crash Sequences and Scenarios

Interdisciplinary efforts have been implemented in defining test scenarios for AV safety evaluation. Many of these efforts agreed that scenario-based tests are essential for AV safety evaluation. AV testing frameworks proposed by programs such as PEGASUS and HumanDrive have included scenario-based testing modules (60, 61). Defining a set of representative test scenarios needs a strong support from a comprehensive understanding of historical traffic crashes, which not only includes static crash features, but also dynamic features and causations.

2.2.1 Crash Sequences

Sequence of Events

An example of a crash sequence of events from the National Highway Traffic Safety Administration (NHTSA) Fatality Analysis Reporting System (FARS) database is as follows:

2017 Crash #390629

Encoded sequence: [v1/64]-[v1/59]-[v1/69]-[v1/12]-[v2/63]-[v2/59]

Short description: [v1 ran off road – left]-[v1 hit traffic sign]-[v1 re-entering highway]
-[v1 in transport]-[v2 ran off road – right]-[v2 hit traffic sign]

It describes a crash case involving two vehicles. One vehicle (v1) ran off road from the left side (crossed median), hit a traffic sign, and re-entered the roadway to continue driving. The other vehicle (v2) ran off road from the right side and hit a traffic sign. Compared with crash contributing factors, crash sequence of events presents a more complete picture of a crash with additional temporal and spatial information. Compared with surrogate safety measures (e.g., from vehicle video or sensor records), crash sequence of events describes crash dynamics in a more natural and concise manner.

The concepts and analytical methods of sequence have been developed and applied in fields such as bioinformatics and sociology, with the sequence of events concepts and analyses especially widely applied in sociology (64–69). In sociology, sequences of events are used to study the effects of change and development in people’s lifetime on the societal outcomes (64). In social sequence analysis, two structural assumptions about sequence are applied (64):

- *Certain social elements are stochastically related to each other, such that they appear in similar positions vis-à-vis each other, even across unrelated actors.*
- *Even though different actors may experience social elements in a different order, similar types of sequence patterns emerge.*

Similar assumptions are applicable to traffic crash sequence of events, but maybe more limited than social sequence, which has wider possible outcome than crash sequences. For

crash sequence of events, the order of events (actions) may weigh heavier in defining the outcome, compared with social sequence. In this research, the assumptions for crash sequence of events are:

- Certain pre-crash events (actions) are stochastically related to each other, such that they appear in similar positions vis-à-vis each other, even across unrelated actors.
- Similar types of sequence patterns emerge from different actors' experience of traffic crashes, although differences between specific sequences exist.

The importance of finding causes to crashes through investigating the progression of crash events has been long advocated by traffic safety researchers (70–75). The 1979 Tri-Level Study of the Causes of Traffic Accidents found that 50% of the 2,000 crashes studied were caused by more than one factor (76). Sequence of events is important information for traffic crash investigation. However, there are few studies applying sequence analysis methods on traffic crashes, partially due to the insufficiency in crash sequence of events data, and partially due to traffic safety researchers' unfamiliarity with sequence analysis methods. In 2011, the National Transportation Safety Board (NTSB) recommended that sequence of events to be included in national crash databases (77). Since then, with several updates, more crash sequence of events data was included in the NHTSA crash databases. However, at state, regional, or municipal-level crash databases, crash sequence of events data is not widely available yet.

Analysis of crash sequences can be used to develop sophisticated characterization of crashes. Such a crash characterization considers dynamics in crashes, rather than

categorizing crashes into broad groups (e.g., head-on, rear-end, sideswipe, angle, and hit fixed-object) through conventional crash characterization (15). Complicated crash causations can also be disentangled through investigation of crash sequence of events, which is also helpful in identifying effective prevention strategies (70).

Prior studies putting an emphasis on crash sequence of events analysis were mostly carried out by Kun-Feng Wu and colleagues (15, 70, 78–80). Before Wu, Krull et al. published a study in 2000, investigating the effects of rollovers and events sequence on single vehicle crash injury (81). Krull et al. used a brief description to summarize the sequence of rollover and another (preceding or succussing) hit object event. Wu started their crash sequence of events research with screening naturalistic driving study (NDS) video and vehicle kinematic data (70, 78, 79). Crashes and near crashes were identified for precursor event analysis. In a 2016 study, Wu et al. used sequence analysis methods, including optimal matching and clustering, on FARS data to group similar crashes and model crash severity outcomes (15). In a 2018 study, Wu et al. identified groups of motorcycle crashes with high crash risks, based on harmful event, crash type, and crash sequence of events data from the National Automotive Sampling System – General Estimates System (NASS-GES) and FARS databases (80). These previous studies proved that crashes can be grouped based on their similarities in sequence of events, and that different groups of sequences did lead to significant differences in crash outcomes.

Causations

Causations in traffic crashes (or “accidents” as a preferred term in other disciplines) have long been pursued by safety researchers, with multiple causation theories developed since the 1940s (73–75, 82). These theories of causations are not limited to traffic crashes.

As traffic crashes and other fields' accidents (e.g., industrial, chemical, or nuclear engineering) have similarities, some theories are also applicable to traffic crashes. Major theories of accident causation include (83, 84):

- The Domino Theory Developed by H.W. Heinrich
- Human Factors Theory
- Accident/Incident Theory
- Epidemiological Theory
- Systems Theory
- The Energy Release Theory by Dr. William Haddon, Jr.
- Behavior Theory

Accident causation theories developed from attribution to mechanical and structural failures in the first technological age; to human behaviors and errors in the second age; to a socio-technical systemic view in the third age; to a recognition of organizational and cultural factors in the fourth age, and to the fifth age's focus on complexity and uncertainty, aiming at create safety well before the occurrence of failures and harms (85).

Sequence of events has been used to study traffic crash causations since as early as the 1970s (86). Snyder and Knoblauch set up a framework of behavioral sequence to model the progression of actions in motor vehicle-pedestrian crashes, as shown in Figure 6 (87).

The model has two main components:

- The function/event sequence: search, detection, evaluation, decision, and action.
- The influencing/predisposing factors: driver, pedestrian, vehicle, and environmental factors.

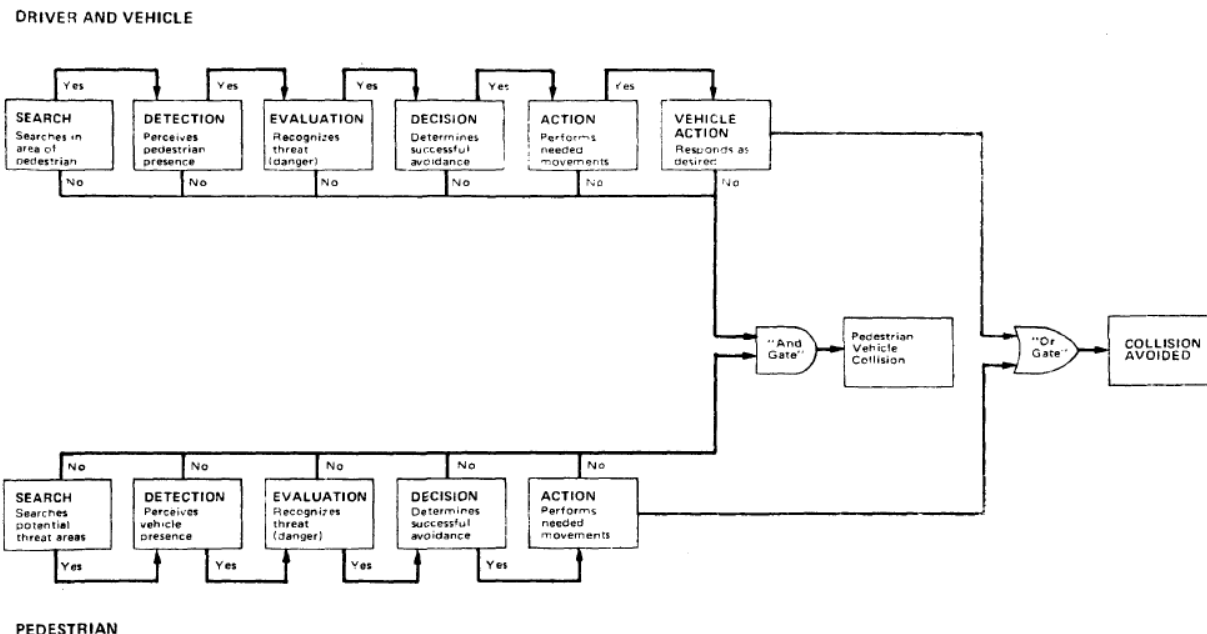


Figure 6 Generalized function/event sequence (87)

To understand crash causations with the Snyder and Knoblauch model, the mechanism and magnitude effects of influencing/predisposing factors on the function/events need to be understood (87). Crash reports were reviewed in Snyder and Knoblauch’s study. Combinations of factors were identified, with their frequencies calculated. Several causal patterns of motor vehicle-pedestrian crashes were then summarized and illustrated using a flowchart-type visualization, as shown in Figure 7.

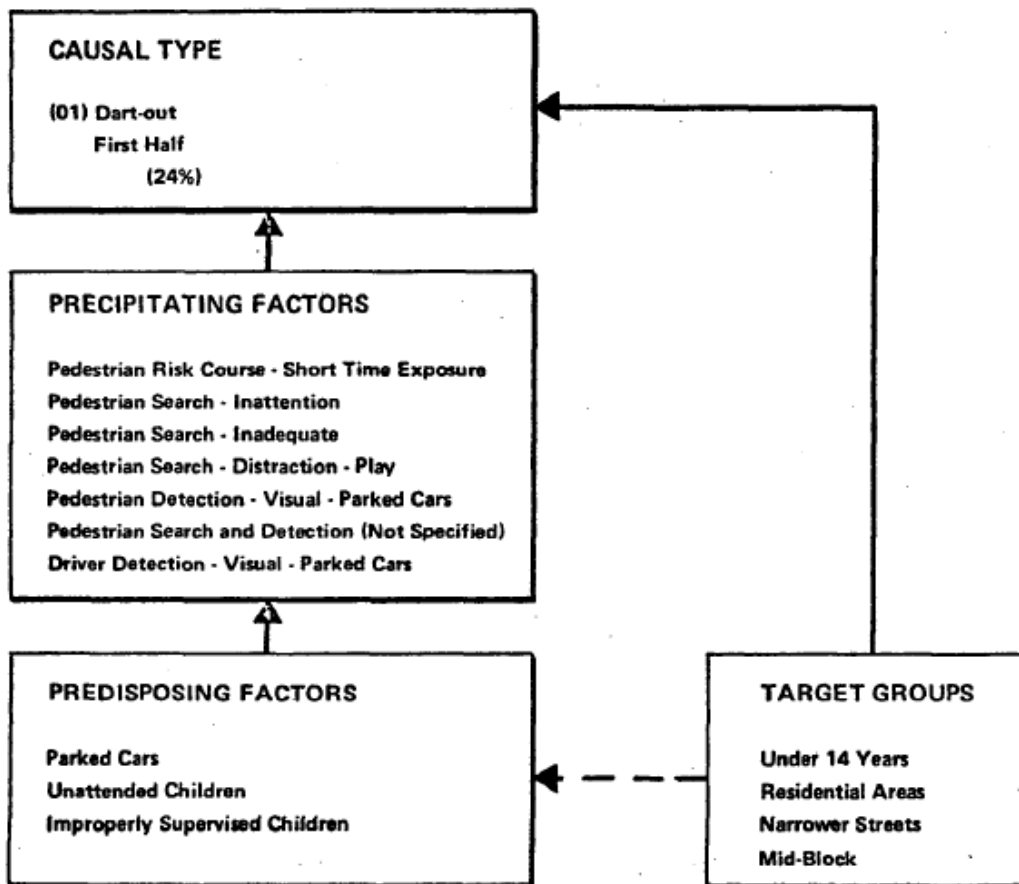


Figure 7 Illustration of causal pattern "dart-out first half" (87)

Cross and Fisher also pointed out that crashes are caused by multiple factors and events. With that assumption, a conceptual model as shown in Figure 8 was proposed (88). Three types of factors: operator, vehicle, and environmental factors are considered potential causal factors. Combinations of factors lead to some function failures, and with critical actions, affecting the terminal event.

In addition to proposing a framework to model crash causations, Cross and Fisher also suggested that a structured representation of crash causations should be developed and used as a basis for crash classification (88). Rather than using arbitrarily chosen

descriptive variables, a way that can capture similarities in crash-generation process should be applied to classify crashes (88). Cross and Fisher reviewed crashes reports of motor vehicle-bicycle crashes, and carried out interviews of motor vehicle operators and bicyclists. Motor vehicle-bicycle crashes were classified into 6 classes including a total of 36 mutually exclusive types. Each type of crashes has subtypes, with frequencies calculated, and illustrated as an example shown in Figure 9.

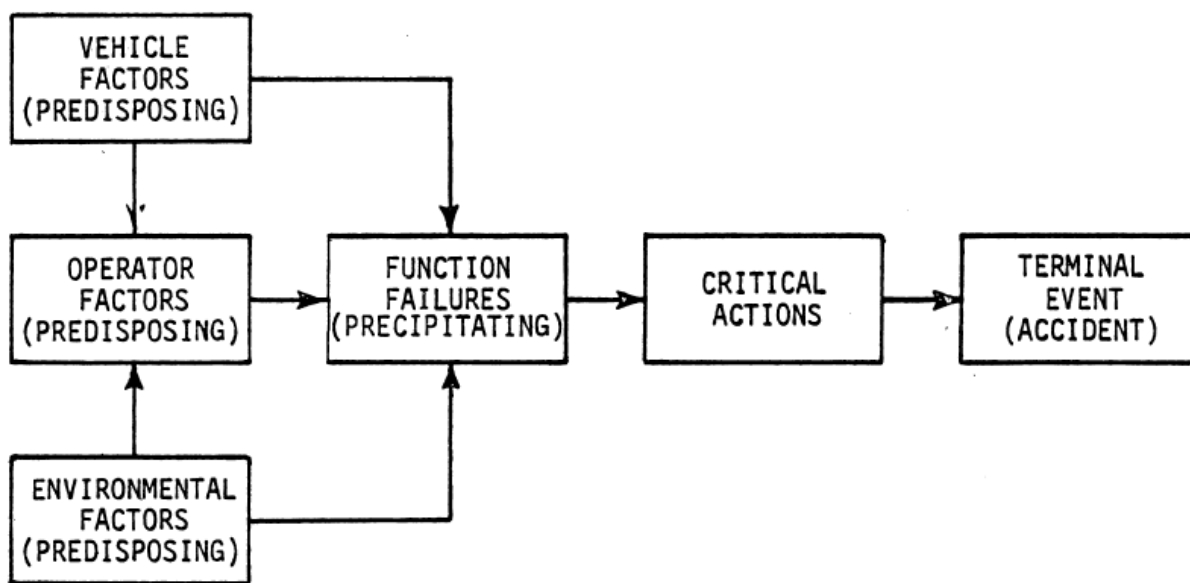


Figure 8 Conceptual model of the crash generation process (88)

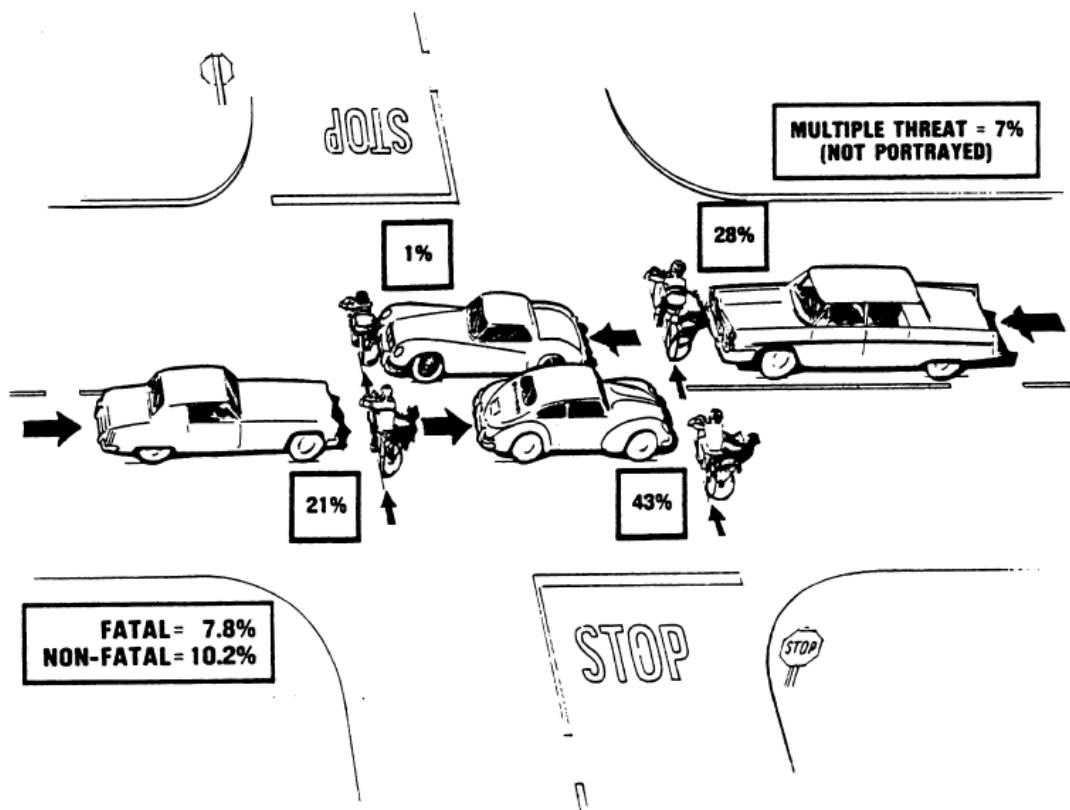


Figure 9 Illustration of a type of motor vehicle-bicycle crash, “bicycle rideout – intersection controlled by sign” (88)

More crash sequence of events data is becoming available from police-report-based crash databases and multi-media and sensor data sources such as naturalistic driving studies (NDS) databases. With abundant data, more in-depth quantitative crash causal analysis can be done. Davis pointed out the necessity of crash causal analysis in making engineering decisions, and discussed forensic inference, statistical and Bayesian probabilistic methods for crash causal analysis, with possible support from simulation (89, 90). Davis et al. later developed a framework for modeling causations in traffic conflicts and crashes (91). However, from a theoretical framework to a practical application requires a

further understanding of the mechanism in evasive actions' progression transferring to conflicts, and the probability of certain conflicts can transfer into crashes with different severity levels(91).

2.2.2 Scenarios for Automated Vehicle Testing

Definition

Ulbrich et al. provided definitions for the terms of “scene”, “situation”, and “scenario”, in the context of automated driving (10). The definitions are as follows:

- *Scene: A scene describes a snapshot of the environment including the scenery and dynamic elements, as well as all actors' and observers' self-representations, and the relationships among those entities. Only a scene representation in a simulated world can be all-encompassing (objective scene, ground truth). In the real world it is incomplete, incorrect, uncertain, and from one or several observers' points of view (subjective scene).*
- *Situation: A situation is the entirety of circumstances, which are to be considered for the selection of an appropriate behavior pattern at a particular point of time. It entails all relevant conditions, options and determinants for behavior. A situation is derived from the scene by an information selection and augmentation process based on transient (e.g. mission-specific) as well as permanent goals and values. Hence, a situation is always subjective by representing an element's point of view.*
- *Scenario: A scenario describes the temporal development between several scenes in a sequence of scenes. Every scenario starts with an initial scene. Actions & events as*

well as goals & values may be specified to characterize this temporal development in a scenario. Other than a scene, a scenario spans a certain amount of time.

Traffic crash sequence of events describes the order of actions in time, thus fits into the definition of scenario.

Frameworks

The PEGASUS project defines three levels of AV test scenarios corresponding to three phases of the ISO 26262 standard “Road Vehicles – Functional Safety” (92). As shown in Figure 10, for the concept phase, functional (descriptive) scenarios are needed; for the system development phase, logical scenarios are needed; and for the test phase, concrete scenarios should be defined. Traffic crash sequence of events fits into all three stages of scenario design, depending on the granularity of event data. As the data sources for crash sequence of events in this research will be historical crash databases and reports, analyses done in this research will support the development of descriptive scenarios. Logical and concrete scenarios should be derived from descriptive scenarios, but would require granular vehicle video, GPS, and sensor data, from sources such as NDS. Menzel et al. gave an example of three levels of scenarios of a car following a truck, as shown in Figure 11.

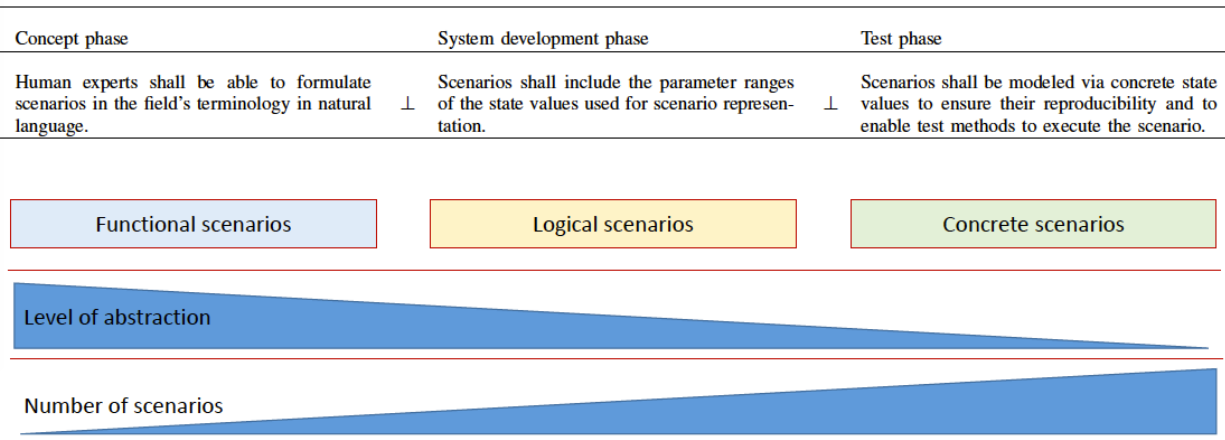


Figure 10 Levels of scenarios (92)

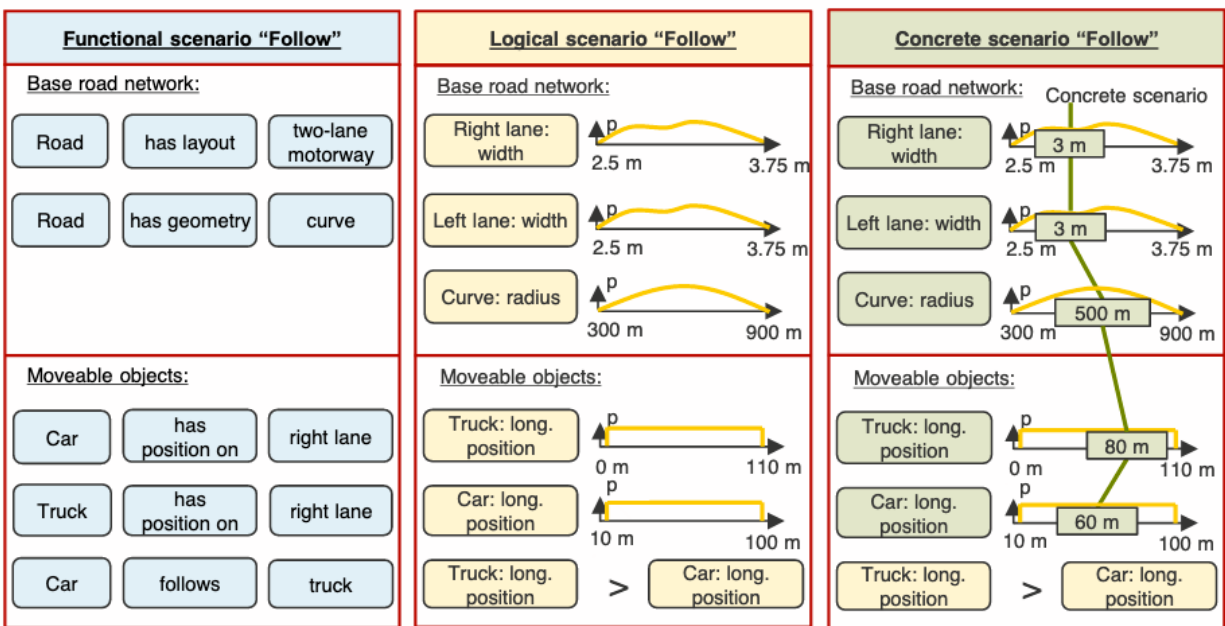


Figure 11 Example of three levels of scenarios (92)

To categorize the variables needed for scenario development, PEGASUS proposed a layered model, as shown in Figure 12. The original layered model consists of 5 layers, including road geometry, road furniture and rules, temporal modifications and events,

moving objects, and environmental conditions. A 6th layer, digital information, was later added to the PEGASUS layered model, to capture the connectivity features of future transportation systems (93). There are also other types of layered model describing the hierarchy of variables for scenario design. An example is a model used by Xia et al., which had 3 layers of “influence factors” (94). Influence factors were first categorized as environmental, road, traffic, and vehicle dynamics. Further, the environmental factors were categorized into weather, time, and light change. Under the weather factor, detailed measures of weather type and extent were included.

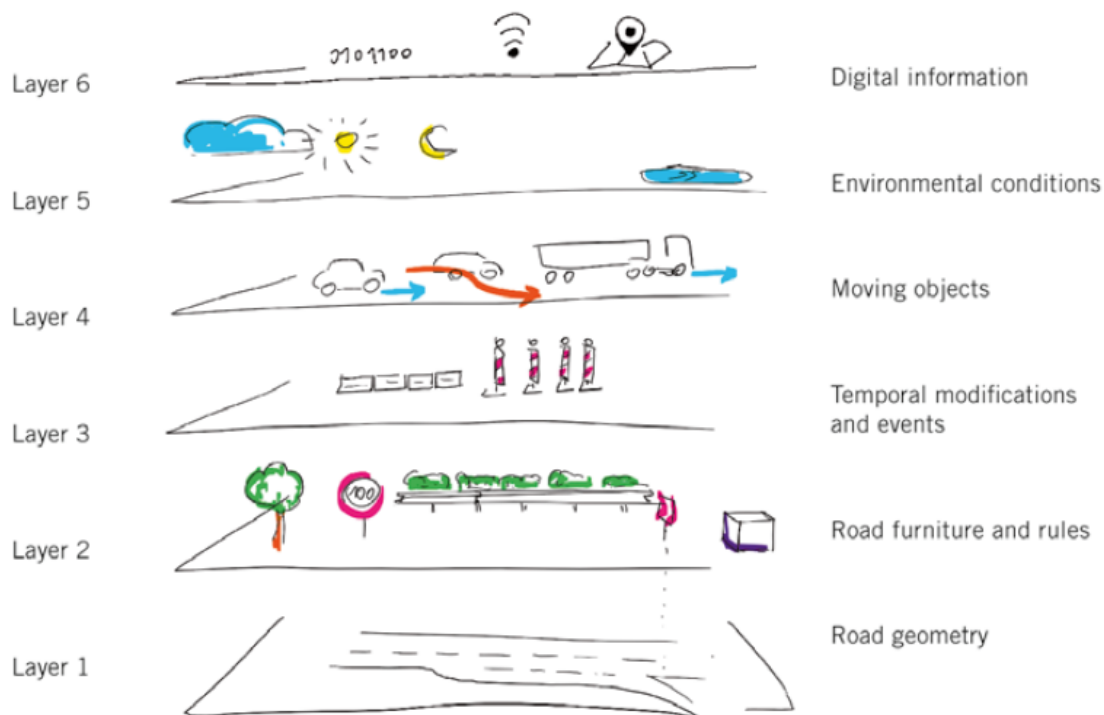


Figure 12 Layered model of variables describing an AV test scenario (93)

NHTSA proposed a framework for ADS testable scenarios (95). However, based on Ulbrich’s definition of “scene”, “situation”, and “scenario”, the NHTSA framework is more

focused on the capabilities of ADS to handle “situations” rather than “scenarios” (10). The NHTSA framework of testable scenarios consists of four components:

- Tactical maneuver behaviors
- ODD elements
- OEDR capabilities
- Failure mode behaviors

An example of such a scenario was provided by NHTSA, as shown in Table 3. This framework was built on the assumption that, given a situation, an AV knows which tactical maneuver to take to appropriately handle the situation. A test would then be focused on examining whether the AV can efficiently carry out the maneuver under such a situation.

Table 3 Example of an NHTSA ADS scenario descriptor (95)

Scenario Elements	Example
Tactical Maneuver Behaviors	Perform lane change/low-speed merge
ODD Elements	Arterial roadway type
	Asphalt roadway surface
	Lane markers
	Straight, flat
	72 kph (45 mph) speed limit
	Nominal traffic
	Clear, dry weather
	Daylight
	...
OEDR Behaviors	Detect and respond to relevant adjacent vehicles (frontal, side, rear)
Failure Mode Behaviors	N/A

Descriptive Scenarios

To identify descriptive (or semantic-level) scenarios for evaluation of AV or driving assistance systems, previous studies have been applying various categorization methods to

crash databases. An effort to develop a library of descriptive crash scenarios was a crash typology developed by the NHTSA in 2007, for crash avoidance research (14). The NHTSA pre-crash scenarios typology was developed using the 2004 NASS-GES data. Thirty-seven pre-crash scenarios were developed to form the typology, through reviewing the GES variables and codes, summarizing frequencies, and estimating crash costs. The 2007 NHTSA pre-crash scenarios typology provides a highly abstract set of functional scenarios, which describes the major events, contributing factors, and limited information about traffic control/facility type, prior to the occurrence of crashes involving at least one light vehicle. Based on crash frequency, economic cost, and functional years loss measures, the top scenarios were defined as:

- Control loss without prior vehicle action
- Lead vehicle stopped
- Road edge departure without prior vehicle maneuver
- Vehicle(s) turning at non-signalized junctions
- Straight crossing paths at non-signalized junctions
- Lead vehicle decelerating
- Vehicles(s) not making a maneuver – opposite direction

The NHTSA pre-crash scenario topology was updated and applied in later NHTSA research on crash avoidance systems and safety applications based on vehicle-to-vehicle communications (96–98). Except for the NHTSA NASS-GES, other national crash databases such as the Crashworthiness Data System (CDS), FARS, and National Motor Vehicle Crash Causation Survey (NMVCCS) have been used to obtain historical crash data to develop

scenarios of specific crash types (e.g., lane departure, opposite-direction road departure, pedestrian, and cyclist) for field or simulation-based testing of corresponding crash-avoidance systems (99–104).

A few recent studies applied clustering techniques on historical crash data to develop crash scenarios for the testing of vehicle technologies such as automatic emergency braking (AEB) systems (11–13, 62, 63). Conventional approaches to obtain scenarios for vehicle safety testing, such as the NHTSA 2007 pre-crash scenario approach, is to classify crashes into homogeneous groups, usually by summarizing crash frequencies based on individual variables such as contributing factors, pre-crash conditions, and manner of collision (e.g., rear-end, sideswipe). The most frequent crash groups were then selected as representative scenarios. However, real-life vehicle safety performance is usually overlooked when using scenarios developed using such a method, which ignores crash dynamics, progression, and causations (12). Crash reports and databases provide rich information of crash dynamics, progression, and causation, which can be extracted using advanced data mining methods.

Lenard et al. used clustering to identify typical pedestrian crash scenarios for the testing of AEB systems, with data of 9,360 motor vehicle – pedestrian crashes from the UK’s “On-The-Spot” (OTS) project database (62). Crashes clustered into 22 groups as representative scenarios. Nitsche et al. developed pre-crash scenarios for road intersections, with a data set of 1,056 HDV intersection crashes from the OTS database (11). Thirteen clusters were found for three-legged intersections, and six clusters were found for four-legged intersections. Sander and Lubbe also used clustering to find representative intersection pre-crash scenarios for AEB testing, and evaluated the

performance of clustering-identified scenarios in terms of predicting the AEB effectiveness, compared with that of randomly-sampled crash cases (12). Sander and Lubbe concluded that the clustering-identified scenarios may not appropriately represent intersection crashes, which are highly diverse due to variation in features such as vehicle kinematics. In a later collaborative study by Sui et al., Sander's methodology was applied again to develop car-to-two-wheeler scenarios for AEB testing, but with kinematics variables added (13). Although the kinematics variables only provide one movement state of the vehicles (e.g., go straight, turn left, turn right, or others) and a first contact point on the car.

These previous studies show that clustering is suitable for characterizing historical HDV crash data and develop descriptive scenarios, but more details about crash dynamics, progression, and causations are needed for a better crash characterization that reflect different pre-crash patterns to be useful in guiding the setup of simulated scenarios for AV testing. A prior effort made by Aust attempted to define crash scenarios based on detailed pre-crash information (105). Aust proposed a nested model consisting of a macroscopic layer of pre-crash scenarios and a layer of driver self-reported crash causation data (from questionnaire surveys). Nested model is a useful tool to map causations to crash characteristics, but driver survey data tends to have issues in validity and accuracy.

Detailing of Scenarios and Implementation in Simulation

With a well-defined set of functional scenarios, to develop more concrete scenarios for simulation-based AV testing, NDS and Field Operational Tests (FOTs) are essential data sources. A number of previous studies have carried out case studies to identify safety-critical events and scenarios using NDS data (106–108) and FOT data (109, 110). Several studies are more methodological, focusing on approaches to mine real-life driving data to

obtain scenarios for AV testing (111–113). Some studies proposed methods for selection of logical and concrete scenarios based on criteria such as criticality, complexity, or some form of cost (114–117). For implementation of developed scenarios in simulation-based testing, several previous studies have proposed automatic test scenario generation frameworks (118–125). These proposed frameworks apply techniques such as stochastic sampling, optimization, and neural networks, with a goal to generate a variety of scenarios that are challenging for AVs (e.g., difficult for AVs' motion planning). The automatically generated scenarios are vehicle-trajectory-based, with safety performance described by surrogate measures such as time-to-collision (TTC).

Chapter 3 A Methodology for Traffic Crash Sequence Analysis: Impact of Event Encoding and Dissimilarity Measures

3.1 Introduction

Characterization of traffic crashes has been emphasized by traffic safety researchers for studying crash patterns and identifying safety countermeasures, with the objectives of mitigating crash injury severity and ultimately preventing crashes (15, 80, 126–128). Early and recent studies have stressed the importance of considering pre-crash information in crash characterization for better insights about crash progression and causations (15, 80, 81, 86–88, 100, 101). The use of crash sequence data has been proved to be effective in developing a crash typology that reflects the progression of crashes and correlates with crash injury severity outcomes (15, 80). Crash sequences are sets of chronologically ordered pre-crash and crash events, which are usually extracted from police crash reports and are available in United States' national-level crash databases such as the Crash Report Sampling System (CRSS) and Fatality Analysis Reporting System (FARS) of the National Highway Traffic Safety Administration (NHTSA).

Large sets of crash sequences can be analyzed using sequence analysis methods adapted from fields such as bioinformatics and sociology (15). The fundamental task of sequence-based crash characterization is to compare and differentiate sequences, which involves a variety of measures and techniques. In biological and social sciences, sequence analysis methods have been developed, studied, and applied for more than 40 years (64, 65, 129, 130). Domain knowledge is needed for sequence analysis because the definition and formation of sequences are domain-specific. Adaptation of sequence analysis to traffic

crash study is still very recent, so there are no existing guidelines for processing sequence data or analyzing sequences.

Sequence analysis follows a procedure of data processing, measuring sequence dissimilarity, and clustering (15, 64). The data processing step, which includes interpretation of source information and encoding of sequences, is highly dependent on domain knowledge. However, few studies have discussed techniques of transferring crash report information into sequences, and none has compared different sequence encoding schemes for crash sequence analysis (16). The core of sequence analysis is measuring sequence dissimilarity, which is also the basis of sequence clustering. Various sequence dissimilarity measures have been developed for application in biological, computer, and social sciences (64, 65, 130–133). A comparison of dissimilarity measures is needed to adapt the most appropriate ones to crash sequence analysis, but no study has made such an effort yet.

The objective of this chapter is to fill the gap in crash sequence analysis by introducing a methodology for crash sequence analysis. This chapter demonstrates and compares the application of crash sequence analysis for various use cases through different sequence encoding schemes and dissimilarity measures with a case study using data from the NHTSA CRSS database.

The rest of this chapter is organized as follows. A literature review follows this section and introduces the basics, needs and the few prior applications of sequence analysis in the study of traffic crashes. The techniques of crash sequence data processing and dissimilarity measuring are then introduced, followed by a case study of single-vehicle

interstate highway crash sequences. The case study demonstrates how different sequence encoding schemes and dissimilarities were compared to adapt to a specific crash sequence data set and multiple use cases. The chapter is concluded with summaries of key findings, contributions, limitations, and directions for future work.

3.2 Literature Review

The importance of characterizing crashes through investigating the progression of crash events has been advocated by early traffic safety researchers (74, 87, 88). For the purpose of systematic investigation of crash loss-reduction options, Haddon Jr. proposed a model that described traffic crashes as a chronological chain of precrash, crash, and post-crash stages, with human, vehicle, and environmental factors coded into each stage (74). In Snyder and Knoblauch's study on motor vehicle-pedestrian crashes and in Cross and Fisher's study on motor vehicle-bicycle crashes, sequences of crash participants' actions were embedded into conceptual models describing crash generation processes (87, 88). Characterization of crashes in the early studies were completed by manually summarizing crash report information, which was not capable enough to analyze long and complicated sequence of events.

Sequence analysis methods such as sequence alignment were developed and applied for purposes such as protein or DNA sequence characterization, codes and error control, text and speech processing, and the study of social phenomena (65, 133). Early foundations of sequence comparison methods were set up by researchers such as Hamming and Levenshtein who proposed measures of sequence dissimilarity, as well as Needleman and Wunsch who developed the Optimal Matching (OM) algorithm to efficiently

find sequence dissimilarity (129, 131, 132). Promoted by Abbot in the 1980s, sequence analysis based on the OM methods started to gain popularity in the sociology field (65, 66, 68). Ever since, variants of OM and new sequence analysis methods have been developed to address various data conditions and study needs (134–139). Sequence dissimilarity measures have been compared, using real or simulated data, for applications in biological and social science research (130, 140–143).

Kun-feng Wu and colleagues were the first to apply sequence analysis methods in traffic crash study (15, 80). In their study to identify the relationship between crash sequences and crash injury outcomes, Wu et al. introduced the basics of sequence analysis to the traffic safety community, and applied OM and clustering to characterize a set of fatal single-vehicle run-off-road crash sequences from the NHTSA FARS database (15). The characterization of crash sequences was validated by estimating the agreement between the sequence clustering results and a benchmark crash typology provided by FARS. The sequence clustering results were then used as a variable in crash injury severity estimation, which showed significant correlations between sequence-derived crash types and crash injury outcomes. Wu et al.'s study confirmed that informative and meaningful crash characterization could be derived from crash sequence analysis, and safety countermeasures could be identified to target crashes and injuries from a crash sequence perspective. In another study by Wu et al., built a risk matrix based on motorcycle crash sequences and identified sequences leading to the highest injury risks (80).

Sequence analysis methods have been applied to study patterns in automated vehicle (AV) crashes (16). Sequences of events were extracted from the original crash reports and were encoded using self-developed encoding scheme and procedure.

Subsequence frequencies and event transition rates were analyzed to explore subsequence-level patterns. OM method was applied to characterize the AV crash sequences into seven types. Based on the associations of sequence types with environmental conditions and crash outcomes, a scenario-based AV safety evaluation framework was proposed to embed crash-sequence-derived test scenarios. Through investigating AV crash sequences, a preliminary procedure for crash sequence analysis was proposed. However, more efforts are needed to develop a comprehensive methodology for crash sequence research. As Wu et al. pointed out, further investigations are needed for crash sequence encoding, dissimilarity measuring and clustering techniques, which is the focus of this chapter.

3.3 Crash Sequence Analysis Methodology

A methodology of crash sequence analysis is proposed here. The methodology consists of three steps – data processing, sequence encoding, and sequence comparison. Common sources of crash sequence data are crash databases and crash reports. Other data sources such as traffic surveillance videos or naturalistic driving study (NDS) data also offer granular information that can be transferred to sequence of events but would require the help from techniques such as video processing. This chapter focuses on developing a methodology for text-based crash sequence data obtained from crash databases and crash reports. The crash sequence analysis methodology is illustrated in Figure 13 and briefly described as follows. A detailed demonstration of the methodology is provided in Section 3.7 Case Study.

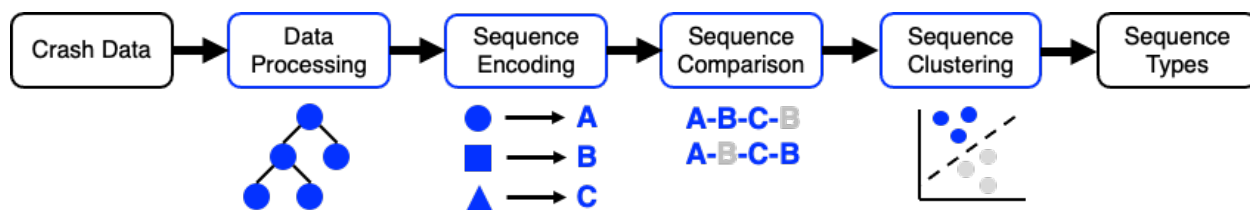


Figure 13 Procedure of crash sequence analysis

Data processing: Depending on the purpose and focus of an analysis, crash attributes other than sequence of events (e.g., time of day, facility type, demographics of persons involved in crash) are used as criteria for subsetting the crash data.

Sequence encoding: Pre-crash and collision events are extracted from crash reports or obtained from databases. Events are encoded into simple representations based on their meanings. Events considered similar in nature can be consolidated and encoded with the same representation. For example, in some analysis, “hitting a utility pole” and “hitting a tree” can both be classified as “hitting a fixed object” and encoded as “HFO”. Depending on the use cases, the same sequence can be encoded differently. Events are assembled into sequences following a specific order, usually chronologically.

Sequence comparison: A sequence is compared with all the other sequences in the sequence space using a dissimilarity measure. There are multiple ways to define the dissimilarity measure and the different measures can be compared to select the optimal one for specific analysis.

Sequence clustering: Sequences are characterized through clustering based on their dissimilarities. Clustering techniques such as hierarchical and k-medoids clustering are widely used.

3.4 Data Processing

To demonstrate the crash sequence analysis methodology, the 2016-2018 NHTSA CRSS data were used in this chapter's research. The CRSS is a United States national-level database archiving a sample of police-reported crashes involving all types of road users, with all levels of severity (144). Each year, the CRSS gathers a sample of about 50,000 crashes, representing about more than six million police-reported crashes. Each sample is assigned a weight showing how many crashes are represented and need to be accounted for in statistical estimations.

There are four levels of data in CRSS, the crash level, vehicle level, person level, and event level. Data files can be linked through case, vehicle, and person identification numbers, as shown in Figure 14. The crash level data files include crash characteristics and environmental conditions. Multiple data files at the vehicle level consist of information about motor vehicles (in-transport, parked or working) involved in the crashes, drivers of motor vehicle in-transport, and pre-crash vehicle and environmental conditions. The person level data files provide information about drivers, passengers in motor vehicles, as well as non-motorists (e.g., pedestrians and bicyclists) involved in the crashes. The event level data files archive non-harmful and harmful events that occurred in the crashes in the form of sequences. For each crash, a sequence of events is recorded based on the police crash report narratives and diagrams. A corresponding vehicle number is provided for each

event, but across different crash cases, the consistency of vehicle numbering is not ensured. The CRSS User Guide states that “all vehicles (motor vehicles in-transport as well as parked/working vehicles) are sequentially ordered starting with 1”, but no detailed information is provided about the rules for vehicle numbering. However, vehicle numbering is not a concern for single-vehicle crash sequences.

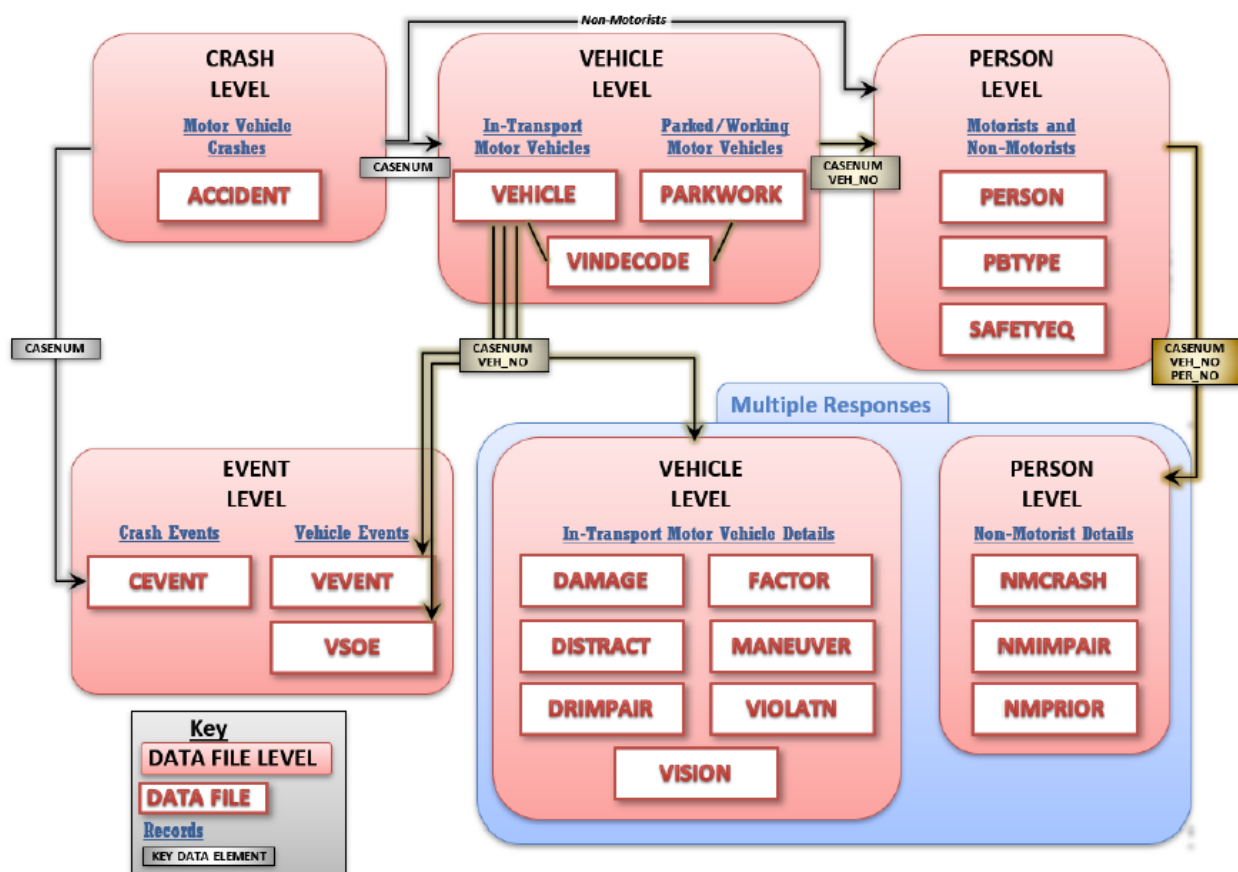


Figure 14 Linkage structure of the CRSS data files (144)

The CRSS database (and many other national or state-level crash databases) consists of a large and comprehensive sample of crashes. Analyzing sequences using an entire database regardless of the basic characteristics of crashes would likely generate

results that are complex and difficult to interpret, or lose important, detailed information if the data was overly summarized. Breaking down a database into smaller subsets and focusing only on crashes relevant to specific analysis purpose is recommended.

Many variables can be used to subset crashes for sequence analysis. Depending on the purpose of analysis and the amount of available data, fewer or more variables can be used to categorize crashes in simpler or more detailed ways. For example, using combinations of the following four variables, different subsets of crash data can be obtained to satisfy different analysis purposes.

- Facility type: intersections, interchanges, or road segments
- Functional classification (or related factors): principal arterial, minor arterial, collector, or local (related factors include number of travel lanes, speed limit, and others)
- Area type: rural or urban/suburban
- Number of participants involved in a crash: single-vehicle or multi-vehicle

Figure 15 illustrates a tree structure formed by the above-mentioned four variables for crash data subsetting.

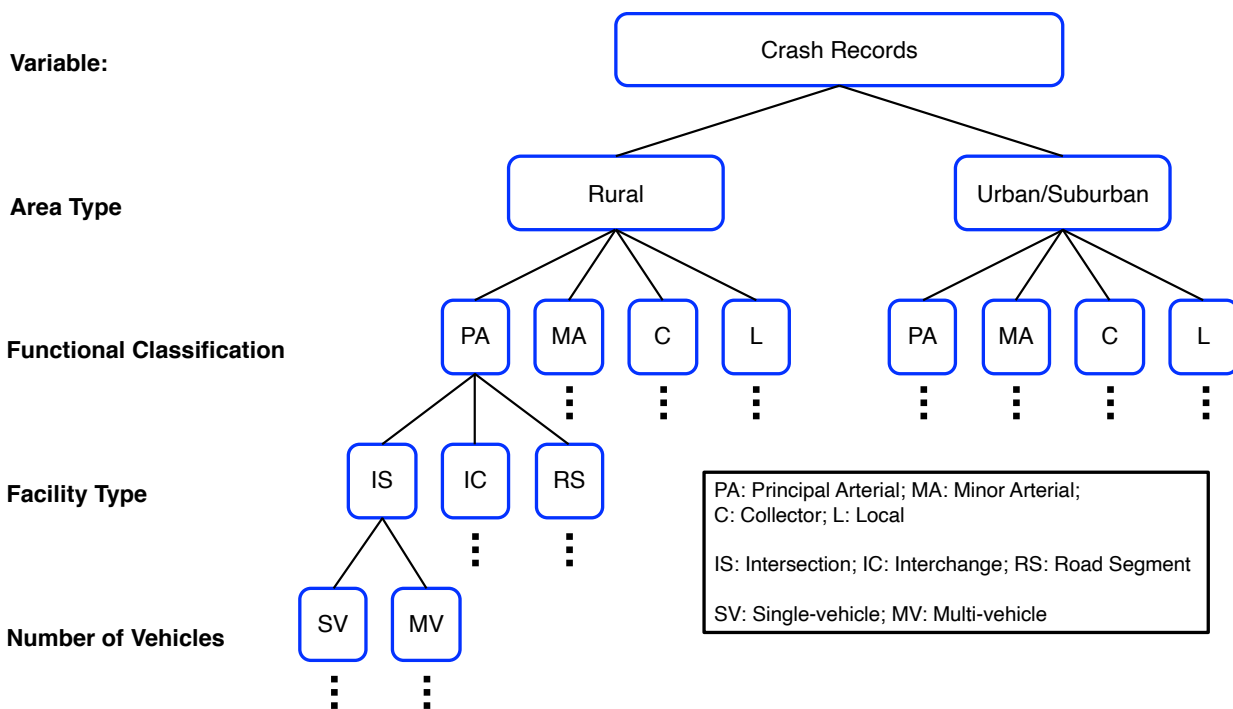


Figure 15 Tree-structured subsetting of crash data

3.5 Crash Sequence Encoding

Before crash sequences can be analyzed, they should be encoded. The raw information of crash sequences in crash reports is usually in the form of text narratives. Encoding is converting the text narratives of pre-crash and crash events into a series of event labels ordered by time. Event labels should be consistent, with very similar events (depending on semantic similarities and domain knowledge) encoded with the same label. Temporal information in the text narratives should also be analyzed to determine the order of events in sequences. When only raw crash reports are available, the text narratives can be processed by using natural language processing (NLP) tools, or manually by following a

procedure of “text narratives → short phrases → labels” to ensure consistency in the encoding (16).

CRSS (and similar national-level crash databases like FARS) has developed a set of sequence encodings, but alternative encodings may be needed to accommodate the needs of different use cases. Four cases where crash sequence of events may be of use are:

- To use as an intermediate variable or base structure in the analysis of crash causations (86, 88).
- To use as a basis for the generation of scenarios for safety testing of automated vehicles (AVs) and advanced driver assistance system (ADAS) (16).
- To use as an explanatory variable in the modeling of crash injury severity (15, 80).
- To estimate frequencies of different types of crashes.

Crash causation analysis, in transportation engineering, is conducted primarily for the purpose of identifying proper countermeasures to prevent crashes or mitigate crash loss. When encoding crash sequences for causation analysis, emphasis should be given to the actions and interactions involved in the pre-crash phase, with as much detail included as possible, rather than the objects or road users involved in the crash.

For developing test scenarios for the safety evaluation of AVs or ADAS, the role of crash sequences is to support crash characterization and development of typology. Sequence-supported crash typology, compared with characterization of crashes by manner of collision, provides more information about chronological orders of events and relative locations of parties involved, which are essential for recreating (real-world or simulated)

scenarios for vehicle testing. Encoding crash sequences for the purpose of test scenario generation requires a clear definition of participating vehicles' initial states. Each multi-vehicle crash can be used to generate multiple scenarios by placing the ego vehicle (i.e., the vehicle being tested) in the positions of multiple participating vehicles. When encoding multi-vehicle crashes, the assignment of vehicle IDs should be based on the pre-crash roles of vehicles. Participating vehicles are classified as either an initiator or responder in a crash (145). For single-vehicle crash scenarios, the ego vehicle is placed in the position of the only participating vehicle.

For crash injury severity modeling, more information about collisions with other road users, animals, or objects are required than for other use cases. Take fixed-object collisions as an example, both the physical characteristics and the number of fixed objects a vehicle collided with are important information for modeling crash severity outcomes. As previous studies have shown, differences in road user and object characteristics such as dimensions, weight, and hardness affect crash injury severity differently (81, 146–151). Also, a vehicle colliding with multiple fixed objects is expected to sustain higher risk of severe damage than a vehicle colliding with a single object. When encoding sequences for the modeling of crash injury severity, fixed objects should be categorized to differentiate their potential injury risks based on physical characteristics. Fixed objects should also be differentiated from live animal, pedestrians, bicyclists, and other road users.

3.6 Crash Sequence Dissimilarity Measures

There are many ways to measure dissimilarities between sequences, which are used as the basis to compare sequences and characterize them into distinctive groups. Studer

and Ritschard reviewed, proposed, and compared more than 20 sequence dissimilarity measures for sociological study of life course sequences (130, 142). Not all dissimilarity measures used in social sequence analysis are adaptable to traffic crash sequences, due to the following characteristics of crash sequences:

- Different crashes have different sequence lengths, so dissimilarity measures requiring same sequence lengths (e.g., Hamming distance) are not applicable.
- Crash sequences consist of events or actions that usually do not repeat consecutively. Therefore, spell duration (the length of a consecutively repeated element), although considered in some social sequence dissimilarity measures, is not applicable to the measuring of crash sequence dissimilarity.
- The structures of crash sequences differ as the number of participants in crashes differ. Therefore, the analyses of single-vehicle crash sequences and multi-vehicle crash sequences may require different types of dissimilarity measures.

In this chapter, two rounds of selection were carried out to compare the dissimilarity measures. In the first round, nine dissimilarity measures considered potentially suitable for crash sequence analysis were selected from a list of measures introduced in the Studer and Ritschard study (130, 142). The nine dissimilarity measures were then compared, with five selected as the most suitable ones for crash sequence analysis. The five selected dissimilarity measures were used to demonstrate the crash sequence analysis methodology in a case study of interstate highway single-vehicle crashes.

According to Studer and Ritschard, sequence dissimilarity can be measured based on element distributions, counts of common attributes, or edit distances (130, 142). The

nine candidate dissimilarity measures for crash sequence analysis are organized by those three categories and are described below.

3.6.1 Distances between element distributions

The frequency of element occurrence can be used to describe a sequence, as proposed by Deville and Saporta (130, 142, 152). If a sequence, x , with the alphabet $\{A, B, C, D\}$, is "ABCADD", it can be described as a vector of the four elements' occurrence frequencies: $(2/6, 1/6, 1/6, 2/6)$. The dissimilarity between a pair of sequences x, y , is then calculated as a Euclidean (EUCLID) distance, d_E , or a Chi-square (CHI2) distance, d_C .

$$d_E(x, y) = \sqrt{\sum_{i=1}^I (p_{i|x} - p_{i|y})^2} \quad [1]$$

$$d_C(x, y) = \sqrt{\sum_{i=1}^I \frac{(p_{i|x} - p_{i|y})^2}{p_i}} \quad [2]$$

where $p_{i|x}$ denotes the occurrence frequency of element i in sequence x , and p_i the overall frequency of element i in the entire sequence space. Element-distribution-based Euclidean and Chi-square distances are sensitive to element frequency and sequence length but are not sensitive to the order of elements. Comparing the EUCLID and the CHI2 measures, CHI2 gives more weight to rare cases than common cases because it divides the squared frequency difference by the overall proportion of occurrence rate of an element (130). An occurrence of a rare element would likely increase the distance significantly.

3.6.2 Measures based on counts of common attributes

Length of longest common subsequences

The length of longest common subsequence (LCS) is the number of elements in the LCS of a pair of sequences. A subsequence is a set of elements following the order in the parent sequence but are not necessarily adjacent (64). For example, the two sequences, "ABCD" and "ACB", have common subsequences: "A", "B", "C", "AB", and "AC". The LCS are "AB" and "AC", both with a length of 2. An LCS-based dissimilarity is measured as:

$$d_{LCS}(x, y) = |x| + |y| - 2l_{LCS}(x, y) \quad [3]$$

where $|x|$ and $|y|$ are the lengths of sequences x and y , and $l_{LCS}(x, y)$ is the length of the LCS between x and y . Therefore, the LCS-based dissimilarity between "ABCD" and "ACB" is $(4 - 2) + (3 - 2) = 3$.

LCS between long sequences are more difficult to identify than that between the example sequences. Finding LCS is a classic computer science problem and is usually achieved by using dynamic programming algorithms (153). LCS-based dissimilarity measure tends to be sensitive to the order of the most frequent elements and the frequency of those elements in sequences.

Number of matching subsequences

Another dissimilarity measure in this category is based on the number of matching subsequences (NMS) (130, 134, 135, 142). The NMS between a pair of sequences, x and y , is represented by:

$$A_{NMS}(x, y) = \sum_{u \in S(x, y)} emb_x(u) emb_y(u) \quad [4]$$

where u is a subsequence in the set of distinct common subsequences $S(x,y)$, $emb_x(u)$ is the count of times that u is embedded in x , and similarly for $emb_y(u)$. The NMS-based dissimilarity is measured as:

$$d_{NMS}(x, y) = \sqrt{S^2(x) + S^2(y) - 2A_{NMS}(x, y)} \quad [5]$$

where $S(x)$ is the count of subsequences of x , and similarly for $S(y)$.

The NMS-based dissimilarity measure is Euclidean, and is moderately sensitive to sequencing (130, 142). This measure is also expected to be sensitive to sequence lengths, as in the case of crash sequences with unequal lengths.

3.6.3 Edit distances

Optimal matching

Unlike the abovementioned dissimilarity measures, edit distances of dissimilarity are ad hoc and based on specific aspects of sequence differences (130, 142). Edit distances define several operations that can be used to transform one sequence to another, with costs assigned to the operations. The dissimilarity between the two sequences is then calculated as the smallest transformation cost needed, using optimal matching (OM) methods (64–66, 130, 133, 142).

OM is a widely used method for sequence alignment in computer science, bioinformatics, and social science, for uses such as speech/text comparisons, DNA sequence alignments, and life course studies (64, 65, 68, 130, 133, 142). Multiple types of operations can be applied in OM, including substitutions, deletions and insertions (or indels), compression and expansions, and transpositions (or swaps) (130, 133, 142). Substitutions and indels are the most commonly used operations.

A formal mathematical expression of OM dissimilarity between a pair of sequences, x and y , is (130, 136, 142):

$$d_{OM}(x, y) = \min_j \sum_{i=1}^{\ell_j} \gamma(T_i^j) \quad [6]$$

where ℓ_j denotes the transformations needed to turn sequence x into y ; $\gamma(T_i^j)$ is the cost of each elementary transformation T_i^j (e.g., indel or substitution).

Here, an example of two sequences “ABCD” and “ACB” is used to illustrate the principles of transformation operation costs and OM. Table 4 shows two ways to align the two sequences, with different combinations of operations. An indel costs d and a substitution costs s . Alignment 1 used 1 deletion and 2 insertions, costing $3d$. Alignment 2 used 1 deletion and 1 substitution, costing $d + s$. There are many other ways to align “ABCD” and “ACB”, with different costs. The OM method applies the Needleman-Wunsch algorithm and returns the minimum alignment cost as the dissimilarity between the two sequences (129).

Table 4 Sequence alignment costs

Sequence 1	A	B	C	D		
Sequence 2	A	C	B			
Alignment 1						
Sequence 1	A		B	C	D	
Sequence 2	A	C/	B	\emptyset	\emptyset	
Cost		d		d	d	$= 3d$
Alignment 2						
Sequence 1	A	B	C	D		
Sequence 2	A	\emptyset	C	<u>B</u>		
Cost		d		s		$= d+s$

Note: Insertion is marked with \emptyset ,
 Deletion is marked with slash/,
 Substitution is marked with underline

Basic optimal matching costs.

Basic OM cost scheme assigns constant costs to indels and substitutions. Measures such as Levenshtein, Levenshtein II, and Hamming use constant costs but allow operations differently (64, 131, 132).

- Levenshtein distance allows indels and substitutions.
- Levenshtein II distance allows only indels (equivalent to LCS).
- Hamming distance allows only substitutions and is only applicable to sequences with the same length.

There have been criticisms of OM on the transformation operations' lack of real-world meanings (64, 68, 69, 130, 142, 154, 155). However, more recent research has developed methods to define substitution and indel costs that sophisticatedly consider the real-world contexts of sequences (130, 137, 142).

Studer and Ritschard summarized that for social sequence of events, indels in OM represent time shifts and substitutions represent mismatches remaining after the time shifts (130, 142). However, for traffic crash sequences, indels are more of a representation of addition or removal of certain actions/events/objects, and substitutions represent replacement of certain actions/events/objects with others. Under this general principle of OM for crash sequences, the choice of indel and substitution costs should reflect real-world meanings and difficulties of conducting operations of additions, removals, and replacements. In this chapter, three cost schemes (Levenshtein, data-driven substitution costs, and localized indel costs) were compared.

The basic Levenshtein distance uses a substitution cost of 2 and an indel cost of 1. Replacements are not treated differently, so a replacement ($A \rightarrow B$) costs the same with ($A \rightarrow C$).

Data-driven optimal matching costs.

A commonly used data-driven method is to set up substitution cost based on the element transition rates (TRATE) observed from the sequence data set. The transition rate between two elements, A and B, is calculated as the probability of element A followed by element B in all cases that element A appears in the observed element space (64, 130):

$$P(AB) = P(B_{p+1}|A_p) = \frac{n(AB)}{n(A)} \quad [7]$$

where $n(AB)$ is the count of times that AB appears at consecutive positions (p and $p+1$) following that order; and $n(A)$ is the count of times A appears in the element space. The TRATE-based symmetrical substitution cost is calculated as (130):

$$\gamma_{tr}(A, B) = 2 - P(AB) - P(BA) \quad [8]$$

By subtracting the transition rates from the basic Levenshtein substitution cost of 2, the TRATE-based substitution cost considers easier substitutions between elements appear adjacent in observed sequences, compared with those that do not appear adjacent.

Although commonly used, the TRATE-based substitution cost is considered questionable due to its equating of high transition to high similarity (130). In the case of crash sequences, it is common for an event of a certain type followed by another event of a very different type. For example, a driver's maneuver (e.g., steering, braking) may be followed by a collision event by a high chance, and the TRATE between such maneuver and

collision events may be high. However, treating the maneuver and the collision event as similar events is contrary to common sense. Therefore, a TRATE-based substitution cost may not suit the analysis of crash sequences. Based on the work by Rousset et al., Studer and Ritschard proposed a substitution cost based on the probability of events sharing a common future (130, 156). The idea is that if the data shows a high probability of two events, A and B, sharing a same event, C, over a (user defined) position lag of q, then the A and B is considered highly similar and the substitution cost between A and B should be low. A mathematical expression of such a substitution cost is:

$$\gamma_{sf}(A, B) = \sum_{C \in \Sigma} \frac{(P(C_{+q}|A) - P(C_{+q}|B))^2}{\sum_{F \in \Sigma} P(C_{+q}|F)} \quad [9]$$

where Σ is the space of all possible elements, and $P(C_{+q}|F)$ is the probability of F followed by C over q positions. For crash sequences, this data-driven substitution cost based on shared future is more suitable compared with the TRATE-based substitution cost.

Localized optimal matching

Hollister developed a Localized OM (LOM) that calculates indel costs based on whether the element being inserted or deleted is the same with its adjacent elements (130, 137). An indel of an element different (or more dissimilar) to its adjacent elements is considered more difficult and should be of a higher cost. In OMloc, the indel cost of inserting an element U between elements A and B is calculated as:

$$c_l(U|A, B) = e\gamma_{\max} + g \frac{\gamma(A,U) + \gamma(B,U)}{2} \quad [10]$$

where γ indicates the substitution cost, with γ_{\max} being the maximum substitution cost, $\gamma(A,U)$ being the substitution cost between A and U, and $\gamma(B,U)$ between B and U. The

parameters, e and g , control the weights assigned to the two components of the function, 1) the maximum substitution cost and 2) the average of distances between the element to be inserted and its neighbors. The first component is a global cost for conducting an indel, and the second component is a local cost considering the difficulty of transition from adjacent actions/events/objects to the inserted one (or vice versa). The e and g parameters are set by researchers based on their domain knowledge and research needs. One of the goals of using OMloc is to avoid using a pair of indels in lieu of a substitution (violating the triangle inequality) (137). Therefore, the values of e and g should satisfy $2e + g \geq 1$.

3.6.4 Summary of dissimilarity measures

Properties

Nine dissimilarity measures were selected for a first round of comparison for compatibility with crash sequence analysis. The measures, cost schemes, and properties are summarized in Table 5. Cost schemes are specific to OM dissimilarity measures and were used in the case study. The summary of measure properties was based on findings from Studer and Ritschard (130, 142). Four properties are introduced here – metric, Euclidean, element dependency, and context.

Metric property.

The metric property means that the measured dissimilarities satisfy mathematical conditions of distances:

- Minimality: $d(x, y) \geq d(x, x) = 0$
- Non-negativity: $d(x, y) > 0$
- Symmetry: $d(x, y) = d(y, x)$

- Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

These mathematical conditions form a concept of distance in the physical world, such as in a road network where there is no one-way road and detours cannot be considered as shortcuts. The metric property is important also because it is required for many calculations and applications (130, 139). TRATE-based OM and LOM measures do not ensure the triangle inequality, but all the other seven dissimilarity measures in Table 5 satisfy all mathematical conditions of metrics. With triangle inequality not satisfied, there would possibly be some substitution cost replaceable by two other substitutions with a lower total cost, which makes a detour equivalent to a shortcut.

Euclidean property.

The Euclidean property is the prototypical example of metric property, and it ensures dissimilarity measures satisfy all the metric mathematical conditions. With the Euclidean property, data objects can be plotted in an Euclidean space, and the distances between them can be visualized using multidimensional scaling (130). Since the Euclidean property requires position-wise element matching, LCS and OM-based measures do not satisfy this property. Euclidean property is not necessary for use in crash sequence analysis, especially with sequences obtained from crash reports and crash databases, which cannot guarantee position-wise correspondence (i.e., the timing and duration of events are not given, only the order).

Element dependency property.

The element dependency property means that different elements in sequences are allowed to be treated differently when calculating dissimilarities (130). Rules of element

dependency are defined based on domain knowledge or using data-driven approaches. OM measures using variable substitution and/or indel costs are element dependent.

Context property.

The context property means that the relationships between elements (rather than only the individual elements) are considered when calculating sequence dissimilarities. The NMS, shared-future-based OM, and LOM measures consider context in sequences. TRATE-based OM does not have the context property because, although TRATE is used to estimate the ease of substitution, only individual elements are considered and compared in the OM process.

Table 5 Dissimilarity measures

Dissimilarity Measure	Cost Scheme	Properties			
		Metric	Euclidean	Element Dependent	Context
Euclidean	“EUCLID”	n/a	✓	✓	
Chi-square	“CHI2”	n/a	✓	✓	
Longest Common Subsequence	“LCS”	n/a	✓		
Number of Matching Subsequences	“NMS”	n/a	✓	✓	✓
Optimal Matching	“OMlev”	Levenshtein (substitution cost = 2, indel = 1)	✓		
	“OMtr”	TRATE-based substitution cost, indel = 1		✓	
	“OMsf”	Shared-future-based substitution cost, indel = 1	✓	✓	✓
Localized Optimal Matching	“LOMtr”	TRATE-based substitution cost, indel cost parameters: e = 0.0~0.4, g = 0.8~0.2		✓	✓
	“LOMsf”	Shared-future-based substitution cost, indel cost parameters: e = 0.0~0.4, g = 0.8~0.2	✓	✓	✓

Sensitivities

Intuitively, for the comparison of crash sequences, the most important attributes determining dissimilarities are the distinct elements and the order of elements (i.e., what happened and what happened first). Depending on the encodings, events sometimes repeat in crash sequences (e.g., hitting multiple fixed objects), so event frequency is another potential attribute contributing to the dissimilarity between crash sequences. Table 6 shows the sensitivity of dissimilarity measures to the distinct elements, order, and frequency attributes of sequences, based on Studer and Ritschard's findings (130, 142).

Table 6 Sensitivity of dissimilarity measures to sequence attributes

Dissimilarity Measure		Sensitive to Attributes		
		Distinct elements	Order	Frequency
Euclidean	"EUCLID"	H	L	H
Chi-square	"CHI2"	H	L	H
Longest Common Subsequence	"LCS"	H	M	M
Number of Matching Subsequences	"NMS"	H	M	M
Optimal Matching	"OMlev"	H	M	M
	"OMtr"	H	H	M
	"OMsf"	H	H	M
Localized Optimal Matching	"LOMtr"	H	H	M
	"LOMsf"	H	H	M

Note: "L" = Low sensitivity, "M" = Moderate sensitivity, "H" = High sensitivity

All measures listed in Table 6 are sensitive to distinct elements. OM measures based on TRATE, shared future, and LOM measures are the most sensitive to element order in sequences than the other measures. LCS, NMS, and OM with Levenshtein costs are moderately sensitive to element order. Euclidean and Chi-square measures are the most sensitive to element frequency as those measures are solely based on element frequencies.

LCS, OM measures with Levenshtein costs or TRATE-based costs, and LOM measures are moderately sensitive to element frequency.

Since the EUCLID and CHI2 measures are not sensitive to the order of elements, these two measures are not suitable for crash sequence analysis. Being not widely applicable, the EUCLID and CHI2 measures were not included in the case study. As LCS is equivalent to OM with Levenshtein II costs, its performance was proved to be very similar to OM dissimilarity measures (130). NMS has been shown to perform very differently to EUCLID, CHI2, LSC, and OM-based measures, with near-zero correlations (141). Also, NMS was found to generate results that were difficult to interpret due to its limited applicability to some specific cases (141). For example, the NMS dissimilarity between two sequences “ABC” and “FGH” (9.9) is smaller than that between “ABC” and “AFBC” (16.1), even though the latter pair shares some subsequences, and the former pair does not share any element. OM dissimilarity measures are widely applicable and easily interpretable. Therefore, the case study focused on comparing the five OM-based dissimilarity measures using the CRSS crash sequence data.

Sequence analysis in the case study was carried out using R (157). All OM-based dissimilarity measures listed in Table 5 were available in the R sequence analysis package, “TraMineR”, which was used as the primary tool for crash sequence analysis in this chapter (158). Other major R packages used in the analysis include “ade4” (for Mantel test), “WeightedCluster” (for weighted k-medoids clustering), and “mclust” (for adjusted Rand index calculation) (159–161).

3.7 Case Study: Single-Vehicle Crashes on Interstate Highways

This case study demonstrates the techniques and procedure for analyzing sequences of single-vehicle crashes on the United States' interstate highways. CRSS data files ACCIDENT (crash-level data), VEHICLE (vehicle-level data), and CEVENT (event-level data) from 2016-2018 were used to form a combined crash-level data set consisting of event sequences and all the needed variables. The three data files are linked using crash case identification numbers and vehicle numbers.

Single-vehicle crashes on interstate highways were filtered from the parent 2016-2018 CRSS data set, using conditions (variables and values) listed in Table 7. The resulting data set consists of 2,676 observations, representing a weighted total of 385,484 crashes during the 2016-2018 period.

Table 7 Conditions for obtaining data for case study from CRSS

Variable (Data Level)	Value	Description of Condition
VE_FORMS (Crash)	= 1	Only one vehicle-in-transport involved in crash.
INT_HWY (Crash)	= 1	Crash occurred on interstate highway.
PVH_INVL (Crash)	= 0	No parked/working vehicles involved.
WRK_ZONE (Crash)	= 0	No work zone at crash location.
ALCHL_IM (Crash)	≠ 1	No alcohol-related crash.
BDYTYP_IM (Vehicle)	< 50	Only automobile, utility vehicles or light trucks* involved.
TOW_VEH (Vehicle)	= 0	No vehicle trailing involved.
BUS_USE (Vehicle)	= 0	No bus involved.
SPEC_USE (Vehicle)	= 0	No special use vehicles involved.
EMER_USE (Vehicle)	= 0	No emergency use vehicles involved.

Note: * Light trucks with Gross Vehicle Weight Rating (GVWR) ≤ 10,000 LBS

CRSS provides pre-crash event data in the VEHICLE data file via variables PCRASH1 (pre-event movement), PCRASH2 (critical event pre-crash), and PCRASH3 (attempted

avoidance maneuver). The CEVENT data file provides a series of harmful and non-harmful events that occurred in the crashes, through the SOE (sequence of events) variable, ordered chronologically. Sequences analyzed in this case study were formed by combining the PCRASH1~3 variables and the SOE variable. Based on the CRSS definitions, PCRASH1~3 described “what a vehicle was doing just prior to the critical precrash event”, “what made the vehicle's situation critical”, and “what was the corrective action made, if any, to this critical situation”, happened before the vehicle’s SOE events. As single-vehicle crashes only had one vehicle-in-transport involved, the final sequence structure was formed as:

(PCRASH1 event) – (PCRASH2 event) – (PCRASH3 event) – (SOE events)

Lengths (number of elements) of the 2,676 crash sequences ranged from 4 to 12, as shown in Figure 16. The average length was 5.6.

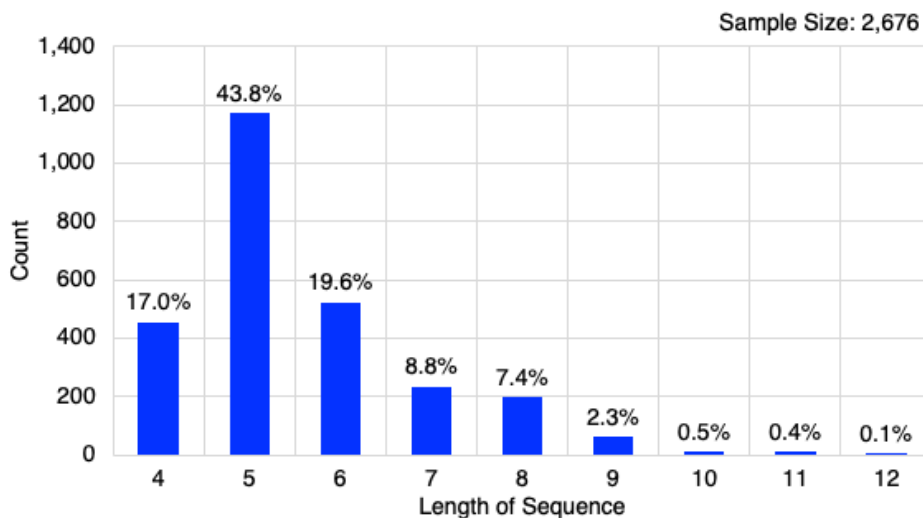


Figure 16 Sequence lengths

3.7.1 Encoding schemes

Encoding schemes serve the intended uses of crash sequence analysis. In this case study, three different encoding schemes were compared. In the single-vehicle interstate crash data set, the original CRSS encoding scheme (OE) had 123 distinct event types. Two other encoding schemes (named Encoding Schemes 1 and 2, or ES1 and ES2) were developed. The process of developing encoding schemes is illustrated in Figure 17. Events of similar nature were classified into higher-level categories (applicable to PCRASH2 and SOE in CRSS), and were consolidated into the same types. ES1 consisted of 59 event types and ES2 consisted of 30 types. Following the order of OE → ES1 → ES2, the level of details decreased, and the level of abstraction increased. As a result of applying different sequence encoding schemes on the single-vehicle interstate crash data set, the OE formed 1,535 distinct sequences, ES1 formed 1,105 distinct sequences, and ES2 formed 550 distinct sequences.

Details of the encoding schemes are presented in Table A - 1 in the Appendix. The example in Table 8 shows the logic of event consolidation. Roadside fixed object collision events were encoded in the three schemes, providing different levels of details about the collision events. In the original CRSS encoding scheme, there were 30 types of fixed object collisions (XF). In Scheme 1, the 30 fixed object collision types were grouped into three categories based on their propensities toward injury risks (ranked XFA, XFB, and XFC, from low to high, based on object hardness and potential contact area). In Scheme 2, the 3 fixed object collision types were all grouped into one category of collision with a roadside object (XF).

	PCRASH Events	+	SOE Events	
Consolidate ↓	Original Encoding Scheme PCRASH 1 <i>Movement before recognition of critical event</i> 16 Types PCRASH 2 <i>Pre-crash critical event</i> 40 Types PCRASH 3 <i>Attempted avoidance maneuver</i> 12 Types		NON-HARMFUL (NH) 9 Types NON-COLLISION HARMFUL (NCH) 8 Types COLLISION WITH NF OBJ (XO) 7 Types COLLISION WITH FIXED OBJ (XF) 30 Types UNKNOWN / NOT REPORTED (N) 1 Type	Total: 123
	Encoding Scheme 1 PCRASH 1, 2, & 3 <i>Scrambled and consolidated</i> 41 Types		NH: 8 NCH: 2 N: 1 XO: 4 XF: 3 Consolidated 18 Types	Total: 59
	Encoding Scheme 2 PCRASH 1, 2, & 3 <i>PCRASH 2 events further consolidated</i> 25 Types		NH, NCH, N, XO, XF <i>Use five categories</i> 5 Types	Total: 30

Figure 17 Process of developing encoding schemes

These three encoding schemes can be selected to meet the needs of the four use cases: 1) crash causation analysis, 2) test scenario generation, 3) crash injury severity modeling, and 4) crash frequency estimation. For crash causation analysis, we consider the OE to be the most appropriate encoding scheme, as it offers the most detailed information of pre-crash events and driver actions, which are essential for analyzing crash causes. For test scenario generation, OE and ES1 are both appropriate because they provide enough details of the pre-crash events for recreating pre-crash scenarios. For crash injury severity modeling, ES1 is considered to be more appropriate than the other two encoding schemes. In OE, there are 55 types of SOE events, with 30 different fixed object collision types. Too many types of SOE events would increase the difficulty for the sequence comparison process to identify patterns. Also, in the OE, the SOE events are not categorized based on

their potential risks in causing injuries. ES2 consolidated collision events into five types and is more suitable for estimating crash frequencies.

Table 8 Example of encoding schemes

Original	Scheme 1	Scheme 2	CRSS Description
1v23	XFB	XF	23 Bridge Rail (Includes Parapet)
1v24	XFA	XF	24 Guardrail Face
1v25	XFB	XF	25 Concrete Traffic Barrier
1v26	XFB	XF	26 Other Traffic Barrier
1v30	XFC	XF	30 Utility Pole/Light Support

3.7.2 Comparison of dissimilarity measures

As mentioned, different dissimilarity measures have different properties and sensitivities, to analyze a specific set of sequences, the most appropriate dissimilarity measure should be used. By comparing the dissimilarity measures, we can understand the measures' correlations and their performance in clustering the sequences. In this section, all five OM based dissimilarity measures were used on the three encoding schemes and compared. The comparison consisted of two parts:

- A Mantel test of correlations between dissimilarity matrices
- A comparison of agreement between clustering results to a benchmark crash typology using Adjusted Rand Index (ARI)

The purposes of these two analyses were to demonstrate how the most appropriate dissimilarity measures were selected for the clustering of interstate single-vehicle crash sequences.

Mantel tests

Mantel test is a correlation test for distance matrices, first developed by Nathan Mantel (141, 143, 162, 163). Using each dissimilarity measure, we obtained an n-by-n dissimilarity matrix of crash sequences, with n equals to the sample size of 2,676. Each dissimilarity matrix had $n(n - 1)/2$ elements. Five dissimilarity matrices were obtained from applying the five dissimilarity measures. For LOM measures, the e and g parameters generating the optimal Adjusted Rand Index (ARI) were used. ARI is introduced in detail in Section 7.2.2. The Mantel test was then used to calculate the correlation between each pair of dissimilarity matrices by calculating the correlation between the two sets of $n(n - 1)/2$ matrix elements. The correlation was calculated multiple times, with random permutations on one of the two matrices' columns and rows, to estimate a significance for the Mantel test result.

The Mantel test result ranges from -1 to 1, with a higher absolute value meaning a higher correlation between the two dissimilarity matrices. The results of Mantel correlation among the five OM dissimilarity measures, with three encoding schemes, are shown in Table 9. The results were also color coded with white being the average correlation, darker red meaning a higher above-average correlation, and darker blue meaning a lower below-average correlation.

Table 9 Mantel test results

(a) Original Encoding

	OMlev	OMtr	OMsf	LOMtr	LOMsf
OMlev		0.98	0.63	0.98	0.78
OMtr	0.98		0.66	0.98	0.81
OMsf	0.63	0.66		0.74	0.90
LOMtr	0.98	0.98	0.74		0.83
LOMsf	0.78	0.81	0.90	0.83	

(b) Encoding Scheme 1

	OMlev	OMtr	OMsf	LOMtr	LOMsf
OMlev		0.98	0.64	0.97	0.77
OMtr	0.98		0.65	0.98	0.78
OMsf	0.64	0.65		0.73	0.91
LOMtr	0.97	0.98	0.73		0.80
LOMsf	0.77	0.78	0.91	0.8	

(c) Encoding Scheme 2

	OMlev	OMtr	OMsf	LOMtr	LOMsf
OMlev		0.98	0.68	0.98	0.80
OMtr	0.98		0.70	0.98	0.81
OMsf	0.68	0.70		0.68	0.86
LOMtr	0.98	0.98	0.68		0.81
LOMsf	0.80	0.81	0.86	0.81	

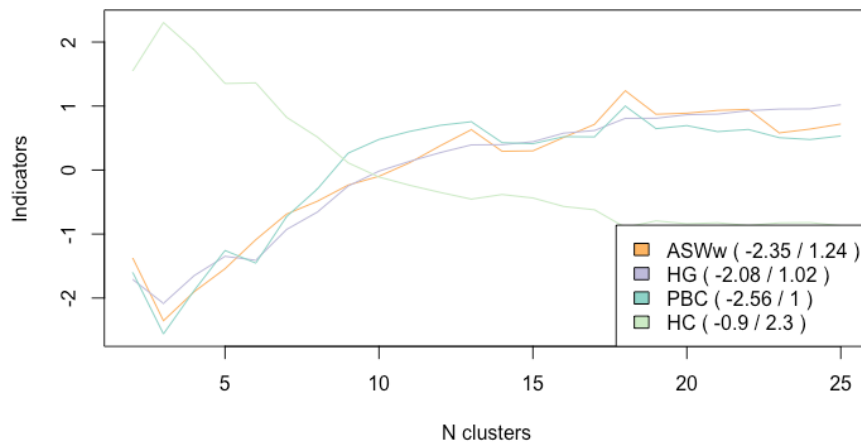
Mantel correlation matrices presented very similar patterns across the three encoding schemes. The results showed that OMlev, OMtr, and LOMtr generated highly positively correlated dissimilarity matrices. OMsf and LOMsf were highly correlated. LOMsf also had moderately high correlations with OMlev, OMtr, and LOMtr. Therefore, based on the Mantel correlations, the five dissimilarity measures could be categorized as the following two groups: Group 1: OMlev, OMtr, and LOMtr; Group 2: OMsf and LOMsf.

Adjusted Rand Index

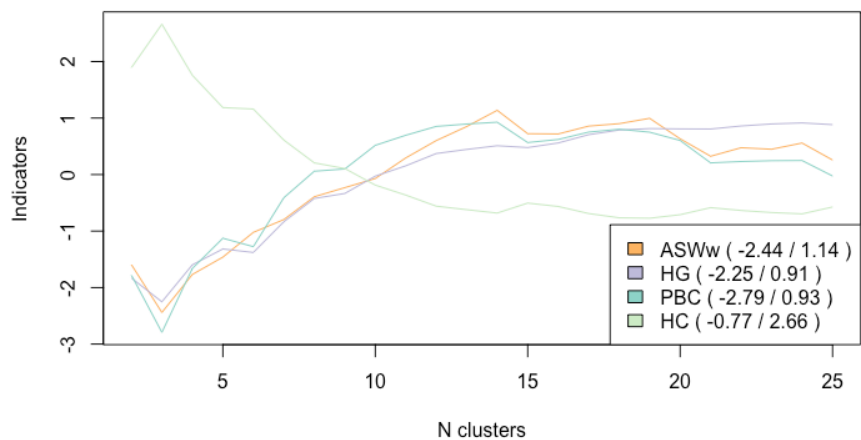
Clustering is commonly used to identify patterns and characterize sequences (64). The dissimilarity matrices are the basis of sequence clustering. For demonstration, the k-medoids clustering was used in this case study. K-medoids is a widely used technique for clustering with categorical data (such as sequences) (16). Because the CRSS crash cases were weighted to represent the population, a weighted k-medoids algorithm was applied to consider sequence weights (160).

To compare the performance of dissimilarity measures in clustering, identical clustering settings were used on all five OM dissimilarity matrices to generate clustering results, which were compared with the same benchmark. The CRSS crash typology (derived from the variable "ACC_TYPE") was used as a benchmark in this chapter. As defined by CRSS, the ACC_TYPE variable "identifies the attribute that best describes the type of crash this vehicle was involved in based on the First Harmful Event and the precrash circumstances". Although crash sequences conveyed more information than the First Harmful Event and precrash circumstances, the ACC_TYPE typology was considered feasible as a benchmark for dissimilarity measure comparison because this typology could partially reflect the sequence patterns.

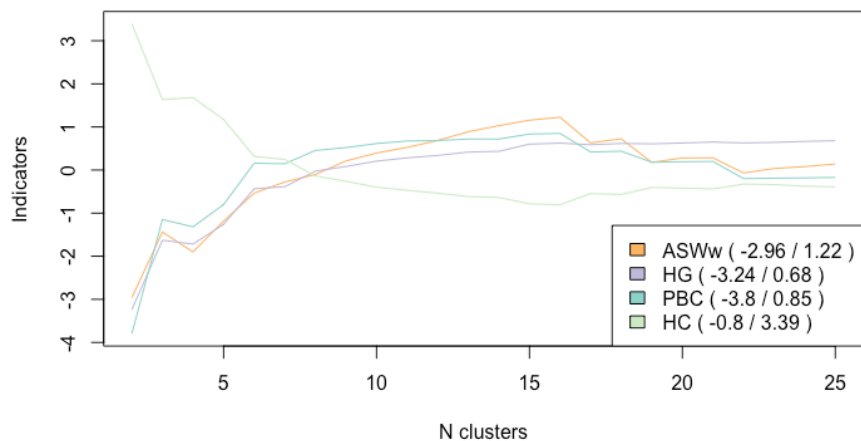
The ACC_TYPE was recoded into 15 types (see Table A - 2 and Table A - 3 in the Appendix for details). To match that number, the target cluster number, k, was also set to 15 for the weighted k-medoids clustering. To confirm that k=15 was an appropriate value, for each combination of encoding scheme and dissimilarity measure, a range of k values were tested with clustering quality indices plotted. An example of clustering quality measures for the OMlev measure under three encoding schemes are shown in Figure 18.



(a) Original Encoding



(b) Encoding Scheme 1



(c) Encoding Scheme 2

Figure 18 Clustering quality of the OMlev measure

Four clustering quality indices were used for evaluation, the Weighted Average Silhouette Width (ASWw), Hubert's Gamma (HG), Point Biserial Correlation (PBC), and Hubert's C (HC) (160). An optimal k value would generate maximum ASWw, HG, and PBC (all range from -1 to 1), and minimum HC (ranges from 0 to 1). The charts in Figure 18 present standardized values of clustering quality indices for easier identification of good k values, and k=15 was shown to ensure satisfactory clustering quality.

To measure the agreement between clustering results and the benchmark (CRSS crash type), the Adjusted Rand Index (ARI) was used (143, 164). With a set of clustering results, Y, and the crash type grouping benchmark, X, a contingency table could be written as:

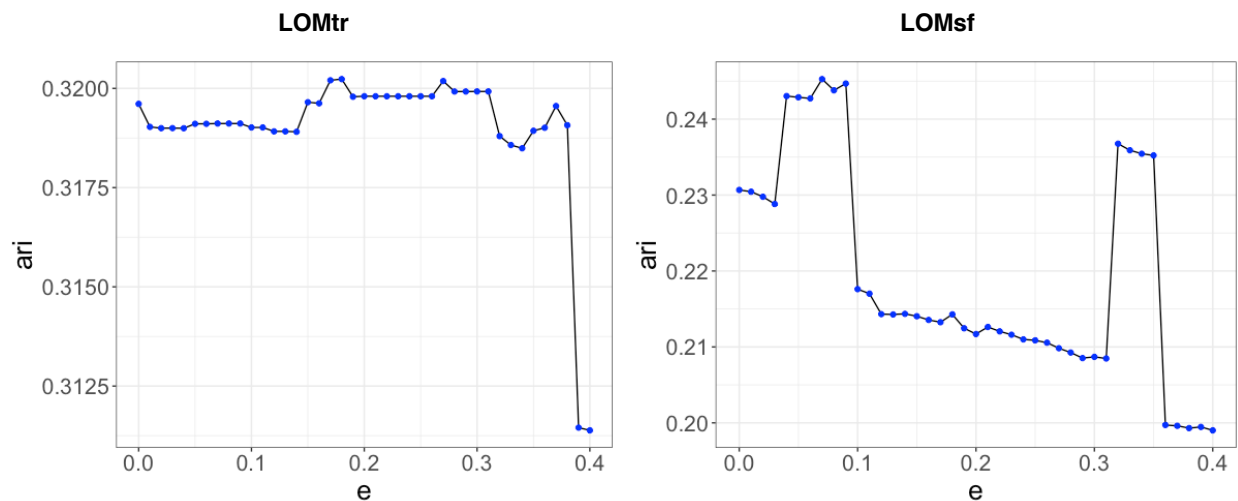
	Y ₁	Y ₂	...	Y _s	Sum
X ₁	n ₁₁	n ₁₂	...	n _{1s}	a ₁
X ₂	n ₂₁	n ₂₂	...	n _{2s}	a ₂
...
X _r	n _{r1}	n _{r2}	...	n _{rs}	a _r
Sum	b ₁	b ₂	...	b _s	

The ARI was then calculated as:

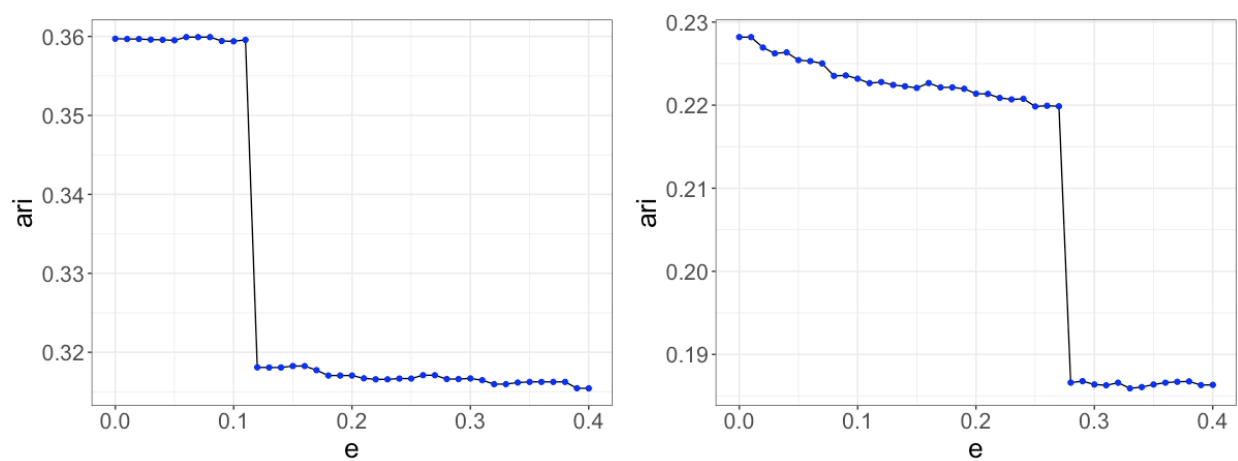
$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}} \quad [11]$$

where n_{ij} = the number of sequences assigned to both groups X_i and Y_j , with $1 \leq i \leq r$ and $1 \leq j \leq s$; $a_i = \sum_{j=1}^s n_{ij}$; and $b_j = \sum_{i=1}^r n_{ij}$. The ARI ranges from -1 to 1. An ARI of 0 means a random agreement and the two groupings can be treated as independent. An ARI of 1 means the two groupings are identical (143, 165).

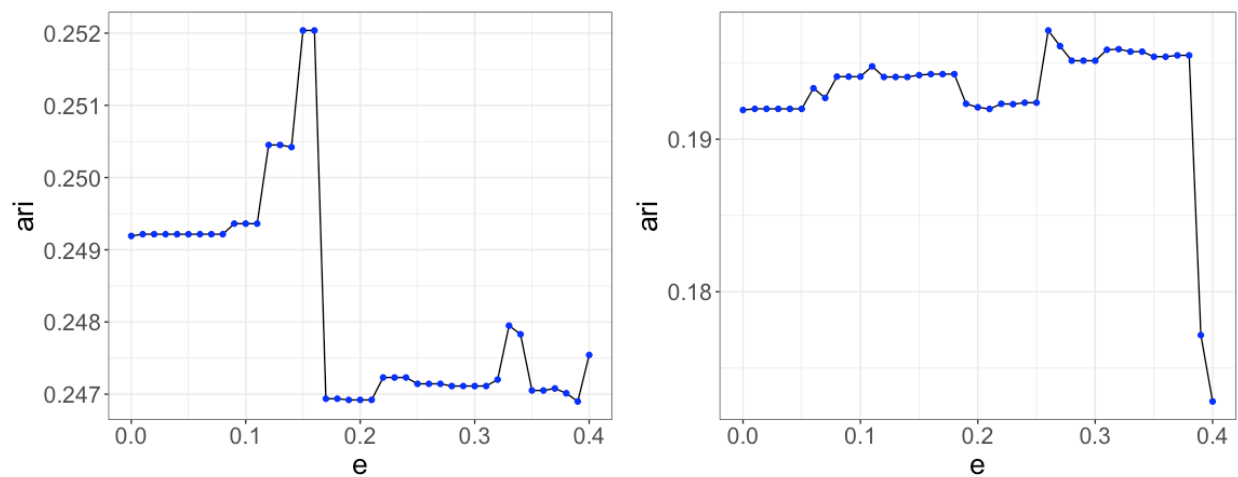
For LOM measures (LOMtr and LOMsf) with parameters to adjust indel costs (see Equation 10), a sensitivity analysis was conducted to identify changes in ARI with changes in indel cost parameters and find the parameter values for the optimal ARIs. The results of sensitivity analysis were plotted as shown in Figure 19. Indel cost parameter e (the weight on the maximum substitution cost) values ranging from 0 to 0.4 with an increment of 0.01 were tested. The corresponding parameter g (the weight on the average of substitution costs between inserted element and its adjacent elements) values were set based on the relationship $g = 1 - 2e$. Based on the plots, we found that in general, a very small e value (e.g., around 0.1) would lead to a good ARI value. However, a sensitivity test would be helpful to identify the optimal parameter settings for specific combination of sequence encoding scheme and dissimilarity measure.



(a) Original Encoding



(b) Encoding Scheme 1



(c) Encoding Scheme 2

Figure 19 ARI sensitivity to the LOM parameter e

Comparing all combinations of sequence encoding schemes and dissimilarity measures, ES1 and LOMtr measure obtained the largest ARI, as shown in Table 10. Under each encoding scheme, the best performing dissimilarity measures were LOMtr for the OE and ES1, and OMlev for ES2. We also observed that the simple OMlev measure performed generally well across all encoding schemes.

Table 10 Comparison of ARIs

Measure	ARI		
	Original Encoding	Encoding Scheme 1	Encoding Scheme 2
OMlev	0.313	0.355	0.265
OMtr	0.313	0.316	0.248
OMsf	0.193	0.171	0.184
LOMtr*	0.319	0.359	0.249
LOMsf*	0.218	0.223	0.194

Note: * Showing best results from sensitivity analysis.

In the study of single-vehicle roadside departure crash sequences by Wu et al., a Cohen's Kappa statistic was used to measure the agreement between clustering results and the original crash typology in FARS, and a 0.28 Kappa value was deemed satisfactory (15). The ARI is equivalent to Cohen's Kappa (166). Therefore, an ARI about 0.28 or larger was considered satisfactory in this case study. Regardless of the absolute values of ARI, for the purpose of comparison, emphasis was given to the relative values of ARI obtained using different dissimilarity measures under different encoding schemes.

Results from the Mantel tests and clustering performance evaluation together showed that highly correlated dissimilarity measures performed similarly in clustering. In this interstate single-vehicle case study, Group 1 dissimilarity measures performed better

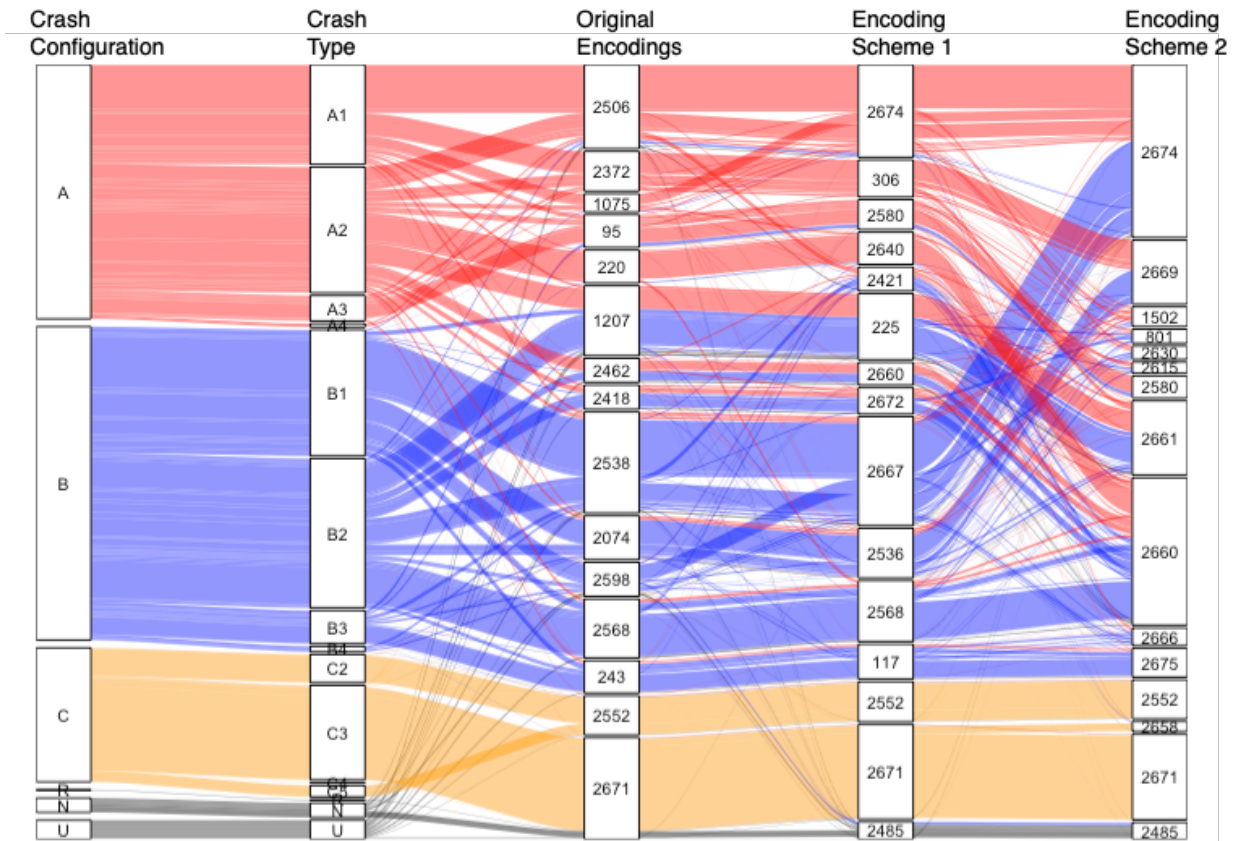
and yielded higher ARIs than Group 2 measures. In Group 1, OMlev is the simplest dissimilarity measure that needs the least effort (i.e., lowest computational cost) for sequence comparison and clustering. With similar performances, OMlev is considered preferable for this case study.

3.7.3 Sequence clustering results

Using the OMlev dissimilarity measure, a clustering on a sample of 2,676 interstate highway single-vehicle crash sequences was carried out, with three encoding schemes (the OE, ES1, and ES2). The sequences, under each encoding scheme, were divided into 15 clusters. To examine the characteristics of the clusters, compare the results from sequence clustering to the crash typologies provided by CRSS (see Table A - 2 and Table A - 3 in the Appendix for details), and understand what the clusters convey, the results were visualized using an alluvial diagram, as shown in Figure 20.

The diagram shows how the crashes were characterized by 1) the CRSS crash configuration (CC), 2) CRSS crash type (CT), 3) sequence clustering with original encodings (OE), 4) sequence clustering with Encoding Scheme 1 (ES1), and 5) sequence clustering with Encoding Scheme 2 (ES2). In the diagram, crashes were illustrated as alluvia and color-coded based on the CRSS crash configuration, the “flows” of the alluvia show how the crashes were re-grouped as we switch from one characterization method to another. For each characterization method, the grouping of crashes was shown as a stack of boxes, with the height of box indicating the proportion of that group in the total sample. To identify crash groups, in the CT and CC stacks, boxes were labeled with IDs of crash categories. In

the clustering generated stacks – OE, ES1, and ES2, boxes were labeled with IDs of cluster medoids.



Note: Labels in CC and CT indicate the categories. Please see the Appendix for details.

Numbers in OE, ES1, and ES2 indicate the IDs of cluster medoids (sequences numbered from #1 to #2676).

Figure 20 Alluvial diagram of clustering results with the OMlev measure

The alluvial diagram shows that the A type (classified as right roadside departure in CC) and the B type (classified as left roadside departure in CC) crashes were re-grouped extensively when sequence clustering was applied, and mixed into different clusters.

However, the C type (forward impact in CC) were not mixed with the other types as much in the sequence clustering process. The study by Wu et al. on roadside departure crashes explained the reason for such an observation (15). Roadside departure crashes in CRSS

were classified based on pre-crash conditions and the first harmful event, thus, many details in crash progression were not considered. For example, a crash classified as a right roadside departure may include events of a left roadside departure, returning to roadway, a right roadside departure, and hitting roadside objects. Sequences of events record such detailed stories of crashes, and the characterization based on sequence of events is expected to generate different results from the CRSS crash typology.

In addition to showing how the original CC and CT were regrouped, the alluvial diagram also shows how the different encoding schemes affect the clustering results. Regrouping of sequences happened from OE to ES1, and was more extensive from ES1 to ES2. To further present the differences, more detailed sequence clustering results with the three encoding schemes are shown in Table 11. Each sub-table shows the clusters, their medoids, and the percentages of crashes they represent. For each cluster, a description of the representative sequence is provided. A representative sequence is the dominant sequence in a cluster, making up the largest proportion. Representative sequences were extracted by sorting sequences in each cluster. Note that a representative sequence does not cover all sequences in a cluster but is the vast majority of them.

Table 11 Sequence clustering results with the OMlev measure**(a) Original Encoding**

Cluster #	%	Representative Sequence Description		
		Pre-crash Events	Avoidance Maneuver	Collision Events
2506	11	Moving Straight-RORR	Avoidance Unknown	Hit Fixed Object and/or Rollover
2372	5	Negotiating Curve-RORR	Avoidance Unknown	Hit Fixed Object and/or Rollover
1075	2	Moving Straight-RORR	No Avoidance	Hit Fixed Object and/or Rollover
95	4	Moving Straight-RORR	Steering R	Hit Fixed Object and/or Rollover
220	4	Moving Straight-Speeding-RORR	Avoidance Unknown	Hit Fixed Object and/or Rollover
1207	9	Negotiating Curve-Speeding-RORL	Avoidance Unknown	Hit Fixed Object and/or Rollover
2462	3	Moving Straight-Poor Surface-ROR	Avoidance Unknown	Hit Fixed Object and/or Rollover
2418	3	Moving Straight-Tire Issue-ROR	Avoidance Unknown	Hit Fixed Object and/or Rollover
2538	14	Moving Straight-RORL	Avoidance Unknown	Hit Fixed Object and/or Rollover
2074	6	Negotiating Curve-RORL	Avoidance Unknown	Hit Fixed Object and/or Rollover
2598	5	Moving Straight-RORL	No Avoidance	Hit Fixed Object and/or Rollover
2568	8	Moving Straight-Speeding-RORL	Avoidance Unknown	Hit Fixed Object and/or Rollover
243	4	Moving Straight-Other in Lane-ROR	Steering L	Hit Fixed Object and/or Rollover
2552	5	Moving Straight-Object in Lane	Avoidance Unknown	Hit Object
2671	14	Moving Straight-Animal/Ped/Bike in Lane	Avoidance Unknown	Hit Animal/Ped/Bike

Note: % of weighted sample total of 385,484 crashes.

RORR = run-off-road right; RORL = run-off-road left; ROR = run-off-road (right or left).

Steering R = steering right; Steering L = steering left.

(b) Encoding Scheme 1

Cluster #	%	Representative Sequence Description		
		Pre-crash Events	Avoidance Maneuver	Collision Events
2674	13	Moving Straight-RORR	Avoidance N	Hit Fixed Object and/or Rollover
306	5	Negotiating Curve-RORR	Avoidance N	Hit Fixed Object and/or Rollover
2580	4	Moving Straight-Other Vehicle in Lane	Steering R	Hit Fixed Object and/or Rollover
2640	4	Moving Straight-Speeding-RORR	Avoidance N	Hit Fixed Object and/or Rollover
2421	3	Moving Straight-Control Loss Other-ROR	Avoidance N	Hit Fixed Object and/or Rollover
225	9	Negotiating Curve-Speeding-ROR	Avoidance N	Hit Fixed Object and/or Rollover
2660	3	Moving Straight-Control Loss-ROR	Avoidance N	Hit Fixed Object and/or Rollover
2672	4	Moving Straight-Equip Failure-ROR	Avoidance N	Hit Fixed Object and/or Rollover
2667	15	Moving Straight-RORL	Avoidance N	Hit Fixed Object and/or Rollover
2536	7	Negotiating Curve-RORL	Avoidance N	Hit Fixed Object and/or Rollover
2568	8	Moving Straight-Speeding-RORL	Avoidance N	Hit Fixed Object and/or Rollover
117	5	Moving Straight-Speeding-Other Vehicle in Lane-RORL	Steering L	Hit Fixed Object and/or Rollover
2552	5	Moving Straight-Object in Lane	Avoidance N	Hit Object
2671	13	Moving Straight-Animal/Ped/Bike in Lane	Avoidance N	Hit Animal/Ped/Bike
2485	2	Moving Straight-Other/Unknown Event	Avoidance N	Hit Fixed Object and/or Rollover

Note: % of weighted sample total of 385,484 crashes; Avoidance N = avoidance unknown / no avoidance.

(c) Encoding Scheme 2

Cluster #	%	Representative Sequence Description		
		Pre-crash Events	Avoidance Maneuver	Collision Events
2674	23	Moving Straight-Event to This Vehicle	Avoidance N	Hit Fixed Object and/or NCH
2669	9	Negotiating Curve-Event to This Vehicle	Avoidance N	Hit Fixed Object and/or NCH
1502	3	Moving Straight/Negotiating Curve-Event to This Vehicle	Steering R	Hit Fixed Object and/or NCH
801	2	Moving Straight-Control Loss	Brake	Hit Fixed Object and/or NCH
2630	2	Changing Lanes/Merging-Event to This Vehicle	Avoidance N	Hit Fixed Object and/or NCH
2615	2	Changing Lanes/Merging-Control Loss	Avoidance N	Hit Fixed Object
2580	3	Moving Straight/Negotiating Curve-Vehicle/Animal/Ped in Lane	Steering R	Hit Fixed Object
2661	10	Negotiating Curve-Control Loss	Avoidance N	Hit Fixed Object and/or NCH
2660	20	Moving Straight-Control Loss	Avoidance N	Hit Fixed Object and/or NCH
2666	2	Moving Straight/Changing Lanes-Event to This Vehicle	Steering L	NH-Hit Fixed Object
2675	4	Moving Straight-Vehicle/Animal in Lane	Steering L	NH-Hit Fixed Object
2552	5	Moving Straight/Negotiating Curve-Object in Lane	Avoidance N	Hit Object
2658	1	Negotiating Curve-Animal/Ped/Bike in Lane	Avoidance N	Hit Animal/Ped/Bike
2671	12	Moving Straight-Animal/Ped/Bike in Lane	Avoidance N	Hit Animal/Ped/Bike
2485	2	Moving Straight-Other/Unknown Event	Avoidance N	NCH-Hit Fixed Object

Note: % of weighted sample total of 385,484 crashes.

NCH = Non-collision harmful event; NH = Non-harmful event; Avoidance N = avoidance unknown / no avoidance.

With the original CRSS encoding scheme, crash sequences were characterized based on detailed pre-crash events, generating sequence types that are suitable for crash causation analysis and scenario generation for AV and ADAS testing. With ES1, sequences were characterized based on consolidated event categories, keeping enough details in pre-crash events to differentiate actions and collision events to reflect potential injury risks. Sequence types generated with ES1 are suitable for scenario generation and crash injury severity modeling. With ES2, sequences were categorized based on further consolidated event categories, generating sequence types suitable for analyzing crash frequencies.

To further interpret the sequence clustering results and demonstrate the effectiveness of the sequence analysis methodology, the sequences assigned to each cluster

under OE was reviewed manually with details summarized as shown in Table 12. The 15 sequence clusters were further categorized into three types, 1) the run-off-road (ROR) crashes (13 clusters), 2) crashes with objects on the road (1 cluster), and 3) crashes with animals, non-motorized road users, or non-collision harmful events (1 cluster). For each type, sequence clusters were summarized in a sub-table of Table 12. In each sub-table, clusters are labeled with corresponding medoid IDs and simplified representative sequences. Each sub-table also shows the number of sequences in each cluster, major crash categories (based on interpretation of sequences), a mis-match ratio calculated as the percentage of sequences assigned to the clusters but did not belong to any of the major crash categories, and other information such as the percentage of crashes happened on curve roads and the percentage of crashes involving rollovers.

For each ROR crash sequence cluster, a mismatch ratio was calculated as the proportion of non-ROR crash sequences in the cluster. The sequence analysis was able to correctly differentiate the directions of ROR. For crash sequences with both ROR to the right and ROR to the left, the sequence analysis was able to sort out the order of events and assign sequences to correct clusters by putting more weight to the direction of the first happened ROR event. Sequences with rollover events occurred in all the ROR clusters, with different percentages. Twelve out of 13 ROR crash sequence clusters of the sequence clusters yielded a mismatch ratio in the range of 0% and 4.7%. The only cluster with a mismatch ratio of 15% was Cluster #2598. The mismatched sequences in Cluster #2598 were crashes with non-motorized road users or non-collision harmful events (e.g., fell out of vehicle, fire or explosion) where the driver did not make any avoidance maneuver.

Table 12 Detailed interpretation of sequence clustering results with OE**(a) Run-Off-Road (ROR) Crashes**

Cluster	Total	ROR Category				Mismatch (Non-ROR)		Sequences with Rollover	
		R	L	R+L	L+R	Count	%	Count	%
2506-ST-RORR	366	310	1	28	15	7	1.9	138	37.7
2372-C-RORR	135	114	1	12	6	2	1.5	23	17.0
1075-ST-RORR-NA	76	69	0	4	3	0	0.0	22	28.9
95-ST-RORR-SR	109	89	8	8	3	1	0.9	19	17.4
220-ST-SPD-RORR	111	107	0	1	2	1	0.9	35	31.5
2462-ST-PoorSurf-ROR	75	39	26	2	5	1	1.3	8	10.7
2418-ST-FlatTire-ROR	85	26	53	2	3	0	0.0	27	31.8
2538-ST-RORL	398	0	318	7	67	0	0.0	118	29.6
2074-C-RORL	169	1	143	0	25	0	0.0	44	26.0
1207-C-SPD-RORL	236	65	127	12	13	11	4.7	48	20.3
2598-ST-RORL-NA	133	0	102	2	7	20	15.0	20	15.0
2568-ST-SPD-RORL	204	7	178	2	12	2	1.0	30	14.7
243-ST-OtherInRd-SL	134	9	108	3	10	4	3.0	32	23.9

Note: ST = Moving straight; C = Negotiating curve; SPD = Speeding.

NA = No avoidance; SR = Steering right; SL = Steering left.

PoorSurf = Poor road surface; FlatTire = Flat tire; OtherInRd = Other vehicle encroached into road.

In ROR Category: R = Right; L = Left; R+L = First right then left; L+R = First left then right.

(b) Crashes with Objects

Cluster	Total	Category		Mismatch		Curve Road		With Rollover	
		Object	Object-Related	Count	%	Count	%	Count	%
2552-Object	112	106	5	1	0.9%	10	8.9	7	6.3

(c) Crashes with Animals/Non-motorized Road Users/Non-collision Harmful (NCH) Events

Cluster	Total	Category				Mismatch		Curve Road		
		Animal	Animal-Related	Pedestrian	Pedalcyclist	NCH	Count	%	Count	%
2671-Animal	333	263	2	31	15	16	6	1.8	19	5.7

For the two non-ROR crash clusters, Cluster #2552 (crashes with objects) and Cluster #2671 (crashes with animals or others), the mismatch ratios were calculated, respectively, as the proportion of non-object (or object-related) crash sequences and the proportion of crash sequences that did not fall into the categories of animal, animal-related, non-motorized road user, or non-collision harmful events. For both clusters, the mismatch

ratios were lower than 2%. Crashes with objects had 6.3% sequences with rollover events, but no rollover occurred in crashes with animals or others.

The detailed interpretation of sequence clustering results showed that the sequence analysis methodology was effective in identifying patterns in crash sequences, with low mismatch ratios. Also, with OE, the comparison and clustering of sequences were based more on crash causes rather than the outcomes, confirming that OE is more appropriate for the analysis of crash causations than the other two encoding schemes. For example, some crashes originally classified as ROR were caused by objects on the road or animal on the road, were assigned into clusters of object or animal-related crashes rather than ROR crashes. Also, by using OE to characterize crashes based on detailed pre-crash events and clearly identify crash causes, we can accurately recreate pre-crash scenarios for the testing of AVs and ADAS.

3.8 Conclusions

Crash sequence analysis supports crash characterization, which has importance in enhancing understanding of crashes and identifying safety countermeasures (15, 74, 77, 87, 88). Sequence analysis methods have recently been adapted to traffic crash study and proved to be effective in characterizing crashes and providing new insights (15, 16, 80). However, there are various techniques and details in adapting sequence analysis to traffic crash study, which have not been systematically presented or studied.

In this chapter, a methodology for crash sequence analysis was introduced. The methodology consists of steps including data processing, sequence encoding of crashes, dissimilarity measuring, and clustering. Crash sequence encoding and sequence

dissimilarity measures are emphasized in this chapter, because encodings are highly domain specific and dissimilarity measures are the foundations of sequence clustering. Through a case study using real-world crash sequence data, this chapter presented the effects of different encoding schemes and dissimilarity measures on crash characterization. Also, the procedure of selecting appropriate encoding schemes and dissimilarity measures for different use cases was demonstrated.

The findings from the case study suggest that the choice of encoding schemes and dissimilarity measures depends on the purpose of crash analysis. As a contribution, this chapter developed and demonstrated a methodology that is ready to be applied to select the most appropriate technique for future crash sequence studies. The findings also suggest that encoding schemes can be developed to be more or less abstract to accommodate different use cases of crash sequence analysis. Dissimilarity measures considered feasible for crash sequence analysis were introduced and compared. To select the most appropriate dissimilarity measures, the differences in measure properties and sensitivities should be considered in addition to clustering performances.

The findings were supported by a case study of single-vehicle crashes on United States' interstate highways. Three encoding schemes were used to match the needs of crash causation analysis, scenario generation for AV and ADAS safety testing, and crash frequency/injury severity modeling. The properties of nine dissimilarity measures were compared to select five OM measures for further comparison in the numerical analysis with a sample of 2,676 crashes (representing a weighted total of 385,484 crashes) from the 2016-2018 NHTSA CRSS database. The case study results showed that for the sequence analysis of interstate single-vehicle crashes, the OMlev (i.e., Levenshtein Distance) was the

preferable dissimilarity measure due to its overall good clustering performance and low computational cost. The original CRSS sequence encoding scheme was suitable for crash causation analysis and scenario generation for testing of AVs and ADAS, the more abstract Encoding Scheme 1 was suitable for crash scenario generation and crash injury severity modeling, and the most abstract Encoding Scheme 2 was suitable for analyzing collision frequencies. Detailed interpretation of sequence clustering results validated the effectiveness of the proposed sequence analysis methodology.

The purpose of this chapter was to develop and demonstrate a crash sequence analysis methodology that is applicable to various sources of crash sequence data. The range of dissimilarity measures compared in this chapter could also serve as a general list of candidates for crash sequence analysis. Considering that the case study was carried out on a specific subset of a national-level crash database, the specific Mantel test and clustering performance (i.e., ARI) results may not be necessarily generalizable. Apart from using an existing crash typology, clustering benchmarks can be obtained in multiple other ways. Two examples are 1) expert-opinion-based clustering and 2) crash classifications using criteria other than sequences (e.g., most harmful events, manner of collision, and injury severity) (167). Decisions on the choice of dissimilarity measures and evaluation benchmarks should be made according to data availability and specific analytical needs.

Chapter 4 Automated Vehicle Crash Sequences: Patterns and Potential Uses in Safety Testing

4.1 Introduction

Improving traffic safety is one of the primary motivations for developing automated vehicles (AVs). Apart from safer roads, AVs are predicted to bring other potential benefits such as improved mobility, better accessibility, lower energy consumption, and more efficient supply chains (18). Safety is prioritized as the top U.S. Government Automated Vehicle Technology Principle (18). Private and public organizations are taking efforts to ensure AV safety by extensively testing the vehicles on both closed courses and public roads (53, 54). In 2019, more than 1,400 automated vehicles were tested by more than 80 organizations across 36 U.S. states and Washington, D.C. (55).

Since 2014, the California Department of Motor Vehicles (DMV) has required all permit-holding organizations that test AVs on California public roads to submit AV collision reports (43). Prior to January 2020, two hundred and thirty-three (233) AV crashes were reported. In 168 of the reported cases, the AV was in automatic driving mode before disengagement or collision.

AV crash reports are a valuable data source for research to understand AV crash patterns. Prior explorations of California AV crashes provided some insights in:

- AV crash distribution by features such as manner of collision, AV-testing organization, year, and time of day
- Contributing factors of AV crashes and disengagements

- AV safety performance, measured by crash frequency per unit distance, compared with conventional (human-driven) vehicles

Crash information reported from AV field testing is useful in AV test scenario design. For example, Waymo is using data from field testing in developing challenging scenarios for closed-course and simulation-based AV testing (168).

AV collision reports provide much more information than what has been used in previous studies. Sequence of events, which can be extracted from the crash report narratives, consists of information of chronologically ordered events happened in the crash. Many analytical methods commonly used in studying genome sequences can be used to characterize crash sequences. Compared with summarizing crashes with manners of collision or contributing factors, a crash characterization based on sequence of events better captures the crash progression characteristics. Differences in events or actions, and the order of events or actions can lead to different crash outcomes (15, 70).

The primary objective of this chapter is to investigate the patterns in sequence of events in AV crashes. The secondary objective of this chapter is to discuss potential uses of crash sequences in scenario design for AV safety testing. Sequences of events were extracted from 168 AV collision reports' text narratives and analyzed using sequence analysis methods. Crash sequences, in combination with variables describing crash outcomes and variables describing the environment, can be used to design abstract semantic scenarios (11–13). A scenario-based AV testing framework, with crash sequence embedded as a core component, is proposed at the end of this study.

The contributions of this chapter are two-fold. This chapter adds to existing literature on California AV crashes and provides new insights by investigating AV crashes using sequence of events analysis. Beyond empirical findings, this chapter points out the practical application of crash sequences with a discussion on their potential uses in AV test scenario design.

4.2 Literature Review

Previous explorations of California AV crashes and disengagements started from 2015, when only limited data was available for aggregated statistical analyses. With increased AV road tests in recent years, more crashes and disengagements were reported to the California DMV. More recent studies were able to provide insights into the relationship among crashes, disengagements, and contributing factors by finding patterns in AV crashes and disengagements. Various analytical methods were used in previous studies. Some examples are statistical summary and tests, regression, classification trees, hierarchical Bayesian modeling, text mining, and clustering.

4.2.1 Patterns in AV Crashes

AVs crashes occurred mostly in the counties of Santa Clara and San Francisco, since the major AV testing organizations Waymo and Cruise carry out their AV testing in those two counties, respectively (169–171). AVs are tested on various types of roads including freeways/expressways, arterials, collectors, and local roads. AV crashes occurred on all roadway functional classes, with most crashes (60%) on arterial roads (171). Intersections are hotspots for AV crashes (48, 169, 171–173). Rear-end crashes were found to be the most common (60%) type of AV crashes (169–175). Most (60%-80%) AV crashes occurred

at a low relative speed between the AV and a second-party vehicle, usually below 10 mph (48, 169). The number of AV crashes is positively correlated with testing mileage, both periodically and cumulatively (169, 174).

Two studies used clustering techniques to group AV crashes (170, 172). Alambeigi et al. grouped 167 AV crashes based on themes in the description section of AV collision reports (172). Alambeigi et al. identified five themes:

- Driver-initiated transition crashes
- Sideswipe crashes during left-side overtaking
- Rear-end crashes with vehicle stopped at an intersection
- Rear-end crashes with vehicle in a turn lane
- Crashes with oncoming traffic

Das et al. grouped 151 AV crashes into six clusters, based on crash attributes provided directly by the collision reports (170). These clusters are:

- Two-vehicle non-injury crashes with unknown values in multiple attributes
- Single-vehicle non-injury crashes with unknown values in multiple attributes
- Injury crashes under poor lighting conditions during turning maneuvers or straight movement
- Single-case outlier cluster
- Two and multi-vehicle crashes with unknown values in multiple attributes
- Crashes with AV stopped and adverse weather conditions

4.2.2 Patterns in AV Disengagements

AV disengagements and test mileage data were aggregated and analyzed in previous studies. The cumulative disengagement number was positively correlated with cumulative test mileage (48, 176). Based on the mode of initiation, disengagements were classified into three types: automated, manual, and planned. Disengagement types were based on whether the disengagement was initiated by AVs, test operators, or as a part of a planned fault injection campaign (48, 169, 174). Before 2018, not all AV testing organizations reported disengagements with a clear differentiation between initiation modes. In a 2014-2016 sample of Waymo reported disengagements, about half were manually initiated, and the other half were automatically initiated (48). In the most recent study of California AV disengagements by Boggs et al., 25% of the sampled disengagements were initiated by human safety operators (177). The monthly automatic disengagement number was highly correlated with the monthly manual disengagement number, indicating test operators' trust in the AVs' capability to navigate through risks (174).

In terms of factors causing disengagement, the California DMV does not provide predefined categories, so different categorizations were used in previous studies to analyze the data (48, 173, 174, 176-178). Overall, system issues, including AVs' perception, planning, and decision-making, caused over half of the disengagements. In the Boggs et al. study, system issues were reported to have caused 89% of the disengagements, with a breakdown into control discrepancy (7%), hardware and software discrepancy (26%), perception discrepancy (21%), and planning discrepancy (35%) (177).

Disengagement reaction time was available in the 2015-2017 disengagement reports, and was another important information studied in past literature (48, 174, 179). The California DMV defined disengagement reaction time as "the period of time elapsed from when the autonomous vehicle test driver was alerted of the technology failure, and the driver assumed manual control of the vehicle" (48). The average disengagement reaction time was estimated to be 0.83 s – 0.87 s (48, 174, 179). The average disengagement reaction time is between the average automobile brake reaction time of 1.13 s and the average motorcycle brake reaction time of 0.60 s (179). Disengagement reaction time is expected to increase with test operators becoming more comfortable and gaining trust in the AV's handling of risky situations on roads (48, 179).

4.2.3 Relationship Among AV Crashes, Disengagements, and Contributing Factors

Many more disengagements than AV crashes have occurred with some disengagements followed by crashes. Banerjee et al. found in the 2014-2016 data that 23% of AV crashes involved disengagements, but a very small fraction (0.8%) of AV disengagements led to crashes (48). Favarò et al., using the 2014-2017 data, found that 1 in every 178 (0.5%) disengagements led to a crash (176).

Previous studies used various statistical modeling methods to evaluate the relationship between contributing factors and AV crash outcomes described by crash type, manner of collision, and severity (171, 175, 180). Leilabadi and Schmidt found that adverse road surface conditions were significantly associated with a higher severity of AV damage. Also, 80% of the crashes when AV was in automated driving mode were identified as hit-and-run crashes, in which 40% were rear-end and 40% were sideswipe crashes. Wang and

Li explored the mechanism of contributing factors affecting AV crash severity and manner of collision, and found that:

- Injury crashes (study included cases from outside California) happened when an AV was in automated driving mode and was responsible for the crashes
- All intersection crashes were rear-end, while roadway segment crashes with AVs in automated driving mode were angle or sideswipe

Boggs et al. focused on the factors affecting rear-end AV crashes and injury AV crashes, and found that:

- AVs in automated driving mode were more likely to get involved in rear-end crashes (without specifying if AVs were rear-ended or AVs rear-ended others) than AVs in manual driving mode or were disengaged from automatic driving mode
- In a mixed land use environment, AV crashes were more likely to be rear-end
- Higher speeds of second-party vehicles, no marked centerline, and non-clear weather were more likely to be associated with injury AV crashes

Two studies evaluated the relationship between various contributing factors and features of AV disengagements (173, 177). Wang and Li explored factors leading to disengagements in different stages of an AV's operation (perception, planning, and control phases) and disengagements with different take-over time (divided into two groups with a threshold of 0.5 s). Major findings of Wang and Li were:

- For an AV, a larger number (> 5) of radar sensors and a more appropriate number (3-4) of LiDAR sensors would lead to fewer disengagements

- Disengagements on local roads and freeways were associated with a shorter take-over time

Boggs et al. focused on factors affecting the mode of disengagement initiation (manual or automatic), and found that:

- Planning discrepancy, software/hardware issues, and environmental/other road user issues all significantly increased the probability of an automatic initiation to different extents
- More automatic initiations occurred as time progressed month by month

4.2.4 Safety Performance of AVs Compared with Conventional Vehicles

AV crash records were compared with conventional vehicle (human-driven vehicle) crash records, by evaluating crash rates (in crashes per mile driven), injury rates (injuries per mile driven), crash types, and crash severity (45–48). Conventional vehicle crash records were obtained from databases such as the National Highway Traffic Safety Administration (NHTSA) Fatality Analysis Reporting System (FARS), National Automotive Sampling System-General Estimates System (NASS-GES), and the Federal Highway Administration (FHWA) Strategic Highway Research Program Naturalistic Driving Study (SHRP2 NDS) database. With a small sample size of AV crashes (fewer than 20 cases), comparative studies could not reach an agreement nor a definite conclusion on whether AVs perform better than conventional vehicles in terms of safety (45–47). In a more recent study by Banerjee et al. California AV collision reports from 2014-2016 (42 crashes) and NHTSA 2015 motor vehicle crash data were used for an AV-human driver performance comparison. Banerjee et al. claimed that current AVs are 15 to 4,000 times worse than

human drivers in terms of safety performance measured by crashes per cumulative mile driven (48).

4.2.5 Sequence of Events in Traffic Crashes

The Tri-Level Study of the Causes of Traffic Accidents found that 50% of the 2,000 crashes studied were caused by more than one factor (76). Sequence of events is important information for traffic crash investigation, and was recommended by the National Transportation Safety Board (NTSB) to be included in national crash databases (77). Crash progression patterns can be discovered through crash sequence analysis and are helpful in identifying effective prevention strategies (70). Sequence analysis was developed in bioinformatics to analyze genome sequences, and is also applied in social sciences (64, 66). Genome sequence analysis methods is applicable to sequence of events to study traffic crash patterns (15, 64). Wu et al. used sequence analysis on Fatality Analysis Reporting System (FARS) data to group similar crashes and model crash severity outcomes (15, 80). A similar type of crash sequence data can be extracted from the California AV collision reports and analyzed using sequence analysis methods. Through sequence analysis, frequent pre-collision events can be identified, the stochastic relationships between events can be evaluated, and whole sequences can be classified into types that represent distinctive crash progression characteristics. Sequence pattern information, together with other crash attributes and environmental (both man-made and natural) condition variables, can be used in designing representative AV test scenarios.

4.3 Data

4.3.1 AV Crashes

AV collision reports from 2014 to 2019 were obtained from the California DMV. “Report of Traffic Accident Involving an Autonomous Vehicle (OL 316)” was the required form for AV testing organizations to submit. The form was updated in 2017, adding information such as weather, lighting, and road surface conditions. All AV collision reports were archived by the California DMV and are publicly available online. A total of 233 reports were gathered and manually reviewed, with key information transferred into a spreadsheet. The 168 reports of crashes where the AV was in automatic driving mode before disengagement or collision were used for analysis in this study. Table 13 presents a summary of several data elements of those reports.

AV crashes in California increased each year during 2015-2019. Data show that 43% of AV crashes occurred in 2019, due to the fact that 43% (2.58 million) of AV testing miles were driven in 2019 during the five-year period. Over 76% of AV crashes happened during the months of May-November. Also, 71% of AV crashes happened during daytime (7:00 am – 6:00 pm). Ninety-five percent of the AV crashes took place in San Francisco, Mountain View, and Palo Alto, where most AV testing was carried out in California.

Most (98%) of the 168 AV crashes occurred between AVs and other road users, and 2% were single-vehicle crashes. AV crashes involving bicyclists, e-scooters, pedestrians, and skateboarders were 8% of all crashes. AV crashes involving motorcycles were 5% of all crashes. AV crashes mostly (73%) took place at intersections (including ramp terminals),

followed by roadway segment (26%) or in a parking lot (2%). Intersections where AV crashes occurred were primarily signal controlled, as stated in 51% of AV collision reports.

Table 13 Summary of data from California AV collision reports

Field	Count	Percentage	Field	Count	Percentage
Year			Month		
2015	9	5%	1	5	3%
2016	12	7%	2	10	6%
2017	24	14%	3	9	5%
2018	50	30%	4	8	5%
2019	73	43%	5	16	10%
City			6	19	11%
Fremont	1	1%	7	19	11%
Los Altos	5	3%	8	19	11%
Milpitas	1	1%	9	14	8%
Mountain View	42	25%	10	25	15%
Palo Alto	17	10%	11	17	10%
San Francisco	100	60%	12	7	4%
Sunnyvale	2	1%	Time of Day		
AV Testing Organization			Day	119	71%
Apple	1	1%	n/a	6	4%
Cruise	95	57%	Night	43	26%
Google Auto (Waymo)	22	13%	Second Party Type		
Jingchi (WeRide)	1	1%	Bike	6	4%
Lyft	1	1%	Bus	3	2%
Pony.AI	2	1%	Car	139	83%
UATC (Uber)	1	1%	E-scooter	3	2%
Waymo	41	24%	Motorcycle	8	5%
Zoox	4	2%	n/a	4	2%
Facility Type			Pedestrian	2	1%
Road Segment	43	26%	Skateboarder	1	1%
Intersection	122	73%	Truck	1	1%
Parking lot	3	2%	Van	1	1%

(Table 13 continued)

Field	Count	Percentage	Field	Count	Percentage
Traffic Control			AV Mode		
AWSC	11	7%	Automatic	127	76%
Crosswalk Sign	1	1%	Automatic-Manual	41	24%
n/a	52	31%	AV Yielding		
Signal	86	51%	No	109	65%
Stop Sign	10	6%	Yes	59	35%
TWSC	1	1%	Severity		
Xwalk Flashing Light	1	1%	Injury	20	12%
Yield Sign	6	4%	Non-Injury	148	88%
Manner of Collision			Second Party Movement		
Broadside	12	7%	Backing	3	2%
Hit Object	3	2%	Changing Lanes	16	10%
Other	3	2%	Entering Traffic	2	1%
Rear End	112	67%	Making Left Turn	10	6%
Sideswipe	37	22%	Making Right Turn	15	9%
Vehicle/Pedestrian	1	1%	Merging	5	3%
AV Movement			n/a	6	4%
Changing Lanes	8	5%	Other	2	1%
Making Left Turn	11	7%	Other Unsafe Turning	2	1%
Making Right Turn	8	5%	Parked	1	1%
Merging	1	1%	Passing Other Vehicle	8	5%
Passing Other Vehicle	1	1%	Proceeding Straight	92	55%
Proceeding Straight	53	32%	Slowing/Stopping	6	4%
Slowing/Stopping	17	10%			
Stopped	69	41%			

Rear-end (62%) and sideswipe (21%) were the two most common manners of collision in AV crashes. Injuries, without differentiating between minor or serious, were reported in 12% of the crashes. Disengagements were reported in 24% of the crashes. In 35% of the cases, AVs were yielding to another road user before a collision took place. In more than half of the crash cases, AVs were stopped (41%) or slowing down (10%). AVs

were proceeding straight in 32% and turning in 12% of the crashes. Second-party road users were proceeding straight in 55% and turning in 15% of the crashes.

Of the 168 AV crashes, 37% and 57% were reported by Waymo and Cruise, respectively. During 2015-2019, Waymo and Cruise ran the most (4.1 million) and second most (1.4 million) public-road AV testing miles (in automatic driving mode), respectively. Waymo and Cruise's AV mileages accounted for 67% and 23% of the total mileage (6.2 million) reported by all AV-testing organizations that reported AV crashes in 2015-2019. Table 14 lists AV test mileages and crash rates (per million miles) by organization, sorted by mileage in ascending order.

Table 14 2015-2019 AV test mileages and crash rates by organization

Organization	Test Mileage	Mileage Share	Crashes	Crashes per Million Miles
Waymo	4,122,878	68.6%	63	15
Cruise	1,420,360	23.6%	95	67
Pony.AI	192,642	3.2%	2	10
Zoox	100,023	1.7%	4	40
Apple	88,283	1.5%	1	11
Lyft	42,931	0.7%	1	23
UATC (Uber)	26,899	0.4%	1	37
Jingchi (WeRide)	19,067	0.3%	1	52
<i>All</i>	<i>6,013,083</i>	<i>100.0%</i>	<i>168</i>	<i>28</i>

Waymo and Cruise yielded crash rates of 15 and 67 crashes per million miles tested. Comparatively, the 2018 passenger car crash rate in United States was approximately 4.4 crashes per million vehicle miles traveled (27). Note that the AV crash rates are not reflective of automatic driving systems' true safety performance, as there were still human operator interventions involved in the studied AV crashes. Without complete information from testing automatic driving systems without human intervention, we cannot make a true comparison

between automatic driving systems and human drivers. However, the information from currently available AV crash reports is to some extent useful in measuring the evolution of AV technology.

4.3.2 AV Disengagements in Crashes

AV disengagement reports from 2015 to 2019 were obtained from California DMV. “Annual Report of Autonomous Vehicle Disengagement (OL 311R)” was the required form for AV-testing organizations to submit. The form was updated in 2017, before which no uniform format for disengagement reporting existed. Disengagement reports provide information such as a summary of AV test mileages, number of disengagements, dates, locations (highway or street), and a description of disengagement causes.

Information from the disengagement reports were matched to the AV crash records which involved disengagements. A breakdown of different causes for AV disengagements followed by a crash is shown in Table 15. Categories of causes were created based on descriptions provided by different AV testing organizations’ reports, as uniform terms for describing disengagement causes were not provided and are not required by the California DMV. Not all disengagements in AV collision reports were in AV disengagement reports. For such cases, causes were summarized based on the collision report descriptions.

Table 15 Causes of disengagements involved in AV crashes

Disengagement Cause	Count	Percentage
Operator precaution	19	46%
Reckless road user	16	39%
Unwanted movement	3	7%
Planned	2	5%
Operator error	1	2%
<i>Total</i>	<i>41</i>	<i>100%</i>

Of the 41 disengagements in AV crashes, 19 (46%) were initiated by an operator out of precaution, and 16 (39%) were a reaction to a nearby reckless road user. The rest of the disengagements were initiated due to unwanted AV movements (3, 7%), operator error (1, 2%), or were planned tasks for tests (2, 5%). None of the 41 disengagements were caused by vehicle system (perception, hardware or software) problems. It could not be determined whether or not all reckless-road-user-caused disengagements were initiated by an operator or an AV itself, as the information was not clearly stated in all AV disengagement and collision reports.

4.3.3 AV Crash Sequences

An AV crash sequence consists of events such as actions or collisions. A collision can happen between an AV and an object, or between an AV and another road user. In some crash cases, multiple road users were involved. For crash sequences used in this study, AV was denoted as “v1”; a second-party road user (vehicle, bicyclist, pedestrian, or others) that collided with the AVs, was denoted as “v2”; and a third-party road user that interacted (but may or may not have collided) with the AVs or second-party road users, was denoted as “v3”.

Events were extracted from text narratives in the AV collision reports, which were reviewed and summarized manually. AV crash sequence lengths ranged from 2 to 5 events, with an average of 2.8. Events were first recorded using short phrases such as “v1 stop” and “v2 pass v1 on right”. Ordering of events in sequences was based on temporal information provided by the text. When recording events, consistency was maintained in the use of short phrases. Each short phrase went through a second round of review and

was encoded with a label, which was a combination of English alphabet letters and/or Arabic numerals. To further enhance consistency, phrases describing similar events (based on our understanding of traffic crashes and judgement) were encoded with the same label. For example, “v2 run stop sign” and “v2 run red light” were encoded with the same label, “V2”, since both phrases describe a second-party road user’s violation of traffic control at an intersection. Following this procedure of “text narratives → short phrases → labels”, we converged to a set of 35 labels for the encoding of 497 events, which made up the 168 AV crash sequences. Of the 35 different labels, 14 denoted actions initiated by AVs; 14 denoted actions initiated by second-party road users/objects; and 7 denoted actions initiated by third-party road users. Detailed encodings are listed in Table 16.

Table 16 Event encodings

Label	Short Phrase	Count	Label	Short Phrase	Count
A1	v1 accelerate/proceed	39	PL1	v1 pass v3 on left	1
A2	v2 accelerate/proceed	2	PL2	v2 pass v1 on left	13
B1	v1 back up	1	PL3	v3 pass v1 on left	1
B2	v2 back up	1	PR2	v2 pass v1 on right	7
D1	v1 decelerate	32	PR3	v3 pass v1 on right	1
D2	v2 decelerate	1	R1	v1 make right turn	9
D3	v3 decelerate	1	R2	v2 make right turn	2
DG	v1 disengage	41	S1	v1 stop	76
DT	v1 detect v2	2	S2	v2 stop	2
L1	v1 make left turn	11	SA2	v2 stop and proceed	2
L2	v2 make left turn	9	V2	v2 run stop sign/red light	5
L3	v3 make left turn	1	X12	v1 contact v2	7
ML1	v1 merge left	6	X1O	v1 hit object	3
ML2	v2 merge left	16	X21	v2 contact v1	155
ML3	v3 merge left	6	X32	v3 contact v2	1
MR1	v1 merge right	7	XO1	object/person contact v1	3
MR2	v2 merge right	12	Y	v1 yield	15
MR3	v3 merge right	6			

Table 17 gives an example of two crash event sequences. A sequence consists of one or more elements, each of which represents a pre-collision or collision event. A subsequence is a set of chronologically ordered elements (following the order in sequence but not necessarily adjacent) that appears in a larger sequence (64). A subsequence that consists of consecutive elements is called a substring, or an n-gram, with n referring to the number of elements in the subsequence. For example, Sequence 1 and Sequence 2 in Table 17 both have a subsequence “S1-X21”, with two elements, “S1” and “X21”. Sequence 1 and Sequence 2 both have a substring, or 2-gram, “PR2-X21”, with two elements, “PR2” and “X21”. Elements, subsequences, and whole sequences were all analyzed in this study to understand patterns in AV crashes.

Table 17 Example of crash event sequences

Sequence	Element 1	Element 2	Element 3	Element 4
Sequence 1	S1	PR2	X21	
Sequence 2	S1	A1	PR2	X21

4.4 Methodology

As mentioned, the primary objective of this chapter was to identify AV crash sequence patterns. Sequences were analyzed at three levels: the element level, subsequence level, and whole sequence level. Element-level analysis investigated the basic components of the sequence and the components’ weight in the entire element space. Subsequence-level analysis investigated the stochastic relationships between elements. Whole-sequence-level analysis investigated the similarities and dissimilarities between sequences, which were used to identify groups or classes of sequences. In the context of a traffic crash sequence study, we were interested in identifying frequently occurring events,

quantifying the interconnections between events, as well as classifying crash progressions. In this study, AV crash sequence analysis focused on identifying patterns from these three aspects: 1) most frequent events and event transitions; 2) disengagements' role in AV crash sequences; and 3) characterization of AV crash sequences.

Descriptive analysis was used to summarize frequencies of events and subsequences in AV crash sequences. Stochastic patterns in event transitions were presented by a transition matrix. To characterize crash sequences, a cluster analysis was carried out.

Following the sequence analysis, a discussion is presented at the end of this chapter, about potential uses of AV crash sequences in scenario-based AV safety testing. In the discussion, a cross-tabulation analysis was used between crash sequence groups, other AV crash attributes, and environmental condition variables. In this section, concepts and methods used in AV crash sequence analysis are introduced in detail.

4.4.1 Transition Matrix

A transition matrix shows the probability of transition between every pair of adjacent positions in all sequences (64). The size of a transition matrix is $k \times k$, where k is the number of elements in the element universe. The rows of a transition matrix indicate elements where transitions are from, and the columns indicate elements where transitions are to. A transition matrix, denoted as P , has the form shown in Figure 21. Cell P_{AB} contains the probability, $p(AB)$, that element A is followed by element B in all cases that element A appears in the element universe, which is calculated as:

$$p(AB) = p(B_p|A_{p-1}) = \frac{n(AB)}{n(A)} \quad [12]$$

where $n(AB)$ = number of times that 2-gram AB appears; and

$n(A)$ = number of times that element A appears.

		Element at position p		
		A	B	C
Element at position $p-1$	A	$p(AA)$	$p(AB)$	$p(AC)$
	B	$p(BA)$	$p(BB)$	$p(BC)$
	C	$p(CA)$	$p(CB)$	$p(CC)$

Figure 21 Form of transition matrix P (64)

The sum of probabilities in each row is 1. Note that $p(AB)$ does not capture both the probability $p(B_p|A_{p-1})$ (the conditional probability that B appears given that A has just appeared) and $p(A_{p-1}|B_p)$ (the conditional probability that A appeared just before given that B appears) (64). Transition matrix P does not show $p(A_{p-1}|B_p)$, which should be calculated as:

$$p(A_{p-1}|B_p) = \frac{n(AB)}{n(B)} \quad [13]$$

where $n(B)$ = number of times that element B appears.

4.4.2 Measuring Sequence Dissimilarity and Optimal Matching

To compare and group AV crashes based on sequences, the dissimilarity (or distance) between sequences needs to be measured. A common approach for sequence comparison is optimal matching (OM), which is widely used in genome sequence and social sequence analysis (64, 133).

When comparing two sequences, the distance between them is defined by the “cost” to transform a sequence to the other. This transformation is called “alignment”, and “cost”

is measured by the number of different operations needed to complete the alignment.

There are multiple ways to align two sequences. For example, Table 18 shows two of the multiple ways that can be taken to align the two sequences in Table 17. The operations applied include insertion, deletion, and substitution. The cost of insertion or deletion is denoted by “d” and is called “indel” cost. The cost of substitution is denoted by “s”. There are other more intricate operations for sequence alignment, but indels and substitutions are commonly used and have been previously used for traffic crash sequence analysis (15). The sum of cost for sequence alignment is a measure of dissimilarity (or distance) between two sequences. In our Table 18 example, the first alignment method uses both indels and substitutions and costs $2s+d$, while the second alignment method only uses indels and costs d . Depending on the selection of operations and setting of operation costs, there are different sequence distance metrics (64). Three basic and commonly used ones are:

- Levenshtein distance that uses both indels and substitutions
- Levenshtein II distance that uses only indels
- Hamming distance that only considers substitutions

Operation costs can also be set based on the needs of analysis and the properties of sequences. In previous analysis of traffic crash sequences, the Levenshtein distance was used as the measure of dissimilarity (15). In this analysis of AV crash sequences also, the Levenshtein distance was used.

Table 18 Example of ways to align two sequences

Sequence 1	S1	PR2	X21		
Sequence 2	S1	A1	PR2	X21	
Alignment 1					
Sequence 1	S1	<u>PR2</u>	<u>X21</u>	∅	
Sequence 2	S1	A1	PR2	X21	
<i>Cost</i>	0	s	s	d	= 2s+d
Alignment 2					
Sequence 1	S1		PR2	X21	
Sequence 2	S1	A1	PR2	X21	
<i>Cost</i>	0	d	0	0	= d

Note: Insertion is marked with ∅;
 Deletion is marked with ~~strikethrough~~; and
 Substitution is marked with underline.

As there can be multiple ways of aligning a pair of sequences which generate different distance values, the alignment that generates the smallest distance value should be found, and the smallest distance should be used as the measure of dissimilarity between that pair of sequences (15, 64). An OM procedure finds the dissimilarity between every pair of sequences in a sequence space. The Needleman-Wunsch algorithm is a classic OM algorithm which is widely used in bioinformatics to align sequences and find sequence dissimilarities (129). For two sequences, A and B, an empty matrix L, of size $\text{length}(A)+1$ by $\text{length}(B)+1$ is created. Based on a set of indel and substitution costs (e.g., for Levenshtein distance, indel and substitution costs are both 1), the Needleman-Wunsch algorithm fills matrix L and returns the smallest alignment cost (distance) between sequences A and B. Pseudocode of the Needleman-Wunsch algorithm is as follows (15, 129).

Algorithm Needleman-Wunsch(A, B)

```

# initialize
L <- matrix of size length(A)+1 * length(B)+1
d <- indel cost
s <- substitution cost
# fill the cells of L
for i = 0 to length(A)
  L(i,0) <- d*i
for j = 0 to length(B)
  L(0, j) <- d*j
for i = 1 to length(A)
  for j = 1 to length(B) {
    insert <- L(i, j-1) + d
    delete <- L(i-1, j) + d
    substitute <- L(i-1, j-1) + s
    L(i, j) <- max(insert, delete, substitute)
  }
# smallest alignment cost (distance)
return L(length(A), length(B))

```

The dissimilarity between every pair of sequences in the studied sample was calculated using the Needleman-Wunsch algorithm. A dissimilarity matrix was formed and used in cluster analysis as the basis for clustering similar sequences.

4.4.3 Cluster Analysis

There are various cluster analysis methods and different ones can produce different clustering results. The k-medoids method was selected for this sequence clustering because the k-medoids method works well with categorical data (such as sequences) and is robust against outliers (11). Most commonly, the k-medoids method is implemented by the partitioning around medoids (PAM) algorithm developed by Kaufmann and Rousseeuw (181). The PAM algorithm asks for the number of demanded clusters, k ($k \leq$ sample size), and greedily finds k points from the sample set (denoted as X) as medoids (denoted as M) to form clusters. In this analysis, X is in the form of a dissimilarity matrix. The objective of the algorithm is to minimize a cost measured by the sum of distances between each x ($\in X$)

to its assigned cluster medoid $m \in M$). Pseudocode of PAM algorithm is as follows (11, 181, 182).

Algorithm PAM(X, k)

```

# build
choose k points  $M \subset X$ 
for all  $x \in X$ 
    assign  $x$  to  $X_i$  if  $x$  is closest to  $m_i$ 
calculate cost
# swap
repeat
for all  $m \in M$ 
    for all  $x \in X$  and  $x \notin M$  {
        swap  $x$  with  $m$ ; calculate cost
        if (cost decreases) keep  $x$  and  $m$ 
        else do not swap
    }
until (cost does not change)

```

The PAM algorithm was applied to the 168 AV crash sequences, with the k value ranging from 2 to 10. A measure for evaluating quality of clustering used for this analysis is called the “silhouette”, which describes how well a data point lies within its own cluster compared to other clusters (183). Silhouette width is calculated as (183):

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad [14]$$

where s_i = silhouette width;

a_i = average dissimilarity of object i to all other objects of A (the cluster that i is assigned to); and

$b_i = \min_{C \neq A} d_{i,C}$, with $d_{i,C}$ = average dissimilarity of i to all objects of C (any cluster that i is not assigned to).

Silhouette width is between -1 and 1, with a higher value meaning a better clustering.

When there is only one object in a cluster, the object's silhouette width is 0.

The average silhouette widths from sequence clustering with different k values are plotted in Figure 22(a). The most appropriate k values were selected based on an evaluation of both overall and cluster-wise average silhouette widths. The overall average silhouette width should be as close to 1 as possible, with each cluster's silhouette width being larger than 0.1. Also, k should be preferably small, for easy cluster interpretation and to avoid "overfitting". Changes in cluster size and cluster average silhouette width are shown in Table 19. When $k = 7$, we could obtain a relatively small number of clusters, with a large enough average silhouette width and better cluster average silhouette widths than obtained using other k values. Therefore, $k = 7$ was used. Detailed silhouette plot for clustering with $k = 7$ is shown in Figure 22(b).

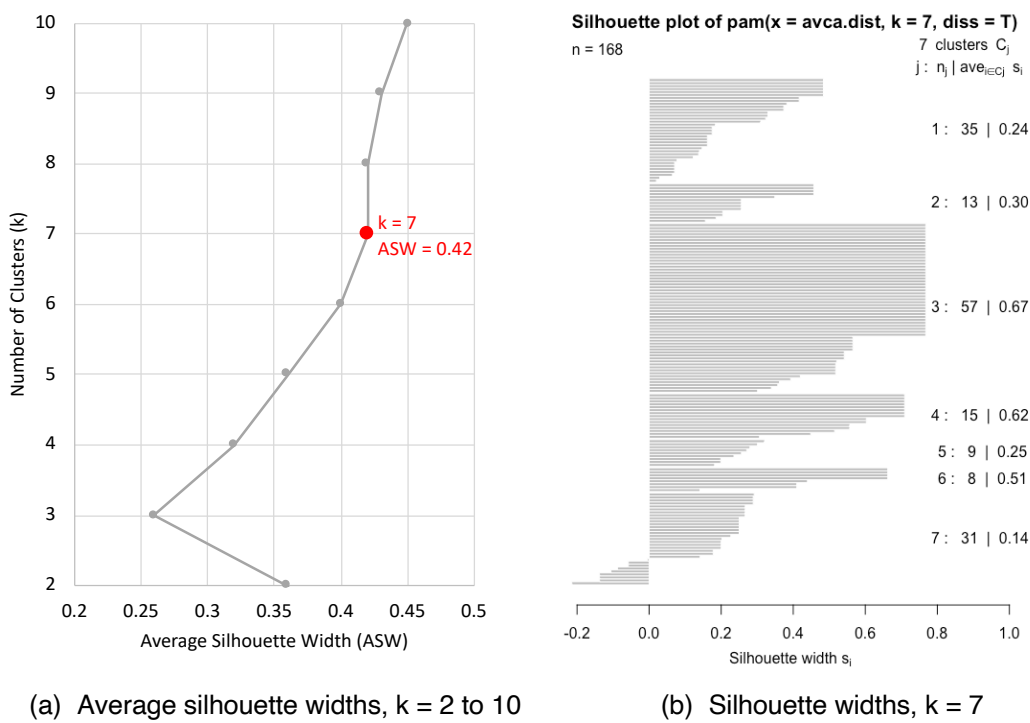


Figure 22 Silhouette widths

Table 19 Cluster size and cluster average silhouette width

Number of Clusters: “k”	Cluster Size		Cluster Avg. Silhouette Width		Avg. Silhouette Width
	Min	Max	Min	Max	
2	51	117	0.14	0.45	0.36
3	48	63	-0.05	0.74	0.26
4	15	57	-0.03	0.74	0.32
5	8	57	0.02	0.73	0.36
6	8	57	0.04	0.62	0.4
7	8	57	0.14	0.67	0.42
8	7	55	0.11	0.7	0.42
9	7	55	0.16	0.69	0.43
10	5	53	0.11	0.73	0.45

4.5 Results

4.5.1 Most Frequent Subsequences

To grasp overall patterns in crash sequences, the 15 most representative subsequences, as listed in Table 20, were investigated. The results showed that 92% of AV crash sequences ended with AV hit by a second-party road user. In 40% of the crash sequences, the AV stopped and was hit by a second-party road user. Disengagement was an event in 24% of the AV crash sequences. AV hit by a second-party road user following disengagement appeared in 19% of the crash sequences. Colliding right after the AV started moving (21%) or the AV started decelerating (19%) are two other common subsequences, which indicates crash cases where AVs were possibly violating expectancies of other road users. Other top-15 subsequences include AVs' yielding and second-party road users' merging actions.

Table 20 Top 15 most frequent subsequences

Rank	Subsequence	Description	Count	%
1	(X21)	(AV hit by 2 nd party)	155	92%
2	(S1)	(AV stops)	71	42%
3	(S1)-(X21)	(AV stops) then (AV hit by 2 nd party)	68	40%
4	(DG)	(AV disengaged)	41	24%
5	(A1)	(AV accelerates/proceeds)	38	23%
6	(A1)-(X21)	(AV accelerates/proceeds) then (AV hit by 2 nd party)	36	21%
7	(D1)	(AV decelerates)	32	19%
8	(D1)-(X21)	(AV decelerates) then (AV hit by 2 nd party)	32	19%
9	(DG)-(X21)	(AV disengaged) then (AV hit by 2 nd party)	32	19%
10	(D1)-(DG)	(AV decelerates) then (AV disengaged)	16	10%
11	(D1)-(DG)-(X21)	(AV decelerates) then (AV disengaged) then (AV hit by 2 nd party)	16	10%
12	(ML2)	(2 nd party merges left)	16	10%
13	(ML2)-(X21)	(2 nd party merges left) then (AV hit by 2 nd party)	15	9%
14	(Y)	(AV yields)	15	9%
15	(Y)-(X21)	(AV yields) then (AV hit by 2 nd party)	15	9%

4.5.2 Transitions to and from Disengagement

A 35 by 35 transition matrix was obtained. Since transitions to and from disengagements (DG) were of the most interest, relevant results are illustrated in Figure 23. Note that transition rates in the left column do not add up to 100%, but the ones in the right column add up to 100%, because of the reason explained previously in the Transition Matrix part of the Methodology section. Figure 23 helps identify the preceding and succeeding events of AV disengagements. Disengagements were initiated generally in two types of situations:

- AVs or human operators detected reckless actions of second or third-party road users
- Human operator felt uncomfortable with some driving maneuvers made by AVs

Examples of possible reactions to reckless actions are that 100% of “second-party road user decelerates” (D2) events were followed by AV disengagements (DG), and 50% of “third-party road user merging right” (MR3) events were followed by AV disengagements (DG). Possibly out of operators’ discomfort with AV’s actions, 44% of “AV deceleration” (D1) events and 43% of “AV merging right” (MR1) events were followed by disengagements.

While none of the studied disengagements were able to help avoid collisions (as these disengagements were all in crash sequences), 68% of them were followed by an immediate collision rather than being followed by certain other actions before collisions. Immediately after 51% of disengagement events, second-party road users hit the AVs. Following 10% and 7% of disengagement events, the AVs hit second-party vehicles or objects, respectively. In the other 32% of cases, there was still time for AVs or second-party road users to take some actions before the collision.

	AV disengaged		
	[-> DG]	[DG ->]	
AV decelerates [D1 ->]	44%	51%	[-> X21] 2 nd party hits AV
2 nd party decelerates [D2 ->]	100%	10%	[-> X12] AV hits 2 nd party
2 nd party makes left turn [ML2 ->]	31%	7%	[-> X10] AV hits objects
3 rd party makes left turn [ML3 ->]	33%	7%	[-> MR1] AV merges right
AV merges right [MR1 ->]	43%	5%	[-> D1] AV decelerates
2 nd party merges right [MR2 ->]	17%	5%	[-> ML2] 2 nd party merges left
3 rd party merges right [MR3 ->]	50%	5%	[-> V2] 2 nd party runs stop sign/red light
2 nd party passes AV from left [PL2 ->]	15%	2%	[-> B1] AV backs up
AV stops [S1 ->]	1%	2%	[-> L2] 2 nd party makes left turn
2 nd party runs stop sign/red light [V2 ->]	20%	2%	[-> ML1] AV merges left
AV yields [Y ->]	27%	2%	[-> MR2] 2 nd party merges right

Figure 23 Transition rates from preceding events to disengagement and from disengagement to succeeding events

4.5.3 Sequence Characterization

Cluster analysis resulted in crash sequences being clustered into 7 groups, as shown in Table 21. Patterns in crash sequences within each group and differences between groups were easily identified. Characteristics of each sequence group were summarized as follows.

- Group 1 as “Disengage-Deceleration”, with a representative subsequence of “D1-DG” (AV deceleration followed by disengagement)
- Group 2 as “Hesitation”, with a representative subsequence of “S1-A1-S1” (AV stops, proceeds, and stops again)
- Group 3 as “Stop”, with a representative subsequence of “S1-X21” (AV hit by a second party after it stops)
- Group 4 as “Yield”, with a representative subsequence of “Y-X21” (AV hit by a second party after it yields to the second party or a third party)
- Group 5 as “Hit Others”, with representative subsequences of “DG-X10” and “DG-X12” (AV disengagement followed by AV hitting a second party)
- Group 6 as “Left Turn”, with a representative subsequence of “L1-X21” (AV was hit while making left turn)
- Group 7 as “Moving-Unexpected”, with a representative subsequence of “A1-X21” (AV was hit while proceeding/accelerating)

Table 21 Clusters of AV crash sequences

"Disengage-Deceleration"		"Hesitation"		"Stop"	
Group 1	Count	Group 2	Count	Group 3	Count
D1-DG-X21	6	S1-A1-S1-X21	4	S1-X21	38
D1-X21	4	S1-A1-X21	4	S1-PR2-X21	5
ML2-DG-X21	3	S1-A1-D1-X21	1	S1-ML2-X21	3
DT-D1-DG-V2-X21	2	S1-A1-S1-A1-X21	1	D1-S1-L2-X21	1
ML3-D1-DG-X21	2	S1-A1-S2-A2-X21	1	D1-S1-X21	1
MR3-D1-X21	2	S1-A1-SA2-X32-X21	1	PL1-PL2-S1-X21	1
D1-DG-L2-X21	1	S1-S2-A1-A2-X21	1	R1-S1-XO1	1
D1-ML2-X21	1	<i>Total</i>	<i>13 (8%)</i>	S1-DG-X21	1
D1-MR1-DG-X21	1			S1-L1-X21	1
D1-PL2-DG-X21	1			S1-ML1-X12	1
D1-PL2-X21	1			S1-PL2-X21	1
DG-B1-B2-SA2-X21	1			S1-R1-X21	1
L1-D1-DG-X21	1			S1-R2-X21	1
ML1-D1-DG-ML2-X21	1			S1-XO1	1
ML1-MR3-DG-MR1-X21	1			<i>Total</i>	<i>57 (34%)</i>
ML3-D1-X21	1				
ML3-DG-X21	1				
MR2-DG-D1-X21	1				
MR3-DG-MR1-ML2-X21	1				
PL2-MR2-DG-X21	1				
PL3-D1-X21	1				
V2-D1-DG-X21	1				
<i>Total</i>	<i>35 (21%)</i>				

(Table 21 continued)

"Yield"		"Hit Others"		"Moving-Unexpected"	
Group 4	Count	Group 5	Count	Group 7	Count
Y-X21	8	DG-X1O	2	A1-X21	6
R1-Y-X21	2	A1-V2-DG-X12	1	A1-MR2-X21	3
Y-DG-X21	2	ML1-MR1-DG-X12	1	R1-X21	3
Y-DG-D1-MR2-X21	1	ML2-DG-X12	1	A1-L2-X21	2
Y-DG-ML2-X21	1	ML3-DG-ML1-X12	1	A1-ML2-X21	2
Y-PL2-X21	1	MR2-MR1-DG-X1O	1	A1-PL2-X21	2
<i>Total</i>	<i>15 (9%)</i>	MR3-DG-MR1-X12	1	R1-ML2-X21	2
		PL2-MR2-D2-DG-X12	1	A1-ML2-DG-MR2-X21	1
		<i>Total</i>	<i>9 (5%)</i>	A1-PL2-DG-X21	1
				A1-PL2-MR2-X21	1
				A1-PR2-X21	1
				A1-PR3-ML3-S1-X21	1
		"Left Turn"		A1-R2-X21	1
		Group 6	Count	A1-V2-X21	1
		L1-L2-X21	4	A1-XO1	1
		L1-X21	2	L1-A1-X21	1
		L1-L2-PL2-MR2-X21	1	ML1-D3-MR1-PR2-X21	1
		L1-L3-MR3-D1-X21	1	PL2-MR2-X21	1
		<i>Total</i>	<i>8 (5%)</i>	<i>Total</i>	<i>31 (18%)</i>

Comparing the sizes, Group 1 consists of 35 sequences (21% of all 168 sequences), Group 2 has 13 (8%), Group 3 has 57 (34%), Group 4 has 15 (9%), Group 5 has 9 (5%), Group 6 has 8 (5%), and Group 7 has 31 (18%) sequences. Disengagements appeared concentratedly in Groups 1, 4, and 5. In Group 1's 35 crash sequences, disengagement appeared in 25 sequences. All disengagements following AV's yielding action were clustered in Group 4. All Group 5 sequences consisted of a disengagement event before AV colliding into an object or a second-party road user. In terms of other types of actions or maneuvers, stopping was mostly seen in Groups 2 and 3; yielding was seen in Group 4; merging and passing actions were mostly seen in Groups 1, 5, and 7; and left turning action

was mostly seen in Group 6. Graph illustrations of the seven sequence patterns are shown in Figure 24. There were multiple different types of second-party road users involved in the 168 crash sequences, but to compactly present the sequence patterns, a motor vehicle was used in the illustrations to represent all types of second-party road users. For each sequence pattern, three panels of figures (from left to right) were used to illustrate the chronologically ordered events.

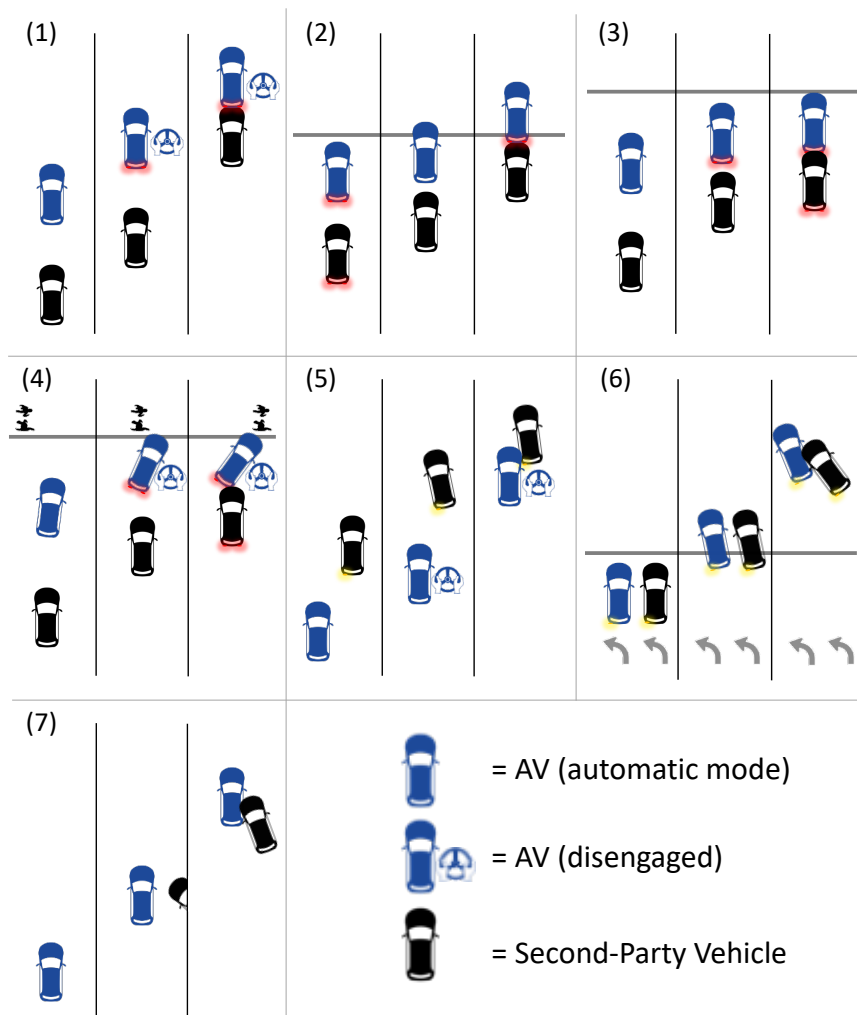


Figure 24 Graph illustrations of AV crash sequence patterns

4.5.4 Cross-tabulation between Sequence Group and Other Variables

A crash event sequence presents the progression of scenes with interactions between an AV and its surrounding moving objects (10). In addition to moving object dynamics provided by crash sequences, multiple variables such as weather, lighting, road surface, road geometries, traffic control, and traffic conditions, need to be considered in developing test scenarios for AVs (11, 12). Cross-tabulation analysis and Chi-square tests were carried out between sequence groups and several other variables that describe crash outcomes and environmental conditions. The purpose of cross-tabulation analysis is to evaluate the association between sequence groups and those other variables.

Figure 25 and Figure 26 illustrate results from a cross-tabulation between sequence group and two crash outcome measures, crash severity and manner of collision, respectively. The results showed that some Group 1, 2, 3, 6, and 7 sequences led to injuries. Comparing the distribution of crash severity in sequence groups, we found that Group 1 had the highest proportion (20%) of crash sequences that ended with injuries. Groups 6 and 7 both had 13% of crash sequences that ended with injuries. Group 3 had 12% of crash sequences that ended with injuries. The Chi-square test result (Chi-squared = 5.69, p-value = 0.47) showed that there is no significant association between sequence group and crash severity. However, after regrouping the sequence groups, with Groups 1, 2, 3, 6, and 7 in a new group, and Groups 4 and 5 in another new group, the Chi-square test result (Chi-squared = 3.78, p-value = 0.08) showed that there is a more significant association between new sequence groups and crash severity.

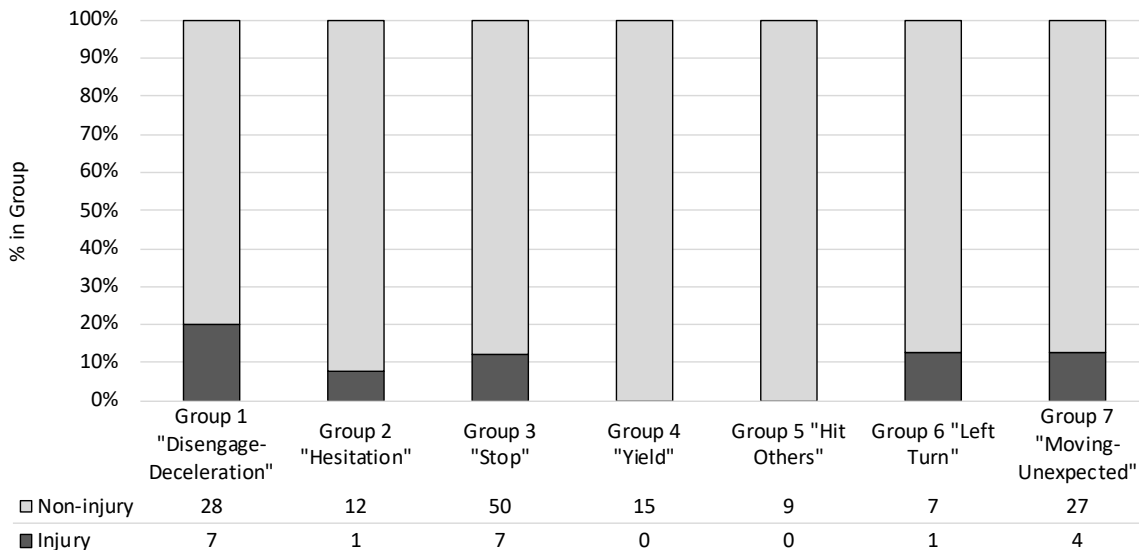


Figure 25 Crash severity distribution by sequence group

Manners of collision distributed differently across different sequence groups. Rear end, sideswipe, and broadside are the three most frequent manners of collision. Sequences in Groups 1, 3, 5, and 7 led to a larger variety of manners of collision, with 4-5 types in each group. A Chi-square test between manners of collision and sequence groups showed a significant association (Chi-squared = 83.73, p-value = 0.00). Different sequence groups led to different compositions of collision manners.

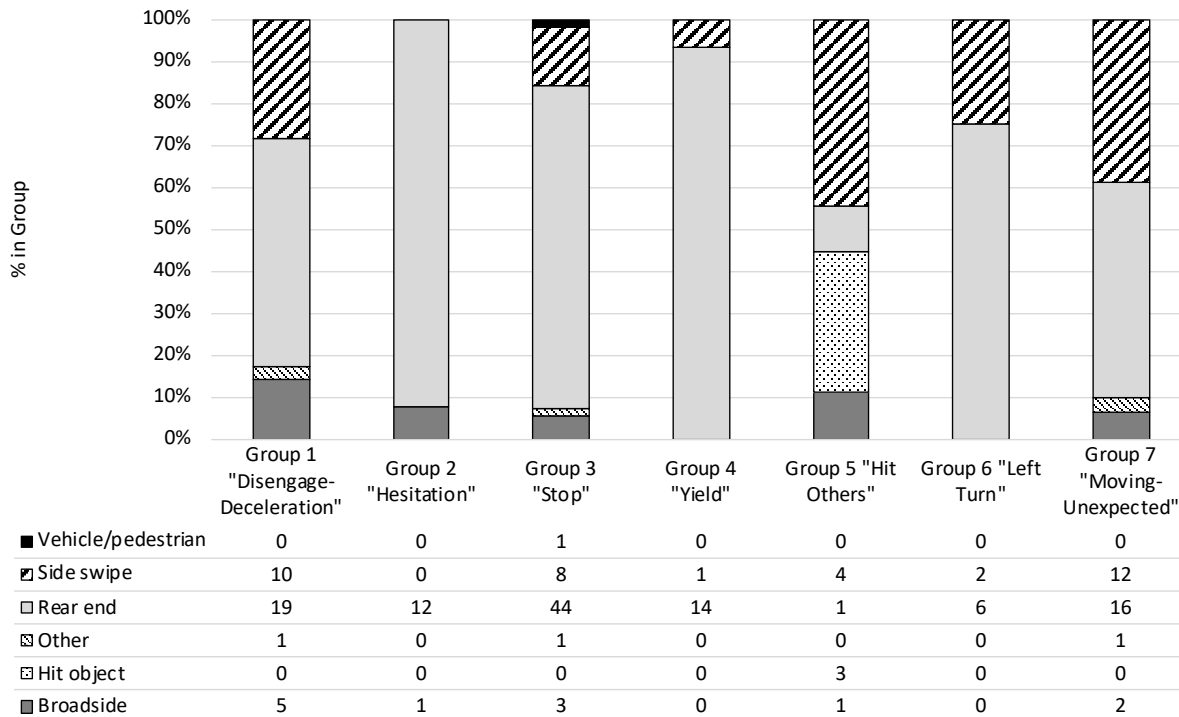


Figure 26 Manner of collision distribution by sequence group

Figure 27 and Figure 28 illustrate results from cross-tabulations between crash sequence groups and variables describing environmental conditions including facility type and time of day. Crash sequence groups distributed differently across facility types. The intersections (including ramp terminals) category had the largest variety of sequence groups. Groups 2 and 6 sequences only took place at intersections. The Chi-square test result (Chi-squared = 43.86, p-value = 0.00) confirmed that there is a significant difference in sequence group distribution across facility type.

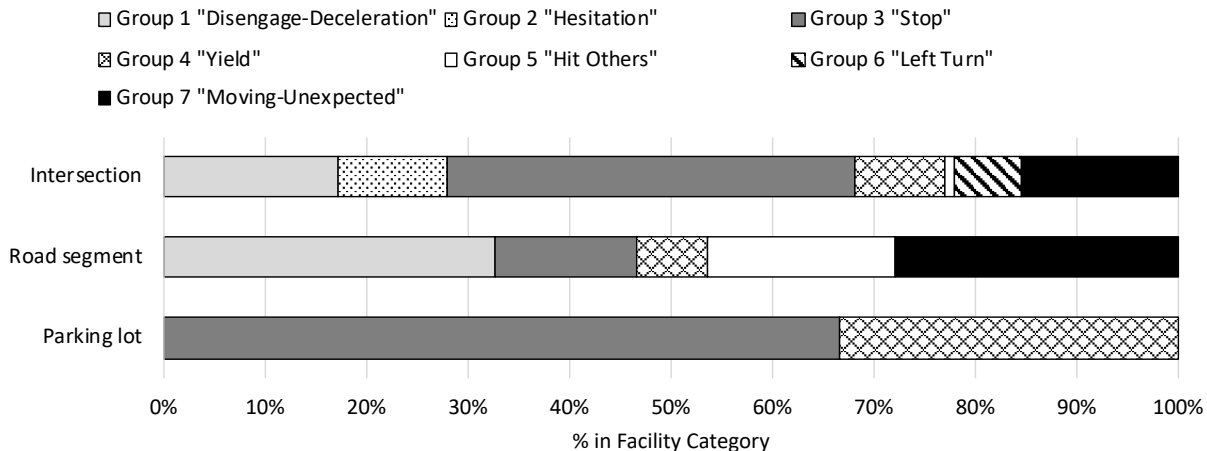


Figure 27 Sequence group distribution by facility type

Time of day is a variable closely related to weather, lighting, and traffic conditions. Based on a visual check, the three most frequently observed AV crash sequence groups were Groups 1, 3, and 7 for both daytime and nighttime. Group 2 crash sequences were only observed during daytime but not nighttime. A Chi-square test result (Chi-squared = 9.22, p-value = 0.16) did not show a significant difference in sequence group distributions between daytime and nighttime.

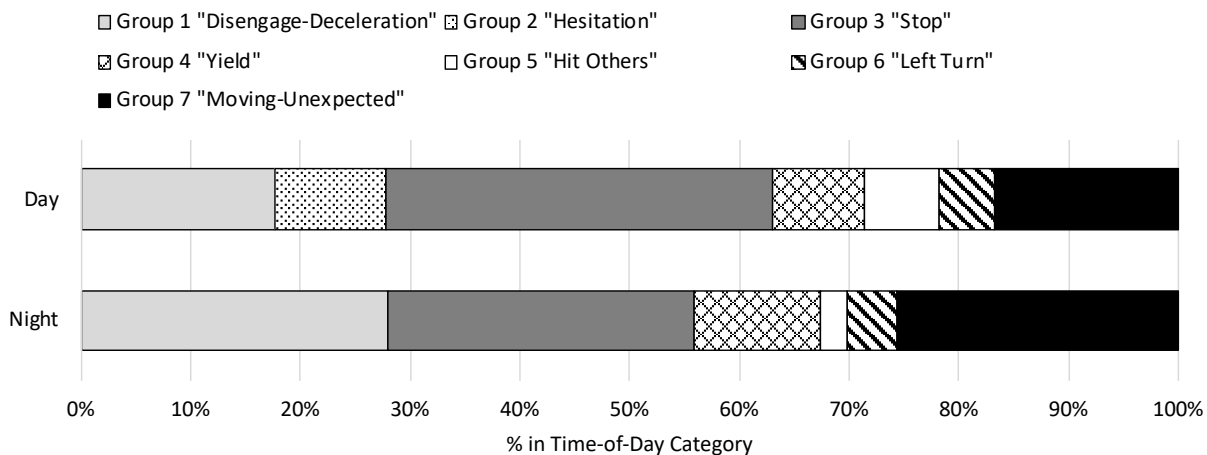


Figure 28 Sequence group distribution by time of day

As AV technology is rapidly developing, AV crash sequence patterns are expected to change over the years. Therefore, for the consideration of designing AV test scenarios, a cross-tabulation was carried out between sequence groups and the years, with results illustrated in Figure 29. Based on a visual check, the distributions of sequence groups varied across the years from 2015 to 2019. A Chi-square test result (Chi-squared = 42.40, p-value = 0.01) showed a significant difference of sequence group distribution across the years.

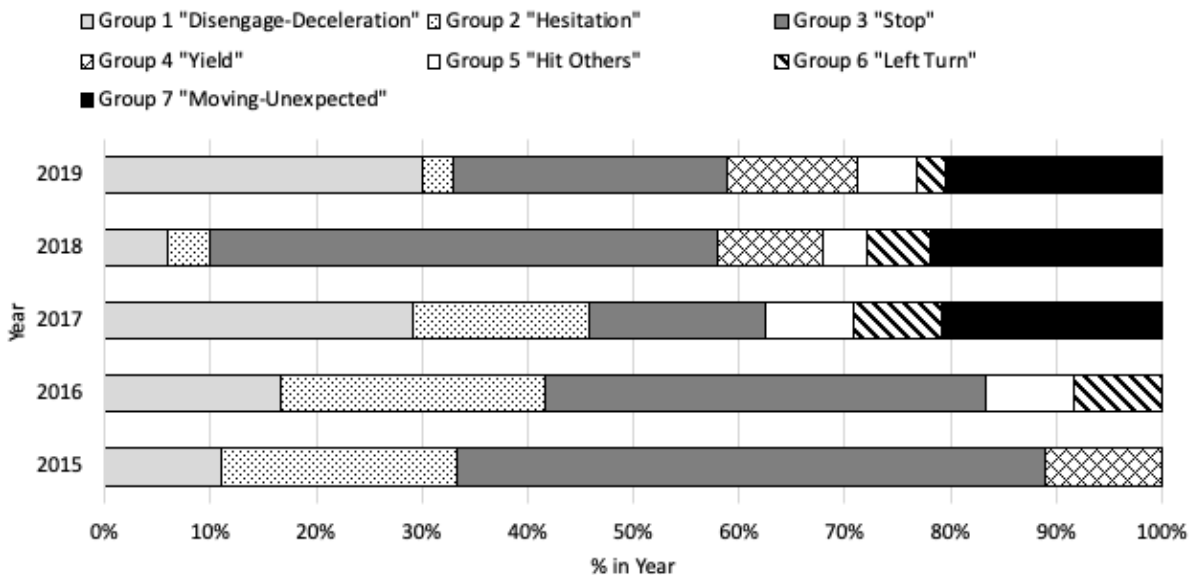


Figure 29 Sequence group distribution by year

4.6 Discussion

With an analysis of California AV crash sequences, this chapter summarized the most frequent events and subsequences, estimated transition probabilities between events, identified cohort groups of sequences, and evaluated the association between sequence

groups and variables measuring crash outcome and environmental conditions. The analysis results led to the following findings:

- The most representative subsequence of California AV crashes was “collision following AV stop”.
- Disengagements were observed in 24% of AV crash sequences. Disengagements were mostly initiated due to operator precaution and detection of other road users’ reckless behavior. Disengagements in the studied AV crash sequences were mostly followed by an immediate collision with other road users or objects, not leaving enough time for the human operator to take further actions.
- AV crash sequences were clustered into seven groups. Each sequence group has a representative subsequence, presenting unique characteristics of AV crash progression.
- AV crash sequence groups were significantly associated with variables measuring crash outcomes and describing environmental conditions, indicating that scenarios described by combinations of event sequences and environmental condition variables can lead to various crash outcomes.

The sequence analysis results revealed patterns in AV crash sequences, which provided information about AV crash progression and form distinctive cohort groups. Sequence groups were shown to lead to different crash outcomes and were associated with environmental condition variables. Events such as disengagement, with its preceding and succeeding events, are unique to AV operations and worthy of consideration in designing AV test scenarios. Also, AV crash sequence patterns changed across the years. As AV technology develops, new crash sequence patterns should be accommodated in AV testing.

A scenario-based AV safety testing framework was developed with sequence of events embedded as a core component. Figure 30 provides a simplified illustration of the framework. A more comprehensive framework can be built based on this one with additional details. This framework consists of a test scenario setup module and a performance evaluation module, which are both built around sequence of events. According to Koopman and Fratrik, AV safety evaluation should be able to validate factors from a four-dimensional validation space with the axes of {Operational design domain (ODD), Object and event detection and response (OEDR), Maneuvers, Fault Management} (8). Our proposed framework captures all these factors through modeling actions and interactions with sequence of events.

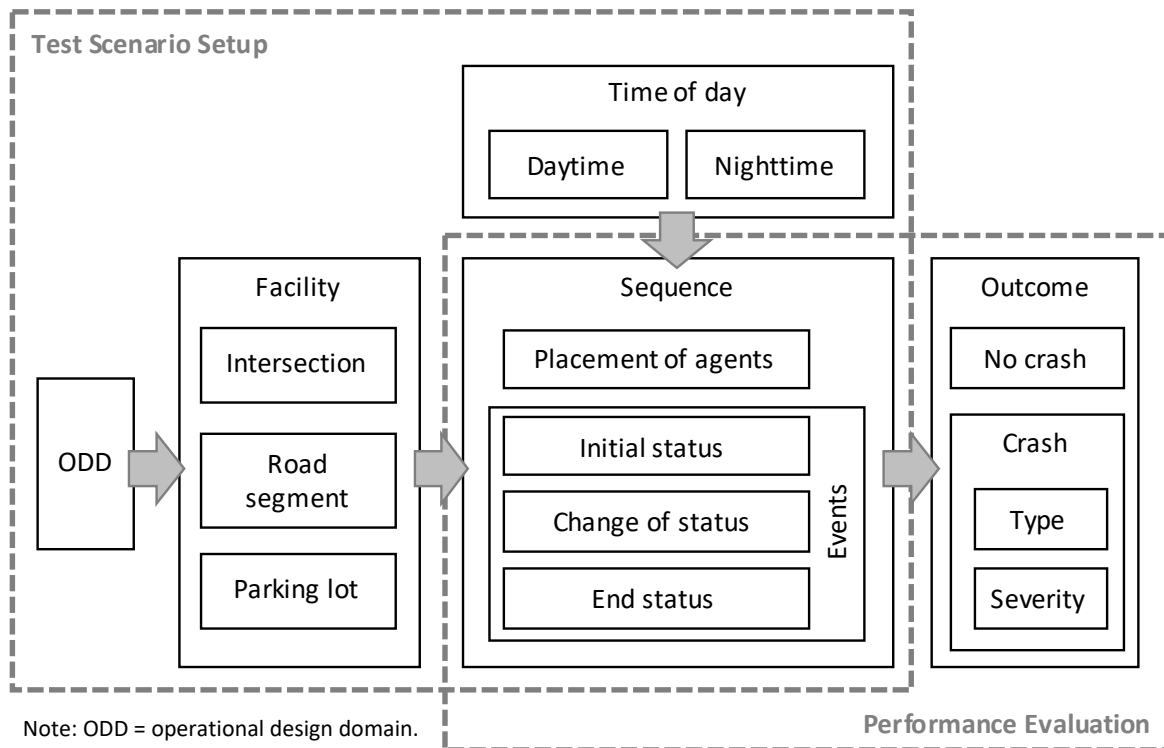


Figure 30 AV safety testing framework with sequence of events embedded

An ODD describes the specific domain in which an automated driving system is designed to properly operate. With a determined ODD, variables such as types of roadways, ranges of speed, time of day, and limits of weather are determined (184). Within the limits of a certain ODD, AV test scenarios can be developed. A scenario is described by variables including road geometries, roadside objects and rules, temporary modifications and events, moving objects, natural environmental conditions, and digital information (93). Crash sequences can be used to encode interactions between moving objects. The crash sequence patterns generated from this study provide a semantic-level description of moving object interactions. Together with the roadway features and environmental condition variables, crash sequence patterns form the basic structure of a test scenario. On the foundation of semantic-level scenarios, concrete scenarios can be generated by defining parameter ranges and specific values for each event and action in the sequences, as well as for each roadway and environmental condition variables (92, 185). Such a process requires additional microscopic AV operations and incidents data, which can be collected from AV field operational tests or NDS databases. To ensure efficient data collection, crash sequence patterns such as the set found in this study, can also be used as a guideline to identify frequent and rare crash cases for data collection.

The following is a procedure to set up test scenarios and evaluate AV safety performance using the proposed framework:

- Based on a determined ODD, a type of facility of interest is selected as the base environment of a test scenario. As many design characteristics of the facility should be considered as possible.

- Environmental condition factors such as time of day, lighting, and weather, are set up on the basis of the selected facility, to create various conditions for testing.
- Moving objects (AVs and other road users) are deployed in the test environment. Representative sequence patterns obtained from sequence analysis of historical crashes can be used to guide the setup of where and when the moving objects appear in the test environment, as well as the interactions between moving objects.
- The moving objects then interact in the test environment and generate crash outcomes, measured by variables such as crash rate, manner of collision, or injury severity. Surrogate safety measures for conflicts are alternative options to describe outcomes. After repeated tests, AVs' safety performance is evaluated based on the test-generated crash/safety outcomes.

4.7 Conclusions

As AV development and testing expand, safety evaluation of such vehicles needs to catch up. Through the analysis of 168 AV crash sequences, this chapter identified patterns in AV crash sequences, which led to a discussion on potential uses of crash sequences in AV safety testing. The conclusion is that crash sequence patterns capture the characteristics of AV crash progression and should be useful in generating AV test scenarios.

Compared with previous studies exploring California AV crash and disengagement patterns, this chapter's research investigated AV crashes and disengagements from a different perspective, sequence of events leading to a crash. Patterns in crash sequences were identified, with AV crash sequences clustering into seven distinctive cohort groups.

Cross-tabulation analysis showed that sequence groups are significantly associated with variables measuring crash outcomes and describing environmental conditions. AV crash sequences can be used in generating semantic-level AV test scenarios. Based on the findings, an AV safety testing framework was proposed with sequence of events embedded as a core component.

In addition to the contribution in discovering AV crash sequence patterns, this chapter showed the value of crash sequence analysis, and reemphasized the importance of collecting crash sequence data. Although the importance of crash sequence of events was stressed by NTSB, reporting such information was not required by the California DMV, nor were any guidelines provided for including crash sequence information in text narratives (43, 77). Crash sequence information was buried in the narratives of AV crash reports. In addition to descriptive summary of crash report data, this chapter carried out a more in-depth analysis, which helped us discover more informative patterns in AV crashes than two very recent studies using the same data source of crash report text narratives (170, 172). With crash sequences, we were also able to further analyze the relationship between AV disengagements and crashes, and better understand the role of disengagement in crashes that happened during AV field tests. Recent studies of AV disengagements focused on finding contributing factors to disengagements rather than evaluating the connection between disengagements and crashes (173, 177).

Limitations of this chapter's research are primarily in the use of crash reports filled out by different AV-testing organizations and submitted to the California DMV. The events were classified, with event sequences developed manually by one researcher to maintain consistency. The recording of events in crash sequences was based on crash text narratives

and the researcher's comprehension of such narratives. Consistency in crash sequences was enhanced through a two-phase encoding process and by having one researcher perform this task.

For future work, a similar analysis can be carried out using more AV crash data as they become available. Improved encoding and sequence analysis methods will also be used. With crash sequence data available in historical human-driven vehicle crash databases, a comparative study can be carried out between patterns in AV crash sequences and human-driven vehicle crash sequences. It is strongly recommended that federal and state transportation agencies require AV testing organizations to share microscopic, event-level data of AV disengagements and crashes that occur during public-road tests and make the data available to safety researchers. Detailed data would enable a much more informed AV testing and evaluation process, bring transparency to public-road AV testing, and enhance public trust in AVs.

Chapter 5 Intersection Two-Vehicle Crash Scenario Specification for Automated Vehicle Safety Evaluation Using Sequence Analysis and Bayesian Networks

5.1 Introduction

Scenario-based testing is an essential part of automated vehicle (AV) safety evaluation, and generating challenging scenarios is critical for scenario-based testing (7, 186). Historical crash data consists of challenging scenarios faced by human drivers and is a good data source for developing test scenarios that may also be challenging for AVs (11, 14, 145). As advanced driver assistance systems (ADAS) and automated driving systems (ADS) are developed to replace human drivers partially or fully, it is reasonable to expect AVs to have capabilities of handling human-driver-faced challenging scenarios and mitigating crash outcomes.

Prior efforts in developing test scenarios using historical crash data have developed characterization of crashes to be used as representative scenarios for the evaluation of ADAS or ADS, based on summarizing and mining patterns in crash attributes (11–14, 187). The end products from prior efforts – representative scenarios – lack considerations of crash progression, dynamics, and mechanisms, which are important information to distinguish crashes and their outcomes (15, 16).

The objective of this chapter is to propose a crash-data-based scenario generating procedure that improves crash characterization and scenario specification by employing crash sequence analysis and Bayesian network modeling. Crash sequence analysis generates informative crash sequence types that describe crash dynamics and progression.

Bayesian network modeling provides a model of crash mechanisms. The two analyses together enable the specification of crash scenarios depicted by the actions and interactions of moving participants, human factors, physical environmental conditions, and crash outcomes.

In this chapter, the scenario generating procedure was used to develop scenarios of intersection two-vehicle crashes, based on 2016-2018 crash data from the Crash Report Sampling System (CRSS) of the United States National Highway Traffic Safety Administration (NHTSA). Following this section, the rest of this chapter is organized as: Section 5.2. A literature review summarizing key literature in scenario generation using historical crash data, crash sequence analysis, and application of Bayesian networks in crash analysis. Section 5.3. Description of the CRSS data, special data preparation for scenario development, crash sequence data processing techniques, and other crash attributes used in modeling. Section 5.4. Explanation of methods employed in the scenario generating procedure. Section 5.5. Results of crash characterization based on sequences, Bayesian network modeling of variable dependencies, and a demonstration of scenario specification process. Section 5.6. Discussion and conclusions.

5.2 Literature Review

5.2.1 Scenario Generation Using Historical Crash Data

Historical crash data has been used to generate scenarios for different purposes, such as general vehicle safety testing, evaluation of crash avoidance systems, ADAS, and ADS (14, 99, 100, 62, 105, 104, 11–13, 187, 145). Some prior research developed a comprehensive set of scenarios that cover as many crashes and operational design domains

(ODDs) as possible, using a national-level crash database and summarizing crash attributes (14). Most prior studies focused on some specific types of crashes and extracted or characterized crash scenarios using national-level crash databases (11–13, 62, 99, 100, 104). Several studies proposed systematic methodologies to characterize crashes and specify representative scenarios (11, 105, 187). A most recent Waymo LLC study used municipal-level crash report census data to reconstruct crash scenarios in simulations for AV safety evaluation (145).

5.2.2 Crash Sequence Analysis

Crash sequences are chronologically ordered events happened during the pre-crash and crash periods. Crash sequence analysis has a similar purpose with sequence analysis in biological and sociological research, focusing on identifying representative components of sequences, the similarity (or dissimilarity) between sequences, and the relationship between sequences and potential outcomes (15).

Prior studies on crash sequence analysis are limited. Krull et al. found significant relationships between the order of rollover and fixed-object collision events and single-vehicle crash injury severity outcomes (81). Wu et al. adapted sequence analysis methods from biological and sociological research on fatal single-vehicle run-off-road crash data, characterized crashes and found significant relationship between crash sequence types and crash injury severity outcomes (15). In Chapter 3, a methodology for crash sequence analysis was developed to help select the most appropriate sequence analysis techniques (e.g., sequence encodings, dissimilarity measures) for different use cases in studying crashes. A case study of interstate single-vehicle crash characterization was used to

demonstrate the effectiveness and usefulness of the proposed methodology. In Chapter 4 and its related journal paper, sequence analysis was used to characterize California AV crash sequences, found significant association between crash sequence types and multiple crash attributes, and proposed a scenario-based AV evaluation framework with crash sequences as the core component (16).

5.2.3 Bayesian Networks for Crash Analysis

Bayesian networks are graphical models known for their use in revealing variable dependencies and causal relationships, and are widely applied in artificial intelligence, medical, and genetic research (188). Prior applications of Bayesian networks in traffic crash research focused on crash forensics, estimating effects of contributing factors on crash outcomes, and predicting crash outcomes (89–91, 189–194). With large samples of crash data, Bayesian networks were proved to be effective in identifying complex interrelationships among multiple crash attributes and crash outcomes (193, 194).

5.3 Data

Historical crash data (2016-2018) from the NHTSA CRSS database was used in this study. The CRSS is a United States crash database with crash data extracted from nationally sampled crash reports (144). A total of about 150,000 crash observations were included in the 2016-2018 CRSS crash data sets, representing more than 18 million police-reported crashes.

The CRSS database organizes data into crash, vehicle, person, and event levels. The four levels of data can be linked through unique IDs for crashes, vehicles, and persons.

Details about the CRSS database has been discussed in Chapter 3. In this chapter, the crash, vehicle, and event level data sets were used.

5.3.1 Subsetting

Since intersection two-vehicle crashes were the focus of this chapter, a subset of crashes of the 2016-2018 CRSS data was obtained by applying the criteria listed in Table 22. The criteria ensured a data set consisting of only the “common” intersection two-vehicle crashes, with “special” crashes (e.g., with emergency vehicles, alcohol-related) or crashes at locations with special configurations (e.g., work zone) excluded. It was concerned that the nature of those “special” crashes could lead to drastically different sequences of events and outcomes from the “common” crashes, and a separate analysis would be needed develop relevant scenarios for “special” crashes. In this chapter, the focus was on generating scenarios for the “common” crashes. By applying the criteria, a data set consisting of 39,850 observations was obtained, representing about 5.9 million crashes.

Table 22 Subsetting criteria

Variable (Data Level)	Value	Description of Criterion
VE_TOTAL (Crash)	= 2	Exactly two vehicles involved in crash.
VE_FORMS (Crash)	=2	Exactly two vehicle-in-transport involved in crash.
PVH_INVL (Crash)	= 0	No parked/working vehicles involved.
RELJCT2_IM (Crash)	= 2 or 3	Crash was at intersection or intersection-related.
WRK_ZONE (Crash)	= 0	No work zone at crash location.
ALCHL_IM (Crash)	≠ 1	No alcohol-related crash.
BDYTYP_IM (Vehicle)	< 50	Only automobile, utility vehicles or light trucks* involved.
TOW_VEH (Vehicle)	= 0	No vehicle trailing involved.
BUS_USE (Vehicle)	= 0	No bus involved.
SPEC_USE (Vehicle)	= 0	No special use vehicles involved.
EMER_USE (Vehicle)	= 0	No emergency use vehicles involved.

Note: * Light trucks with Gross Vehicle Weight Rating (GVWR) ≤ 10,000 LBS.

5.3.2 Numbering Crash Participants

Each intersection two-vehicle crash has two participating motor vehicles. The CRSS numbered the two vehicles were as Vehicle 1 and Vehicle 2 “sequentially”, without specifying the ordering basis (144). Before analyzing the data and develop crash scenarios, crash participating vehicles were renumbered based on their intents (implied by initial positions and trajectories) in crashes. By renumbering the participating vehicles, scenarios generated using the crash data were ensured to have consistent vehicle alignments. In simulation-based tests with a two-vehicle crash scenario, the automated driving system (ADS) being tested would be aligned as one of the two participating vehicles (145). The ADS would be first aligned as Vehicle 1, with Vehicle 2 as an adversarial agent, and then be aligned as Vehicle 2, with Vehicle 1 as an adversarial agent. In that case, a scenario can be fully utilized by testing the ADS on both participants’ roles in a two-vehicle crash.

To renumber the participating vehicles, information about vehicles intents was obtained from the PC23 Crash Type Diagram of CRSS, as shown in Figure 31. For two-vehicle crashes, there are 5 high-level categories including “Same Trafficway Same Direction”, “Same Trafficway Opposite Direction”, “Change Trafficway Vehicle Turning”, “Intersect Paths”, and “Miscellaneous”. The 5 high-level categories split into 10 configurations (denoted as using letters from “D” to “M”), which split again into more specific crash types (with participating vehicle’s intent numbered from “20” to “99”). The rules of vehicle renumbering are shown in Table 23. Take the “68-69” (left turn meets through) combination in crash configuration “J” as an example, the vehicle with an intent “68” (left turn) is numbered as Vehicle 1, and the vehicle with intent “69” (through) was numbered as Vehicle 2.

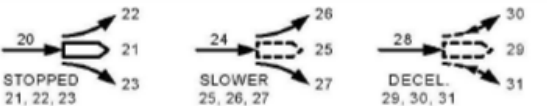
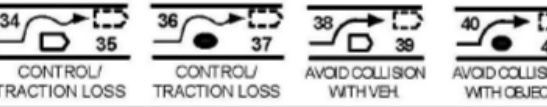
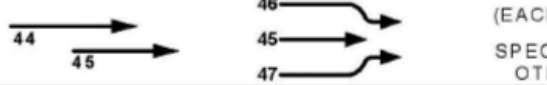
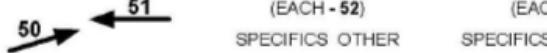
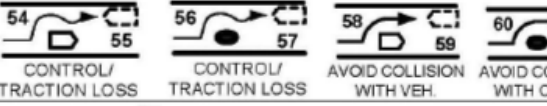
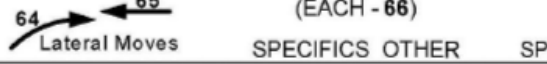
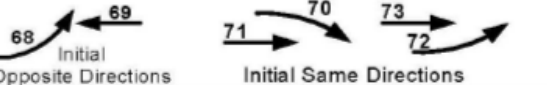
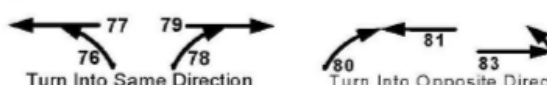
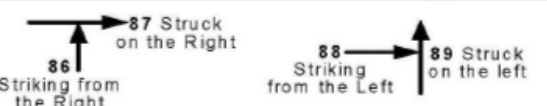
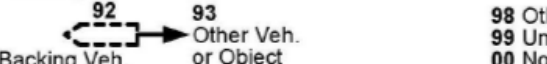
II Same Trafficway Same Direction	D Rear End		(EACH - 32) SPECIFICS OTHER	(EACH - 33) SPECIFICS UNKNOWN
	E Forward Impact		(EACH - 42) SPECIFICS OTHER	(EACH - 43) SPECIFICS UNKNOWN
	F Angle, Sideswipe		(EACH - 48) SPECIFICS OTHER	(EACH - 49) SPECIFICS UNKNOWN
III Same Trafficway Opposite Direction	G Head-On		(EACH - 52) SPECIFICS OTHER	(EACH - 53) SPECIFICS UNKNOWN
	H Forward Impact		(EACH - 62) SPECIFICS OTHER	(EACH - 63) SPECIFICS UNKNOWN
	I Angle, Sideswipe		(EACH - 66) SPECIFICS OTHER	(EACH - 67) SPECIFICS UNKNOWN
IV Change Trafficway Vehicle Turning	J Turn Across Path		(EACH - 74) SPECIFICS OTHER	(EACH - 75) SPECIFICS UNKNOWN
	K Turn Into Path		(EACH - 84) SPECIFICS OTHER	(EACH - 85) SPECIFICS UNKNOWN
V Intersect Paths	L Straight Paths		(EACH - 90) SPECIFICS OTHER	(EACH - 91) SPECIFICS UNKNOWN
VI Misc.	M Backing, Etc.		98 Other Accident Type 99 Unknown Accident Type 00 No Impact	

Figure 31 CRSS two-vehicle crash types (144)

Table 23 Vehicle renumbering

CC	IF	CTID=	VN=	CC	IF	CTID=	VN=	CC	IF	CTID=	VN=	CC	IF	CTID=	VN=
D		20	1	F		44	1	J		68	1	L		86	1
		21	2			45	2			69	2			87	2
		22	2			46	1			70	1			88	1
		23	2			47	1			71	2			89	2
		24	1			48	1			72	1			90	1
		25	2			49	1			73	2			91	1
		26	2	G		50	1			74	1	M		92	1
		27	2			51	2			75	1			93	2
		28	1			52	1	K		76	1			98	1
		29	2			53	1			77	2			99	1
		30	2	H		58	1			78	1				
		31	2			59	2			79	2				
		32	1			62	1			80	1				
		33	1	I		64	1			81	2				
E		38	1			65	2			82	1				
		39	2			66	1			83	2				
		42	1			67	1			84	1				
										85	1				

Note: CC = Crash configuration; CTID = Vehicle's ID in crash type;
VN = Vehicle number in crash sequence.

5.3.3 Encoding Sequences

The crash sequences were formed using four CRSS variables, PCRASH1 (pre-event movement), PCRASH2 (critical event pre-crash), and PCRASH3 (attempted avoidance maneuver) in the vehicle level data set VEHICLE, as well as the SOE (sequence of events) variable from the event level data set CEVENT. The PCRASH1~3 variables describe “what a vehicle was doing just prior to the critical precrash event”, “what made the vehicle's situation critical”, and “what was the corrective action made, if any, to this critical situation” (144). The SOE variable records series of harmful and non-harmful events occurred in the crashes, in chronological order.

The PCRASH1~3 and SOE events were combined following the rule illustrated in Figure 32. If the first event in SOE was a Vehicle 1 event, then the PCRASH1~3 events of Vehicle 1 were inserted before the PCRASH1~3 events of Vehicle 2, and all PCRASH events were inserted before SOE. Vice versa, if the first event in SOE was a Vehicle 2 event, then the PCRASH1~3 events of Vehicle 2 were inserted before the PCRASH1~3 events of Vehicle 1, and all PCRASH events were inserted before SOE. Lengths (number of events) of the sequences ranged from 7 to 16, as shown in Table 24. With over 92% of the sequences having 7 events, the average length was 7.2.

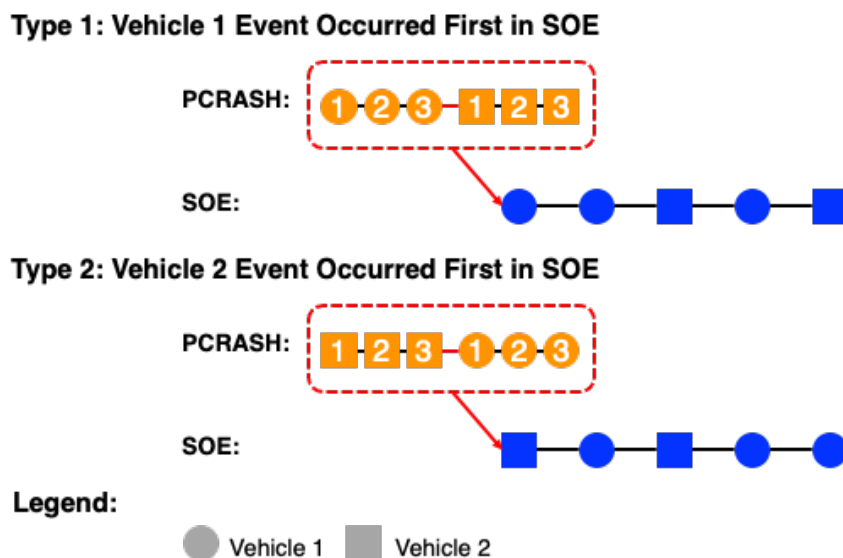


Figure 32 Sequence structure

CRSS has a total of 160 categories of pre-crash and collision events, including 20 in PCRASH1, 57 in PCRASH2, 14 in PCRASH3, and 69 in SOE. For two vehicles, the total number would be 320. With a length of 16, there would be theoretically 320^{16} possible sequences. Identifying patterns in sequences would be difficult with too many event

categories. Therefore, in this chapter's research, the events were encoded by consolidating events that are similar in nature. With the encodings, the total number of event categories became 71, including 14 in PCRASH1, 29 in PCRASH2, 11 in PCRASH3, and 17 in SOE.

Details of event encodings are in Table A - 4 of the Appendix.

Table 24 Sequence lengths

Length	Count
7	36,924
8	844
9	1327
10	423
11	194
12	87
13	41
14	6
15	1
16	3
Average: 7.2 Total: 39,850	

5.3.4 Other Crash Attributes

To specify crash scenarios, variables describing crash outcomes, human factors, and environmental conditions were needed, in addition to the crash sequence patterns. The crash outcomes help identify scenarios with more severe injuries and fatalities, which may be of more interest in testing AVs. The human factors and environmental conditions help depict ODDs which consist of other moving objects and static surroundings. Crash outcome variables used in this chapter are summarized in Table 25. Variables of human factors and environmental conditions are summarized in Table 26.

More than half (52.5%) of the sample crashes ended with a maximum injury severity of no apparent injury, less than a half (46.9%) with injury to different extents, and

0.4% with fatality. In terms of manner of collision, angle crashes made up 47.0% of the sample crashes, front-to-rear crashes made up 38.7%, and others made up 14.3%. The crash outcomes reflect a motor vehicle's exposure to crash risks. When setting up test scenarios for AV safety evaluation, the exposure can be adjusted through selecting a combination of crash sequence types.

In this chapter, the human factors variables were derived from the "D22 Speeding Related" and "D24 Related Factors – Driver Level" variables of the CRSS VEHICLE data file. The focus was on the most frequent driver factors involved in the sample crashes, including speeding, careless driving, did not see, reckless driving, and improper control. Environmental condition variables included were urbanicity, time of day, lighting condition, weather, type of intersection, speed limits, road surface condition, and traffic control device. For test scenario specifications, the human factors and environmental conditions describe the ODDs.

Table 25 Crash outcomes

Variable	%	Variable	%
Maximum severity (maxsev)		Manner of collision (moc)	
No apparent injury	52.5%	Angle	47.0%
Possible injury	27.7%	Front-to-rear	38.7%
Suspected minor injury	11.5%	Sideswipe, same direction	7.0%
Suspected serious injury	7.2%	Front-to-front	5.3%
Fatal	0.4%	Sideswipe, opposite direction	1.0%
Injured, severity unknown	0.5%	(Other)	1.0%

Note: Sample size 39,850.

Table 26 Human factors and environmental conditions

Variable	%	Variable	%	Variable	%
Driver speeding (speeding)		Urbanicity (urbrur)		Speed limits (spdlim)	
N+N	90.5%	Urban	80.2%	45+45	20.7%
Y+N	5.7%	Rural	19.8%	35+35	18.0%
U+N	2.2%	Time of day (tod)		Unknown	13.4%
N+Y	0.6%	Day	80.7%	40+40	10.4%
N+U	0.5%	Night	19.3%	25+25	7.7%
U+U	0.4%	Lighting condition (light)		30+30	6.2%
(Other)	0.1%	Daylight	77.1%	(Other)	23.5%
Careless driving (careless)		Dark-Not Lighted	3.6%	Road surface condition (surcon)	
N+N	90.3%	Dark-Lighted	15.3%	Dry	83.2%
N+Y	1.0%	Dawn	1.2%	Wet	13.6%
Y+N	8.5%	Dusk	2.5%	Unknown	1.1%
Y+Y	0.2%	Dark-Unknown Lighting	0.3%	Snow	0.9%
Driver did not see (didnotsee)		(Other)	0.0%	Ice/frost	0.5%
		Weather (weather)		Non-trafficway or driveway access	0.4%
N+N	99.0%	Clear	73.2%	(Other)	0.2%
N+Y	0.1%	Cloudy	15.9%	Traffic control devices (tcd)	
Y+N	0.8%	Rain	9.1%	Signal+Signal	49.1%
Y+Y	0.1%	Snow	1.4%	No TCD+No TCD	19.7%
Reckless driving (reckless)		Fog, Smog, Smoke	0.2%	Sign+No TCD	11.7%
N+N	96.7%	Sleet or Hail	0.1%	Sign+Sign	7.8%
N+Y	0.3%	(Other)	0.2%	No TCD+Sign	5.1%
Y+N	3.0%	Type of intersection (typint)		Unknown+Unknown	2.3%
Y+Y	0.0%	(Other)	0.7%	(Other)	4.3%
Improper control (impropctrl)		3-Legged	22.0%		
N+N	99.5%	Unknown	21.8%		
N+Y	0.0%	4-Legged	55.6%		
Y+N	0.5%				

Note: Sample size 39,850.

Labels with "+" indicate conditions of Vehicle 1 on the left side and Vehicle 2 on the right side.

In speeding, etc.: U = Unknown; Y = Yes; N = No.

5.4 Methodology

This chapter employed a procedure of two steps to specify crash scenarios and their ODDs. The first step was characterization of crash sequences using sequence analysis, and the second step was ODD specification based on a Bayesian network modeling of relationships among crash sequence types, outcomes, human factors, and environmental

conditions. Sequence types and ODDs make up scenarios. Crash sequence types describe what could happen between moving objects in the scenarios. ODDs illustrate what the surroundings could be like in the scenarios.

5.4.1 Crash Sequence Comparison and Clustering

In sequence analysis, the encoded crash sequences were compared and grouped. The basis of sequence comparison and clustering is the measure of dissimilarity, which quantifies the difference between each pair of sequences (64, 130, 142). Optimal matching (OM) based dissimilarity measures have been widely used in bioinformatic and sociological research to for gene sequence or life course sequence analysis (64–66, 130, 133, 142). The Chapter 3 research also found OM based dissimilarity measures to be appropriate for crash sequence analysis. In Chapter 3, a methodology was also proposed to select the optimal encoding schemes and dissimilarity measures for different crash analysis use cases and found the Levenshtein distance to be an overall good choice for measuring crash sequence dissimilarity.

The OM takes two sequences and align them. The alignment of sequences involves several operations such as substitutions, deletions/insertions (or indels), compression and expansions, and transpositions (or swaps) (13, 18, 30). The Levenshtein distance uses only substitutions and indels. The mathematical expression of OM based dissimilarity between a pair of sequences, x and y , is:

$$d_{OM}(x, y) = \min_j \sum_{i=1}^{\ell_j} \gamma(T_i^j) \quad [15]$$

where ℓ_j denotes the transformations needed to turn sequence x into y ; and

$\gamma(T_i^j)$ is the cost of each elementary transformation T_i^j (e.g., indel or substitution).

An example of sequence alignment using Levenshtein distance is shown in Table 27. There are multiple ways to align two sequences, “ABCD” and “ACB”. Using a substitution cost of s and an indel cost of d , the two ways of alignment shown in Table 27 yielded different total costs. The OM then applies a greedy algorithm to return the minimum alignment cost as the dissimilarity between the two sequences (129).

Table 27 Sequence alignment

Sequence 1	A	B	C	D
Sequence 2	A	C	B	
Alignment 1				
Sequence 1	A		B	C D
Sequence 2	A	C/	B	\emptyset \emptyset
Cost		d	d	d = $3d$
Alignment 2				
Sequence 1	A	B	C	D
Sequence 2	A	\emptyset	C	<u>B</u>
Cost		d	s	= $d+s$

Note: Insertion is marked with \emptyset ,
 Deletion is marked with slash/,
 Substitution is marked with underline

Using the Levenshtein distance, a dissimilarity matrix for a set of crash sequences can be calculated. The matrix is of size $n*n$, where n is the number of sequences in the set. Each element in the matrix indicates the dissimilarity between a pair of sequence. The dissimilarity matrix can then be plug into a clustering algorithm to characterize crash sequences as distinctive types.

For clustering, a weighted k-medoid method was employed in this chapter. K-medoid clustering has been applied in prior studies for crash characterization due to its good performance with categorical data and robustness against outliers (11, 16). The

weighted k-medoid clustering algorithm used in this chapter was developed by Studer, accommodating sampling weights (provided by CRSS data sets) in the clustering (160). The sequence dissimilarity calculation and sequence clustering were completed using R and the libraries “TraMineR” and “WeightedCluster” (157, 158, 160).

To characterize the intersection two-vehicle crashes, the sequence comparison and clustering were done under the existing CRSS crash configuration (CC) classification. The distribution of CC is shown in Table 28. Sequence comparison and clustering was done for each CC category. By conducting sequence analysis under the existing CC classification, representations were kept for rarer crash types, which would otherwise be overlooked if all 39,850 crash sequences were analyzed in a lump sum. Those rarer crash types are also useful for developing potentially challenging test scenarios.

Table 28 Distribution of CRSS intersection two-vehicle crash configurations

Category	CC	Description	Count	%	Weighted Count	%
Same Trafficway, Same Direction	D	Rear End	14,784	37.10%	2,300,603	39.08%
	E	Forward Impact	8	0.02%	1,143	0.02%
Same Trafficway, Opposite Direction	F	Angle, Sideswipe	1,692	4.25%	301,033	5.11%
	G	Head-On	153	0.38%	14,178	0.24%
	H	Forward Impact	9	0.02%	586	0.01%
Change Trafficway, Vehicle Turning	I	Angle, Sideswipe	117	0.29%	18,069	0.31%
	J	Turn Across Path	7,932	19.90%	1,095,637	18.61%
Intersect Paths Miscellaneous	K	Turn Into Path	6,777	17.01%	1,010,868	17.17%
	L	Straight Paths	7,160	17.97%	938,248	15.94%
	M	Backing, Etc.	1,218	3.06%	207,224	3.52%
<i>Total</i>			<i>39,850</i>	<i>100.00%</i>	<i>5,887,588</i>	<i>100.00%</i>

5.4.2 Bayesian Network Modeling

Bayesian network modeling has been used in prior studies to evaluate factors affecting crash type and injury severity (89, 90, 195, 91, 196, 189–191, 193, 194). Bayesian

networks are directed acyclic graphs (DAGs), with nodes denoting variables and directed edges denoting the dependencies between variables (197). The strengths of influences between variables are measured by conditional probabilities. If the graph has variables x_1, \dots, x_n , and S_i as the set of parents of x_i , an estimated conditional probability is then $P'(x_i|S_i)$. The following joint probability distribution exists for the graph:

$$P(x_1, \dots, x_n) = \prod_i P'(x_i|S_i) \quad [16]$$

The construction of a Bayesian network consists of two steps (189, 190):

- Structure learning: determine selection of variables (nodes) and determine the dependencies or independencies between nodes, to form a DAG.
- Parameter learning: based on the determined DAG, estimate a conditional probability table for each node to quantify relationship between nodes.

In this chapter, the structure learning was completed using a hill climbing algorithm, which finds the network structure with the highest Akaike information criterion (AIC) score. The AIC score used in the R library, “bnlearn”, is calculated as the classic definition rescaled by -2 (198):

$$AIC = \ln(\hat{L}) - 2k \quad [17]$$

where \hat{L} = the maximum likelihood of the model; k = the number of estimated parameters. Therefore, a higher AIC score means a better Bayesian network model. The learned network structure was adjusted based on the authors’ domain knowledge in traffic crashes. Multiple networks structures were generated and compared to determine a most appropriate one for ODD specification, based on the conditional probability table. The

structure and parameter learning were completed using R and the “bnlearn” library (157, 198). Bayesian networks were visualized using the “Rgraphviz” library (199).

A determined Bayesian network can reveal the dependencies among variables including sequence types (developed from the sequence analysis), crash outcomes (manner of collision and injury severity), human factors, and environmental conditions, using graph visualizations. The network can also help identify the sequence types likely yielding serious crash outcomes and the ODD settings for specific sequence types by querying for conditional probabilities in the network. Scenarios can be defined using combinations of sequence types and ODD settings, and can be used to help render simulation tests for AV safety evaluation.

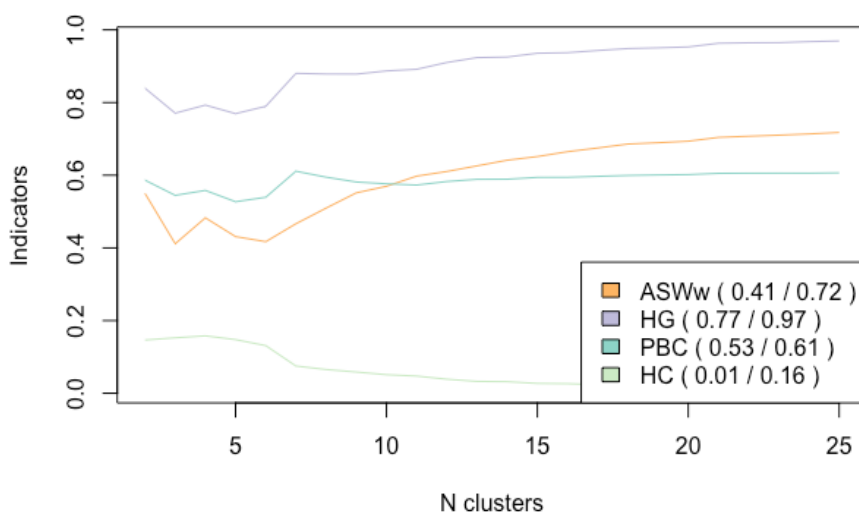
5.5 Results

This section presents the results of 1) crash sequence characterization from sequence analysis and 2) variable dependencies from Bayesian network modeling. An example of scenario specification based on the crash sequence type and Bayesian network is provided at the end of this section.

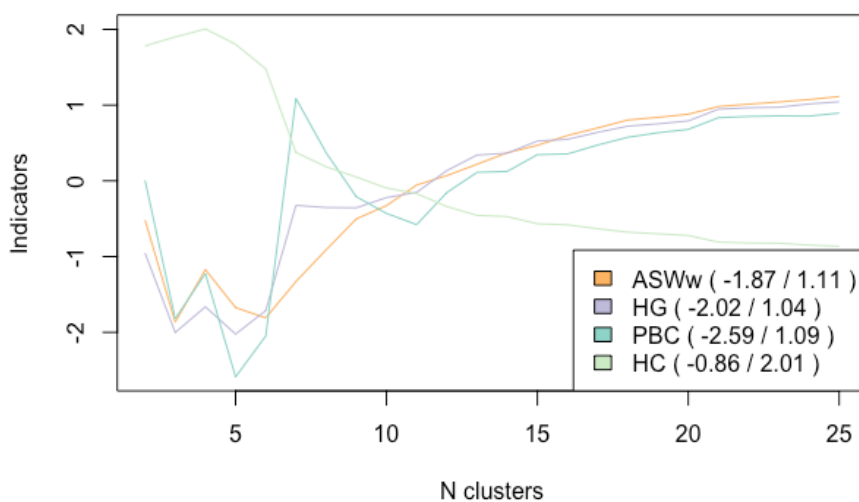
5.5.1 Sequence Types

As mentioned, sequence clustering was carried out for each CC category using Levenshtein distance and weighted k-medoid clustering. To measure the quality of clustering and determine the appropriate number of clusters, clustering quality indices including the Weighted Average Silhouette Width (ASWw), Hubert’s Gamma (HG), Point Biserial Correlation (PBC), and Hubert’s C (HC) were calculated. A range of k values (2 to 25) were used to cluster the sequences, and the indices were plotted for comparison.

The plots in Figure 33 illustrate the clustering quality indices for the clustering of CC D crash sequences. An optimal k value would give us maximum ASWw, HG, and PBC (all range from -1 to 1), and a minimum HC (ranges from 0 to 1). Based on Figure 33, k = 12 was chosen as the number of clusters for CC D sequence clustering. The same procedure of plotting quality indices and determining the appropriate k value was used for all CC categories.



(a) Original values



(b) Standardized values

Figure 33 Clustering quality indices for CC D (rear end) sequences

The clustering results are summarized in Table 29. Only the sequence representing the most sequences in each cluster is presented. More detailed results showing the top three representative sequences are shown in Table A - 5 of the Appendix. CC D Same Trafficway, Same Direction – Rear End (39% of all) crashes were characterized as 12 sequence types, CC J – Change Trafficway, Vehicle Turning – Turn Across Path (18% of all) crashes were characterized as 3 sequence types, CC K – Change Trafficway, Vehicle Turning – Turn Into Path (17% of all) crashes were characterized as 9 sequence types, and CC L – Intersect – Straight Paths (16% of all) crashes were characterized as 14 sequence types.

Interpretation of the representative sequences is shown in Table 30 for easier understanding of the sequence types. For example, the representative sequence of Type e3 was “1ST-1OES-1BR-2ST-2OIS-2NA-1XV-1ROR-1XF-1NCH”, which was interpreted as “v1 moving straight-other encroached into lane SD (brake and turned right) → v2 moving straight (no)”, meaning Vehicle 1 and Vehicle 2 were both moving straight along the same direction. Some other vehicle encroached into Vehicle 1’s lane, making Vehicle 1 brake and steer right. Vehicle 1 then collided into the rear of Vehicle 2, which did not make any maneuver to avoid the collision. Following the collision, Vehicle 1 ran off the road, hit a fixed object, and suffered a non-collision harmful event.

Table 29 Sequence clustering results

Category	CC	Type	Representative Sequence	%
Same Trafficway, Same Direction	D-Rear End	d1	1ST-1OIS-1B-2B-2OIS-2N-1XV	0.7
		d2	1R-1OIS-1N-2S-2OIS-2NA-1XV	0.8
		d3	1C-1OIS-1N-2S-2OIS-2NA-1XV	1.4
		d4	1ST-1OIS-1N-2B-2OIS-2N-1XV	5.4
		d5	1A-1OIS-1N-2S-2OIS-2NA-1XV	2.0
		d6	2S-2OIS-2NA-1ST-1OIS-1N-2XV	0.7
		d7	1ST-1OIS-1B-2S-2OIS-2NA-1XV	3.5
		d8	1ST-1OIS-1N-2ST-2OIS-2N-1XV	1.8
		d9	1ST-1OIS-1N-2S-2OIS-2N-1XV	1.2
		d10	1B-1OIS-1N-2S-2OIS-2NA-1XV	1.5
		d11	1ST-1OIS-1NA-2S-2OIS-2NA-1XV	2.9
		d12	1ST-1OIS-1N-2S-2OIS-2NA-1XV	17.1
E-Forward Impact F-Angle, Sideswipe	E-Forward Impact F-Angle, Sideswipe	e3	1ST-1OES-1BR-2ST-2OIS-2NA-1XV-1ROR-1XF-1NCH	0.0
		f1	2ST-2ELL-2N-1ST-1OES-1N-2XV	0.2
		f2	1E-1OIS-1N-2S-2OIS-2NA-1XV	0.4
		f3	2ST-2OES-2N-1E-1ELL-1N-2XV	0.3
		f4	1E-1ELL-1N-2S-2OES-2NA-1XV	0.5
		f5	1E-1ERL-1NA-2ST-2OES-2NA-1XV	0.2
		f6	2ST-2ELL-2N-1S-1OES-1NA-2XV	0.2
		f7	1ST-1ERL-1N-2ST-2OES-2N-1XV	0.3
		f8	1E-1ELL-1N-2ST-2OES-2N-1XV	1.2
f9	1E-1ERL-1N-2ST-2OES-2N-1XV	1.7		
Same Trafficway, Opposite Direction	G-Head-On	g1	1ST-1ELL-1N-2ST-2OEO-2N-1XV	0.2
		g2	1ST-1OET-1L-2S-2OEO-2NA-1XV	0.0
	H-Forward Impact	h1	1ST-1OEO-1L-2ST-2OEO-2N-1CM-1XV	0.0
	I-Angle, Sideswipe	i1	1ST-1ELL-1N-2S-2OEO-2NA-1XV	0.2
i2		2ST-2OEO-2N-1ST-1LCF-1N-2XV-1ROL-1XF	0.1	

(Table 29 Continued)

Category	CC	Type	Representative Sequence	%		
Change Trafficway, Vehicle Turning	J-Turn Across Path	j1	1R-1R-1N-2ST-2OES-2N-1XV	1.5		
		j2	1L-1L-1N-2ST-2OEO-2N-1XV	9.2		
		j3	2ST-2OEO-2N-1L-1L-1N-2XV	7.8		
	K-Turn Into Path	k1	1L-1ELL-1N-2S-2OEO-2NA-1XV	0.9		
		k2	2ST-2ST-2N-1L-1OET-1N-2XV	1.4		
		k3	2ST-2OEO-2N-1L-1L-1N-2XV	3.9		
		k4	1L-1L-1N-2ST-2OEO-2N-1XV	3.1		
		k5	1L-1L-1N-2S-2OEO-2NA-1XV	0.4		
		k6	1R-1R-1N-2ST-2OES-2N-1XV	3.0		
		k7	2ST-2OES-2N-1R-1R-1N-2XV	0.9		
		k8	1L-1OET-1N-2ST-2ST-2N-1XV	0.5		
		k9	1L-1L-1N-2ST-2OES-2N-1XV	3.2		
		Intersect Paths	L-Straight Paths	l1	2ST-2OET-2N-1ST-1ST-1N-2XV	0.2
				l2	1A-1ST-1N-2ST-2OET-2N-1XV	0.5
l3	1ST-1OET-1N-2A-2ST-2N-1XV			0.4		
l4	1ST-1OET-1N-2ST-2OET-2N-1XV			0.3		
l5	1ST-1OET-1B-2ST-2ST-2N-1XV			0.2		
l6	2ST-2ST-2N-1ST-1OET-1N-2XV			1.5		
l7	1ST-1ST-1N-2ST-2OET-2N-1XV-2ROR-2XF			0.2		
l8	1ST-1ST-1NA-2ST-2OET-2NA-1XV			0.2		
l9	1ST-1ST-1NA-2ST-2OET-2N-1XV			0.7		
l10	1ST-1OET-1N-2ST-2ST-2NA-1XV			0.6		
l11	1ST-1ST-1N-2ST-2ST-2N-1XV			0.7		
l12	1ST-1OET-1N-2ST-2ST-2N-1XV			3.6		
l13	1ST-1ST-1N-2ST-2OET-2N-1XV			6.6		
l14	1ST-1ST-1B-2ST-2OET-2N-1XV			0.3		
Miscellaneous	M-Backing, Etc.	m1	1BU-1BU-1N-2S-2OIR-2NA-1XV	1.5		
		u	1U-1U-1N-2ST-2OEO-2N-1XV	2.0		

Table 30 Interpretation of representative sequences

Type	Interpretation
d1	v1 moving straight (decelerated) → v2 decelerating
d2	v1 turning right → v2 stopped (no)
d3	v1 negotiating curve → v2 stopped (no)
d4	v1 moving straight → v2 decelerating
d5	v1 accelerating → v2 stopped (no)
d6	v2 stopped (no) → v1 moving straight (no)
d7	v1 moving straight (decelerated) → v2 stopped (no)
d8	v1 moving straight → v2 moving straight
d9	v1 moving straight → v2 stopped
d10	v1 decelerating → v2 stopped (no)
d11	v1 moving straight (no) → v2 stopped (no)
d12	v1 moving straight → v2 stopped (no)
e3	v1 moving straight-other encroached into lane SD (brake and turned right) → v2 moving straight (no)
f1	v2 moving straight-encroached into lane on left → v1 moving straight
f2	v1 changing lane → v2 stopped (no)
f3	v2 moving straight → v1 changing lane-encroached into lane on left
f4	v1 changing lane-encroached into lane on left → v2 stopped (no)
f5	v1 changing lane-encroached into lane on right (no) → v2 moving straight
f6	v2 moving straight-encroached into lane on left → v1 stopped (no)
f7	v1 moving straight-encroached into lane on right → v2 moving straight
f8	v1 changing lane-encroached into lane on left → v2 moving straight
f9	v1 changing lane-encroached into lane on right → v2 moving straight
g1	v1 moving straight-encroached into lane on left → v2 moving straight
g2	v1 moving straight (steering left) → v2 stopped (no)
h1	v1 moving straight (steering left) → v2 moving straight
i1	v1 moving straight-encroached into lane on left → v2 stopped (no)
i2	v2 moving straight-other encroached into lane → v1 moving straight-speeding

Note: v1 = Vehicle 1; v2 = Vehicle 2. See Section 5.3.2 Numbering Crash Participants for details.

SD = same direction.

The arrow symbol, “→”, means that the vehicle left of the arrow collided into the vehicle right of it.

Content in the parentheses is crash avoidance maneuver, no parenthesis means action unknown.

(Table 30 Continued)

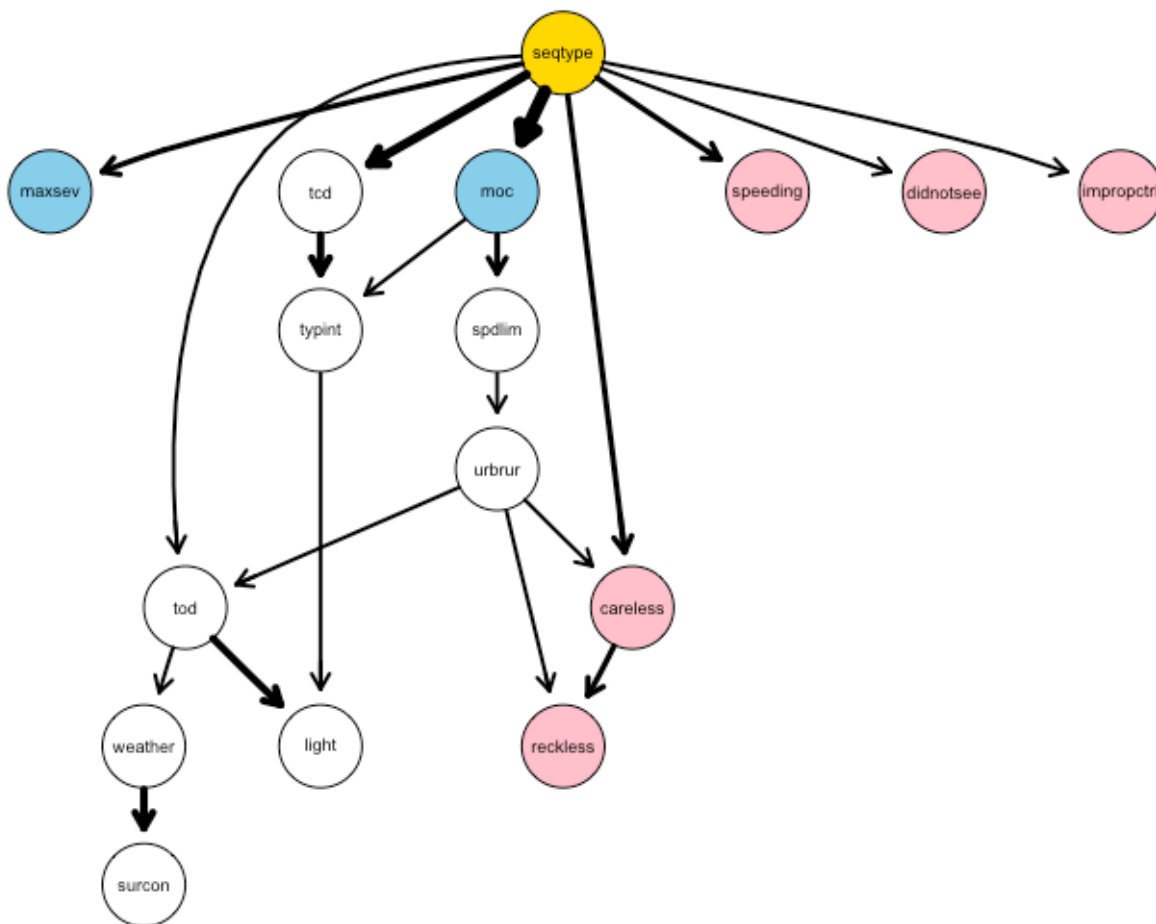
Type	Interpretation
j1	v1 turning right → v2 moving straight-other encroached into lane SD
j2	v1 turning left → v2 moving straight-other encroached into lane OD
j3	v2 moving straight-other encroached into lane OD → v1 turning left
k1	v1 turning left → v2 stopped-other turning into OD (no)
k2	v2 moving straight → v1 turning left-other encroached on cross street
k3	v2 moving straight-other encroached on OD → v1 turning left
k4	v1 turning left → v2 moving straight-other encroached into lane OD
k5	v1 turning left → v2 stopped-other turning into OD (no)
k6	v1 turning right → v2 moving straight-other encroached into lane SD
k7	v2 moving straight-other turning into lane SD → v1 turning right
k8	v1 turning left-other encroached into lane from cross street → v2 moving straight
k9	v1 turning left → v2 moving straight-other encroached into lane SD
l1	v2 moving straight-other encroached into lane CS → v1 moving straight
l2	v1 accelerating/start moving → v2 moving straight-other encroached into lane CS
l3	v1 moving straight-other encroached into lane CS → v2 accelerating/start moving
l4	v1 moving straight-other encroached into lane CS → v2 moving straight-other encroached into lane CS
l5	v1 moving straight-other encroached into lane CS (brake) → v2 moving straight
l6	v2 moving straight → v1 moving straight-other encroached into lane CS
l7	v1 moving straight → v2 moving straight-other encroached into lane CS
l8	v1 moving straight (no) → v2 moving straight-other encroached into lane CS (no)
l9	v1 moving straight (no) → v2 moving straight-other encroached into lane CS
l10	v1 moving straight-other encroached into lane → v2 moving straight (no)
l11	v1 moving straight → v2 moving straight
l12	v1 moving straight-other encroached into lane CS → v2 moving straight
l13	v1 moving straight → v2 moving straight-other encroached into lane CS
l14	v1 moving straight (brake) → v2 moving straight-other encroached into lane CS
m1	v1 backing up → v2 stopped-other backing (no)
u	v1 making U-turn → v2 moving straight-other encroached into lane OD

Note: SD = same direction; OD = opposite direction; CS = cross street.

5.5.2 Variable Dependencies

Using a hill climbing algorithm with AIC as the criterion for model selection, a Bayesian network was learned to illustrate the relationships among sequence types, crash outcomes, human factors, and environmental condition variables. The network is shown in Figure 34. Each node represents a variable, and the directed arcs represent dependencies among variables. The weight of an arc represents its strength, measured by the potential change in AIC score which would be caused by removing the arc from the network (i.e., the difference between the network's AIC score with and without the arc). If the change in AIC is negative, that means removing the arc harmed the network by losing information. Therefore, a more negative difference indicates a higher arc strength (i.e., stronger relationship between two variables).

The Bayesian network shows that sequence type had strong relationships with crash outcome variables – maximum injury severity (“maxsev”) and manner of collision (“moc”). Sequence type was also directly or indirectly associated with human factors and environmental conditions. Detailed arc strengths are listed in Table 31. Five strongest direct links with sequence type were manner of collision (moc), traffic control device (tcd), maximum injury severity (maxsev), speeding, and careless driving (careless).



Note: Crash outcomes are in blue; human factors are in pink; environmental conditions are in white.

Figure 34 Bayesian network generated from hill climbing learning

The variable pointed by an arc is dependent on the variable on the other end of the arc. Note that the arcs pointing from crash sequence type to crash outcomes indicated expected dependencies, supported by prior studies' findings that the order of events and actions taken during the pre-crash and crash periods directly affect manner of collision and injury severity (15). Human factors and environmental conditions were suggested by some prior studies to affect crash types and crash outcomes, with arcs in Bayesian networks pointing from human factors and environmental conditions to crash types and crash

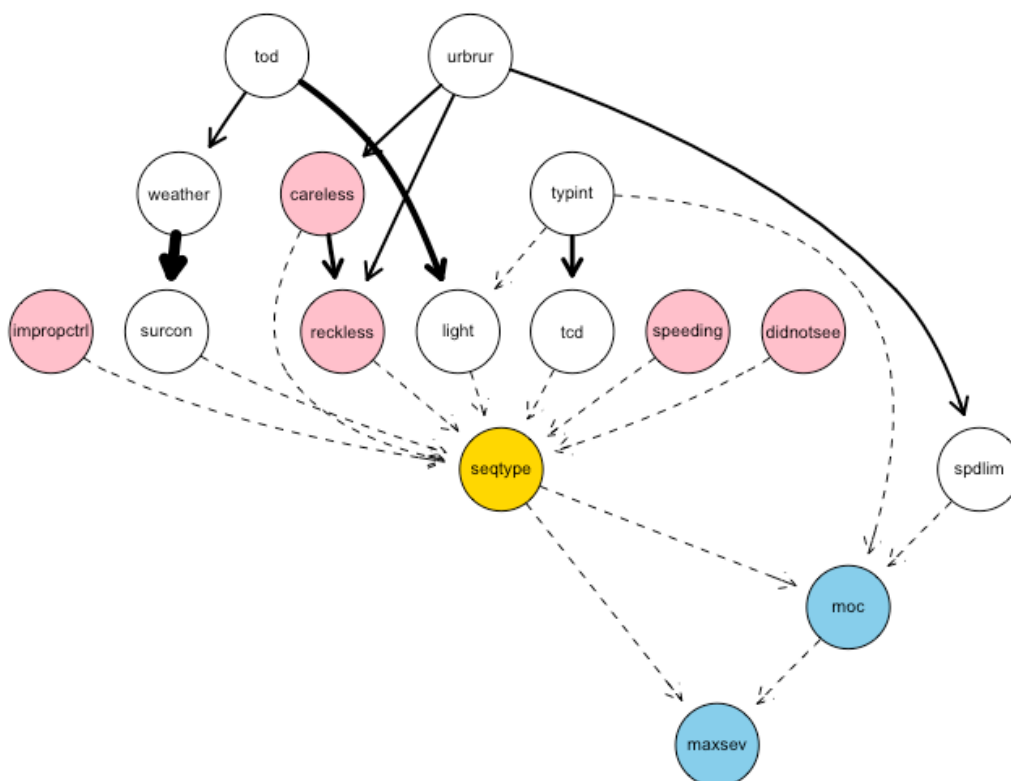
outcomes (190, 191). However, some other studies indicated that crash data did not support such findings and presented Bayesian networks with arcs pointing from crash types and outcomes toward crash types and outcomes (193, 194). In the algorithmically learned network shown in Figure 34, arcs pointed from sequence type and manner of collision to human factors and environmental conditions.

Table 31 Arc strengths

Arc	From	To	Strength (change in AIC)
seqtype → moc	Sequence Type	Manner of Collision	-31,693.6
seqtype → tcd	Sequence Type	Traffic Control Device	-12,347.0
weather → surcon	Weather	Road Surface Condition	-10,261.1
tod → light	Time of Day	Lighting Condition	-8,899.4
tcd → typint	Time of Day	Type of Intersection	-2,897.5
careless → reckless	Careless Driving	Reckless Driving	-2,627.5
moc → spdlim	Manner of Collision	Speed Limit	-2,283.0
seqtype → maxsev	Sequence Type	Maximum Injury Severity	-2,065.7
seqtype → speeding	Sequence Type	Speeding	-1,536.6
seqtype → careless	Sequence Type	Careless Driving	-1,106.3
spdlim → urbrur	Speed Limit	Urbanicity	-920.4
moc → typint	Manner of Collision	Type of Intersection	-553.2
seqtype → didnotsee	Sequence Type	Looked But Did Not See	-198.1
seqtype → tod	Sequence Type	Time of Day	-131.9
seqtype → improprctl	Sequence Type	Improper Control	-139.8
urbrur → reckless	Urbanicity	Reckless Driving	-71.2
typint → light	Type of Intersection	Lighting Condition	-65.2
urbrur → careless	Urbanicity	Careless Driving	-4.7
urbrur → tod	Urbanicity	Time of Day	-3.8
tod → weather	Time of Day	Weather	-53.6

An alternative Bayesian network was constructed to test the hypothesis that human factors and environmental conditions affect sequence types, as shown in Figure 35. In the alternative network, arcs were manually added to point from human factors and environmental conditions to sequence types, and the arc strengths (variable dependencies)

were weak. Therefore, the hypothesis about sequence type being dependent on human factors and environmental conditions was not supported by the data. Here, the data and network learning algorithm were trusted, and the Figure 34 network was determined to be selected as the final Bayesian network for scenario specification.



Note: Dashed arcs indicate weak strengths.

Figure 35 Alternative Bayesian network

To confirm the relationships between sequence types and crash outcomes, as well as among sequence types, human factors, and environmental conditions, structural stability of the local network was tested by developing partial Bayesian networks as shown in Figure 36 and Figure 37.

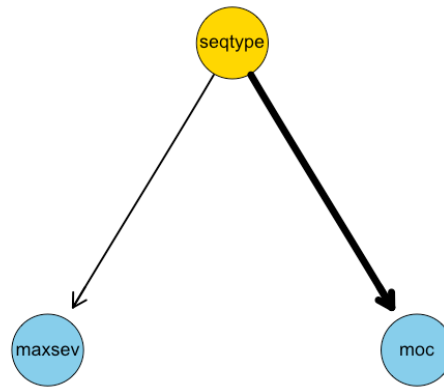


Figure 36 Bayesian network of sequence types and crash outcomes

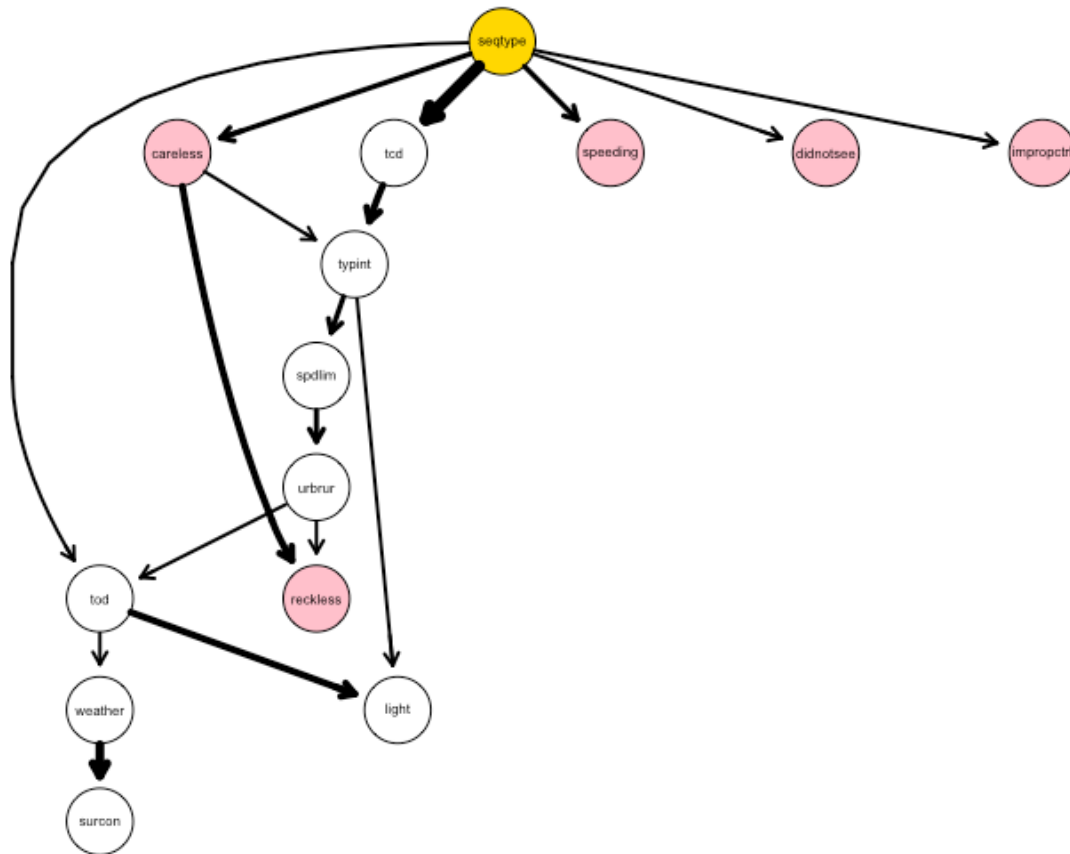


Figure 37 Bayesian network of sequence types, human factors, and environmental conditions

Figure 36 shows that the relationships between sequence types and crash outcomes did not change after removing all other variables from the original network (in Figure 34). Figure 37 shows that the local network structure changed only slightly on type of intersection (“typint”) and speed limit (“spdlim”) after crash outcome variables were removed from the original network. Therefore, the local network structures were stable, confirming the relationships between sequence types and other variables.

5.5.3 Scenario Specification

The final Bayesian network in Figure 34 is of two uses in specifying test scenarios. First, crash sequences of certain injury severity levels can be selected based on the dependencies between sequence types and crash outcomes. Second, after sequence types of interest are determined, their associated ODD attributes can be specified based on the dependencies among sequence types, human factors, and environmental conditions. The scenario specification can be done by querying the conditional probability table generated from the final Bayesian network. The process is demonstrated here with an example.

If we would like to specify scenarios for some intersection two-vehicle crashes that resulted in fatalities, we can first query the final Bayesian network to obtain the distribution of sequence types that resulted in fatalities, as shown in Figure 38. Using the “bnlearn” library in R, the distribution was generated based on Monte Carlo particle filters (198). The query was ran with 1,000 replications and obtained the average counts of sequences. The results showed that sequence types j2, j3, k3, l7, l12, and l13 were the most frequent fatal crash sequences.

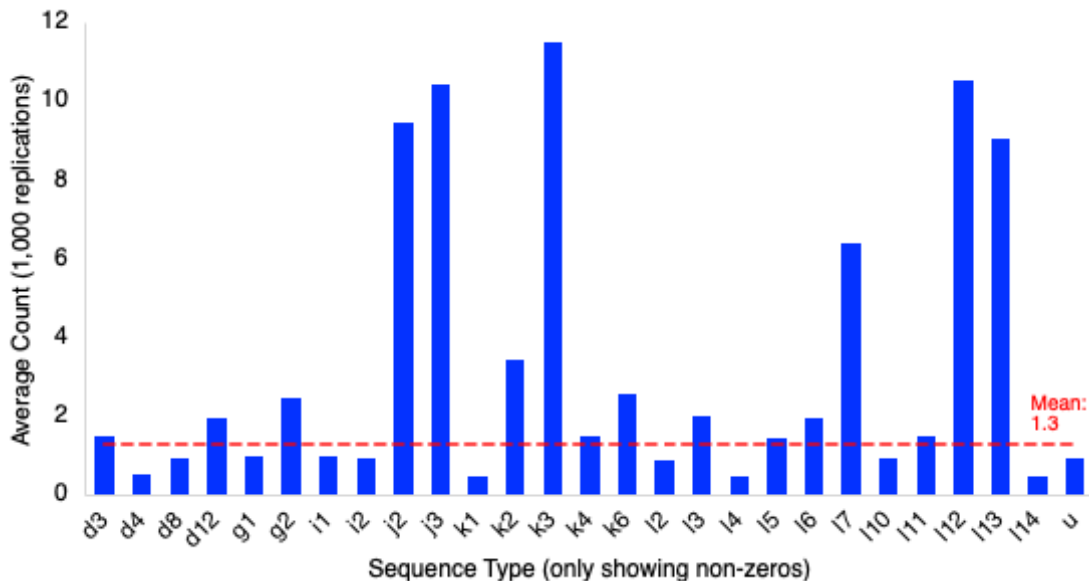


Figure 38 Distribution of sequence types resulting in fatalities

If we would like to specify ODDs for a sequence type, for example, k3, we can query the Bayesian network again for distributions of human factor and environmental condition variables using “seqtype = k3” as a criterion. Two examples are demonstrated here. Table 32 shows query results for the distribution of intersection type and traffic control device (TCD) in k3 crashes. Table 33 shows query results for the distribution of drivers’ speeding behavior and time of day in k3 crashes. The two tables are color coded to show large values in darker red, smaller values in darker blue, and values close to the mean in white. The query results showed that k3 crashes happened most frequently at 4-legged intersections with sign control on minor approaches, 3-legged intersections with sign control on minor approaches, and at 4-legged intersections with signal control on all approaches. Speeding was not related to 95% of k3 crashes. More k3 crashes happened in daytime than in

nighttime, but the proportion of speeding-related crashes in all k3 crashes were the same 2% regardless of daytime or nighttime.

Table 32 Distribution of intersection type and TCD in k3 crashes

TCD	Type of Intersection			
	4-Legged	3-Legged	Other	Unknown
No TCD+No TCD	40.5	45.9	0.6	21.8
No TCD+Sign	20.0	4.0	0.1	1.6
No TCD+Signal	1.8	0.3	0.0	0.3
Other+No TCD	0.2	0.3	0.0	0.0
Other+Other	0.5	0.6	0.0	0.4
Sign+No TCD	217.4	184.9	4.8	30.0
Sign+Other	3.3	5.3	0.1	1.4
Sign+Sign	29.6	5.8	0.8	7.9
Sign+Signal	1.9	0.7	0.0	0.3
Sign+Unknown	8.9	7.1	0.0	9.2
Signal+No TCD	3.0	0.5	0.0	0.4
Signal+Sign	0.4	0.1	0.0	0.0
Signal+Signal	128.2	15.7	0.8	22.2
Unknown+No TCD	1.8	4.2	0.1	3.8
Unknown+Sign	2.7	0.6	0.0	2.4
Unknown+Signal	1.6	0.4	0.0	0.5
Unknown+Unknown	2.8	1.1	0.1	5.1

Note: Average values of 1,000 replications.

Labels with "+" indicate conditions of Vehicle 1 on the left side and Vehicle 2 on the right side.

Table 33 Distribution of speeding behavior and time of day in k3 crashes

Time of Day	Speeding					
	N+N	N+U	U+N	U+U	N+Y	Y+N
Day	671.4	6.6	7.6	5.3	9.6	2.8
Night	146.4	1.4	1.6	1.2	2.0	0.7

Note: Average values of 1,000 replications.

With the information obtained from Table 32 and Table 33, several ODDs can be specified for the testing of k3 crash sequence type. For example, "a 4-legged intersection

with sign control on the minor approaches at daytime”. More complex queries can also be run to obtain more comprehensive descriptions of ODDs.

Given an ODD, the distribution of sequence types can also be obtained by querying the Bayesian network, and carry out tests accordingly. For example, Table 34 shows the query results of sequence type distribution at signal-controlled intersections. The results are color-coded to show the within-category (4-legged or 3-legged) distribution, with darker red indicating a higher proportion and darker blue indicating a lower proportion within category. Most frequently occurred crash sequence types for both 4-legged and 3-legged signalized intersections are i1, d12, j2, and j3. Crashes can be sampled based on this distribution and used to reconstruct scenarios in a simulation environment for AV testing.

Table 34 Distribution of sequence types at signal-controlled intersections

Seq Type	Type of Intersection		Seq Type	Type of Intersection		Seq Type	Type of Intersection	
	4-Legged	3-Legged		4-Legged	3-Legged		4-Legged	3-Legged
d1	35.7	6.9	f8	57.0	6.1	l1	13.5	1.7
d2	35.8	6.9	f9	90.9	9.8	l2	7.3	0.9
d3	45.7	8.8	g1	17.5	2.1	l3	3.5	0.4
d4	257.8	49.7	g2	13.2	1.9	l4	23.2	2.8
d5	162.8	31.6	h1	1.8	0.2	l5	8.9	1.0
d6	43.1	8.1	i1	2,176.0	263.3	l6	90.3	11.0
d7	237.7	45.3	i2	129.1	15.8	l7	14.5	1.7
d8	104.5	20.2	j1	93.1	10.7	l8	10.5	1.3
d9	72.2	13.9	j2	958.6	115.6	l9	25.2	3.0
d10	98.7	18.8	j3	816.9	98.9	l10	15.8	1.9
d11	194.5	37.3	k1	29.9	3.9	l11	98.5	12.0
d12	1,171.2	225.9	k2	194.1	23.4	l12	226.1	27.7
e3	1.1	0.2	k3	128.2	15.5	l13	453.4	55.5
f1	14.2	1.6	k4	48.3	6.0	l14	22.2	2.7
f2	23.3	2.7	k5	9.6	1.3	m1	54.9	10.4
f3	14.8	1.7	k6	119.5	14.5	u	98.0	12.7
f4	39.6	4.4	k7	33.7	4.3			
f5	7.0	0.8	k8	61.0	7.4			
f6	9.7	1.1	k9	89.2	10.9			
f7	16.9	1.8						

Note: Average values of 1,000 replications.

5.6 Discussion and Conclusions

This chapter presents a procedure to generate crash scenarios for AV safety testing. The method consists of two steps, 1) characterization of crashes encoded by sequences of events using sequence analysis techniques, and 2) specification of scenarios based on a Bayesian network modeling dependencies among crash sequence types, crash outcomes, and variables depicting ODDs. The procedure was demonstrated using 2016-2018 intersection two-vehicle crash data from the NHTSA CRSS database.

This chapter has two major findings. First, the intersection two-vehicle crashes were characterized as 55 types based on their patterns in sequences of events. The 55 crash sequence types offered more information about crash progression to the original CRSS configuration and helped identify rare crash types which would otherwise be overlooked. Second, the dependencies among crash sequence types, crash outcomes, human factors, and environmental conditions were shown in a Bayesian network. Sequence types were found to be the core of the network and have direct effects on crash outcomes. The dependencies of human factors and environmental conditions on sequence types are useful in specifying ODDs for crash sequence types and identifying distributions of crash sequence types for specific ODDs.

The contribution of this chapter is that it offers a methodology to systematically generate a crash scenario library based on national-level crash databases. Such a library would offer a comprehensive set of crash sequence types and ODDs, making it an appropriate guide for developing simulation-based tests for AV safety evaluation.

Therefore, such a scenario library would complement scenario generating methodologies developed based on vehicle kinematic data sources such as naturalistic driving data.

This chapter is subject to limitations in crash sequence data source. More detailed crash event data than what was obtained from the CRSS database (or any other current national or state-level crash database in the United States) were not available, so crash sequences used in the case study were limited to the level of details provided by CRSS. More ODD attributes would benefit the Bayesian network modeling by providing more information. However, a more complex network would require more effort to interpret. For future work, data sources with more detailed crash sequence data would be sought and used to generate more detailed crash characterization. Also, more sophisticated Bayesian network modeling techniques would be explored.

Chapter 6 Conclusions

6.1 Summary of Dissertation

Motivated by an urgent need in developing scenario-based safety verification of an emerging transportation technology – automated vehicles (AVs), this dissertation research developed a methodology for crash sequence analysis which was applied to generate test scenarios for AV safety evaluation, with a focus on evaluating the operational safety of AVs with SAE Level 3 or Level 4 automation. Level 3 and Level 4 AVs are being tested on public roads, closed courses, and in simulations, with a near-term expectation to share the roads and face similar challenges with human drivers. Using historical crash data, from national-level human-driven vehicle crash databases and the California AV collision report archives, this dissertation applied crash sequence analysis followed a scenario-generating procedure consisting of two steps – 1) characterization of crash sequences and 2) specification of dependencies between crash sequences and other crash attributes. To achieve the research objective, three studies were completed.

The first study set up the methodological foundation for this dissertation by developing a methodology to analyze crash sequences. The study further investigated the sequence encoding and dissimilarity measuring techniques to accommodate various needs in traffic crash analysis. The proposed crash sequence analysis methodology consisted of steps of 1) crash data processing, 2) sequence encoding, 3) sequence comparison, and 4) sequence clustering. The properties of nine dissimilarity measures were compared, and five Optimal Matching (OM) based measures were selected for further comparison in a case study of interstate highway single-vehicle crash sequences, with three different encoding

schemes. The five measures were categorized into two groups based on correlations between dissimilarity matrices. The optimal dissimilarity measures were identified for each encoding schemes based on the agreements with a benchmark crash typology. The case study results demonstrated the effectiveness and usefulness of the proposed crash sequence analysis methodology. This study has been written in the form of a manuscript which has been submitted to Accident Analysis and Prevention and is currently under review.

The second study applied crash sequence analysis methods to the California AV crash data, identified representative scenarios in crashes during AV field tests, and proposed a framework for scenario-based AV safety evaluation. In the study, sequence of events data was extracted from California AV collision reports and used to investigate patterns and how they may be used to develop AV test scenarios. The study evaluated 168 AV crashes (with AV in automatic driving mode before disengagement or collision) reported to the California DMV from 2015 to 2019. The analysis of subsequences showed that the most representative pattern in AV crashes was “collision following AV stop”. Analysis of event transition showed that disengagement, as an event in 24% of all studied AV crash sequences, had a transition probability of 68% to an immediate collision. Cluster analysis characterized AV crash sequences into seven groups with distinctive crash dynamic features. The cross-tabulation analysis showed that sequence types were significantly associated with crash outcomes and environmental conditions. The study concluded that crash sequences are useful for developing AV test scenarios and proposed a scenario-based AV safety testing framework with sequence of events embedded as a core

component. This study has been written in the form of a manuscript which has been published in *Accident Analysis and Prevention*.

The third study applied crash sequence analysis methodology to generate scenarios of intersection two-vehicle crash. A procedure of sequence analysis and Bayesian network modeling of sequence-ODD relationships was followed. Crash data was obtained from the 2016-2018 NHTSA CRSS database. Participating vehicles were specifically renumbered based on their intents in the crashes. Crash sequences were encoded to include detailed pre-crash events and concise crash events considering the use in scenario generation. Based on sequence patterns, the crashes were characterized as 55 types. A Bayesian network model was developed to depict the interrelationships among crash sequence types, crash outcomes, human factors, and environmental conditions. Scenarios were specified by querying the Bayesian network's conditional probability table. Distributions of ODD attributes (e.g., driver behavior, weather, lighting condition, intersection geometry, traffic control device) could be specified based on conditions of sequence types. Also, distribution of sequence types could be specified on specific crash outcomes or combinations of ODD attributes. This study has been written in the form of a manuscript which will be submitted to the *Journal of Safety Research*.

6.2 Contribution

The scholarly contribution of this dissertation is two-fold. First, it contributes to enhancing the understanding of traffic crash progression and causations by developing a first-of-its-kind crash sequence analysis methodology that is applicable to multiple use

cases. Second, it contributes to the efficient testing of AVs by applying crash sequence analysis to generate scenarios for AV safety evaluation.

The findings from this dissertation will further benefit traffic safety research and the development of an AV safety validation program. Knowledge of crash sequences will help further research in identifying appropriate safety countermeasures. A comprehensive test scenario library will speed up large-scale AV safety testing with the help of simulation. With effective and efficient testing, AVs will be deployed to public roads with safe certification.

6.3 Limitations

This dissertation is primarily subject to limitations in the crash sequence data source. The NHTSA CRSS crash database and California AV collision reports were the two data sources used in this dissertation. The two data sources each had their advantages in offering crash sequence data, making them appropriate sources for this dissertation research. The CRSS offered a comprehensive sample of crashes covering different ODDs. The California AV collision reports provided a unique sample of AV crashes otherwise cannot be obtained. Apart from the advantages, limitations of the two data sources are as follows.

1. The CRSS database did not provide detailed sequence of events in its SOE variable in the CEVENT data file, where crash events were archived. For the intersection two-vehicle crashes studied in Chapter 5, a majority of the SOE records had only one event of “collision with motor-vehicle-in-transport”. Also, detailed pre-crash events (e.g., crash avoidance maneuver) were not included in

the SOE variable. By combining the SOE with PCRASH1~3 variables, this limitation was addressed to some extent. However, further research is needed to determine appropriate sequence structures when combining multiple variables, especially for multi-vehicle crashes.

2. Reporting of crash sequence data was not required by the California DMV. No guideline was provided for including crash sequence information in text narratives. In Chapter 4's study, crash sequences were manually extracted from the narratives of AV collision reports. To ensure consistency in the extracted crash sequences, the narratives were interpreted by one researcher and a two-phase encoding process was conducted.

6.4 Future Directions

For future work, limitations in crash sequence data sources could be addressed by exploring the potential in large-scale automatic extraction of crash sequence data from crash reports using natural language processing techniques. Sequence analysis methods can also be applied to naturalistic driving data and video surveillance data for traffic conflict and crash study. Other than exploring other data sources, three more research directions to derive from this dissertation are listed as follows.

Validation of scenarios. A preliminary thought is to use driving simulator experiments to validate the scenarios. Since scenarios are developed based on historical human-driven vehicle crashes, when human drivers are tested using those scenarios, we should expect the experiments to yield a failure rate (e.g., crash rate) close to a pre-set failure rate.

AV safety performance metrics and benchmarks. The development of AV aims to create a driver with artificial intelligence (i.e., ADS) to replace human drivers partially or fully. To evaluate ADS, we would naturally want to compare the ADS with human drivers. We need to set up a set of safety performance metrics and benchmarks to match our expectations for ADS functionalities, but at the same time being realistic. Several questions involved in this research topic are:

- What performance metrics to use? Specifically, functional metrics related to the perception, reaction, planning, and control processes; and outcome measures such as crash/conflict frequency and crash injury severity.
- What benchmark do we choose? Do we use the performance of an average human driver as a benchmark, a so-called expert driver, or some other benchmark to be manually defined?

Some theoretical studies are needed to answer the above questions. Once we have the answers, we can carry out case studies with driving simulator experiments to evaluate ADS. Using scenario and simulation-based tests, experiments can be carried out (e.g., using driving simulators) to measure and compare the safety performances of ADS and human drivers.

Applications of the crash sequence analysis methodology and scenario-based testing techniques. Apart from research on testing AVs, the crash sequence analysis methodology and scenario-based testing techniques can be applied to a variety of other research topics. Two use cases are 1) evaluation of bus crash sequences and testing of bus collision avoidance systems, and 2) evaluation of non-motorized user (pedestrian, bicycle,

and e-scooter crashes) related crash sequences and testing of Internet-of-Things (IoT) applications for non-motorized user safety.

References

1. Levinson, D. Climbing Mount Next: The Effects of Autonomous Vehicles on Society. *Minn. J. L. Sci. & Tech.*, Vol. 16, 2015.
2. SAE. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles* -. Publication J3016_202104. Society of Automotive Engineers, 2021.
3. Waymo LLC. *Waymo Safety Report*. 2021.
4. General Motors. *General Motors 2018 Self-Driving Safety Report*. 2018.
5. Singh, S. *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey*. Publication DOT HS 812 506. National Highway Traffic Safety Administration, 2018.
6. Webb, N., D. Smith, C. Ludwick, T. Victor, Q. Hommes, F. Favaro, G. Ivanov, and T. Daniel. Waymo's Safety Methodologies and Safety Readiness Determinations. *arXiv preprint arXiv:2011.00054*, 2020.
7. Koopman, P., and M. Wagner. Challenges in Autonomous Vehicle Testing and Validation. *SAE International Journal of Transportation Safety*, Vol. 4, No. 1, 2016, pp. 15–24. <https://doi.org/10.4271/2016-01-0128>.
8. Koopman, P., and F. Fratrick. How Many Operational Design Domains, Objects, and Events? Presented at the SafeAI 2019: AAAI Workshop on Artificial Intelligence Safety, 2019.
9. Koopman, P., and M. Wagner. Autonomous Vehicle Safety: An Interdisciplinary Challenge. *IEEE Intelligent Transportation Systems Magazine*, Vol. 9, No. 1, 2017, pp. 90–96. <https://doi.org/10.1109/MITS.2016.2583491>.
10. Ulbrich, S., T. Menzel, A. Reschka, F. Schuldt, and M. Maurer. Defining and Substantiating the Terms Scene, Situation, and Scenario for Automated Driving. Presented at the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 2015.
11. Nitsche, P., P. Thomas, R. Stuetz, and R. Welsh. Pre-Crash Scenarios at Road Junctions: A Clustering Method for Car Crash Data. *Accident Analysis & Prevention*, Vol. 107, 2017, pp. 137–151. <https://doi.org/10.1016/j.aap.2017.07.011>.
12. Sander, U., and N. Lubbe. The Potential of Clustering Methods to Define Intersection Test Scenarios: Assessing Real-Life Performance of AEB. *Accident Analysis & Prevention*, Vol. 113, 2018, pp. 1–11. <https://doi.org/10.1016/j.aap.2018.01.010>.
13. Sui, B., N. Lubbe, and J. Bärghman. A Clustering Approach to Developing Car-to-Two-Wheeler Test Scenarios for the Assessment of Automated Emergency Braking in China Using in-Depth Chinese Crash Data. *Accident Analysis & Prevention*, Vol. 132, 2019. <https://doi.org/10.1016/j.aap.2019.07.018>.
14. Najm, W. G., J. D. Smith, and M. Yanagisawa. *Pre-Crash Scenario Typology for Crash Avoidance Research*. United States. National Highway Traffic Safety Administration, 2007.
15. Wu, K.-F., C. P. Thor, and M. N. Ardiansyah. Identify Sequence of Events Likely to Result in Severe Crash Outcomes. *Accident Analysis & Prevention*, Vol. 96, 2016, pp. 198–207. <https://doi.org/10.1016/j.aap.2016.08.009>.
16. Song, Y., M. V. Chitturi, and D. A. Noyce. Automated Vehicle Crash Sequences: Patterns and Potential Uses in Safety Testing. *Accident Analysis & Prevention*, Vol. 153, 2021, p. 106017. <https://doi.org/10.1016/j.aap.2021.106017>.

17. Koopman, P. The Heavy Tail Safety Ceiling. 2018, p. 2.
18. NHTC, and USDOT. *Ensuring American Leadership in Automated Vehicle Technologies: Automated Vehicles 4.0*. United States Government, 2020.
19. Anderson, J. M., N. Kalra, K. D. Stanley, P. Sorensen, C. Samaras, and O. A. Oluwatola. *Autonomous Vehicle Technology: A Guide for Policymakers*. Rand Corporation, Santa Monica, CA, 2014.
20. Milakis, D., B. van Arem, and B. van Wee. Policy and Society Related Implications of Automated Driving: A Review of Literature and Directions for Future Research. *Journal of Intelligent Transportation Systems*, Vol. 21, No. 4, 2017, pp. 324–348. <https://doi.org/10.1080/15472450.2017.1291351>.
21. AAA. *Consumer Sentiment on Automated Vehicles*. American Automobile Association, 2020.
22. AAA. *Automated Vehicle Survey - Phase IV*. American Automobile Association, 2019.
23. Norton, P. D. *Fighting Traffic: The Dawn of the Motor Age in the American City*. Mit Press, 2011.
24. FHWA. *Highway Statistics 2018 - STATE MOTOR-VEHICLE REGISTRATIONS*. Federal Highway Administration, 2019, p. mv1.
25. Hu, P. *Summary of Travel Trends 2001 National Household Travel Survey*. Publication ORNL/TM-2004/297, 885762. 2005, p. ORNL/TM-2004/297, 885762.
26. FHWA. *Travel Profile: United States 2017 National Household Travel Survey*. https://nhts.ornl.gov/assets/2017_USTravelProfile.pdf.
27. NHTSA. *Police-Reported Motor Vehicle Traffic Crashes in 2018*. Publication DOT HS 812860. National Highway Traffic Safety Administration, 2020.
28. Xu, J. Mortality in the United States, 2018. No. 355, 2020, p. 8.
29. CDC. Data Brief 355. Mortality in the United States, 2018 Data Tables. https://www.cdc.gov/nchs/data/databriefs/db355_tables-508.pdf#2.
30. CDC. Road Traffic Injuries and Deaths—A Global Problem. *Centers for Disease Control and Prevention*. <https://www.cdc.gov/injury/features/global-road-safety/index.html>. Accessed Nov. 11, 2020.
31. Bratland, D. Annual US Miles Traveled (Blue), Traffic Fatalities per Billion Vehicle Miles Traveled (Red), per Million People (Orange), Total Annual Deaths (Light Blue), VMT in 10s of Billions (Dark Blue) and Population in Millions (Teal), from 1921 to 2017. By Dennis Bratland - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=66179446>.
32. Hedlund, J. *Autonomous Vehicles Meet Human Drivers: Traffic Safety Issues for States*. Governors Highway Safety Association, 2017.
33. Nyholm, S., and J. Smids. Automated Cars Meet Human Drivers: Responsible Human-Robot Coordination and the Ethics of Mixed Traffic. *Ethics and Information Technology*, 2018. <https://doi.org/10.1007/s10676-018-9445-9>.
34. USDOT. Average Age of Automobiles and Trucks in Operation in the United States. <https://www.bts.gov/content/average-age-automobiles-and-trucks-operation-united-states>. Accessed Nov. 11, 2020.
35. Gordon, J. Autonomous Cars and Human Drivers: Can They Share the Road? *Telematics.com*, Dec 06, 2017.

36. Ferris, R. Manual Transmission Cars Are Disappearing, but Purists Prefer to Drive a Stick Shift. *CNBC*. <https://www.cnn.com/2020/04/15/manual-transmission-cars-are-disappearing-but-purists-prefer-to-drive-a-stick-shift.html>. Accessed Nov. 11, 2020.
37. Quach, K. Waymo Presents ChauffeurNet, a Neural Net Designed to Copy Human Driving. https://www.theregister.com/2018/12/12/waymo_presents_chauffernet/. Accessed Sep. 19, 2020.
38. University of Leeds ITS. Mimicking Human Driving in Autonomous Vehicles | Institute for Transport Studies | University of Leeds. <https://environment.leeds.ac.uk/transport/news/article/165/mimicking-human-driving-in-autonomous-vehicles>. Accessed Sep. 19, 2020.
39. Brown, B. The Social Life of Autonomous Cars. *Computer*, Vol. 50, No. 2, 2017, pp. 92–96. <https://doi.org/10.1109/MC.2017.59>.
40. Brown, B., and E. Laurier. *The Trouble with Autopilots: Assisted and Autonomous Driving on the Social Road*. New York, NY, USA, 2017.
41. Wang, J., J. Lu, F. You, and Y. Wang. *Act like a Human: Teach an Autonomous Vehicle to Deal with Traffic Encounters*. 2018.
42. Millard-Ball, A. Pedestrians, Autonomous Vehicles, and Cities. *Journal of Planning Education and Research*, Vol. 38, No. 1, 2018, pp. 6–12. <https://doi.org/10.1177/0739456X16675674>.
43. California DMV. Adopted Regulations for Testing of Autonomous Vehicles by Manufacturers. *State of California Department of Motor Vehicles - Testing of Autonomous Vehicles*. <https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/testing>. Accessed May 31, 2020.
44. California DMV. Autonomous Vehicle Collision Reports. *California DMV*. <https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports/>. Accessed Sep. 20, 2020.
45. Schoettle, B., and M. Sivak. *A Preliminary Analysis of Real-World Crashes Involving Self-Driving Vehicles*. Publication UMTRI-2015-34. The University of Michigan Transportation Research Institute, 2015.
46. Blanco, M., J. Atwood, S. M. Russell, T. Trimble, J. A. McClafferty, and M. A. Perez. *Automated Vehicle Crash Rate Comparison Using Naturalistic Data*. Virginia Tech Transportation Institute, 2016.
47. Teoh, E. R., and D. G. Kidd. Rage against the Machine? Google’s Self-Driving Cars versus Human Drivers. *Journal of Safety Research*, Vol. 63, 2017, pp. 57–60. <https://doi.org/10.1016/j.jsr.2017.08.008>.
48. Banerjee, S. S., S. Jha, J. Cyriac, Z. T. Kalbarczyk, and R. K. Iyer. Hands Off the Wheel in Autonomous Vehicles?: A Systems Perspective on over a Million Miles of Field Data. Presented at the 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Luxembourg City, 2018.
49. NTSB. *Collision Between a Car Operating with Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida, May 7, 2016*. Publication HAR1702. National Transportation Safety Board, 2017.
50. NTSB. *Collision Between a Sport Utility Vehicle Operating with Partial Driving Automation and a Crash Attenuator, Mountain View, California, March 23, 2018*. Publication HAR-20/01. National Transportation Safety Board, 2020.

51. NTSB. *Collision Between Car Operating with Partial Driving Automation and Truck-Tractor Semitrailer, Delray Beach, Florida, March 1, 2019*. Publication HAB-20/01. National Transportation Safety Board, 2020.
52. NTSB. *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018*. Publication HAR-19/03. National Transportation Safety Board, 2019.
53. NHTSA. Automated Vehicles Transparency and Engagement for Safe Testing Initiative. *AV TEST Initiative*. <https://www.nhtsa.gov/automated-vehicles-safety/av-test>. Accessed Jun. 18, 2020.
54. Koopman, P., and M. Wagner. Toward a Framework for Highly Automated Vehicle Safety Validation. Presented at the WCX World Congress Experience, 2018.
55. Etherington, D. Over 1,400 Self-Driving Vehicles Are Now in Testing by 80+ Companies across the US. TechCrunch, Jun 11, 2019.
56. Fogarty, K. How Many Test Miles Make A Vehicle Safe? Semiconductor Engineering, Aug 06, 2019.
57. NHTSA. FMVSS. <https://www.nhtsa.gov/laws-regulations/fmvss>. Accessed Nov. 14, 2020.
58. Kalra, N., and S. M. Paddock. *Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?* Publication RR-1478-RC. RAND Corporation, 2016.
59. Etherington, D. Waymo Has Now Driven 10 Billion Autonomous Miles in Simulation. TechCrunch, Jul 10, 2019.
60. Winner, H., K. Lemmer, T. Form, and J. Mazzega. PEGASUS—First Steps for the Safe Introduction of Automated Driving. In *Road Vehicle Automation 5* (G. Meyer and S. Beiker, eds.), Springer International Publishing, Cham, pp. 185–195.
61. Mira, H., and R. Hillman. *Test Methods for Interrogating Autonomous Vehicle Behaviour – Findings from the HumanDrive Project*. 2020.
62. Lenard, J., A. Badea-Romero, and R. Danton. Typical Pedestrian Accident Scenarios for the Development of Autonomous Emergency Braking Test Protocols. *Accident Analysis & Prevention*, Vol. 73, 2014, pp. 73–80. <https://doi.org/10.1016/j.aap.2014.08.012>.
63. Nitsche, P., R. H. Welsh, A. Genser, and P. D. Thomas. A Novel, Modular Validation Framework for Collision Avoidance of Automated Vehicles at Road Junctions. Presented at the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, 2018.
64. Cornwell, B. *Social Sequence Analysis: Methods and Applications*. Cambridge University Press, 2015.
65. Abbott, A. Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, Vol. 16, No. 4, 1983, pp. 129–147. <https://doi.org/10.1080/01615440.1983.10594107>.
66. Abbott, A. Sequence Analysis: New Methods for Old Ideas. *Annual Review of Sociology*, Vol. 21, No. 1, 1995, pp. 93–113. <https://doi.org/10.1146/annurev.so.21.080195.000521>.
67. Pearson, W. R., and D. J. Lipman. Improved Tools for Biological Sequence Comparison. *Proceedings of the National Academy of Sciences*, Vol. 85, No. 8, 1988, pp. 2444–2448. <https://doi.org/10.1073/pnas.85.8.2444>.

68. Abbott, A., and A. Tsay. Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. *Sociological Methods & Research*, Vol. 29, No. 1, 2000, pp. 3–33. <https://doi.org/10.1177/0049124100029001001>.
69. Wu, L. L. Some Comments on “Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect.” *Sociological Methods & Research*, Vol. 29, No. 1, 2000, pp. 41–64. <https://doi.org/10.1177/0049124100029001003>.
70. Wu, K.-F., and C. P. Thor. Method for the Use of Naturalistic Driving Study Data to Analyze Rear-End Crash Sequences. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2518, No. 1, 2015, pp. 27–36. <https://doi.org/10.3141/2518-04>.
71. Chapman, A. L. An Epidemiological Approach to Traffic Safety. *Public Health Reports*, Vol. 69, No. 8, 1954, pp. 773–775.
72. Kleinbaum, D. G., L. L. Kupper, and H. Morgenstern. *Epidemiologic Research: Principles and Quantitative Methods*. John Wiley & Sons, 1982.
73. Haddon Jr, W. The Changing Approach to the Epidemiology, Prevention, and Amelioration of Trauma: The Transition to Approaches Etiologically Rather than Descriptively Based. *American journal of public health and the Nations health*, Vol. 58, No. 8, 1968, pp. 1431–1438.
74. Haddon Jr, W. A Logical Framework for Categorizing Highway Safety Phenomena and Activity. *Journal of Trauma and Acute Care Surgery*, Vol. 12, No. 3, 1972, pp. 193–207.
75. Haddon, W. Energy Damage and the Ten Countermeasure Strategies. *Human Factors*, Vol. 15, No. 4, 1973, pp. 355–366. <https://doi.org/10.1177/001872087301500407>.
76. Treat, J. R., N. S. Tumbas, S. T. McDonald, D. Shinar, R. D. Hume, R. E. Mayer, R. L. Stansifer, and N. J. Castellan. Tri-Level Study of the Causes of Traffic Accidents: Final Report. Volume I: Casual Factor Tabulations and Assessments: (488172008-001). American Psychological Association, , 1979.
77. NTSB. *Safety Recommendation H-11-028*. Publication H-11-28. National Transportation Safety Board, 2011.
78. Wu, K.-F., and P. P. Jovanis. Crashes and Crash-Surrogate Events: Exploratory Modeling with Naturalistic Driving Data. *Accident Analysis & Prevention*, Vol. 45, 2012, pp. 507–516. <https://doi.org/10.1016/j.aap.2011.09.002>.
79. Wu, K.-F., and P. P. Jovanis. Defining and Screening Crash Surrogate Events Using Naturalistic Driving Data. *Accident Analysis & Prevention*, Vol. 61, 2013, pp. 10–22. <https://doi.org/10.1016/j.aap.2012.10.004>.
80. Wu, K.-F., L. Sasidharan, C. P. Thor, and S.-Y. Chen. Crash Sequence Based Risk Matrix for Motorcycle Crashes. *Accident Analysis & Prevention*, Vol. 117, 2018, pp. 21–31. <https://doi.org/10.1016/j.aap.2018.03.022>.
81. Krull, K. A., A. J. Khattak, and F. M. Council. Injury Effects of Rollovers and Events Sequence in Single-Vehicle Crashes. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1717, No. 1, 2000, pp. 46–54. <https://doi.org/10.3141/1717-07>.
82. Heinrich, H. W. Industrial Accident Prevention. A Scientific Approach. *Industrial Accident Prevention. A Scientific Approach.*, No. Second Edition, 1941.
83. Duffy, S. F. Section 03 - Theories of Accident Causation. Cleveland, OH, , 2007.
84. DeCamp, W., and K. Herskovitz. The Theories of Accident Causation. In *Security Supervision and Management*, Elsevier, pp. 71–78.

85. Pillay, M. Accident Causation, Prevention and Safety Management: A Review of the State-of-the-Art. *Procedia Manufacturing*, Vol. 3, 2015, pp. 1838–1845. <https://doi.org/10.1016/j.promfg.2015.07.224>.
86. Porter, R. J., S. Himes, A. Musunuru, T. Le, K. Eccles, K. Peach, I. Tasic, M. Zlatkovic, K. Tatineni, and B. Duffy. *Understanding the Causative, Precipitating, and Predisposing Factors in Rural Two-Lane Crashes*. United States. Federal Highway Administration, 2018.
87. Snyder, M. B., and R. L. Knoblauch. *Pedestrian Safety: The Identification of Precipitating Factors and Possible Countermeasures. Volume 1: Final Report*. United States. National Highway Traffic Safety Administration, 1971.
88. Cross, K. D., and G. Fisher. *A Study of Bicycle/Motor-Vehicle Accidents: Identification of Problem Types and Countermeasure Approaches*. National Highway Traffic Safety Administration, 1977.
89. Davis, G. A. Using Bayesian Networks to Identify the Causal Effect of Speeding in Individual Vehicle/Pedestrian Collisions. Presented at the The Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI2001), 2001.
90. Davis, G. A. Bayesian Reconstruction of Traffic Accidents. *Law, Probability and Risk*, Vol. 2, No. 2, 2003, pp. 69–89. <https://doi.org/10.1093/lpr/2.2.69>.
91. Davis, G. A., J. Hourdos, H. Xiong, and I. Chatterjee. Outline for a Causal Model of Traffic Conflicts and Crashes. *Accident Analysis & Prevention*, Vol. 43, No. 6, 2011, pp. 1907–1919. <https://doi.org/10.1016/j.aap.2011.05.001>.
92. Menzel, T., G. Bagschik, and M. Maurer. Scenarios for Development, Test and Validation of Automated Vehicles. Presented at the 2018 IEEE Intelligent Vehicles Symposium, 2018.
93. Sauerbier, J., J. Bock, H. Weber, and L. Eckstein. Definition of Scenarios for Safety Validation of Automated Driving Functions. *ATZ worldwide*, Vol. 121, No. 1, 2019, pp. 42–45. <https://doi.org/10.1007/s38311-018-0197-2>.
94. Xia, Q., J. Duan, F. Gao, Q. Hu, and Y. He. Test Scenario Design for Intelligent Driving System Ensuring Coverage and Effectiveness. *International Journal of Automotive Technology*, Vol. 19, No. 4, 2018, pp. 751–758. <https://doi.org/10.1007/s12239-018-0072-6>.
95. Thorn, E., S. C. Kimmel, M. Chaka, and B. A. Hamilton. *A Framework for Automated Driving System Testable Cases and Scenarios*. United States. Department of Transportation. National Highway Traffic Safety ..., 2018.
96. Perez, M., L. S. Angell, J. Hankey, C. A. Green, M. L. Neurauter, R. K. Deering, and R. E. Llaneras. *Advanced Crash Avoidance Technologies (ACAT) Program-Final Report of the GM-VTTI Backing Crash Countermeasures Project*. United States. National Highway Traffic Safety Administration, 2011.
97. Najm, W. G., S. Toma, and J. Brewer. *Depiction of Priority Light-Vehicle Pre-Crash Scenarios for Safety Applications Based on Vehicle-to-Vehicle Communications*. United States. National Highway Traffic Safety Administration, 2013.
98. Najm, W. G., R. Ranganathan, G. Srinivasan, J. D. Smith, S. Toma, E. Swanson, and A. Burgett. *Description of Light-Vehicle Pre-Crash Scenarios for Safety Applications Based on Vehicle-to-Vehicle Communications*. United States. National Highway Traffic Safety Administration, 2013.

99. Kusano, K. D., and H. C. Gabler. Field Relevance of the New Car Assessment Program Lane Departure Warning Confirmation Test. *SAE International Journal of Passenger Cars - Mechanical Systems*, Vol. 5, No. 1, 2012, pp. 253–264. <https://doi.org/10.4271/2012-01-0284>.
100. Kusano, K. D., and H. C. Gabler. Characterization of Lane Departure Crashes Using Event Data Recorders Extracted from Real-World Collisions. *SAE International journal of passenger cars-mechanical systems*, Vol. 6, No. 2013-01-0730, 2013, pp. 705–713.
101. Kusano, K. D., and H. C. Gabler. Characterization of Opposite-Direction Road Departure Crashes in the United States. *Transportation Research Record*, Vol. 2377, No. 1, 2013, pp. 14–20. <https://doi.org/10.3141/2377-02>.
102. Kusano, K. D., and H. C. Gabler. Comprehensive Target Populations for Current Active Safety Systems Using National Crash Databases. *Traffic Injury Prevention*, Vol. 15, No. 7, 2014, pp. 753–761. <https://doi.org/10.1080/15389588.2013.871003>.
103. Kurt, A., G. Özbilgin, K. A. Redmill, R. Sherony, and Ü. Özgüner. Test Scenarios, Equipment and Testing Process for LDW LDP Performance Evaluation. Presented at the SAE 2015 World Congress & Exhibition, 2015.
104. MacAlister, A., and D. Zuby. Cyclist Crash Scenarios and Factors Relevant to the Design of Cyclist Detection Systems. 2015, p. 12.
105. Aust, M. L. Generalization of Case Studies in Road Traffic When Defining Pre-Crash Scenarios for Active Safety Function Evaluation. *Accident Analysis & Prevention*, Vol. 42, No. 4, 2010, pp. 1172–1183. <https://doi.org/10.1016/j.aap.2010.01.006>.
106. Dozza, M., and N. P. González. Recognising Safety Critical Events: Can Automatic Video Processing Improve Naturalistic Data Analyses? *Accident Analysis & Prevention*, Vol. 60, 2013, pp. 298–304. <https://doi.org/10.1016/j.aap.2013.02.014>.
107. Elrofai, H., D. Worm, and O. Op den Camp. Scenario Identification for Validation of Automated Driving Functions. In *Advanced Microsystems for Automotive Applications 2016* (T. Schulze, B. Müller, and G. Meyer, eds.), Springer International Publishing, Cham, pp. 153–163.
108. Liu, L., X. Zhu, and Z. Ma. Study on Test Scenarios of Environment Perception System under Rear-End Collision Risk. Presented at the WCX World Congress Experience, 2018.
109. Zhou, J., and L. del Re. Identification of Critical Cases of ADAS Safety by FOT Based Parameterization of a Catalogue. Presented at the 2017 11th Asian Control Conference (ASCC), Gold Coast, QLD, 2017.
110. Langner, J., J. Bach, L. Ries, S. Otten, M. Holzapfel, and E. Sax. Estimating the Uniqueness of Test Scenarios Derived from Recorded Real-World-Driving-Data Using Autoencoders. Presented at the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, 2018.
111. de Gelder, E., and J.-P. Paardekooper. Assessment of Automated Driving Systems Using Real-Life Scenarios. Presented at the 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 2017.
112. de Gelder, E., J. Manders, C. Grappiolo, J.-P. Paardekooper, O. O. den Camp, and B. De Schutter. Real-World Scenario Mining for the Assessment of Automated Vehicles. *arXiv:2006.00483 [cs]*, 2020.
113. Ponn, T., M. Breitfuß, X. Yu, and F. Diermeyer. Identification of Challenging Highway-Scenarios for the Safety Validation of Automated Vehicles Based on Real Driving Data. *arXiv:2008.11609 [cs]*, 2020.

114. Amersbach, C., and H. Winner. Defining Required and Feasible Test Coverage for Scenario-Based Validation of Highly Automated Vehicles*. Presented at the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2019.
115. Ponn, T., A. Schwab, F. Diermeyer, C. Gndt, and J. Záhorský. A Method for the Selection of Challenging Driving Scenarios for Automated Vehicles Based on an Objective Characterization of the Driving Behavior. 2019.
116. Feng, S., Y. Feng, C. Yu, Y. Zhang, and H. X. Liu. Testing Scenario Library Generation for Connected and Automated Vehicles, Part I: Methodology. *IEEE Transactions on Intelligent Transportation Systems*, 2020, pp. 1–10.
<https://doi.org/10.1109/TITS.2020.2972211>.
117. Feng, S., Y. Feng, H. Sun, S. Bao, Y. Zhang, and H. X. Liu. Testing Scenario Library Generation for Connected and Automated Vehicles, Part II: Case Studies. *IEEE Transactions on Intelligent Transportation Systems*, 2020, pp. 1–13.
<https://doi.org/10.1109/TITS.2020.2988309>.
118. Tuncali, C. E., T. P. Pavlic, and G. Fainekos. Utilizing S-TaLiRo as an Automatic Test Generation Framework for Autonomous Vehicles. Presented at the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 2016.
119. Tuncali, C. E., G. Fainekos, H. Ito, and J. Kapinski. Simulation-Based Adversarial Test Generation for Autonomous Vehicles with Machine Learning Components. *arXiv:1804.06760 [cs]*, 2019.
120. Mullins, G. E., P. G. Stankiewicz, and S. K. Gupta. Automated Generation of Diverse and Challenging Scenarios for Test and Evaluation of Autonomous Vehicles. Presented at the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, Singapore, 2017.
121. Althoff, M., and S. Lutz. Automatic Generation of Safety-Critical Test Scenarios for Collision Avoidance of Road Vehicles. Presented at the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, 2018.
122. Hallerbach, S., Y. Xia, U. Eberle, and F. Koester. Simulation-Based Identification of Critical Scenarios for Cooperative and Automated Vehicles. *SAE International Journal of Connected and Automated Vehicles*, Vol. 1, No. 2, 2018, pp. 93–106.
<https://doi.org/10.4271/2018-01-1066>.
123. Jenkins, I. R., L. O. Gee, A. Knauss, H. Yin, and J. Schroeder. Accident Scenario Generation with Recurrent Neural Networks. Presented at the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, 2018.
124. Ponn, T., D. Fratzke, C. Gndt, and M. Lienkamp. Towards Certification of Autonomous Driving: Systematic Test Case Generation for a Comprehensive but Economically-Feasible Assessment of Lane Keeping Assist Algorithms: Presented at the 5th International Conference on Vehicle Technology and Intelligent Transport Systems, Heraklion, Crete, Greece, 2019.
125. Tao, J., Y. Li, F. Wotawa, H. Felbinger, and M. Nica. On the Industrial Application of Combinatorial Testing for Autonomous Driving Functions. Presented at the 2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 2019.
126. Abdel-Aty, M., J. Keller, and P. A. Brady. Analysis of Types of Crashes at Signalized Intersections by Using Complete Crash Data and Tree-Based Regression. *Transportation*

- Research Record*, Vol. 1908, No. 1, 2005, pp. 37–45.
<https://doi.org/10.1177/0361198105190800105>.
127. Carrick, G., S. Srinivasan, and S. S. Washburn. Law Enforcement Vehicle Crashes in Florida: Descriptive Analysis and Characterization. *Transportation Research Record*, Vol. 2182, No. 1, 2010, pp. 40–47. <https://doi.org/10.3141/2182-06>.
 128. Klena II, T., and J. Woodrooffe. *Characterization of Commercial Vehicle Crashes and Driver Injury*. SAE Technical Paper, 2011.
 129. Needleman, S. B., and C. D. Wunsch. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology*, Vol. 48, No. 3, 1970, pp. 443–453. [https://doi.org/10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4).
 130. Studer, M., and G. Ritschard. A Comparative Review of Sequence Dissimilarity Measures. 2014. <https://doi.org/10.12682/LIVES.2296-1658.2014.33>.
 131. Levenshtein, V. I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. No. 10, 1965, pp. 707–710.
 132. Hamming, R. W. Error Detecting and Error Correcting Codes. *The Bell System Technical Journal*, Vol. 29, No. 2, 1950, pp. 147–160. <https://doi.org/10.1002/j.1538-7305.1950.tb00463.x>.
 133. Kruskal, J. B. An Overview of Sequence Comparison: Time Warps, String Edits, and Macromolecules. *SIAM Review*, Vol. 25, No. 2, 1983, pp. 201–237. <https://doi.org/10.1137/1025045>.
 134. Elzinga, C. H. Sequence Similarity: A Nonaligning Technique. *Sociological methods & research*, Vol. 32, No. 1, 2003, pp. 3–29.
 135. Elzinga, C. H. Combinatorial Representations of Token Sequences. *Journal of Classification*, Vol. 22, No. 1, 2005, pp. 87–118. <https://doi.org/10.1007/s00357-005-0007-6>.
 136. Yujian, L., and L. Bo. A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 6, 2007, pp. 1091–1095. <https://doi.org/10.1109/TPAMI.2007.1078>.
 137. Hollister, M. Is Optimal Matching Suboptimal? *Sociological Methods & Research*, Vol. 38, No. 2, 2009, pp. 235–264. <https://doi.org/10.1177/0049124109346164>.
 138. Gauthier, J.-A., E. D. Widmer, P. Bucher, and C. Notredame. Multichannel Sequence Analysis Applied to Social Science Data. *Sociological Methodology*, Vol. 40, No. 1, 2010, pp. 1–38. <https://doi.org/10.1111/j.1467-9531.2010.01227.x>.
 139. Elzinga, C. H., and M. Studer. Spell Sequences, State Proximities, and Distance Metrics. *Sociological Methods & Research*, Vol. 44, No. 1, 2015, pp. 3–47. <https://doi.org/10.1177/0049124114540707>.
 140. Blackburne, B. P., and S. Whelan. Measuring the Distance between Multiple Sequence Alignments. *Bioinformatics*, Vol. 28, No. 4, 2012, pp. 495–502. <https://doi.org/10.1093/bioinformatics/btr701>.
 141. Robette, N., and X. Bry. Harpoon or Bait? A Comparison of Various Metrics in Fishing for Sequence Patterns. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, Vol. 116, No. 1, 2012, pp. 5–24. <https://doi.org/10.1177/0759106312454635>.
 142. Studer, M., and G. Ritschard. What Matters in Differences between Life Trajectories: A Comparative Review of Sequence Dissimilarity Measures. *Journal of the Royal*

- Statistical Society. A*, Vol. 179, No. 2, 2016, pp. 481–511.
<https://doi.org/10.1111/rssa.12125>.
143. Kang, W., S. Rey, L. Wolf, E. Knaap, and S. Han. Sensitivity of Sequence Methods in the Study of Neighborhood Change in the United States. *Computers, Environment and Urban Systems*, Vol. 81, 2020, p. 101480.
<https://doi.org/10.1016/j.compenvurbsys.2020.101480>.
 144. NHTSA. *Crash Report Sampling System CRSS Analytical User's Manual 2016-2018*. 2018, p. 305.
 145. Scanlon, J. M., K. D. Kusano, T. Daniel, C. Alderson, A. Ogle, and T. Victor. Waymo Simulated Driving Behavior in Reconstructed Fatal Crashes within an Autonomous Vehicle Operating Domain. 2021, p. 24.
 146. Turner, D. S., and J. W. Hall. *NCHRP Synthesis 202: Severity Indices for Roadside Features*. Transportation Research Board, 1994, p. 68.
 147. Council, F. M., and J. R. Stewart. Severity Indexes for Roadside Objects. *Transportation Research Record*, Vol. 1528, No. 1, 1996, pp. 87–96.
<https://doi.org/10.1177/0361198196152800109>.
 148. Holdridge, J. M., V. N. Shankar, and G. F. Ulfarsson. The Crash Severity Impacts of Fixed Roadside Objects. *Journal of Safety Research*, Vol. 36, No. 2, 2005, pp. 139–147.
<https://doi.org/10.1016/j.jsr.2004.12.005>.
 149. Stigson, H., A. Ydenius, and A. Kullgren. Variation of Crash Severity and Injury Risk Depending on Collisions with Different Vehicle Types and Objects. 2006, p. 12.
 150. Qin, X., K. Wang, and C. E. Cutler. Analysis of Crash Severity Based on Vehicle Damage and Occupant Injuries. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2386, No. 1, 2013, pp. 95–102.
<https://doi.org/10.3141/2386-11>.
 151. Wang, K., and X. Qin. Use of Structural Equation Modeling to Measure Severity of Single-Vehicle Crashes. *Transportation Research Record*, Vol. 2432, No. 1, 2014, pp. 17–25. <https://doi.org/10.3141/2432-03>.
 152. Deville, J.-C., and G. Saporta. Correspondence Analysis, with an Extension towards Nominal Time Series. *Journal of econometrics*, Vol. 22, No. 1–2, 1983, pp. 169–189.
 153. Bergroth, L., H. Hakonen, and T. Raita. A Survey of Longest Common Subsequence Algorithms. Presented at the Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000, 2000.
 154. Levine, J. H. But What Have You Done for Us Lately? Commentary on Abbott and Tsay. *Sociological methods & research*, Vol. 29, No. 1, 2000, pp. 34–40.
 155. Abbott, A. Reply to Levine and Wu. *Sociological methods & research*, Vol. 29, No. 1, 2000, pp. 65–76.
 156. Rousset, P., J.-F. Giret, and Y. Grelet. Typologies de Parcours et Dynamique Longitudinale. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, Vol. 114, No. 1, 2012, pp. 5–34.
<https://doi.org/10.1177/0759106312437142>.
 157. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.
 158. Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller. Mining Sequence Data in R with the TraMineR Package: A User's Guide. <http://mephisto.unige.ch/traminer>.

159. Thioulouse, J., S. Dray, A.-B. Dufour, A. Siberchicot, T. Jombart, and S. Pavoine. *Multivariate Analysis of Ecological Data with Ade4*. Springer, 2018.
160. Studer, M. *WeightedCluster Library Manual: A Practical Guide to Creating Typologies of Trajectories in the Social Sciences with R*. 2013.
161. Scrucca, L., M. Fop, T. B. Murphy, and A. E. Raftery. Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R journal*, Vol. 8, No. 1, 2016, pp. 289–317.
162. Mantel, N. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, Vol. 27, No. 2 Part 1, 1967, pp. 209–220.
163. Guillot, G., and F. Rousset. Dismantling the Mantel Tests. *Methods in Ecology and Evolution*, Vol. 4, No. 4, 2013, pp. 336–344. <https://doi.org/10.1111/2041-210x.12018>.
164. Hubert, L., and P. Arabie. Comparing Partitions. *Journal of Classification*, Vol. 2, No. 1, 1985, pp. 193–218. <https://doi.org/10.1007/BF01908075>.
165. Yeung, K. Y., and W. L. Ruzzo. Details of the Adjusted Rand Index and Clustering Algorithms Supplement to the Paper “An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data.” *Bioinformatics*, Vol. 17, No. 9, 2001, pp. 763–774.
166. Warrens, M. J. On the Equivalence of Cohen’s Kappa and the Hubert-Arabie Adjusted Rand Index. *Journal of Classification*, Vol. 25, No. 2, 2008, pp. 177–183. <https://doi.org/10.1007/s00357-008-9023-7>.
167. Hennig, C. *Cluster Validation: How to Think and What to Do*. Oviedo, Spain, Aug, 2016.
168. Schwall, M., T. Daniel, T. Victor, F. Favaro, and H. Hohnhold. Waymo Public Road Safety Performance Data. *arXiv preprint arXiv:2011.00038*, 2020.
169. Favarò, F. M., N. Nader, S. O. Eurich, M. Tripp, and N. Varadaraju. Examining Accident Reports Involving Autonomous Vehicles in California. *PLOS ONE*, Vol. 12, No. 9, 2017. <https://doi.org/10.1371/journal.pone.0184952>.
170. Das, S., A. Dutta, and I. Tsapakis. Automated Vehicle Collisions in California: Applying Bayesian Latent Class Model. *IATSS Research*, 2020. <https://doi.org/10.1016/j.iatssr.2020.03.001>.
171. Boggs, A. M., B. Wali, and A. J. Khattak. Exploratory Analysis of Automated Vehicle Crashes in California: A Text Analytics & Hierarchical Bayesian Heterogeneity-Based Approach. *Accident Analysis & Prevention*, Vol. 135, 2020. <https://doi.org/10.1016/j.aap.2019.105354>.
172. Alambeigi, H., A. D. McDonald, and S. R. Tankasala. Crash Themes in Automated Vehicles: A Topic Modeling Analysis of the California Department of Motor Vehicles Automated Vehicle Crash Database. *arXiv:2001.11087 [stat]*, 2020.
173. Wang, S., and Z. Li. Exploring Causes and Effects of Automated Vehicle Disengagement Using Statistical Modeling and Classification Tree Based on Field Test Data. *Accident Analysis & Prevention*, Vol. 129, 2019, pp. 44–54. <https://doi.org/10.1016/j.aap.2019.04.015>.
174. Dixit, V. V., S. Chand, and D. J. Nair. Autonomous Vehicles: Disengagements, Accidents and Reaction Times. *PLOS ONE*, Vol. 11, No. 12, 2016. <https://doi.org/10.1371/journal.pone.0168054>.

175. Leilabadi, S. H., and S. Schmidt. In-Depth Analysis of Autonomous Vehicle Collisions in California. Presented at the 2019 IEEE Intelligent Transportation Systems Conference - ITSC, Auckland, New Zealand, 2019.
176. Favarò, F., S. Eurich, and N. Nader. Autonomous Vehicles' Disengagements: Trends, Triggers, and Regulatory Limitations. *Accident Analysis & Prevention*, Vol. 110, 2018, pp. 136–148. <https://doi.org/10.1016/j.aap.2017.11.001>.
177. Boggs, A. M., R. Arvin, and A. J. Khattak. Exploring the Who, What, When, Where, and Why of Automated Vehicle Disengagements. *Accident Analysis & Prevention*, Vol. 136, 2020. <https://doi.org/10.1016/j.aap.2019.105406>.
178. Lv, C., D. Cao, Y. Zhao, D. J. Auger, M. Sullman, H. Wang, L. M. Dutka, L. Skrypchuk, and A. Mouzakitis. Analysis of Autopilot Disengagements Occurring during Autonomous Vehicle Testing. *IEEE/CAA Journal of Automatica Sinica*, Vol. 5, No. 1, 2018, pp. 58–68. <https://doi.org/10.1109/JAS.2017.7510745>.
179. Dinges, J. T., and N. J. Durisek. Automated Vehicle Disengagement Reaction Time Compared to Human Brake Reaction Time in Both Automobile and Motorcycle Operation. Presented at the WCX SAE World Congress Experience, 2019.
180. Wang, S., and Z. Li. Exploring the Mechanism of Crashes with Automated Vehicles Using Statistical Modeling Approaches. *PLOS ONE*, Vol. 14, No. 3, 2019. <https://doi.org/10.1371/journal.pone.0214550>.
181. Kaufmann, L., and P. Rousseeuw. Clustering by Means of Medoids. In *Data Analysis based on the L1-Norm and Related Methods*, North-Holland, pp. 405–416.
182. Phillips, J. M. *Mathematical Foundations for Data Analysis*. 2019.
183. Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, Vol. 20, 1987, pp. 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
184. NHTSA. *Automated Driving Systems: A Vision for Safety*. US Department of Transportation, 2017.
185. Menzel, T., G. Bagschik, L. Isensee, A. Schomburg, and M. Maurer. From Functional to Logical Scenarios: Detailing a Keyword-Based Scenario Description for Execution in a Simulation Environment. Presented at the 2019 IEEE Intelligent Vehicles Symposium (IV), 2019.
186. Riedmaier, S., T. Ponn, D. Ludwig, B. Schick, and F. Diermeyer. Survey on Scenario-Based Safety Assessment of Automated Vehicles. *IEEE Access*, Vol. 8, 2020, pp. 87456–87477. <https://doi.org/10.1109/ACCESS.2020.2993730>.
187. Watanabe, H., L. Tobisch, J. Rost, J. Wallner, and G. Prokop. Scenario Mining for Development of Predictive Safety Functions. Presented at the 2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES), Cairo, Egypt, 2019.
188. Pearl, J. *Causality*. Cambridge university press, 2009.
189. Zhu, S., X. Cai, J. Lu, and Y. Peng. Analysis of Factors Affecting Serious Multi-Fatality Crashes in China Based on Bayesian Network Structure. *Advances in Mechanical Engineering*, Vol. 9, No. 6, 2017, p. 1687814017704145. <https://doi.org/10.1177/1687814017704145>.
190. Zou, X., and W. L. Yue. A Bayesian Network Approach to Causation Analysis of Road Accidents Using Netica. *Journal of Advanced Transportation*. Volume 2017, e2525481. <https://www.hindawi.com/journals/jat/2017/2525481/>. Accessed Oct. 1, 2020.

191. Ma, X., Y. Xing, and J. Lu. Causation Analysis of Hazardous Material Road Transportation Accidents by Bayesian Network Using Genie. *Journal of Advanced Transportation*. Volume 2018, e6248105. <https://www.hindawi.com/journals/jat/2018/6248105/>. Accessed Oct. 1, 2020.
192. Zong, F., X. Chen, J. Tang, P. Yu, and T. Wu. Analyzing Traffic Crash Severity With Combination of Information Entropy and Bayesian Network. *IEEE Access*, Vol. 7, 2019, pp. 63288–63302. <https://doi.org/10.1109/ACCESS.2019.2916691>.
193. de Oña, J., G. López, R. Mujalli, and F. J. Calvo. Analysis of Traffic Accidents on Rural Highways Using Latent Class Clustering and Bayesian Networks. *Accident Analysis & Prevention*, Vol. 51, 2013, pp. 1–10. <https://doi.org/10.1016/j.aap.2012.10.016>.
194. Prati, G., L. Pietrantonio, and F. Fraboni. Using Data Mining Techniques to Predict the Severity of Bicycle Crashes. *Accident Analysis & Prevention*, Vol. 101, 2017, pp. 44–54. <https://doi.org/10.1016/j.aap.2017.01.008>.
195. Simoncic, M. A Bayesian Network Model of Two-Car Accidents. *Journal of transportation and Statistics*, Vol. 7, No. 2/3, 2004, pp. 13–25.
196. Fenton, N., and M. Neil. The Use of Bayes and Causal Modelling in Decision Making, Uncertainty and Risk. *CEPIS Upgrade*, Vol. 12, No. 5, 2011, pp. 10–21.
197. Pearl, J. Bayesian Networks: A Model Cf Self-Activated Memory for Evidential Reasoning. 1985.
198. Scutari, M. Learning Bayesian Networks with the Bnlearn R Package. *Journal of Statistical Software*, Vol. 35, No. 1, 2010, pp. 1–22. <https://doi.org/10.18637/jss.v035.i03>.
199. Hansen, K. D., J. Gentry, L. Long, R. Gentleman, S. Falcon, F. Hahne, and D. Sarkar. 2.12.0. *Rgraphviz: Provides Plotting Capabilities for R Graph Objects*. R.

Appendices

Appendix for Chapter 3

Table A - 1 Encoding Schemes

PCRASH1

Original	Scheme 1	Scheme 2	CRSS Description
11p0	N	N	0 No Driver Present/Unknown if Driver Present
11p1	ST	ST	1 Going Straight
11p2	B	B	2 Decelerating in Road
11p3	A	A	3 Accelerating in Road
11p4	A	A	4 Starting in Road
11p5	S	S	5 Stopped in Roadway
11p6	PA	PA	6 Passing or Overtaking Another Vehicle
11p10	R	R	10 Turning Right
11p11	L	L	11 Turning Left
11p12	U	U	12 Making a U-turn
11p13	BU	BU	13 Backing Up (Other Than for Parking Position)
11p14	C	C	14 Negotiating a Curve
11p15	E	E	15 Changing Lanes
11p16	E	E	16 Merging
11p17	CA	CA	17 Successful Corrective Action to a Previous Critical Event
11p98	N	N	98 Other

PCRASH2

Original	Scheme 1	Scheme 2	CRSS Description
			<i>Loss of Control:</i>
12p1	LCS	LC	1 Blow Out/Flat Tire
12p3	LCS	LC	3 Disabling Vehicle Failure (e.g., Wheel Fell Off)
12p4	LCM	LC	4 Non-Disabling Vehicle Problem (e.g., Hood Flew Up)
12p5	LCM	LC	5 Poor Road Conditions (Puddle, Pothole, Ice, etc.)
12p6	LCF	LC	6 Traveling Too Fast for Conditions
12p8	LCO	LC	8 Other Cause of Control Loss
12p9	LCO	LC	9 Unknown Cause of Control Loss
			<i>This Vehicle Traveling:</i>
12p10	ELL	VT	10 Over the Lane Line on Left Side of Travel Lane

Original	Scheme 1	Scheme 2	CRSS Description
12p11	ERL	VT	11 Over the Lane Line on Right Side of Travel Lane
12p12	ELE	VT	12 Off the Edge of The Road on The Left Side
12p13	ERE	VT	13 Off the Edge of The Road on The Right Side
12p14	ED	VT	14 End Departure
12p15	L	VT	15 Turning Left
12p16	R	VT	16 Turning Right
12p19	N	VT	19 Unknown Travel Direction
12p20	BU	VT	20 Backing
12p21	U	VT	21 Making a U-Turn
			<i>Other Vehicle in Lane:</i>
12p50	OIS	OI	50 Other Vehicle Stopped
12p51	OIS	OI	51 Traveling in Same Direction with Lower Steady Speed
12p52	OIS	OI	52 Traveling in Same Direction while Decelerating
12p59	OIN	OI	59 Unknown Travel Direction of The Other Motor Vehicle in Lane
			<i>Other Vehicle Encroaching into Lane:</i>
12p60	OES	OE	60 From Adjacent Lane (Same Direction)-Over Left Lane Line
12p61	OES	OE	61 From Adjacent Lane (Same Direction)-Over Right Lane Line
12p62	OEO	OE	62 From Opposite Direction Over Left Lane Line
12p63	OEO	OE	63 From Opposite Direction Over Right Lane Line
12p64	OES	OE	64 From Parking Lane/Shoulder, Median/Crossover, Roadside
12p66	OET	OE	66 From Crossing Street, Across Path
12p74	OES	OE	74 From Entrance to Limited Access Highway
12p78	OEN	OE	78 Encroaching by Other Vehicle – Details Unknown
			<i>Pedestrian in Lane:</i>
12p80	PII	PI	80 Pedestrian in Road
12p81	PIA	PI	81 Pedestrian Approaching Road
			<i>Pedalcyclist in Lane:</i>
12p83	BII	BI	83 Pedalcyclist/Other Non-Motorist in Road
12p85	BIN	BI	85 Pedalcyclist Or Other Non-Motorist Unknown Location
			<i>Animal in Lane:</i>
12p87	AII	AI	87 Animal in Road
12p88	AIA	AI	88 Animal Approaching Road
12p89	AIN	AI	89 Animal Unknown Location
			<i>Object in Lane:</i>
12p90	OBI	OB	90 Object in Road
12p91	OBA	OB	91 Object Approaching Road
			<i>Other/Unknown:</i>

Original	Scheme 1	Scheme 2	CRSS Description
12p98	N	N	98 Other Critical Precrash Event
12p99	N	N	99 Unknown

PCRASH3

Original	Scheme 1	Scheme 2	CRSS Description
13p0	N	N	0 No Driver Present/Unknown if Driver Present
13p1	N	N	1 No Avoidance Maneuver
13p5	RB	RB	5 Releasing Brakes
13p6	L	L	6 Steering Left
13p7	R	R	7 Steering Right
13p8	BL	BL	8 Braking and Steering Left
13p9	BR	BR	9 Braking and Steering Right
13p10	A	A	10 Accelerated
13p12	AR	AR	12 Accelerating and Steering Right
13p15	B	B	15 Braking and Unknown Steering Direction
13p16	B	B	16 Braking
13p98	N	N	98 Other Actions
13p99	N	N	99 Unknown/Not Reported

SOE

Original	Scheme 1	Scheme 2	CRSS Description
1v1	RLO	NCH	1 Rollover/Overturn
1v2	NCH	NCH	2 Fire/Explosion
1v3	NCH	NCH	3 Immersion or Partial Immersion
1v5	NCH	NCH	5 Fell/Jumped from Vehicle
1v7	NCH	NCH	7 Other Noncollision
1v8	XP	XO	8 Pedestrian
1v9	XB	XO	9 Pedalcyclist
1v11	XA	XO	11 Live Animal
1v15	XP	XO	15 Non-Motorist on Personal Conveyance
1v16	NCH	NCH	16 Thrown or Falling Object
1v17	XFC	XF	17 Boulder
1v18	XO	XO	18 Other Object Not Fixed
1v19	XFA	XF	19 Building
1v20	XFA	XF	20 Impact Attenuator/Crash Cushion

Original	Scheme 1	Scheme 2	CRSS Description
1v21	XFC	XF	21 Bridge Pier or Support
1v23	XFB	XF	23 Bridge Rail (Includes Parapet)
1v24	XFA	XF	24 Guardrail Face
1v25	XFB	XF	25 Concrete Traffic Barrier
1v26	XFB	XF	26 Other Traffic Barrier
1v30	XFC	XF	30 Utility Pole/Light Support
1v31	XFC	XF	31 Post, Pole or Other Support
1v32	XFC	XF	32 Culvert
1v33	XFB	XF	33 Curb
1v34	XFA	XF	34 Ditch
1v35	XFA	XF	35 Embankment
1v38	XFA	XF	38 Fence
1v39	XFB	XF	39 Wall
1v40	XFC	XF	40 Fire Hydrant
1v41	XFA	XF	41 Shrubbery
1v42	XFC	XF	42 Tree (Standing Only)
1v43	XFB	XF	43 Other Fixed Object
1v44	NCH	NCH	44 Pavement Surface Irregularity (Ruts, Potholes, Grates, etc.)
1v46	XFC	XF	46 Traffic Signal Support
1v48	XFA	XF	48 Snow Bank
1v50	XFC	XF	50 Bridge Overhead Structure
1v52	XFC	XF	52 Guardrail End
1v53	XFB	XF	53 Mail Box
1v57	XFA	XF	57 Cable Barrier
1v58	XFA	XF	58 Ground
1v59	XFB	XF	59 Traffic Sign Support
1v61	EF	NH	61 Equipment Failure (blown tire, brake failure, etc.)
1v63	ROR	NH	63 Ran Off Roadway-Right
1v64	ROL	NH	64 Ran Off Roadway-Left
1v65	CM	NH	65 Cross Median
1v67	AIR	NH	67 Vehicle Went Airborne
1v68	CM	NH	68 Cross Centerline
1v69	RE	NH	69 Re-entering Roadway
1v71	ED	NH	71 End Departure
1v72	NCH	NCH	72 Cargo/Equipment Loss, Shift, or Damage (Harmful)
1v73	XO	XO	73 Object That Had Fallen from Motor Vehicle In-Transport
1v79	RO	NH	79 Ran off Roadway - Direction Unknown

Original	Scheme 1	Scheme 2	CRSS Description
1v91	XO	XO	91 Unknown Object Not Fixed
1v93	XFB	XF	93 Unknown Fixed Object
1v99	N	N	99 Reported as Unknown

Note for Scheme 2: NCH = Non-collision harmful event; NH: non-harmful event;
 XO = hit object (non-fixed); XF = hit fixed object; N = Unknown.

Table A - 2 CRSS Crash Type Diagrams

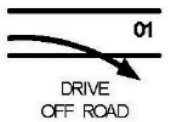
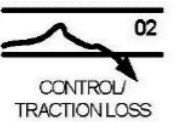
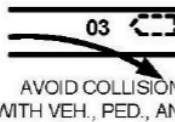
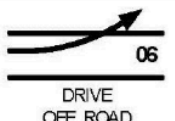

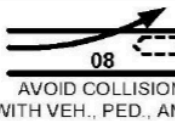

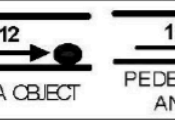
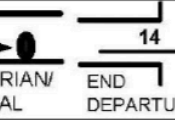



I Single Driver	A Right Roadside Departure	 01 DRIVE OFF ROAD	 02 CONTROL/ TRACTION LOSS	 03 AVOID COLLISION WITH VEH., PED., ANIM.	04 SPECIFICS OTHER	05 SPECIFICS UNKNOWN
	B Left Roadside Departure	 06 DRIVE OFF ROAD	 07 CONTROL/ TRACTION LOSS	 08 AVOID COLLISION WITH VEH., PED., ANIM.	09 SPECIFICS OTHER	10 SPECIFICS UNKNOWN
	C Forward Impact	 11 PARKED VEH.	 12 STA OBJECT	 13 PEDESTRIAN/ ANIMAL	 14 END DEPARTURE	15 SPECIFICS OTHER
VI Misc.	M Backing, Etc.	 92 Backing Veh.		 93 Other Veh. or Object		98 Other Accident Type 99 Unknown Accident Type 00 No Impact

Table A - 3 Crash Type Recoding

Crash Type	Code	Crash Type	Code
00	N	10	B4
01	A1	11*	C1
02	A2	12	C2
03	A3	13	C3
04	A4	14	C4
05	A4	15	C5
06	B1	16	C5
07	B2	92	R
08	B3	98	U
09	B4	99	U

Note: * Type C1 did not appear in the case study sample data set

Appendix for Chapter 5

Table A - 4 Event Encodings

PCRASH1		
Veh1	Veh2	CRSS Original Categories and Description
1A	2A	3 Accelerating in Road, 4 Starting in Road
1B	2B	2 Decelerating in Road
1BU	2BU	13 Backing Up (Other Than for Parking Position)
1C	2C	14 Negotiating a Curve
1CA	2CA	17 Successful Corrective Action to a Previous Critical Event
1E	2E	8 Leaving a Parking Position; 15 Changing Lanes; 16 Merging
1L	2L	11 Turning Left
1N	2N	0 No Driver Present/Unknown if Driver Present; 98 Other; 99 Unknown
1P	2P	9 Entering a Parking Position
1PA	2PA	6 Passing or Overtaking Another Vehicle
1R	2R	10 Turning Right
1S	2S	5 Stopped in Roadway; 7 Disabled or Parked in Travel Lane
1ST	2ST	1 Going Straight
1U	2U	12 Making a U-turn

PCRASH2

Veh1	Veh2	CRSS Original Categories and Description
<i>This Vehicle</i>		
1B	2B	18 This Vehicle Decelerating
1BU	2BU	20 Backing
1ELE	2ELE	12 Off The Edge of The Road on The Left Side
1ELL	2ELL	10 Over The Lane Line on Left Side of Travel Lane
1ERE	2ERE	13 Off The Edge of The Road on The Right Side
1ERL	2ERL	11 Over The Lane Line on Right Side of Travel Lane
1L	2L	15 Turning Left
1LCF	2LCF	6 Traveling Too Fast For Conditions
1LCM	2LCM	2 Stalled Engine; 4 Non-Disabling Vehicle Problem (e.g., Hood Flew Up); 5 Poor Road Conditions (Puddle, Pothole, Ice, etc.)
1LCO	2LCO	8 Other Cause of Control Loss, 9 Unknown Cause of Control Loss
1LCS	2LCS	1 Blow Out/Flat Tire, 3 Disabling Vehicle Failure (e.g., Wheel Fell Off)
1N	2N	19 Unknown Travel Direction; 98 Other Critical Precrash Event; 99 Unknown
1R	2R	16 Turning Right
1ST	2ST	17 Crossing Over (Passing Through) Junction
1U	2U	21 Making a U-Turn
<i>Other Vehicle</i>		
1OEN	2OEN	Encroached Unknown Direction: 73 From Driveway, Intended Path Not Known; 78 Encroaching By Other Vehicle – Details Unknown
1OEO	2OEO	Encroached Opposite Direction: 62 From Opposite Direction Over Left Lane Line; 63 From Opposite Direction Over Right Lane Line; 67 From Crossing Street, Turning Into Opposite Direction; 72 From Driveway, Turning Into Opposite Direction
1OES	2OES	Encroached Same Direction: 60 From Adjacent Lane (Same Direction)-Over Left Lane Line; 61 From Adjacent Lane (Same Direction)-Over Right Lane Line; 64 From Parking Lane/Shoulder, Median/Crossover, Roadside; 65 From Crossing Street, Turning Into Same Direction; 70 From Driveway, Turning Into Same Direction; 74 From Entrance to Limited Access Highway
1OET	2OET	Encroached Across Path: 66 From Crossing Street, Across Path; 71 From Driveway, Across Path
1OIN	2OIN	In Lane: 59 Unknown Travel Direction Of The Other Motor Vehicle in Lane
1OIO	2OIO	In Lane: 54 Traveling in Opposite Direction
1OIR	2OIR	In Lane: 56 Backing
1OIS	2OIS	In Lane Same Direction: 50 Other Vehicle Stopped; 51 Traveling in Same Direction with Lower Steady Speed; 52 Traveling in Same Direction while Decelerating; 53 Traveling in Same Direction with Higher Speed
1OIT	2OIT	In Lane: 55 In Crossover
<i>Animal/Object/Bike/Ped</i>		
1AIA		88 Animal Approaching Road
1AII	2AII	87 Animal in Road
	2BII	83 Pedalcyclist/Other Non-Motorist in Road
	2OBI	90 Object in Road
1PII	2PII	80 Pedestrian in Road

PCRASH3

Veh1	Veh2	CRSS Original Categories and Description
1A	2A	10 Accelerated
1AL	2AL	11 Accelerating And Steering Left
1AR	2AR	12 Accelerating And Steering Right
1B	2B	15 Braking and Unknown Steering Direction; 16 Braking
1BL	2BL	8 Braking And Steering Left
1BR	2BR	9 Braking And Steering Right
1L	2L	6 Steering Left
1N	2N	0 No Driver Present/Unknown if Driver Present; 98 Other Actions; 99 Unknown/Not Reported
1NA	2NA	1 No Avoidance Maneuver
1R	2R	7 Steering Right
1RB	2RB	5 Releasing Brakes

SOE

Veh1	Veh2	CRSS Original Categories and Description
1AIR	2AIR	67 Vehicle Went Airborne
1CARG	2CARG	60 Cargo/Equipment Loss or Shift (non-harmful)
1CM	2CM	65 Cross Median; 68 Cross Centerline
1ED	2ED	71 End Departure
1EF	2EF	61 Equipment Failure (blown tire, brake failure, etc.); 62 Separation of Units Non-collision Harmful Events: 2 Fire/Explosion; 3 Immersion or Partial Immersion; 4 Gas Inhalation; 5 Fell/Jumped from Vehicle; 6 Injured in Vehicle (Non-Collision); 7 Other Noncollision; 16 Thrown or Falling Object; 44 Pavement Surface Irregularity (Ruts, Potholes, Grates, etc.); 51 Jackknife (Harmful to This Vehicle); 72 Cargo/Equipment Loss, Shift, or Damage (Harmful)
1NCH	2NCH	
1RE	2RE	69 Re-entering Roadway
1RLO	2RLO	1 Rollover/Overturn
1RO	2RO	79 Ran off Roadway - Direction Unknown
1ROL	2ROL	64 Ran Off Roadway-Left
1ROR	2ROR	63 Ran Off Roadway-Right
1XB	2XB	Collision: With Pedalcyclist
1XF	2XF	Collision: With Fixed Object
1XO	2XO	Collision: With Other Object Not Fixed
1XP	2XP	Collision: With Pedestrian, Non-Motorist on Personal Conveyance
	2XPV	Collision: With Parked Motor Vehicle
1XV	2XV	Collision: With the Other Motor Vehicle In-Transport

Table A - 5 Intersection two-vehicle crash sequence types

Type	Weighted Count	% in Total	Representative Sequences	% in Type
d1	42,092	0.7%	1ST-1OIS-1B-2B-2OIS-2N-1XV	45%
			1ST-1OIS-1B-2B-2B-2N-1XV	8%
			1B-1OIS-1B-2B-2OIS-2N-1XV	7%
d2	46,241	0.8%	1R-1OIS-1N-2S-2OIS-2NA-1XV	87%
			1R-1OIS-1N-2R-2OIS-2NA-1XV	4%
			1R-1OIS-1N-2R-2R-2NA-1XV	2%
d3	81,464	1.4%	1C-1OIS-1N-2S-2OIS-2NA-1XV	91%
			1C-1OIS-1N-2S-2OIS-2N-1XV	3%
			1C-1OIS-1R-2S-2OIS-2NA-1XV	1%
d4	320,110	5.4%	1ST-1OIS-1N-2B-2OIS-2N-1XV	41%
			1ST-1OIS-1N-2R-2OIS-2N-1XV	5%
			1ST-1OIS-1N-2B-2B-2N-1XV	5%
d5	119,820	2.0%	1A-1OIS-1N-2S-2OIS-2NA-1XV	88%
			1A-1OIS-1N-2S-2OIS-2N-1XV	7%
			1A-1OIS-1N-2A-2OIS-2NA-1XV	1%
d6	39,166	0.7%	2S-2OIS-2NA-1ST-1OIS-1N-2XV	28%
			2B-2OIS-2N-1ST-1OIS-1N-2XV	7%
			2ST-2OIS-2N-1ST-1OIS-1N-2XV	5%
d7	205,907	3.5%	1ST-1OIS-1B-2S-2OIS-2NA-1XV	68%
			1B-1OIS-1B-2S-2OIS-2NA-1XV	11%
			1ST-1OIS-1B-2S-2OIS-2N-1XV	3%
d8	108,359	1.8%	1ST-1OIS-1N-2ST-2OIS-2N-1XV	72%
			1ST-1OIN-1N-2ST-2OIS-2N-1XV	4%
			1ST-1OIS-1B-2ST-2OIS-2N-1XV	4%
d9	72,006	1.2%	1ST-1OIS-1N-2S-2OIS-2N-1XV	86%
			1R-1OIS-1N-2S-2OIS-2N-1XV	3%
			1E-1OIS-1N-2S-2OIS-2N-1XV	2%
d10	87,269	1.5%	1B-1OIS-1N-2S-2OIS-2NA-1XV	87%
			1B-1OIS-1N-2S-2OIS-2N-1XV	5%
			1B-1OIS-1N-2B-2OIS-2NA-1XV	3%
d11	172,901	2.9%	1ST-1OIS-1NA-2S-2OIS-2NA-1XV	52%
			1A-1OIS-1NA-2S-2OIS-2NA-1XV	18%
			1C-1OIS-1NA-2S-2OIS-2NA-1XV	5%
d12	1,005,268	17.1%	1ST-1OIS-1N-2S-2OIS-2NA-1XV	89%
			1E-1OIS-1N-2S-2OIS-2NA-1XV	1%
			1ST-1OIS-1N-2B-2OIS-2NA-1XV	1%

Type	Weighted Count	% in Total	Representative Sequences	% in Type
e3	1,143	0.0%	1ST-1OES-1BR-2ST-2OIS-2NA-1XV-1ROR-1XF-1NCH	17%
			1ST-1OIS-1N-2S-2OEN-2NA-1XV	15%
			1E-1OEN-1N-2ST-2OES-2N-1XV	14%
f1	14,220	0.2%	2ST-2ELL-2N-1ST-1OES-1N-2XV	29%
			2E-2ELL-2N-1ST-1OES-1N-2XV	11%
			2C-2ELL-2N-1C-1OES-1N-2XV	6%
f2	23,919	0.4%	1E-1OIS-1N-2S-2OIS-2NA-1XV	16%
			1ST-1OIS-1N-2S-2OIS-2NA-1XV	11%
			1PA-1OIS-1N-2S-2OIS-2NA-1XV	6%
f3	20,576	0.3%	2ST-2OES-2N-1E-1ELL-1N-2XV	25%
			2ST-2OES-2N-1E-1ERL-1N-2XV	20%
			2ST-2OES-2N-1E-1ERL-1NA-2XV	5%
f4	32,310	0.5%	1E-1ELL-1N-2S-2OES-2NA-1XV	22%
			1E-1ERL-1N-2S-2OES-2NA-1XV	18%
			1ST-1ERL-1N-2S-2OES-2NA-1XV	9%
f5	8,931	0.2%	1E-1ERL-1NA-2ST-2OES-2NA-1XV	39%
			1E-1ELL-1NA-2ST-2OES-2NA-1XV	37%
			1E-1OIS-1NA-2ST-2OES-2NA-1XV	5%
f6	10,150	0.2%	2ST-2ELL-2N-1S-1OES-1NA-2XV	20%
			2E-2OIS-2N-1S-1OIS-1NA-2XV	8%
			2ST-2OIS-2N-1S-1OIS-1NA-2XV	5%
f7	18,779	0.3%	1ST-1ERL-1N-2ST-2OES-2N-1XV	25%
			1ST-1N-1N-2ST-2N-2N-1XV	16%
			1ST-1OES-1N-2ST-2ELL-2N-1XV	8%
f8	72,270	1.2%	1E-1ELL-1N-2ST-2OES-2N-1XV	67%
			1E-1ELL-1NA-2ST-2OES-2N-1XV	8%
			1E-1ELL-1N-2C-2OES-2N-1XV	4%
f9	99,879	1.7%	1E-1ERL-1N-2ST-2OES-2N-1XV	55%
			1E-1ERL-1NA-2ST-2OES-2N-1XV	8%
			1PA-1ERL-1N-2ST-2OES-2N-1XV	4%

Type	Weighted Count	% in Total	Representative Sequences	% in Type
g1	11,260	0.2%	1ST-1ELL-1N-2ST-2OEO-2N-1XV	18%
			1ST-1ELL-1N-2ST-2OEO-2N-1CM-1XV	12%
			1C-1ELL-1N-2C-2OEO-2N-1CM-1XV	7%
g2	2,918	0.0%	1ST-1OET-1L-2S-2OEO-2NA-1XV	10%
			1ST-1N-1N-2ST-2N-2N-1XV	10%
			1C-1LCF-1N-2C-2OEO-2N-1XV	7%
h1	586	0.0%	1ST-1OEO-1L-2ST-2OEO-2N-1CM-1XV	42%
			1ST-1ELL-1N-2ST-2OEO-2N-1XV	12%
			1ST-1ELL-1L-2ST-2OEO-2N-1XV	11%
i1	14,663	0.2%	1ST-1ELL-1N-2S-2OEO-2NA-1XV	10%
			1C-1ELL-1N-2C-2OEO-2N-1XV	7%
			1C-1ELL-1N-2C-2OEO-2N-1CM-1XV	5%
i2	3,405	0.1%	2ST-2OEO-2N-1ST-1LCF-1N-2XV-1ROL-1XF	16%
			1ST-1OIO-1L-2ST-2OIO-2N-1XV	14%
			1ST-1LCO-1N-2ST-2OEO-2N-1XV	8%
j1	89,654	1.5%	1R-1R-1N-2ST-2OES-2N-1XV	15%
			1L-1ERL-1N-2L-2OES-2N-1XV	8%
			1R-1R-1N-2L-2OEO-2N-1XV	7%
j2	544,349	9.2%	1L-1L-1N-2ST-2OEO-2N-1XV	52%
			1L-1L-1N-2ST-2OES-2N-1XV	5%
			1L-1L-1NA-2ST-2OEO-2N-1XV	4%
j3	461,634	7.8%	2ST-2OEO-2N-1L-1L-1N-2XV	51%
			2ST-2ST-2N-1L-1OEO-1N-2XV	6%
			2ST-2ST-2N-1L-1L-1N-2XV	5%

Type	Weighted Count	% in Total	Representative Sequences	% in Type
k1	50,571	0.9%	1L-1ELL-1N-2S-2OEO-2NA-1XV	18%
			1R-1ELL-1N-2S-2OEO-2NA-1XV	17%
			1R-1R-1N-2S-2OEO-2NA-1XV	14%
k2	82,579	1.4%	2ST-2ST-2N-1L-1OET-1N-2XV	61%
			2ST-2ST-2N-1R-1OET-1N-2XV	5%
			2ST-2ST-2B-1L-1OET-1N-2XV	4%
k3	228,645	3.9%	2ST-2OEO-2N-1L-1L-1N-2XV	43%
			2ST-2OES-2N-1L-1L-1N-2XV	10%
			2ST-2ST-2N-1L-1L-1N-2XV	7%
k4	183,944	3.1%	1L-1L-1N-2ST-2OEO-2N-1XV	57%
			1L-1L-1NA-2ST-2OEO-2N-1XV	7%
			1L-1L-1N-2C-2OEO-2N-1XV	5%
k5	21,835	0.4%	1L-1L-1N-2S-2OEO-2NA-1XV	49%
			1L-1L-1NA-2S-2OEO-2NA-1XV	10%
			1L-1L-1N-2S-2OES-2NA-1XV	8%
k6	175,550	3.0%	1R-1R-1N-2ST-2OES-2N-1XV	50%
			1R-1R-1NA-2ST-2OES-2N-1XV	5%
			1R-1R-1N-2C-2OES-2N-1XV	3%
k7	50,321	0.9%	2ST-2OES-2N-1R-1R-1N-2XV	50%
			2ST-2OES-2N-1R-1R-1NA-2XV	6%
			2C-2OES-2N-1R-1R-1N-2XV	5%
k8	26,569	0.5%	1L-1OET-1N-2ST-2ST-2N-1XV	53%
			1R-1OET-1N-2ST-2ST-2N-1XV	6%
			1L-1OET-1N-2ST-2ST-2L-1XV	4%
k9	190,854	3.2%	1L-1L-1N-2ST-2OES-2N-1XV	44%
			1L-1L-1NA-2ST-2OES-2N-1XV	6%
			1L-1L-1N-2ST-2ST-2N-1XV	5%
m1	91,068	1.5%	1BU-1BU-1N-2S-2OIR-2NA-1XV	40%
			1BU-1OIS-1N-2S-2OIR-2NA-1XV	6%
			1BU-1BU-1NA-2S-2OIR-2NA-1XV	6%
u	116,156	2.0%	1U-1U-1N-2ST-2OEO-2N-1XV	6%
			1U-1U-1N-2ST-2OES-2N-1XV	6%
			1ST-1N-1N-2ST-2N-2N-1XV	4%

Type	Weighted Count	% in Total	Representative Sequences	% in Type
I1	12,508	0.2%	2ST-2OET-2N-1ST-1ST-1N-2XV	52%
			2ST-2OET-2N-1A-1ST-1N-2XV	7%
			2ST-2OET-2B-1ST-1ST-1NA-2XV	5%
I2	31,062	0.5%	1A-1ST-1N-2ST-2OET-2N-1XV	71%
			1A-1ST-1N-2A-2OET-2N-1XV	5%
			1A-1ST-1N-2A-2ST-2N-1XV	3%
I3	21,903	0.4%	1ST-1OET-1N-2A-2ST-2N-1XV	63%
			1ST-1OET-1B-2A-2ST-2N-1XV	6%
			1ST-1OET-1N-2A-2OET-2N-1XV	4%
I4	16,065	0.3%	1ST-1OET-1N-2ST-2OET-2N-1XV	85%
			1ST-1OET-1N-2ST-2OET-2N-1XV-2ROR-2XF	2%
			1ST-1OET-1N-2S-2OET-2N-1XV	1%
I5	10,319	0.2%	1ST-1OET-1B-2ST-2ST-2N-1XV	73%
			1ST-1OET-1B-2ST-2ST-2N-1XV-2RLO	5%
			1ST-1OET-1B-2ST-2ST-2N-1XV-2ROL-2XF	4%
I6	88,353	1.5%	2ST-2ST-2N-1ST-1OET-1N-2XV	54%
			2A-2ST-2N-1ST-1OET-1N-2XV	7%
			2ST-2ST-2N-1ST-1ST-1N-2XV	6%
I7	12,516	0.2%	1ST-1ST-1N-2ST-2OET-2N-1XV-2ROR-2XF	55%
			1ST-1ST-1N-2ST-2OET-2N-1XV-2ROR-2XF-2XF	19%
			1ST-1ST-1N-2ST-2OET-2R-1XV-2ROR-2XF	2%
I8	12,597	0.2%	1ST-1ST-1NA-2ST-2OET-2NA-1XV	47%
			1ST-1OET-1NA-2ST-2OET-2NA-1XV	6%
			1ST-1ST-1NA-2ST-2ST-2NA-1XV	5%
I9	43,371	0.7%	1ST-1ST-1NA-2ST-2OET-2N-1XV	56%
			1A-1ST-1NA-2ST-2OET-2N-1XV	16%
			1ST-1ST-1NA-2ST-2OET-2B-1XV	3%
I10	35,840	0.6%	1ST-1OET-1N-2ST-2ST-2NA-1XV	46%
			1ST-1OET-1N-2A-2ST-2NA-1XV	10%
			1ST-1OET-1NA-2ST-2ST-2NA-1XV	9%
I11	40,560	0.7%	1ST-1ST-1N-2ST-2ST-2N-1XV	85%
			1ST-1ST-1N-2ST-2ST-2N-1XV-1ROL-1XF	2%
			1ST-1ST-1B-2ST-2ST-2N-1XV	2%
I12	210,273	3.6%	1ST-1OET-1N-2ST-2ST-2N-1XV	75%
			1ST-1OET-1N-2ST-2ST-2N-1XV-2ROR-2XF	2%
			1ST-1OET-1N-2ST-2ST-2N-1XV-2RLO	2%
I13	385,833	6.6%	1ST-1ST-1N-2ST-2OET-2N-1XV	78%
			1ST-1ST-1N-2A-2OET-2N-1XV	1%
			1ST-1ST-1N-2ST-2OET-2NA-1XV	1%
I14	17,047	0.3%	1ST-1ST-1B-2ST-2OET-2N-1XV	61%
			1ST-1ST-1B-2ST-2OET-2NA-1XV	5%
			1ST-1ST-1B-2ST-2OET-2B-1XV	3%