ADVANCING ROAD SAFETY THROUGH DATA INTEGRATION, LOCATION ACCURACY, AND RISK-BASED ROUTE OPTIMIZATION

by

Lars Skaug

A Dissertation Submitted to the Faculty of

The College of Engineering & Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Florida Atlantic University

Boca Raton, FL

December 2025

ADVANCING ROAD SAFETY THROUGH DATA INTEGRATION, LOCATION ACCURACY, AND RISK-BASED ROUTE OPTIMIZATION

by

Lars Skaug

This dissertation was prepared under the direction of the candidate's dissertation advisor, Dr. Mehrdad Nojoumian, Department of Electrical Engineering and Computer Science, and has been approved by the members of his supervisory committee. It was submitted to the faculty of the College of Engineering & Computer Science and was accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

SUPERVISORY COMMITTEE:

Mehrdad Nojoumian, Ph.D.
Dissertation Advisor

Borivoje Furht, Ph.D.

Taghi Khoshgoftaar, Ph.D.

Dingding Wang, Ph.D.

Hari Kalva, Ph.D.
Chair, Department of Electrical Engineering and Computer Science

Stella Batalama, Ph.D.
Dean, The College of Engineering & Computer Science

Robert W. Stackman, Ph.D.
Dean, Graduate College

Nov 5, 2025
Date

VITA

An immigrant from Norway, Lars moved to the U.S. in 2000. After completing his MBA at Iowa State University, he built a career in data analytics at JM Family Enterprises in Deerfield Beach, where he has spent two decades transforming business insights through data science. He lives in Boca Raton with his wife and two children.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my wife for her patience with me as I worked long nights and weekends to fulfill my dream of obtaining a PhD. Without the invaluable support and encouragement of my advisor, Dr. Mehrdad Nojoumian, I would not have been able to make this dream come true. I would also like to thank my committee for the trust and confidence in my work.

ABSTRACT

Author:	Lars Skaug
Title:	Advancing Road Safety Through Data Integration, Location Accuracy, and Risk-Based Route Optimization
Institution:	Florida Atlantic University
Dissertation Advisor:	Dr. Mehrdad Nojoumian
Degree:	Doctor of Philosophy
Year:	2025

This dissertation advances transportation safety through three interconnected contributions that address fundamental challenges in crash data analysis and application. First, we present a comprehensive survey of road crash analysis methodologies, tracing the evolution from traditional statistical approaches to modern machine learning techniques while identifying critical gaps between research sophistication and practical implementation. This survey reveals that while methodological advances have been substantial, ongoing data quality issues, particularly in location accuracy and completeness, limit the effectiveness of even the most sophisticated analytical approaches.

Building on these findings, we develop a novel system to validate and correct crash location data using multi-modal large language models (LLMs) integrated with geospatial analysis. Our credibility-based scoring system evaluates location accuracy by comparing multiple data sources, analyzing crash narratives and diagrams, and applying spatial validation techniques. Empirical testing on 5,000 Ohio crash reports demonstrates that approximately 20% require location corrections, with our method successfully identifying and correcting these errors through automated post-processing.

The third contribution introduces a risk-aware navigation solution applicable to both human-driven and autonomous vehicles. By integrating historical crash patterns with predicted traffic volumes, we create standardized risk metrics for individual road segments that can be incorporated into routing algorithms. Validation across 56 high-volume commute routes in Ohio demonstrates that safety-prioritized routing reduces crash risk exposure by an average of 22% while increasing travel time by only 9% on average.

Together, these contributions form a cohesive approach to transportation safety from data quality assessment through practical application, demonstrating how advances in data science and analytical methods translate into tangible safety benefits for current transportation systems and emerging navigation technologies.

To my mother who wanted to pursue a PhD herself. She would have been so proud to see her son reach this milestone.

ADVANCING ROAD SAFETY THROUGH DATA INTEGRATION, LOCATION ACCURACY, AND RISK-BASED ROUTE OPTIMIZATION

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 TRANSPORTATION SAFETY AS A PUBLIC HEALTH CHALLENGE

Road traffic crashes constitute a significant global public health burden, resulting in approximately 1.35 million deaths annually worldwide and up to 50 million injuries. Traffic crashes represent the leading cause of death for individuals aged 5-29 years [1]. In the United States, approximately 41,000 fatalities occur annually, generating substantial economic losses. These statistics represent considerable human and economic costs that warrant careful investigation and intervention.

The persistence of elevated crash rates despite advances in vehicle safety technology, infrastructure design, and traffic management systems warrants a more comprehensive approach to safety analysis and intervention. This dissertation examines how contemporary data science methods, artificial intelligence technologies, and navigation systems can be applied to address longstanding challenges in transportation safety research and practice.

1.2 THE CHALLENGE: FROM SOPHISTICATED METHODS TO PRACTICAL SAFETY IMPROVEMENTS

Effective road safety intervention depends on accurate, comprehensive data on crash occurrence, causation, and spatial distribution. Transportation agencies have built extensive crash databases capturing millions of incidents—to support policy decisions, infrastructure investments, and safety interventions across jurisdictions.

However, despite major advances in crash analysis methods—from classical statistical modeling to state-of-the-art machine learning—practical safety improvements have not kept pace. A significant gap persists between analytical sophistication and real-world outcomes, driven by three core limitations.

First, comprehensive solutions that translate advanced methods into real-world applications remain scarce. While researchers continue to develop highly accurate predictive models, practitioners often fall back on basic techniques due to integration challenges and the lack of user-friendly deployment pathways.

Second, fundamental data quality issues limit the effectiveness of any analytical method. Location inaccuracies—affecting as many as one in five crash reports in some jurisdictions [2]—lead to misclassified spatial patterns and misallocated safety investments, with errors compounding through downstream analyses.

Third, even when actionable insights are generated, they rarely influence the decisions of transportation users. Most navigation systems still optimize solely for travel time and distance, implicitly treating all road segments as equally safe. This neglects critical opportunities to deliver safety-relevant guidance in everyday routing decisions.

1.3 RESEARCH APPROACH AND CONTRIBUTIONS

This dissertation addresses critical challenges in crash data analysis through an integrated research agenda designed to answer three core questions:

RQ1: What methodological approaches exist for crash data analysis, and what factors limit their adoption in practice?

RQ2: How can artificial intelligence be used to automatically detect and correct location inaccuracies in crash data?

RQ3: How can enhanced crash data be incorporated into navigation systems to enable risk-aware routing for both human-driven and autonomous vehicles?

These questions form a progression: identifying current limitations, addressing data quality challenges, and leveraging improved data for practical applications. The research methodology combines literature synthesis, algorithm design, and geospatial modeling, with empirical validation using real-world transportation data. Throughout, the work prioritizes practical applicability by leveraging open-source tools and integrating with existing systems—ensuring solutions that are both technically sound and readily deployable. The work produces three key contributions:

1. **A comprehensive taxonomy** that traces the evolution of crash analysis from traditional statistics to machine learning while identifying critical implementation barriers [3].

2. **A novel AI framework** using multi-modal large language models for scalable, automated validation and correction of crash location data [4].

3. **A risk-aware navigation tool** that integrates historical crash data with traffic exposure models to inform safer routing decisions [5].

CHAPTER 2
BACKGROUND AND TECHNOLOGICAL CONTEXT

Transportation safety research has undergone several paradigm shifts since the introduction of motorized vehicles. Early approaches characterized crashes as random events or individual failures, focusing on descriptive statistics and basic causal attribution. The introduction of the Haddon Matrix in 1972 established a systematic framework for categorizing crash factors across temporal phases and contributing elements, shifting focus from individual blame toward preventive strategies.

Subsequent decades introduced epidemiological methods that treated crashes as population health phenomena amenable to intervention. The 1990s brought intelligent transportation systems approaches that emphasized technological prevention methods. The 2000s introduced naturalistic driving studies that provided detailed behavioral data for crash causation analysis. Recent developments have emphasized big data analytics and artificial intelligence applications for predictive modeling and real-time risk assessment.

Despite these advances, enduring challenges remain in translating research insights into practical interventions that demonstrably improve safety outcomes. This dissertation addresses this implementation challenge through frameworks that connect theoretical understanding with deployable solutions.

2.1 CRASH DATA COLLECTION AND MANAGEMENT SYSTEMS

Contemporary crash data collection in the United States begins with law enforcement documentation at incident scenes using standardized reporting forms. These

reports capture multiple attributes including location, temporal factors, environmental conditions, vehicle characteristics, occupant demographics, injury severity, and contributing factors. Information flows through state transportation departments to federal databases including the Fatality Analysis Reporting System (FARS) and the Crash Report Sampling System (CRSS).

This data collection process involves multiple sources of potential error and inconsistency. Officers completing reports under time constraints and adverse conditions must make rapid assessments about complex events. Location data may be estimated rather than precisely measured. Injury severity classifications made at incident scenes often differ from subsequent medical evaluations. Contributing factor identification requires subjective judgments about causal relationships.

These individual inaccuracies aggregate into patterns of bias when data are used for safety analysis. Mislocated crashes may be attributed to incorrect network segments, affecting intervention prioritization. Incorrectly classified injury severities impact economic analysis of safety improvements. Missing or inaccurate vehicle identification prevents accurate assessment of safety technology effectiveness.

International variations in data collection practices add complexity for comparative analysis. While some countries integrate police and hospital reporting systems, others rely primarily on law enforcement data with inherent limitations. This dissertation's methods are designed for applicability across diverse data collection systems while maximizing utility within commonly available data constraints.

2.2 GEOGRAPHIC INFORMATION SYSTEMS IN TRANSPORTATION SAFETY

Geographic Information Systems (GIS) have become essential tools for transportation safety analysis, enabling spatial pattern identification, network analysis, and geographic risk assessment. Contemporary GIS platforms support sophisticated analytical meth-

ods including kernel density estimation for hotspot identification, network analysis for accessibility assessment, and spatial regression for geographic risk modeling.

However, the effectiveness of GIS applications is constrained by data quality limitations. Crash reports reference locations using various methods—street addresses, intersections, mileposts, or GPS coordinates—each with different accuracy characteristics and geocoding challenges. Road networks exist in multiple representations across state transportation databases, OpenStreetMap crowd-sourced data, and commercial navigation systems. Integrating crashes with appropriate network segments requires resolving inconsistencies between these systems while accounting for temporal changes in network topology.

This dissertation advances GIS applications in transportation safety by developing methods that address data quality challenges directly rather than treating them as external constraints. The approaches work with imperfect data while implementing enhancement procedures, creating feedback loops where improved analysis enables better data collection.

2.3 ARTIFICIAL INTELLIGENCE APPLICATIONS IN TRANSPORTATION

Artificial intelligence has demonstrated significant capabilities in transportation applications, including computer vision for object detection, predictive modeling for traffic flow analysis, and optimization algorithms for resource allocation. In crash analysis specifically, machine learning algorithms have shown superior performance for crash likelihood prediction, risk factor identification, and intervention optimization compared to traditional statistical approaches.

However, AI applications in crash data analysis face domain-specific challenges. Unlike applications such as image recognition with extensive labeled datasets, crash data is inherently sparse, with most network segments experiencing no crashes in

typical observation periods. Data imbalance, where severe crashes are rare compared to property damage incidents, complicates model training. Additionally, the critical nature of transportation safety requires interpretable models whose decisions can be validated and explained.

The emergence of large language models (LLMs) represents a significant development in AI capabilities relevant to crash data challenges. These models demonstrate abilities to understand contextual information, resolve linguistic ambiguities, and extract structured information from unstructured text. This dissertation pioneers the application of LLMs to crash data validation, demonstrating how these capabilities can address location accuracy problems that have been unsolvable at scale.

2.4 CONNECTED AND AUTONOMOUS VEHICLE CONTEXT

The development of connected and autonomous vehicles (CAVs) provides additional application context for this research. CAVs represent a fundamental transformation in transportation systems, with potential to eliminate human error factors that contribute to the majority of serious crashes. However, realizing this safety potential requires navigation systems capable of incorporating complex safety considerations into routing decisions.

CAVs also generate extensive high-quality data about road conditions, traffic patterns, and near-miss incidents. The data validation methods developed in this dissertation can be applied to this emerging data stream, creating feedback loops where improved data enables safer automation, which generates enhanced data quality.

2.5 REGULATORY AND POLICY CONTEXT

Transportation safety operates within complex regulatory structures spanning federal, state, and local jurisdictions. In the United States, the National Highway Traffic Safety Administration sets vehicle safety standards, the Federal Highway Administration

guides infrastructure design, and state transportation departments implement specific policies. This multi-level governance creates both opportunities and constraints for safety innovations.

Recent policy developments have created supportive conditions for data-driven safety approaches. The Infrastructure Investment and Jobs Act of 2021 allocated substantial funding for safety improvements with explicit emphasis on data-driven methods. The National Roadway Safety Strategy emphasizes leveraging technology and data for systematic safety improvement. Many state and local jurisdictions have adopted Vision Zero policies that prioritize safety outcomes.

The methods developed in this dissertation align with these policy priorities while maintaining flexibility for different regulatory contexts. The risk assessment solution supports federal safety performance management requirements, state safety planning processes, and local safety initiatives through standardized, quantitative risk metrics that enable evidence-based decision making.

2.6 ETHICAL AND EQUITY CONSIDERATIONS

The development of safety technologies raises ethical questions that require careful consideration. How should navigation systems balance individual travel efficiency against collective safety? What are the equity implications of routing that may redistribute traffic across different communities? How should algorithmic decision-making systems handle scenarios where all available options involve elevated risk?

This dissertation adopts a transparent approach to these ethical challenges by making risk trade-offs explicit and configurable rather than embedding them in algorithmic black boxes. We provide tools to quantify and compare risks while preserving decision authority for users and policymakers. By improving access to safety information, the research enables informed choice rather than imposing predetermined constraints.

2.7 GLOBAL CONTEXT AND APPLICABILITY

While this dissertation's empirical analysis focuses on Ohio data, the methods are designed for global applicability. Road safety represents a worldwide challenge, with low- and middle-income countries experiencing disproportionate crash rates despite having fewer vehicles per capita. These countries often lack extensive data infrastructure available in developed nations, requiring robust methods that function with limited data.

The approaches developed work with commonly available data types: police crash reports and basic traffic measurements. The LLM-based validation methods can adapt to different languages and reporting conventions. The risk assessment accommodates various road classification systems and traffic patterns. This flexibility ensures that safety benefits can extend beyond well-resourced jurisdictions to regions where needs are greatest.

2.8 INTEGRATION OF THEORY AND PRACTICE

A continuing challenge in transportation safety research is the disconnect between theoretical advances and practical implementation. Sophisticated models published in academic journals rarely translate into tools used by practitioners. This implementation disconnect has multiple causes: computational complexity, data requirements, lack of accessible software tools, and insufficient communication between research and practice communities.

This dissertation addresses this disconnect through several strategies. The implementations use open-source tools that are freely available globally. The validation employs real-world data with inherent imperfections rather than cleaned academic datasets. The approaches provide clear implementation pathways showing how agencies can adopt methods incrementally within existing operational constraints.

By emphasizing practical implementation alongside theoretical contribution, this dissertation aims to ensure that methodological advances translate into measurable safety improvements. Each research component includes not only theoretical development but practical demonstration, performance metrics relevant to practitioners, and explicit consideration of deployment requirements.

CHAPTER 3
SYSTEMATIC REVIEW OF METHODS, DATA, AND EMERGING TECHNOLOGIES

Building on the technological context established in Chapter 2, this chapter provides a systematic examination of the methodological approaches available for crash data analysis. Through comprehensive review of the literature, we trace the evolution from traditional statistical methods to contemporary machine learning applications, while identifying critical gaps between research sophistication and practical implementation. This analysis establishes the foundation for understanding both the capabilities and limitations that inform the targeted innovations presented in subsequent chapters.

The analysis reveals that while sophisticated methods exist for crash data analysis, persistent data quality issues—particularly in location accuracy and completeness—limit the effectiveness of even the most advanced analytical techniques. This finding directly motivates the location validation solution developed in Chapter 4 and the risk-aware navigation system presented in Chapter 5.

3.1 INTRODUCTION

Road crashes represent a persistent global health crisis, causing over 1.3 million fatalities and up to 50 million injuries annually [1]. Beyond the immeasurable human suffering, these crashes impose substantial economic costs through medical expenses, lost productivity, and property damage. By 2030, road traffic crashes are projected to become the fifth leading cause of death globally, underscoring the urgent need for evidence-based approaches to road safety improvement.

The staggering impact of road crashes on society has spurred extensive efforts to improve road safety through driver behavior monitoring [6], vulnerable occupant detection [7], vehicle design improvements [8], road infrastructure enhancements [9], traffic laws and their enforcement [10, 11], public awareness campaigns [12, 13], and technological advancements [14, 15]. While these measures can potentially reduce crash frequency and severity, the persistently high toll of road crashes remains unacceptable and call for continued research and innovation in road safety.

3.1.1 The Critical Role of Data-Driven Safety Analysis

The majority of the analysis in this field is based on crash data collected by transportation agencies, law enforcement, hospitals, and insurers. The statistical modeling of these data has long provided the empirical foundation for identifying risk factors, evaluating countermeasures, and developing data-driven safety policies that have demonstrably saved lives. However, the field faces three fundamental challenges that limit the effectiveness of current approaches.

First, data quality issues significantly compromise analysis reliability. From systematic underreporting to spatial inaccuracies, these problems create substantial analytical challenges that require targeted solutions.

Second, methodological fragmentation persists between traditional statistical approaches and emerging machine learning techniques, with limited integration of these complementary analytical schemes.

Third, a research–practice gap continues to separate sophisticated analytical methods from practical implementation in safety management and policy development [16].

3.1.2 Research Contributions and Framework

This systematic review addresses these challenges by focusing on five specific areas that span the complete spectrum from foundational data issues to cutting-edge

technological applications:

1. A comprehensive data quality taxonomy that categorizes quality issues into collection- stage and analysis-stage challenges, providing a structured framework for understanding and addressing data limitations (Section 3.3).

2. A methodological evolution overview that traces the historical development from descriptive crash analysis to sophisticated system-based approaches, demonstrating how traditional statistical methods and emerging AI techniques can be integrated (Section 3.4).

3. Domain-specific intervention synthesis that demonstrates how advanced approaches address real-world safety challenges across infrastructure design, vulnerable road users, and targeted countermeasures (Section 3.5).

4. Evidence-based implementation guidelines that bridge the research–practice gap by translating methodological advances into actionable recommendations for safety management and policy development (Section 3.6).

5. A future-oriented technology roadmap that examines emerging research frontiers in big data analytics, deep learning, real-time prediction systems, and connected/autonomous vehicle safety, identifying pathways for next-generation crash analysis capabilities (Section 5.6).

These contributions are unified by a central organizing principle: methodological sophistication must be balanced with practical applicability to achieve meaningful improvements in road safety outcomes. The framework progresses systematically from foundational data and methodological concerns through specific applications and policy implications, culminating in emerging technologies that will define the future of crash analysis.

This systematic approach serves two purposes: providing researchers and practitioners with a comprehensive methodological reference while identifying pathways for advancing sophisticated analytical approaches that enhance our ability to analyze crash data at scale and translate findings into effective safety interventions.

3.2 METHODS

This systematic review included studies presenting original research on road crash data analysis methodologies using statistical or machine learning techniques. Studies were required to analyze real-world crash data from transportation agencies, law enforcement, hospitals, or insurance records. Both traditional statistical modeling and emerging methodologies including deep learning and real-time prediction systems were included.

Studies were excluded if they were purely descriptive without methodological contributions, focused solely on vehicle engineering without crash data analysis, or analyzed only simulated data without real-world validation. Studies were grouped by methodological approach, data source type, analytical focus, and geographic scope.

3.2.1 Information Sources and Search Strategy

The primary search was conducted using Google Scholar to ensure broad interdisciplinary coverage across transportation engineering, statistics, computer science, and medical research. The search strategy integrated terms related to crash analysis methodologies, data sources, and analytical approaches, with attention to emerging technologies in connected and autonomous vehicle safety.

Additional records were identified through the citation searching of the included studies and relevant review articles. Search terms captured both traditional econometric approaches and advanced machine learning applications in crash analysis.

3.2.2 Selection Process and Data Collection

The search yielded 582 records through the Google Scholar database searching and 60 through citation searching and other sources (total: 642 records). After duplicate removal, 532 unique records underwent title and abstract screening. Of these, 367 were excluded, leaving 165 for full-text assessment. An additional 21 articles were excluded as not relevant, resulting in 144 studies for final synthesis.

3.2.3 Data Items and Study Characteristics

Data extraction captured the methodological approaches (statistical models, machine learning techniques, and spatial–temporal analysis), the data source characteristics (police reports, hospital records, telematics data, and naturalistic driving studies), the performance metrics, the validation procedures, and the geographic/temporal scope. Particular attention was given to model specifications, comparative evaluations, and the handling of common challenges in crash data analysis.

3.3 DATA SOURCES AND QUALITY

3.3.1 Common Sources of Data

The most common source of data for crash analysis are reports filed by Law Enforcement Officers at the scene of a crash. Several researchers have also used hospital records of traffic injuries, and in some countries, this is standard practice, e.g., the Netherlands and Sweden. Insurance records are another, less commonly used source of crash data [17].

Naturalistic driving data, i.e., real-world driving data collected without experimental controls or driver awareness, are also available for limited time periods in a small number of locations. Video of highways [18, 19] and intersections [20] are examples of studies based on naturalistic data. A key challenge with naturalistic data is that

they are expensive to collect and therefore not available for comprehensive analysis across geographies and time.

Another emerging source is telematics—data collected automatically by vehicles or electronic devices installed in vehicles. This type of data is often collected at the initiative of insurance companies in return for insurance premium discounts and provides real-time information about driving behavior and vehicle performance. One provider of telematics data has looked into whether drivers change their behavior to take advantage of discounts [21], but further research should be performed to determine the usefulness of such data at scale.

A significant development in the collection of crash data is the mandate that all new cars sold in the European Union (EU) after 7 July 2024 must be equipped with an Event Data Recorder (EDR), an onboard device that records information about a vehicle's operation before, during, and after a crash. At the time this paper was written, research on EDR data was limited, but preliminary results [22] show promise. Figure 3.1 provides an overview of data sources used in crash research.

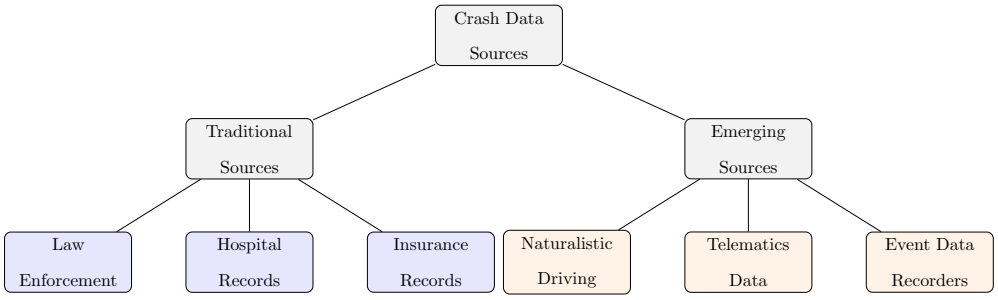


Figure 3.1: Common sources of crash data, categorized by traditional and emerging methods.

3.3.2 Data Quality Challenges in Crash Analysis

Crash data quality issues present significant challenges for road safety research and policy development. These challenges can significantly impact the reliability of crash analysis and the effectiveness of resulting safety interventions. Through extensive

review of the literature and practice, we identify six fundamental categories of data quality issues that affect crash analysis, summarized in Figure 3.2. Each of these challenges requires specific mitigation solutions, and failure to account for them can lead to biased results and ineffective safety recommendations.

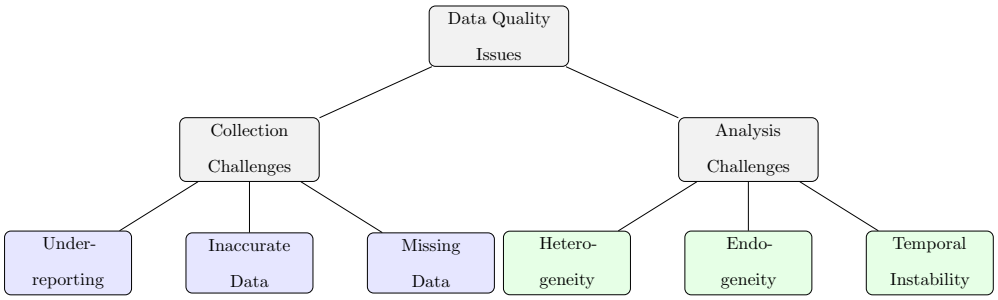


Figure 3.2: Categories of data quality issues in crash analysis, grouped by collection and analysis challenges.

3.3.3 Data Completeness and Accuracy

The most fundamental challenge in crash analysis is underreporting, where a significant number of crashes are not captured in official databases. The extent of underreporting varies by crash type and jurisdiction, with Watson et al. [23] finding that motorcyclists, cyclists, and young people are particularly under-represented in official reports.

Beyond missing cases entirely, the accuracy of reported data presents another significant challenge. Location data prove especially problematic, with Miler et al. [24] finding inaccurate locations in approximately one-third of their studied crash reports. Such inaccuracies can severely impact spatial analysis and the identification of high-risk areas.

3.3.4 Statistical Challenges in Crash Analysis

The analysis of crash data is further complicated by three interrelated statistical challenges: heterogeneity, endogeneity, and temporal instability. Heterogeneity manifests as variations among units of analysis that available data features fail to capture, while

endogeneity occurs when explanatory variables correlate with model error terms, often due to omitted variables or simultaneity.

The relationship between speed enforcement and crash rates illustrates both concepts. Speed enforcement measures, typically implemented in areas with historically high crash rates, create complex cause-and-effect relationships. The varying effectiveness of enforcement across different road segments demonstrates heterogeneity, while the potential reverse causality—where high crash rates might trigger enforcement implementation—

exemplifies endogeneity.

Temporal instability adds another layer of complexity, as relationships between variables can shift across different time scales, from daily patterns to seasonal variations and long-term trends. These statistical challenges require sophisticated methodological approaches, which we will explore in subsequent sections along with practical strategies for improving data quality at the source.

3.3.5 Strategies for Addressing Data Quality Issues

Approaches to addressing data quality issues in crash data can be broadly categorized into two types: (1) strategies to improve data quality at the collection stage, addressing the root causes of data problems, and (2) methodological approaches to identify and correct for data quality issues during analysis. Both approaches are necessary and complementary in improving the reliability of crash analysis.

Improving Data Quality at Collection

A fundamental approach to addressing data quality issues is to improve the initial data collection process. Two key strategies have emerged in this area: the implementation of automated data collection systems and the integration of multiple data sources.

The advent of new technologies offers promising solutions for reducing data quality

issues at their source. Event Data Recorders (EDRs) in the European Union [22] represent a significant advance in automated crash data collection, providing accurate information about crash circumstances without relying on human reporting. Similarly, Imprialou and Quddus [25] advocate for intelligent crash reporting systems to address the frequent misreporting of crucial attributes such as crash location, time, and severity.

Likewise, the systematic integration of data from multiple sources has proven effective in creating more complete crash records. Sweden's STRADA system exemplifies this approach, as illustrated in Figure 4.3, by systematically combining various data sources to improve reporting completeness [26].

Several studies demonstrate the value of this approach:

- Short and Caulfield [17] showed how combining insurance claim data with police and hospital records in Ireland provided a more comprehensive picture of crash incidents.
- Lombardi et al. [27] improved crash injury identification by linking hospital discharge data with state-level crash reports.
- Janstrup et al. [28] demonstrated the benefits of connecting police and medical records for understanding individual crash characteristics.
- Burdett et al. [29] revealed significant discrepancies between law enforcement and medical assessments of injury severity, finding overestimation in 45% to 90% of cases.

Statistical Methods for Addressing Existing Data Issues

When working with historical data or in contexts where improved data collection is not yet feasible, statistical methods can help identify and correct for data quality issues during analysis.

Several methodological advances help address inherent data challenges. Li et al. [30] demonstrated that binary logit models can effectively handle heterogeneous effects of road design features and traffic conditions, while their grouped random-parameter logit models specifically address unobserved heterogeneity among crash units. This work builds on Mannering et al.'s [31] comprehensive framework for dealing with unobserved heterogeneity in crash analysis.

For temporal stability issues, Shabab et al. [32] developed a "mixed spline indicator pooled model" that captures parameter changes over time while incorporating unobserved heterogeneity across severity levels and time periods. Their approach achieved 55–78% prediction accuracy for 2021 using Florida crash data from 2011–2019.

Furthermore, targeted statistical methods can address specific types of data bias. Chang and Mannering [33] developed a nested logit model to correct occupancy overestimation bias, while Yasmin et al. [34] advanced methods for handling endogeneity in transportation safety studies. These approaches demonstrate how statistical techniques can compensate for known data quality issues when analyzing existing datasets.

While both collection-stage improvements and analytical methods are valuable, Imprialou and Quddus [25] note that the full impact of data quality problems on road safety analyses remains incompletely understood. This suggests that the continued development of both approaches—improving data collection and advancing statistical methods—will be necessary for comprehensive improvement in crash data quality.

3.4 METHODOLOGICAL APPROACHES IN CRASH RESEARCH

Over the past century, road safety research has progressed through distinct paradigms, moving from simple descriptions of crash statistics and accident-prone individuals to increasingly complex, system-based approaches. Early studies offered descriptive accounts and basic mathematical models to understand traffic incidents. Later, re-

search began focusing on single causes of accidents, leading to solutions centered on engineering, education, and enforcement. The introduction of the Haddon Matrix in the 1970s [35] introduced more complex interactions of factors into the analysis, while recent years have emphasized comprehensive, system-level approaches, incorporating behavioral theories like risk homeostasis.

Few studies capture the historical trajectory of road safety research but Hagenzieker et al. [36] applied bibliometric techniques to identify past research trends and explore potential future developments in the field. Figure 3.3 shows a timeline based on their research.

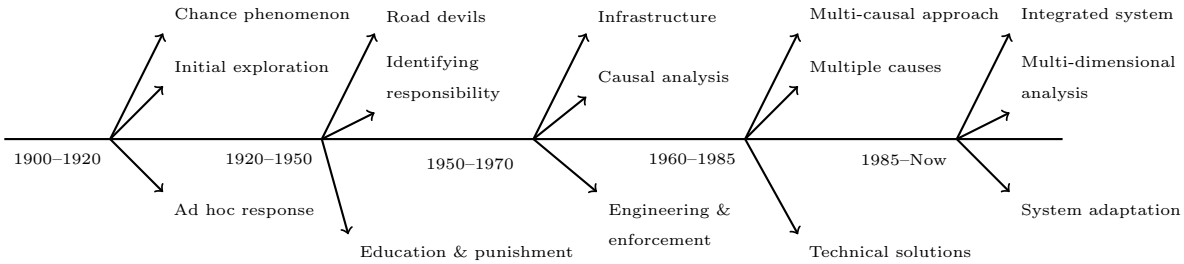


Figure 3.3: The development of crash analysis over time.

Building on these developments in crash research, Table 3.1 synthesizes current methodological approaches, from traditional statistical methods to emerging techniques for connected and autonomous vehicles. The table highlights key characteristics and limitations of each approach, demonstrating the field's increasing methodological sophistication.

Table 3.1: Consolidated analysis of methodological approaches in crash research

Technique	**Key Characteristics**	**References**
Generalized Linear Modeling (GLM)	Identification of three models with varying variables such as exposure, AADT, driveway density, curvature ratio, and roadside hazard rating. Limited to specific road section data; may not generalize to all road types.	[37]
Full Bayes (FB) hierarchical models	FB models better account for spatial correlation, showing higher accuracy in injury crash prediction compared with negative-binomial models. Complexity in implementing FB models at large scale due to computational demand. First-order adjacency improves fit and reduces bias in parameter estimates.	[38, 39]
Bayesian multivariate models	Multivariate Poisson lognormal approach enhances precision in crash-frequency estimates across severity levels. May require extensive data to calibrate effectively.	[40]
Multinomial Generalized Poisson (MGP)	MGP model with error components showed superior fit in analyzing crash frequency and severity together. Spatial exogenous-EMGP model best captures spatial dependencies in crash data. Complexity in interpreting factors contributing to both frequency and severity. Model complexity increases with alternative spatial structures.	[41, 42]

Table 3.1: (continued)

Technique	Key Characteristics	References
Accident modification factors (AMFs)	Curve radius AMFs derived for Texas showed higher crash risks on curves. Variability in intersection data may impact AMF accuracy.	[43]
Statistical and machine learning methods	Nearest-Neighbor Classification (NNC) had the best predictive performance; K-means clustering improved model performance; latent class clustering lowered NNC performance. Results may vary by method.	[44]
Spatial–geographic models	Random-parameter negative-binomial (RPNB) and S-GWPR models. S-GWPR better captures spatial heterogeneity and crash data correlation, improving regional crash modeling; requires high spatial granularity; may not apply to broader regions.	[45]
Statistical modeling	Bivariate negative-binomial spatial models; multilevel models; Full Bayes models; logistic regression; multivariate tobit models; comparative analysis with GLMs.	[46, 47, 37]
Random-parameter models	Account for heterogeneity; handle unobserved elements; incorporate correlated parameters; use instrumental variables.	[48, 34]

Table 3.1: (continued)

Technique	Key Characteristics	References
Surrogate safety measures	Traffic conflict techniques; in-vehicle data analysis; kinetic parameters for risk assessment.	[49, 50]
Injury severity analysis	Ordered probit models; neural networks; multivariate probit models; flexible econometric structures.	[51, 52]
Real-time risk prediction	Bayesian hierarchical models; temporal–spatial dependencies; weather and geometry factors.	[53, 54]
HMM prediction	Time-varying risk maps; real-time assessment.	[55]

3.4.1 Traditional Statistical Foundations

The evolution of crash analysis methodologies began with fundamental statistical approaches that established the empirical foundation for road safety research. Mannering and Bhat [16] provide context for the evolution of statistical methods in highway-accident research, highlighting persistent challenges like unobserved heterogeneity and endogeneity. Lord and Mannering [56] further elaborate on these methodological challenges, particularly in crash-frequency analysis, addressing issues such as data overdispersion, underreporting, and omitted-variable bias.

The foundation of crash analysis often begins with basic statistical approaches. Al-Ghamdi [57] demonstrates the application of logistic regression methodology in analyzing accident severity, identifying location and cause as significant variables. Building on these fundamentals, Jones and Jørgensen [47] show how multilevel modeling frameworks can better account for residual variation across accidents and geographical locations, revealing significant intra-unit correlation in accident outcomes. Cafiso

et al. [37] calibrated comprehensive accident models using extensive road characteristics data, demonstrating the application of Generalized Linear Modeling approaches with variables such as exposure, AADT, driveway density, curvature ratio, and roadside hazard rating.

A significant advancement came with the development of random-parameter models to address unobserved heterogeneity. Anastasopoulos and Mannering [48] explored random-parameter count models for analyzing vehicle accident frequencies, demonstrating ways to account for heterogeneity across various factors. Yasmin et al. [34] presented an econometric framework using instrumental variables to estimate causal effects while controlling for endogeneity. Saeed et al. [58] compared uncorrelated and correlated random-parameter count models, providing methodological guidance for model selection in multilane highway analysis. Anastasopoulos and Mannering [48] further explored these models using Indiana data, revealing significant impacts of pavement condition and geometric features. Anastasopoulos et al. [59] utilized random-parameter tobit regression for urban interstate accident analysis, identifying eleven significant factors affecting accident rates. Aziz et al. [60] applied random-parameter logit models to explain pedestrian injury severity levels in New York City.

Advanced regression techniques addressed specific analytical challenges in crash data. Anastasopoulos et al. [61] applied Tobit regression methodological frameworks to analyze vehicle accident rates, offering novel approaches to understanding crash data by treating accident rates as continuous variables. Castro et al. [62] proposed flexible econometric structures for highway segment analysis. Chen and Jovanis [63] developed variable-selection procedures for crash injury severity analysis. Fundamental issues in statistical modeling were addressed by Bijleveld [64], who addressed statistical issues in the simultaneous analysis of accident-related outcomes, particularly regarding variance–covariance structure estimation. Bhat [65] advanced the field computationally with the Maximum Approximate Composite Marginal Likelihood

(MACML) estimation method. Bhat et al. [66] introduced formulations for count data models with endogenous covariates, applying multinomial discrete-count modeling approaches to address self-selection and simultaneity bias.

3.4.2 Advanced Bayesian and Spatial Methods

The field progressed toward more sophisticated approaches that could better handle the complex spatial and temporal dependencies inherent in crash data. Aguero-Valverde and Jovanis [38] employed Full Bayes (FB) hierarchical modeling frameworks, incorporating spatial and temporal effects for county-level crash analysis. In subsequent work [39], they demonstrated the importance of spatial correlation structures in road crash-frequency models, showing that first-order adjacency structures improve model fit and reduce bias in parameter estimates. Their Bayesian multivariate Poisson lognormal modeling approach [40] shows improvements in precision estimation across severity levels, though requiring extensive data for effective calibration.

Wang et al. [46] advanced the field with bivariate negative-binomial spatial conditional autoregressive models for joint analysis of crashes and violations, enabling the identification of high-risk areas while accounting for spatial relationships. Xu and Huang [45] investigated spatial heterogeneity using random-parameter negative binomial and semi-parametric geographically weighted Poisson regression frameworks, demonstrating that geographically weighted approaches better capture spatial heterogeneity and crash data correlation for regional crash modeling.

Chiou and Fu [41] proposed integrated modeling approaches under the Multinomial Generalized Poisson architecture for the simultaneous analysis of crash frequency and severity. Chiou et al. [42] extended this framework with spatial Multinomial Generalized Poisson models, demonstrating superior performance of spatial modeling approaches, particularly the spatial exogenous-EMGP model for capturing spatial dependencies in crash data. Bonneson and Pratt [43] developed accident modification factors using

cross-sectional data, particularly effective for large roadway systems, with their work on curve radius AMFs showing higher crash risks on curves.

3.4.3 Machine Learning and Data Mining Approaches

As computational capabilities expanded, the field embraced machine learning and artificial intelligence techniques. Iranitalab and Khattak [44] compared statistical and machine learning for crash severity prediction, evaluating various classification algorithms, including Nearest-Neighbor Classification, finding that NNC had the best predictive performance, while K-means clustering improved model performance. Abdelwahab and Abdel-Aty [67] explored neural network frameworks for predicting driver injury severity, demonstrating early applications of artificial intelligence techniques to traffic safety analysis.

Chiou et al. [68] developed frameworks with genetic mining rules using stepwise rule-mining algorithms for crash severity analysis, integrating data mining techniques with traditional statistical modeling and demonstrating the effectiveness of a two-stage mining framework in capturing the joint effects of risk factors. Mahmud et al. [49] developed count data modeling approaches using traffic conflict techniques as surrogate safety measures, providing alternative methods for assessing road safety when crash data are limited. Zhang et al. [50] presented methodological frameworks for identifying crash risk through the coupling of in-vehicle data with kinetic parameters, advancing the integration of real-time vehicle dynamics in safety assessment.

Wu et al. [69] employed gradient boosting decision trees (GBDTs) to address key challenges in crash analysis, particularly the multicollinearity inherent in real-world traffic data. Their approach demonstrates strong predictive accuracy across four crash indicators while ranking 27 influential factors—revealing crucial insights into variable importance that traditional "black-box" machine learning methods obscure. This methodological contribution advances the field by combining robust

statistical performance with interpretability, enabling researchers to identify and understand complex relationships within crash datasets.

3.4.4 Real-Time Prediction and Emerging Technologies

Several recent developments focus on real-time analysis capabilities and applications in emerging vehicle technologies. Li et al. [70] developed hybrid Long Short–Term Memory-Convolutional Neural Network (LSTM-CNN) frameworks for real-time crash risk prediction, combining temporal sequence modeling with spatial feature extraction capabilities and demonstrating the benefits of parallel structures for capturing both long-term dependencies and local features. Zheng et al. [55] presented real-time risk assessment approaches for connected autonomous vehicles, utilizing HMM-based prediction methods and time-varying risk maps for continuous safety monitoring.

Castro et al. [53] incorporated temporal and spatial dependencies in modeling frameworks for urban intersection analysis, developing latent variable representations of count data models to accommodate for spatial and temporal dependence. Ahmed et al. [54] developed frameworks using Bayesian hierarchical models for analyzing crash frequencies with temporal and spatial dependencies on mountainous freeways. Adjenughwure et al. [71] proposed a Monte Carlo-based microsimulation approach to estimating collision probability in real traffic conflicts, with a methodology that can simulate conflicts involving an arbitrary number of vehicles under various initial conditions, using automated detection methods and accounting for variability in driver behavior parameters.

Specialized applications demonstrate their practical utility. Abdel-Aty and Keller [51] investigated factors influencing crash severity at signalized intersections by using ordered probit models. Abay [52] explored pedestrian injury severity by using various disaggregate modeling approaches. Abay et al. [72] presented multivariate probit modeling frameworks for the simultaneous analysis of injury severity and seat belt

use. Geographic-specific applications include work by Altwaijri et al. [73], who examined factors affecting crash severity in Riyadh; Abbas [74], who assessed rural road safety conditions in Egypt; and Jahan et al. [75], who proposed enhanced frameworks to model crash frequency while accommodating zero-crash zones. Infrastructure and environmental factors have been examined by Papadimitriou et al. [76], Wu et al. [77], Moslem et al. [78], and Farooq et al. [79] using a variety of techniques.

3.5 TARGETED SAFETY INTERVENTIONS

Traffic crashes arise from a complex interplay of infrastructure, road user, technological, and environmental factors. Figures 3.4–3.7 provide a structured overview of these domains and the specific topics reviewed in the following subsections.

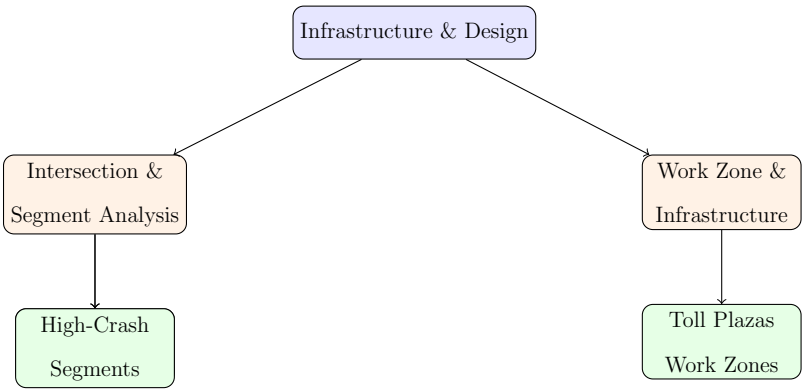


Figure 3.4: Infrastructure and design domain with its subcategories and example topics.

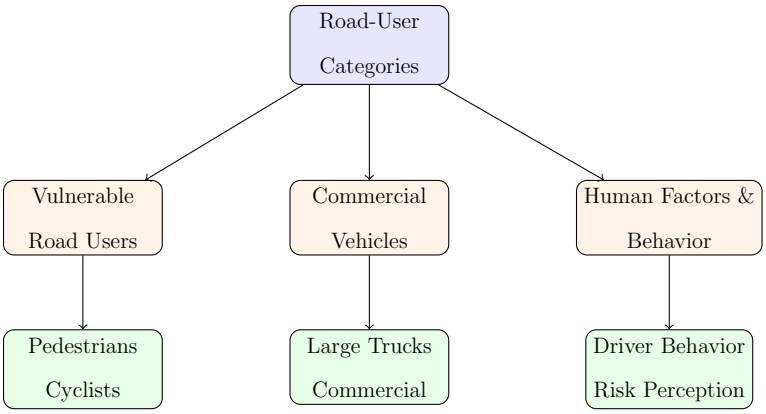


Figure 3.5: Road user categories and associated safety topics.

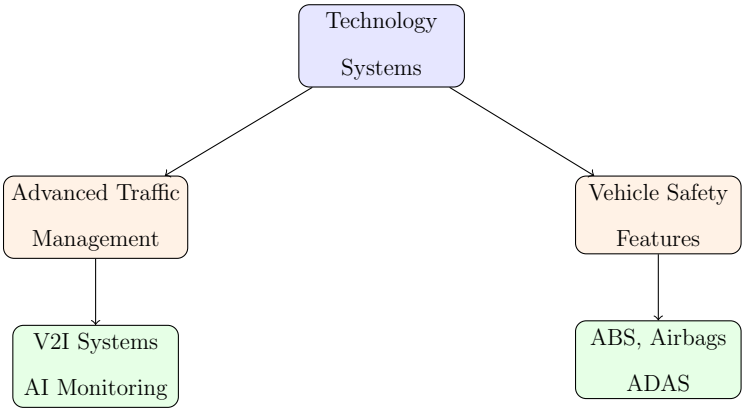


Figure 3.6: Technology-based safety systems: management and in-vehicle features.

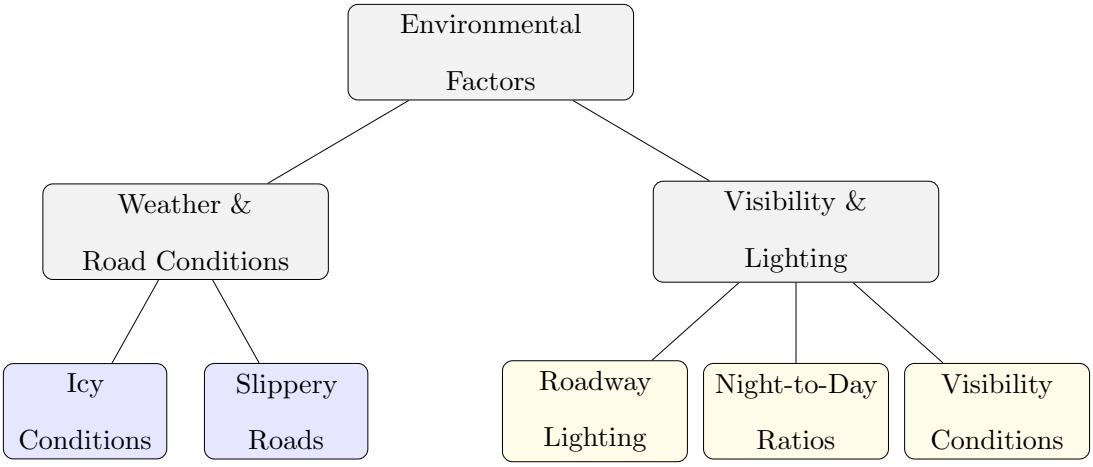


Figure 3.7: Factors affecting crash risk including weather and visibility.

3.5.1 Intersection and Segment-Level Crash Analysis

Retting et al. [80] investigated the characteristics and countermeasures for motor vehicle crashes at stop signs, focusing on four U.S. cities. The study found that stop sign violations—particularly "rolling" stops—accounted for roughly 70% of crashes, with younger and older drivers being disproportionately involved.

Boroujerdian et al. [81] proposed a dynamic wavelet-based model to locate and measure high-crash road segments. Their approach improves identification performance by 25–38% when the analyst seeks the worst 10–20% of roadway length, underscoring the value of precise segment delineation for multi-scale safety analysis.

Amoros et al. [82] compared traffic safety across French counties with Generalized Linear Models, revealing that differences in safety vary jointly with county and road type. Accounting for sub-county socio-economic factors and road-type mix substantially improves explanatory power.

Bonneson and McCoy [83] developed a negative-binomial crash-frequency model for 125 two-way stop-controlled intersections. With a product-of-flows formulation and gamma-distributed mean, the model captures a nonlinear rise in crashes with traffic demand, illustrating how such distributions pinpoint hazardous sites.

These studies demonstrate the complexity of intersection and segment safety analysis, with consistent evidence of nonlinear relationships between traffic volume and crash risk. Research reveals significant methodological diversity, from wavelet-based approaches achieving 25–38% performance improvements to negative-binomial models capturing traffic flow interactions. A critical finding across studies is the importance of spatial and demographic context: both Amoros et al.'s [82] county-level analysis and Retting et al.'s age-specific patterns highlight that location and user characteristics substantially influence crash patterns. The dominance of stop sign violations (70% of crashes) in Retting et al.'s findings, combined with Bonneson and McCoy's traffic demand relationships, suggests that intersection control design must

account for both human behavioral patterns and traffic flow characteristics.

3.5.2 Work Zone Safety and Roadway Infrastructure Factors

Abuzwidah and Abdel-Aty [84] analyzed crash frequency for different toll plaza designs. Compared with traditional plazas, hybrid designs cut crashes by 44.7% and all-electronic toll collection (AETC) by 72.6%. Diverge areas at hybrid plazas, however, exhibit 23% higher risk than merge areas.

Feknssa et al. [85] applied a random-parameter negative-binomial model with heterogeneity in means and variances to freeway ramp crashes. Ramp type, horizontal alignment, truck volume, and interchange geometry all significantly affect crash counts, demonstrating the need for spatially nuanced design standards.

Carson and Mannering [86] found that highway ice warning signs are not, by themselves, a significant crash-reduction factor—although their effects intertwine with location-specific variables that influence ice-related crash frequency and severity.

Anastasopoulos et al. [61] employed Tobit regression to treat crash rates as continuous outcomes on Indiana interstates, identifying pavement condition, geometry, and traffic composition as key predictors.

Chen and Tarko [87] compared two-level random-parameter and fixed-parameter negative-binomial models for work zone safety, showing that fixed-parameter specifications can suffice in certain contexts.

Petegem and Wegman [88] built a network-wide crash prediction model for rural roads. Roads with $\leq$2 m safety zones see 50% more run-off-road crashes, while sharp curves triple run-off-road risk; roadside barriers halve it.

Bhat et al. [66] introduced a count data model with endogenous covariates for urban intersection crashes, finding that crest approaches, frontage-road locations, and flashing-light control substantially increase crash numbers—stressing the value of addressing hidden as well as overt risk factors.

Kwon and Varaiya [89] documented chronic under-utilization and capacity penalties in California high-occupancy-vehicle (HOV) lanes. They argue that improving overall freeway efficiency, rather than expanding HOV networks, is the more cost-effective path.

Infrastructure interventions demonstrate a clear effectiveness hierarchy, with physical design modifications significantly outperforming information-based approaches. Electronic toll collection systems achieve the highest safety benefits (72.6% crash reduction), followed by roadside barriers (50% reduction), while ice warning signs show minimal effectiveness. This pattern suggests that passive infrastructure changes that modify driver behavior through design constraints are more effective than active warning systems requiring driver response. The studies reveal important spatial heterogeneity, with Feknssa et al.'s random-parameter models and Petegem and Wegman's safety zone analysis both emphasizing location-specific factors. Methodologically, the comparison between Chen and Tarko's findings and other studies suggests that while sophisticated random-parameter models often outperform fixed-parameter approaches, the context determines optimal model complexity. The failure of HOV lanes to achieve the intended benefits (Kwon and Varaiya) contrasts sharply with the success of toll plaza modifications, highlighting the importance of design compatibility with actual driver behavior rather than idealized usage patterns.

3.5.3 Vulnerable Road User Safety

Austin and Faigin [90] used travel surveys, crash databases, and an ordered probit model to show that older occupants travel more by passenger car and suffer higher risk in side-impact crashes, heightening fatal and serious-injury odds.

Brude and Larsson [91] demonstrated that even simple exposure models—with motor vehicle and unprotected-user counts—can give "nearly perfect" predictions of pedestrian and cyclist crashes. Risk rises with motor vehicle volume but falls as pedestrian

and cyclist volumes grow; cyclists face roughly double the risk of pedestrians under comparable conditions.

Vulnerable road user safety research reveals both age-related and mode-specific risk patterns that challenge conventional safety approaches. Austin and Faigin's findings on older occupants contrast with Brude and Larsson's "safety in numbers" effect for pedestrians and cyclists, suggesting that vulnerability mechanisms differ substantially between age-based and mode-based classifications. The counterintuitive finding that increased pedestrian and cyclist volumes reduce individual risk contradicts simple exposure-based models and implies that infrastructure and driver behavior adapt to user presence. However, cyclists face double the risk of pedestrians under similar conditions, indicating that mode-specific factors beyond simple exposure influence safety outcomes. These findings suggest that effective vulnerable user protection requires differentiated strategies addressing both demographic vulnerability (age-related) and mode-specific risks.

3.5.4 Large Truck and Commercial Vehicle Safety

Abdel-Aty and Abdelwahab [92] developed nested logit models showing that visibility obstruction by light trucks markedly increases the likelihood of a following passenger car striking them in rear-end crashes—especially when the lead vehicle brakes sharply.

Ballesteros et al. [93] found that pedestrians struck by SUVs or pickup trucks suffer more severe and fatal injuries than those hit by conventional cars; vehicle mass and speed are key drivers, with front-end geometry influencing injury patterns at lower speeds.

Commercial vehicle safety research demonstrates that vehicle size and mass create dual safety challenges: increased crash likelihood through visibility obstruction (Abdel-Aty and Abdelwahab) and increased injury severity when crashes occur (Ballesteros et al.). Both studies highlight the interaction between vehicle design characteristics

and crash dynamics, with visibility obstruction increasing rear-end collision probability while mass and geometry determine injury outcomes in pedestrian crashes. The emphasis on sudden- braking scenarios and front-end geometry suggests that commercial vehicle safety interventions must address both crash prevention through improved visibility and injury mitigation through design modifications. These findings indicate that the growing prevalence of larger vehicles in traffic streams creates compound safety challenges requiring multi-faceted intervention approaches.

3.5.5 Human Factors, Driver Behavior, and Risk Perception

Chang and Yeh [94] identified common and divergent fatality-risk factors for motorcyclists and other drivers, emphasizing seat belt use, speed management, rider risk perception, and low-class roadway quality.

Bédard et al. [95] showed that drivers aged 80+ are five times more likely to die in crashes than those aged 40–49; seat belt use is strongly protective, whereas alcohol effects vary with concentration.

Benfield et al. [96] revealed that anthropomorphizing vehicles (e.g., attributing "agreeable" personalities) can predict aggressive driving as well as or better than driver personality traits.

Bhat and Eluru [97] used a copula-based model to disentangle built-environment and self-selection effects on daily vehicle miles traveled (VMT), finding that traditional Gaussian assumptions understate self-selection's contribution.

Hasan et al. [98] reviewed distracted-driving studies, highlighting links to workload, environment, demographics, and roadway design. They advocate surrogate safety metrics, targeted lighting/lane-marking treatments, and technology-based countermeasures.

Human factor research reveals complex interactions among demographics, psychology, and behavior that challenge simple intervention approaches. Age emerges as

a critical factor across studies, with Bédard et al.'s five-fold mortality increase for drivers 80+ and Chang and Yeh's age-specific patterns for motorcyclists. However, the effectiveness of protective factors varies significantly—seat belt use provides consistent protection across age groups and vehicle types, while alcohol effects vary with concentration and user characteristics. The psychological dimensions explored by Benfield et al. suggest that vehicle anthropomorphism may be as predictive of dangerous behaviors as traditional personality measures, indicating that human–vehicle interaction psychology warrants greater attention in safety interventions. The complexity of built-environment effects (Bhat and Eluru) and multi-faceted nature of distraction (Hasan et al.) underscore that human factor interventions must account for individual differences, environmental context, and technological integration rather than relying on universal behavioral assumptions.

3.5.6 ATMSs: Advanced Traffic Management Systems

Thabit et al. [99] divided modern monitoring and management into four phases—data gathering, transmission, analysis, and application—surveying sensor technologies, 4–6 G and LPWAN communications, and AI-driven analytics for congestion and safety.

De Souza et al. [100] cataloged challenges for traffic management systems, including heterogeneous data, real-time hazard representation, route-choice side effects, and security/privacy in vehicular ad hoc networks.

Mandal et al. [101] presented a deep learning traffic-surveillance suite (Mask R-CNN, YOLO, and Faster R-CNN) that detects queues with 90.5% accuracy and stationary vehicles with an F1 of 0.83, outperforming manual methods.

Milanes et al. [102] prototyped a V2I-based fuzzy-logic controller that dynamically manages headways and speed, prioritizing safety in complex urban layouts; IEEE 802.11p field tests confirm feasibility.

Advanced traffic management system research demonstrates the progression from conceptual frameworks to practical implementation, with significant performance achievements but persistent challenges. Mandal et al.'s deep learning systems achieve impressive detection accuracy (90.5% for queues), while Milanes et al.'s V2I controllers demonstrate real-world feasibility through field testing. However, the challenges cataloged by de Souza et al.—heterogeneous data integration, real-time processing, and security concerns—highlight the gap between laboratory performance and system-wide deployment. The four-phase framework proposed by Thabit et al. provides structure for understanding system complexity, but the practical challenges suggest that each phase presents implementation barriers that may limit overall system effectiveness. The contrast between high-performance individual components and systemic integration challenges indicates that ATMS success depends as much on architectural design and standardization as on component-level performance.

3.5.7 Vehicle Features: ABS, AirBags, and ADAS

Høye [103] found that frontal airbags cut driver fatalities by 22% for belted occupants but provide no net benefit to unbelted drivers, contradicting earlier claims of airbag-induced risk.

Kusano and Gabler [104] evaluated three pre-collision system (PCS) algorithms; the most comprehensive one (FCW + PBA + PB) reduces injury severity by up to 34% and could prevent 3.2–7.7% of rear-end crashes.

Ding et al. [105] compared 1001 SAE-L2 ADAS and 548 SAE-L4 ADS crashes, finding L2 events cluster on highways and L4 in urban areas; low mileage and new-technology generation correlate with lower injury odds.

Vehicle safety technology research reveals important patterns in system effectiveness and user interaction dependencies. The effectiveness of safety systems varies dramatically with user behavior—airbags provide a 22% fatality reduction for belted

drivers but no benefit for unbelted drivers (Høye), illustrating that passive safety systems require complementary protective behaviors. More advanced systems show greater effectiveness, with comprehensive pre-collision systems achieving an injury reduction of up to 34% (Kusano and Gabler) compared with airbags' 22% benefit. However, Ding et al.'s analysis of automated driving systems reveals that the deployment context significantly influences the outcomes, with different automation levels experiencing distinct crash patterns (L2 on highways and L4 in urban areas). The correlation between low mileage and lower injury odds suggests learning effects or selection bias in early adopters. These findings indicate that vehicle safety technology effectiveness depends on the interaction among system sophistication, user behavior, and deployment context rather than technology capabilities alone.

3.5.8 Weather, Environmental, and Temporal Factors

Malin et al. [106] employed Palm probability to relate crash risk to the time spent on a segment; relative risk is the highest in icy rain and on slippery surfaces, with single-vehicle crashes being particularly sensitive.

Bullough et al. [107] linked roadway lighting to night-to-day crash ratios, finding observed ratio drops ($\leq$13%) smaller than the oft-cited 30%, likely because of uncontrolled covariates in earlier studies.

Zhang et al. [108] built a spatial multinomial-logit injury-severity model with real-time weather, identifying vertical grade, visibility, EMS response time, and vehicle type as key factors; spatial correlation improves fit and predictive accuracy.

Environmental factor research demonstrates the complexity of weather and visibility effects on crash risk, with important implications for intervention assessment. Malin et al.'s finding of the highest risk during icy rain conditions, combined with the particular sensitivity of single-vehicle crashes, suggests that environmental interventions must target specific weather–crash type combinations rather than general adverse

conditions. The lighting research by Bullough et al. reveals a significant methodological issue: the observed safety benefits ($\geq$13%) are substantially smaller than commonly cited values (30%), highlighting how uncontrolled covariates can overstate intervention effectiveness. Zhang et al.'s integration of real-time weather data with spatial modeling demonstrates that environmental factors interact with infrastructure characteristics (vertical grade) and emergency response capabilities, suggesting that effective environmental safety interventions require integrated approaches considering multiple interacting factors. The consistent emphasis on spatial correlation across studies indicates that environmental effects vary significantly by location, challenging universal intervention strategies.

3.6 APPLICATIONS AND POLICY IMPLICATIONS

Road safety research has evolved significantly in recent decades, moving from simple before–after studies to sophisticated analytical approaches that combine empirical evidence with advanced statistical methods. This evolution has enabled a more nuanced understanding of safety interventions and their effectiveness, leading to evidence-based policy recommendations. This section examines key applications and methodological advances in road safety analysis, focusing on critical areas of safety intervention evaluation, the identification of high-risk locations, and the development of analytical frameworks for safety performance assessment.

3.6.1 Evidence-Based Safety Interventions

The evaluation of traffic safety measures requires robust methodological approaches to separate true effects from statistical artifacts and confounding factors. Recent studies have employed increasingly sophisticated methods to assess various safety interventions, providing crucial insights for policy development.

Legislative and Behavioral Interventions

Cohen and Einav [109] conducted a landmark study on mandatory seat belt laws by using panel data analysis. Their findings challenge previous assumptions about the magnitude of safety benefits while providing important evidence against the risk compensation hypothesis. Specifically, their analysis shows that while seat belt laws significantly reduce traffic fatalities, the effect is more modest than earlier estimates suggested, and importantly, they found no evidence of compensatory risk-taking behavior among drivers.

Chang and Yeh's [94] comparison between non-motorcycle drivers and motorcyclists revealed common factors as well as risk discrepancies between the two groups. The study concluded that enhancing seat belt use rates, speed management, rider risk perceptions, and road quality improvements are particularly important in reducing the risk of fatality for both groups.

Bédard et al.'s [95] analysis found that drivers aged 80+ are five times more likely to experience fatal injuries compared with those aged 40–49 while confirming the protective effects of seat belts. These findings support age-specific driver assessment and vehicle design policies, highlighting the need for targeted interventions for older drivers.

Infrastructure Modifications and Design Interventions

Infrastructure modifications have been subject to rigorous evaluation with varying degrees of success. Zheng and Sayed [110] demonstrated the effectiveness of smart channel conversions for right-turn lanes, employing time-to-collision metrics and extreme value theory. Their finding of a 34% reduction in severe conflicts, though with limitations regarding merging conflicts, provides valuable guidance for intersection design policies.

Abuzwidah and Abdel-Aty's [84] evaluation of toll plaza designs found that hybrid

toll plazas result in 44.7% fewer crashes than traditional toll plazas, while all-electronic toll-collection systems achieve 72.6% fewer crashes. For hybrid systems, crash risk in diverge areas is 23% higher than in merge areas. These findings provide clear guidance for toll plaza design as they indicate that all-electronic toll plazas are significantly safer.

Petegem and Wegman [88] modeling results found that roads with safety zones of 2 m or less resulted in 50% more run-off-road crashes, while strong curvature increases run-off-road crashes by three times compared with straight roads. Roadside barriers were found to reduce 50% of run-off-road crashes compared with roads with small safety zones. These specific percentages provide quantitative guidance for rural road design standards.

However, not all infrastructure interventions prove effective. Carson and Mannering's [86] statistical analysis showed that ice warning signs were not a significant factor in reducing accident frequency or severity, indicating that this common safety measure may not provide the expected benefits and resources might be better allocated elsewhere.

Vehicle Technology Safety Impacts

Vehicle safety technologies have demonstrated significant benefits when properly implemented. Høye's [103] analysis found that airbags reduce driver fatality for belted drivers by 22%; however, airbags are neither effective nor counterproductive for unbelted drivers. This finding supports the continued promotion of seat belt use alongside airbag deployment.

Kusano and Gabler's [104] evaluation of pre-collision systems found that systems utilizing forward collision warning, pre-crash brake assist, and autonomous pre-crash brake achieved the highest effectiveness, reducing severity by 14–34% and reducing severity for belted drivers by 29–50%. The systems could prevent 3.2–7.7% of rear-end

collisions, supporting policies mandating such technologies.

Ding et al.'s [105] analysis revealed distinct operational patterns between ADAS (SAE Level 2) and ADS (SAE Level 4) vehicles, with ADAS crashes being concentrated on highways and ADS crashes in urban environments. This finding supports targeted testing and deployment strategies for different automation levels.

Intersection and Traffic Control Interventions

Retting et al.'s [80] investigation of motor vehicle crashes at stop signs across four U.S. cities revealed that stop sign violations, particularly when drivers had initially stopped, accounted for about 70 percent of crashes, with younger and older drivers being disproportionately involved. These findings provide specific targets for intervention design and driver education programs.

Bonneson and McCoy's [83] analysis of 125 two-way stop-controlled intersections demonstrated that accident frequency follows a gamma distribution with nonlinear increases relative to traffic demands. This relationship enables the identification of hazardous locations based on traffic volume thresholds.

In an innovative study, Yanmaz-Tuzel and Ozbay [111] utilized Full Bayes analysis to evaluate various road safety countermeasures. Their work not only identified the most effective interventions—including improved road alignment and median barrier installation—but also advanced the methodological framework by demonstrating the advantages of P-LN model structures with hierarchical priors for limited-data scenarios.

Vulnerable Road User Protection Strategies

Austin and Faigin's [90] analysis revealed that older individuals are more likely to be involved in side-impact crashes compared with younger occupants, which significantly increases their fatality and injury risk. This finding supports targeted vehicle design improvements and intersection safety modifications for aging populations.

Brude and Larsson's [91] modeling results show that accident risk involving unprotected road users increases with motor vehicle numbers while decreasing with more pedestrians and cyclists present. Additionally, accident risk is approximately twice as high for cyclists compared with pedestrians under similar traffic conditions. These findings support policies promoting safety in numbers and differentiated protection strategies.

Ballesteros et al.'s [93] investigation in Maryland revealed that pedestrians hit by SUVs and pickup trucks are more likely to suffer severe injuries and fatalities compared with conventional cars, with vehicle weight and speed being significant contributors. These findings support vehicle design regulations and urban speed limit policies.

Commercial Vehicle Safety Interventions

Abdel-Aty and Abdelwahab's [92] analysis demonstrated that the visibility obstruction caused by light truck vehicles significantly increases the probability of rear-end collisions involving regular passenger cars, particularly when the lead vehicle stops suddenly. This supports policies regarding commercial vehicle design standards and following distance regulations.

Chen and Tarko's [87] analysis identified specific safety effects of work zone designs and traffic management features, providing evidence-based guidance for temporary traffic control strategies during construction activities.

Environmental and Weather-Related Interventions

Malin et al.'s [106] analysis showed relative accident risks to be the highest for icy rain and slippery road conditions. The overall relative accident risk is lower on motorways compared with other road types; however, risk under poor weather conditions is higher on motorways. These findings support weather-responsive traffic management

strategies.

Nighttime driving can be more dangerous due to reduced visibility, but Bullough et al.'s [107] found that the crash risk increased about 12%, less than previously assumed, suggesting that lighting improvements provide measurable but modest safety benefits that should be evaluated against costs.

Zhang et al.'s [108] investigation using real-time weather data identified key risk factors including vertical grade, visibility, emergency medical services response time, and vehicle type. These factors provide specific targets for infrastructure improvements and emergency response optimization.

3.6.2 Spatial Analysis and Risk Assessment

The spatial dimension of road safety has emerged as a crucial consideration in both research and practice, leading to new approaches in hotspot identification and network screening.

Methodological Advances in Spatial Analysis

Ziakopoulos and Yannis [112] provided a comprehensive review of spatial analysis methods in road safety, emphasizing the critical role of spatial heterogeneity and dependence in crash risk analysis. Their work establishes a framework for incorporating geographical factors into safety assessments, highlighting the importance of appropriate areal unit selection and Bayesian modeling approaches.

Ryan et al. [113] advanced the field by integrating risk assessment into path planning through an innovative modification of Dijkstra's algorithm. Their approach combines traditional distance optimization with risk exposure metrics while employing self-organizing maps to identify distinct risk groups. This methodology bridges the gap between theoretical risk assessment and practical route planning applications.

Afghari et al. [114] developed an integrated approach to blackspot identification, combining crash count and severity in a joint econometric model. Their weighted risk score methodology, which incorporates both frequency and severity predictions, demonstrates superior performance in identifying locations with high risk of severe injuries. This approach provides a more nuanced tool for prioritizing safety improvements, particularly in contexts where resources are limited.

Boroujerdian et al.'s [81] dynamic modeling approach demonstrated a 25–38% improvement in comparison with existing models for identifying 10–20% of high-crash road segments. This performance improvement has direct implications for resource allocation in safety improvement programs.

Amoros et al.'s [82] comparison of traffic safety across French counties identified significant interactions between county and road type, indicating that differences in safety across counties depend on the road type and vice versa. This finding emphasizes the need for location-specific safety strategies rather than one-size-fits-all approaches.

3.6.3 Safety Performance Functions and Crash Modification Factors

Safety Performance Functions (SPFs) have become the cornerstone of modern traffic safety analysis since their introduction in the Highway Safety Manual (HSM) [115] by the American Association of State Highway and Transportation Officials (AASHTO). These statistical models, which predict the average crash frequency for a given site type under specific conditions, are now widely used by the Federal Highway Administration (FHWA) and infrastructure analysts across the United States, providing transportation agencies with robust tools for identifying high-risk locations and evaluating safety improvements.

The foundation for modern SPF implementation was established by Hauer [116], who demonstrated the Empirical Bayes (EB) method's effectiveness in addressing

two critical challenges: improving precision with limited crash data and mitigating regression-to-mean bias. This methodological breakthrough, combined with the increasing availability of calibrated SPFs and overdispersion parameters, has facilitated the widespread adoption of EB estimation in safety analysis.

The practical implementation of these methods has been demonstrated in various contexts. Powers and Carson [117] developed an accessible Excel-based approach for evaluating safety improvements in Montana's roadway reconstruction projects. Their work highlighted both the method's utility and its constraints, particularly the requirement for three-year aggregated crash data to ensure reliable SPF modeling.

Persaud and Lyon [118] provided crucial validation of the EB methodology, demonstrating its superiority over traditional approaches in before–after safety studies. Their research emphasized the importance of comprehensive data collection and proper analyst training while also identifying potential pitfalls in CMF derivation. They proposed future research directions, including refinements to SPFs and the exploration of Full Bayes (FB) modeling for handling spatial correlations in accident data.

A significant advancement came from Hauer [119], who challenged the conventional assumption of uniform overdispersion parameters. His work showed that shorter road sections were disproportionately affected by this assumption, leading to potentially biased estimates. By proposing length-dependent overdispersion parameters, Hauer improved the accuracy of EB estimates across varying road section lengths.

Elvik [120] conducted a comprehensive assessment of the EB method's performance in observational studies, confirming its position as the leading approach for before–after safety analyses. His research demonstrated that EB estimates based on accident prediction models achieved the highest accuracy among available methods, though noting variations in prediction errors across different techniques.

Recent developments, exemplified by Park et al. [121], have expanded the application of both EB and FB methods to specific safety interventions. Their analysis

of roadside barriers revealed the importance of considering multiple factors in safety assessments, including vehicle characteristics, driver demographics, and environmental conditions. This work demonstrates the evolution of CMFs toward more nuanced, condition-specific applications, reflecting the increasing sophistication of safety analysis methods.

This progression in methodology, from basic prediction models to sophisticated multi-factor analyses, coupled with FHWA's standardization of SPFs, has established a robust framework for evidence-based safety analysis in transportation infrastructure. Future developments will likely focus on incorporating emerging data sources and refining condition-specific applications while maintaining the fundamental principles that have made these methods successful. Table 3.2 summarizes the develpment of SPFs and Crash Modification Factors (CMFs).

Table 3.2: Evolution and applications of SPFs and CMFs in road safety analysis.

Methodology	**Key Contributions and Limitations**
Empirical Bayes (EB)	**Contributions:** Precise estimation in sparse-data settings; corrects regression-to-mean bias. **Limitations:** Requires well-calibrated SPFs and overdispersion parameters [116].
EB for infrastructure assessment	**Contributions:** Post-reconstruction safety evaluation (Montana); Excel-based implementation. **Limitations:** Needs three-year aggregated crash counts [117].
EB methodology validation	**Contributions:** Demonstration of EB's superiority in CMF derivation. **Limitations:** Sensitivity to data quality and underlying EB assumptions [118].

Table 3.2: (continued)

Methodology	Key Contributions and Limitations
Variable overdispersion	**Contributions:** Length-based overdispersion to reduce short-segment bias. **Limitations:** Breaks uniform-parameter assumption; more complex calibration [119].
EB in observational studies	**Contributions:** Lower prediction errors than alternatives; decade-long assessment. **Limitations:** Context-specific performance; data-intensive [120].
Advanced Bayesian methods	**Contributions:** EB vs. Full Bayes comparison; condition-specific CMFs. **Limitations:** Higher computational cost; needs richer data [121].

3.6.4 Economic Analysis, Crash Costs, and Resource Allocation

Bougna et al. [122] completed a quantitative analysis of studies estimating the socio-economic costs of road crashes, highlighting methodological differences between high-income countries (favoring willingness-to-pay method) and lower-income countries (using human capital approach). They conclude that there is potential for high returns on investment in road safety measures and argue that comprehensive cost analysis can bolster support for crash-reduction programs, potentially driving economic development.

Wijnen et al. [123] analyzed road crash cost estimates for 31 European countries, providing an overview of the official monetary valuations. The study found the total costs of road crashes to be 0.4–4.1% of GDP. The valuation of preventing a serious injury was determined to be 2.5–34.0% of the value per fatality and the valuation of preventing a slight injury to be 0.03–4.2% of the value per fatality. The results

reveal that the method of obtaining valuations majorly impacts results, underlining the importance of harmonization of valuation practices.

Wu et al. [69] examined the economic dimensions of road safety in Zhongshan, China, revealing a nonlinear relationship between GDP per capita and crash outcomes. Their analysis demonstrates that economic development initially increases crash risk but reduces crashes beyond approximately RMB 60,000 per capita, when improved economic conditions enable safety investments. The study documents how the government allocation of RMB 546 million for road infrastructure improvements and RMB 7 million for safety education during 2008–2009 resulted in measurable crash reductions, illustrating the potential for strategic economic resource allocation in safety interventions.

Zaloshnja et al. [124] estimated the costs per crash for three crash severity groups within 16 selected crash geometry types and 2 speed limit categories by using police crash reports. The results of the study find the most costly crashes to be non-intersection, fatal or disabling injury crashes on roads with a speed limit of at least 50 mph where there were head-on collisions or human–vehicle collisions. These crashes are estimated at over USD 1.69 and USD 1.16 million per crash, respectively. The study also found run-off-road collisions to make up 34% of total crash costs.

Pirdavani et al. [125] demonstrated the application of zonal crash prediction models to evaluate travel demand management strategies, specifically examining fuel cost increases as a safety intervention. Their analysis of a 20% fuel price increase scenario in Flanders, Belgium, predicted an 11.57% reduction in vehicle kilometers traveled and a corresponding 2.83% decrease in crash frequency, illustrating how economic policies can yield measurable safety benefits through reduced exposure.

3.6.5 Emerging Technology Applications and Connected Vehicle Integration

A challenges when introducing new technology, including connected and autonomous vehicles, is the public's risk assessments and acceptance. Ahmed et al.'s [126] performed a survey of public opinions and found that while 66% and 68% of respondents expect fewer and less severe crashes with autonomous vehicles, significant concerns exist regarding equipment failure (71%), system failures (73%), hacking (68%), and privacy breaches (74%). These findings highlight critical areas requiring attention for successful deployment and public acceptance.

Autonomous Vehicle Crash Patterns and Safety Implications

Bogg et al.'s [127] analysis of California crash data revealed that 61.1% of autonomous vehicle-including accidents were rear-end collisions. Environmental factors, such as mixed land use and proximity to schools, play a significant role in crash propensity. These findings support targeted safety system improvements and deployment strategies, particularly enhanced rear-end collision avoidance through automatic emergency braking systems in conventional vehicles.

Mixed Traffic Flow Dynamics

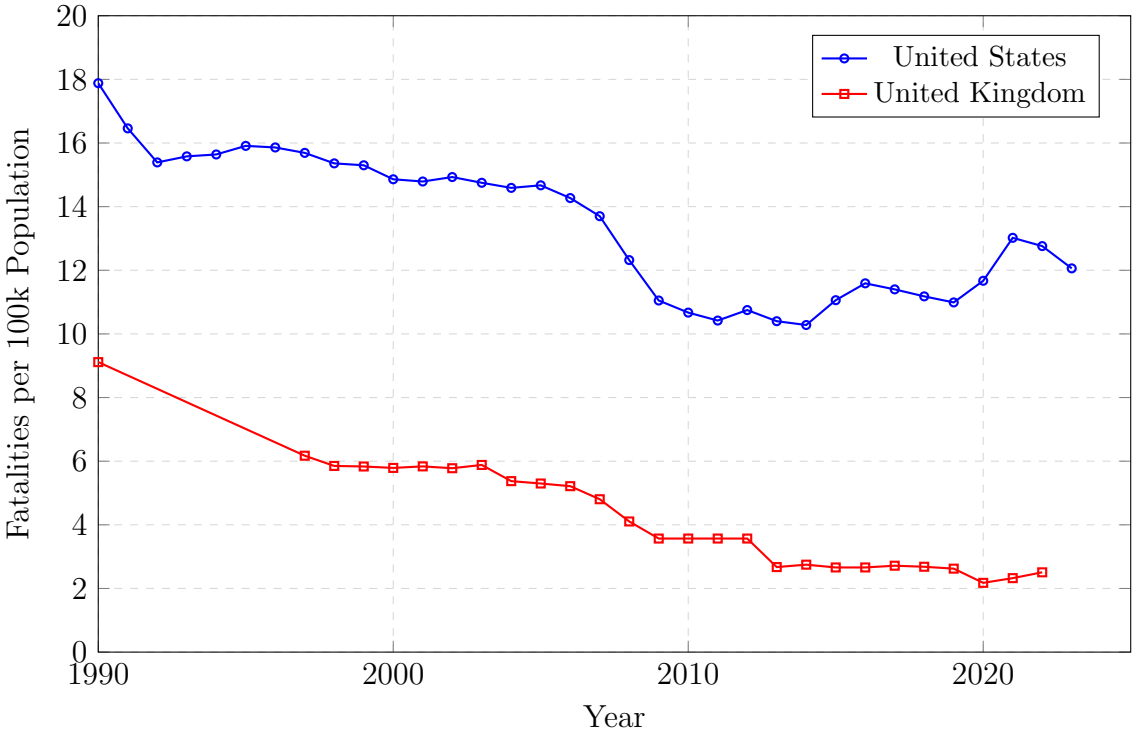
Chang et al.'s [128] analysis revealed that intelligent and connected vehicles can improve mixed traffic flow stability under a critical speed and effectively improve traffic capacity. However, they can degrade stability if the critical speed is exceeded, with this critical speed decreasing as the maximum platoon size increases. These findings have important implications for traffic management policies in mixed autonomous and conventional vehicle environments.

3.6.6 Impact of Interventions

It can be hard to measure the direct impact of individual safety interventions. When seat belts were introduced, for instance, there was no specific date at which all cars had seat belts—there was a gradual transition from optional equipment to mandatory installation, to mandatory usage laws, and finally to widespread compliance. Similar gradual adoption patterns apply to other safety technologies and legal changes, from ABS brakes to drunk-driving laws.

If we take a closer look at recent decades, we can compare traffic fatalities to population to assess the efficacy of various safety interventions and countermeasures. Figure 3.8 presents this comparative analysis for both the United States and the United Kingdom from 1994 to 2022. The longitudinal data reveal divergent trends in road safety outcomes between these nations, despite similar levels of economic development and technological advancement. While both countries have implemented evidence-based safety measures since the 1990s, when U.S. fatality rates peaked at approximately 15.7 per 100,000 inhabitants, the rate disparity remains significant, with the U.S. currently experiencing about 12.9 deaths per 100,000 inhabitants compared with less than 3 in the United Kingdom.

While exposure differences, particularly higher vehicle miles traveled (VMT) in the United States, contribute to this disparity, multivariate analyses suggest that this factor alone does not fully explain the variation. The U.S. built environment, characterized by auto-centric development patterns, creates systematic exposure risks by necessitating motor vehicle use across all demographic groups, including populations that may be more susceptible to crash involvement. Moreover, controlled studies examining fatality rates per vehicle mile traveled (VMT) indicate persistent disparities, suggesting fundamental differences in system design parameters, vehicle fleet characteristics, and regulatory frameworks between the two countries.



Data source: [129]

Figure 3.8: Comparison of motor vehicle fatalities per 100,000 inhabitants between the United States and the United Kingdom (1994–2022).

The divergent trajectories, particularly the recent uptick in U.S. fatalities while the UK rates maintain a downward trend, underscore the critical role of systemic factors and policy interventions in determining road safety outcomes. This empirical evidence suggests that elevated U.S. fatality rates are not merely a function of increased exposure through higher mobility but rather reflect addressable systemic factors. These findings have important implications for the application of countermeasures and the transfer of successful safety interventions between jurisdictions, particularly in the context of emerging analytical methodologies and real-time crash prediction systems.

3.7 EMERGING RESEARCH AREAS AND FUTURE DIRECTIONS

This section reviews emerging research, a summary of which is available in Table 3.3.

Table 3.3: Summary of emerging research areas and key findings.

Research Area	**Key Contributions and Findings**	**References**
Big Data Analytics and Data Mining	Two-stage mining framework integrating 29 mined rules into mixed logit model; identifies seat belt fastening as most critical safety condition; capture of joint effects of risk factors in single-vehicle freeway crashes.	Chiou et al. [68]

Table 3.3: (continued)

Research Area	**Key Contributions and Findings**	**References**
Deep Learning and Advanced AI Applications	Comparative analysis shows simpler models often achieve performance comparable to or better than deep models; random forest models are most effective for crash risk prediction using crowdsourced probe vehicle data.	Huang et al. [130]; Zhang et al. [131]
Real-Time Crash Risk Prediction	Hybrid LSTM–CNN model with parallel structure captures long-term dependencies and local features; achieves AUC 0.93, highest sensitivity, and lowest false alarm rate for urban arterial prediction.	Li et al. [70]
Connected and Autonomous Vehicle Safety	Survey of 584 U.S. respondents: 66–68% expect fewer/less severe crashes; concerns include equipment failure in poor weather (71%), system failures (73%), hacking (68%), and privacy breaches (74%).	Ahmed et al. [126]

Table 3.3: (continued)

Research Area	**Key Contributions and Findings**	**References**
Autonomous Vehicle Crash Pattern Analysis	COOLCAT clustering identifies six AV-including crash clusters from UK STATS19; 61.1% are rear-end collisions; environmental factors like mixed land use and school proximity influence crash propensity.	Esenturk et al. [132]; Bogg et al. [127]
Intelligent Connected Vehicle Traffic Flow	Mixed-traffic analysis shows ICVs improve stability below critical speeds and enhance capacity; stability degrades when critical speed is exceeded; critical speed decreases as maximum platoon size increases.	Chang et al. [128]

3.7.1 Big Data Analytics and Data Mining Techniques

Chiou et al. [68] developed a genetic mining rule model utilizing a stepwise rule-mining algorithm. Their study integrated 29 mined rules into a mixed logit model to identify key safety and risk conditions associated with severe crashes. Their analysis revealed that seat belt fastening was the most critical safety condition, while risk conditions included vehicle type, alcohol use, driver characteristics, time period, road status, and surface condition. These findings demonstrate the effectiveness of a two-stage mining framework in capturing the joint effects of risk factors contributing to single-vehicle crash severity on freeways.

The application of advanced data mining techniques, as exemplified in the study, provides a robust foundation for identifying intricate, multi-dimensional relationships in traffic safety, thereby informing more effective crash prevention strategies.

3.7.2 Deep Learning and Advanced AI Applications

In a study performed in 2020, Huang et al. [130] concluded that while deep models can be effectively applied to traffic data for crash occurrence classification and risk prediction, simpler models can often achieve comparable or even better performance. Specifically, they found that for crash detection, CNNs with dropout outperformed some shallow models and, for crash prediction, deep models showed comparable performance to shallow models.

Building on this, Zhang et al. [131] utilized a state-wide live traffic database that provides crowdsourced probe vehicle data to develop real-time traffic crash prediction models. The crash prediction models use machine learning models to predict crash risk according to pre-crash traffic dynamics and static freeway attributes. The results of the study reveal a significant relationship between rear-end crashes and pre-crash traffic dynamics. Additionally, the study ranks traffic speed factors in terms of feature of importance, finding the speed variance and speed reduction prior to crashes to be most important, both of which are positively related to rear-end crash risk. Random forest models emerged as the most effective among various machine learning approaches, highlighting significant relationships between rear-end crashes and pre-crash traffic dynamics. Key predictive factors included speed variance and reductions prior to crashes, offering actionable insights for traffic safety interventions.

Together, these studies underscore both the potential and limitations of AI-driven methodologies in crash analysis and risk prediction, emphasizing the necessity of aligning model selection with specific data and research objectives.

3.7.3 Integration of Emerging Data Sources and Technologies

Recent studies have shed light on the prevalence and impact of distracted driving in the United States. Cambridge Mobile Telematics (CMT) and Arity, two companies aggregating data from mobile phones and vehicle telematics, provide alarming insights:

- CMT's 2023 report reveals that 34% of all drivers who crash interact with their phone in the minute before the crash [133].
- Arity's 2023 report notes a 30% increase in distracted driving per mile from 2019 to 2023 [134].

These findings contrast with the National Highway Traffic Safety Administration's (NHTSA) 2022 research note, which reports lower percentages of distraction-affected crashes [135]. The discrepancy can be attributed to different methodologies and data sources, highlighting the value of telematics data in supplementing traditional police crash reports.

The studies also reveal interesting patterns in distracted-driving behavior, including seasonal and geographic variations. However, it is important to note potential limitations in the data sample collected by companies like Arity and CMT, such as selection bias and the focus on phone-based distractions.

These findings underscore the urgent need for continued efforts to combat distracted driving through legislation, enforcement, education, and technology-based solutions.

3.7.4 Real-Time Crash Risk Prediction and Proactive Safety Management

Li et al. [70] developed a hybrid Long Short-Term Memory–Convolutional Neural Network (LSTM-CNN) model to predict real-time crash risk on urban arterials. Using a year's worth of traffic, signal, and weather data, they applied SMOTE to address data imbalance. Their parallel LSTM-CNN model outperformed other methods, including sequential LSTM-CNN, LSTM, CNN, XGBoost, and Bayesian Logistic Regression,

achieving the highest AUC of 0.93, the highest sensitivity, and the lowest false alarm rate. The study demonstrated the potential of deep learning in traffic safety prediction, highlighting the benefits of combining LSTM with CNNs in a parallel structure for capturing both long-term dependencies and local features.

3.7.5 Safety Implications of Connected and Autonomous Vehicles

To ensure the effective deployment of autonomous vehicle (AV) technologies, it is crucial to account for both public perception and the underlying safety challenges highlighted by real-world accident data. While Ahmed et al. [126] underscore the public's optimism regarding the potential of AVs to reduce crash frequency and severity, their findings also emphasize significant apprehensions about system failures, cybersecurity risks, and privacy concerns. They analyzed public perceptions of autonomous vehicles (AVs) by applying a grouped random-parameter bivariate probit model with heterogeneity in means. Based on a survey of 584 U.S. respondents, the study found that while 66% and 68% expected fewer and less severe crashes, respectively, significant concerns existed regarding equipment failure in poor weather (71%) and potential crashes due to system failures (73%). Furthermore, 68% of respondents worry about hacking and terrorist attacks, while 74% express concerns about privacy breaches. The study highlights the importance of continuously monitoring these perceptions for effective AV deployment strategies.

In the same vein, Esenturk et al. [132] discussed solutions to traffic safety regarding autonomous vehicles (AVs) through two main objectives: identifying patterns in traffic accidents and developing test scenarios for AVs based on these patterns. The authors analyze the STATS19 accident data, a dataset of 20,000 accidents from the UK, using the COOLCAT clustering algorithm, which is designed for high-dimensional categorical data. This analysis reveals six distinct clusters of traffic accidents, each characterized by unique real-world situations, aiding in the understanding of risk

factors. Additionally, the study employs association rule mining to create non-trivial test scenarios for AVs, addressing the industry's challenge of ensuring safe deployment in risky situations. The findings show the value of clustering techniques and more effective data collection methods to inform safety strategies for emerging vehicle technologies, contributing to safer transportation systems.

These findings by Esenturk et al. [132] underscore the importance of addressing complex accident patterns and developing tailored safety strategies for autonomous vehicles. Expanding on this focus, Bogg et al. [127] transitioned to real-world crash data from California, offering valuable insights into specific collision types and the environmental factors influencing AV-including accidents. Together, these studies highlight the critical need for both predictive safety frameworks and practical interventions to enhance the safe integration of AVs in diverse traffic environments. Their research concludes that while automated vehicles (AVs) in California have accumulated significant mileage, the insights gained from analyzing crash reports reveal critical patterns in AV-including accidents, particularly the high frequency of rear-end collisions (61.1%). The study emphasizes the need for careful consideration of unobserved heterogeneity in crash data, advocating for the use of informative uniform priors in Bayesian models over the traditional uninformative inverse-gamma priors. The findings suggest that environmental factors, such as mixed land use and proximity to schools, play a significant role in crash propensity. Practical implications include the potential for enhanced rear-end collision avoidance through the implementation of automatic emergency braking systems in conventional vehicles, which could lead to improved safety outcomes in mixed traffic scenarios involving both AVs and human-driven vehicles.

Similarly, Chang et al. [128] delved into the dynamics of mixed traffic scenarios, shedding light on how intelligent and connected vehicles (ICVs) influence traffic flow stability and capacity. They analyzed the traffic flow configurations and the spatial

distributions of various types of vehicles when mixed traffic flow is in equilibrium. The study revealed that intelligent and connected vehicles (ICVs) can improve the stability of mixed traffic flow under a critical speed; however, ICVs can degrade stability if the critical speed is exceeded. This critical speed decreases as the maximum platoon size of ICVs increases. Additionally, the results also suggest that ICVs can effectively improve traffic capacity.

Collectively, these studies underscore the transformative potential of connected and autonomous vehicle technologies while emphasizing the necessity of addressing technical and societal challenges for their safe and effective integration into transportation systems. The integration of AV and ICV technologies demands a multidisciplinary approach to ensure their benefits are maximized while mitigating associated risks.

3.8 CONCLUSIONS

This systematic review addressed three fundamental challenges in crash data analysis: data quality issues, methodological fragmentation, and research–practice gaps. Through the comprehensive analysis of methodological evolution from descriptive to system-based approaches (Figure 3.3) and the systematic categorization of data quality challenges (Figure 3.2), we demonstrate how sophisticated analytical approaches can be balanced with practical applicability. That being said, persistently high injury rates in traffic, particularly in the United States (Figure 3.8), reveal limited success in translating research advances into effective countermeasures.

3.8.1 Key Methodological Advancements

The field has witnessed significant evolution in analytical approaches. The progression from fixed-parameter to random-parameter models has improved accounting for unobserved heterogeneity, while hierarchical Bayesian methods have enhanced the incorporation of spatial–temporal correlations. Advanced spatial analysis techniques, including

geographically weighted regression, have revealed geographical patterns in crash occurrences.

Data integration represents another crucial advancement. Combining police reports, hospital records, and insurance claims has addressed underreporting and misclassification issues. This multi-source approach, complemented by surrogate safety measures and traffic conflict analysis, provides alternatives when crash data are limited.

Machine learning and AI applications have uncovered complex, nonlinear relationships in crash data. Real-time crash risk prediction, utilizing streaming sensor and telematics data, enables proactive safety management. Novel severity analysis approaches, including latent class and mixed logit models, have improved injury outcome identification. Big data analytics has opened avenues for discovering previously unknown risk patterns, while advances in addressing endogeneity and self-selection bias have produced more accurate intervention estimates.

Despite these advances, autonomous and mixed traffic solutions emerge as the most promising frontier, alongside continuous data quality improvements.

3.8.2 Future Research Directions

Given their potential for transformative change, autonomous and mixed traffic solutions should be prioritized as primary research directions. Big data availability presents unprecedented analytical opportunities, requiring advanced data mining and machine learning algorithms to extract meaningful patterns and uncover hidden risk factors.

Real-time crash risk prediction represents a prominent frontier. The evolution from Yuan et al.'s LSTM-RNN improvements [136] through Lim et al.'s Temporal Fusion Transformer architecture [137] to Han et al.'s transformer-based approach [138] demonstrates rapid progress. Their 15.69% recall improvement over traditional methods highlights the potential of integrating connected vehicle data for comprehensive risk assessment.

Data integration remains critical. While traditional analysis relies on police reports, emerging approaches leverage connected vehicle and roadside sensor data—resources rarely available in standard records. This mirrors successful practices like Sweden's mandatory hospital reporting system. Additional data streams, including social media, detailed weather information, and expanded telematics data, offer further potential for holistic risk assessment.

Despite social acceptability [139, 140] and trust challenges [141, 142], autonomous and connected vehicles with advanced safety features [143, 144] will require new analytical frameworks for human–autonomous vehicle interactions. Addressing endogeneity and self-selection bias [33, 34] remains crucial to accurate intervention evaluation.

Interdisciplinary research combining crash analysis with behavioral psychology shows promise. McCarty et al. [145] demonstrated that demographic factors explain over 28% of accident rate variance, while Gu et al. [146] revealed how environmental factors create complex causation chains. These findings emphasize the need for comprehensive models accounting for multiple interacting factors—from individual behavior to demographic patterns and environmental conditions.

Success requires pursuing data-driven approaches leveraging technological and methodological advances but also effectively bridging the persistent gap between research sophistication and practical implementation. Only through this integration can the field fulfill its potential to significantly reduce road crashes and save lives worldwide.

CHAPTER 4

LLM-ASSISTED LOCATION VALIDATION: IMPROVING TRAFFIC CRASH DATA ACCURACY

Crash data quality challenges significantly impact safety analyses, from hot-spot identification to real-time crash prediction models. These issues can be broadly categorized into systemic dimensions and specific error types.

4.0.1 Systemic Data Quality Dimensions

1. **Data Completeness**

   Under-reporting issues persist across jurisdictions, varying significantly by crash severity and road user type. Studies indicate that while nearly all fatal crashes are recorded by law enforcement, minor crashes and those involving vulnerable road users such as pedestrians and cyclists suffer from substantial under-reporting. A meta-analysis of reporting in 13 countries further supports this, finding that official statistics often fail to capture injuries comprehensively: approximately 95 percent of fatal injuries, 70 percent of serious injuries (hospitalized cases), 25 percent of slight injuries (treated as outpatients), and just 10 percent of very slight injuries (treated outside hospitals) are reported on average. Reporting levels vary considerably between countries and road user types, with car occupants more frequently reported than cyclists. Notably, single-vehicle bicycle accidents are rarely included in official road accident statistics, highlighting a persistent gap in data completeness that can significantly skew safety analyses and resource allocation decisions.

2. **Location Accuracy**

   The accurate recording of crash locations is fundamental for effective safety analysis and countermeasure implementation. This dimension forms the core focus of our study, as spatial precision directly impacts the ability to identify high-risk areas and implement targeted interventions. Inaccurate location data can lead to misidentification of hazardous locations and ineffective deployment of safety countermeasures, ultimately reducing the effectiveness of road safety programs.

3. **Temporal Accuracy**

   Several aspects of analysis depend on the accurate capture of the time and date of the crash, including any pre-crash conditions. Temporal data enables analysis of crash patterns across different times of day, weather conditions, and seasonal variations. This information is crucial for understanding the role of environmental factors, traffic volumes, and emergency response times in crash outcomes. Accurate temporal data also facilitates the correlation of crashes with specific events, road conditions, or traffic patterns.

4. **Severity Classification**

   The consistent and accurate classification of crash severity is essential for prioritizing safety interventions and allocating resources. Classification systems must account for both immediate injury assessment and delayed onset of symptoms, particularly in cases involving vulnerable road users. Variations in severity classification methods between jurisdictions can complicate cross-jurisdictional comparisons and may affect the allocation of safety resources. Standardized severity scales, such as the KABCO injury scale, help maintain consistency but require proper training and application by law enforcement personnel.

4.0.2 Error Types and Sources

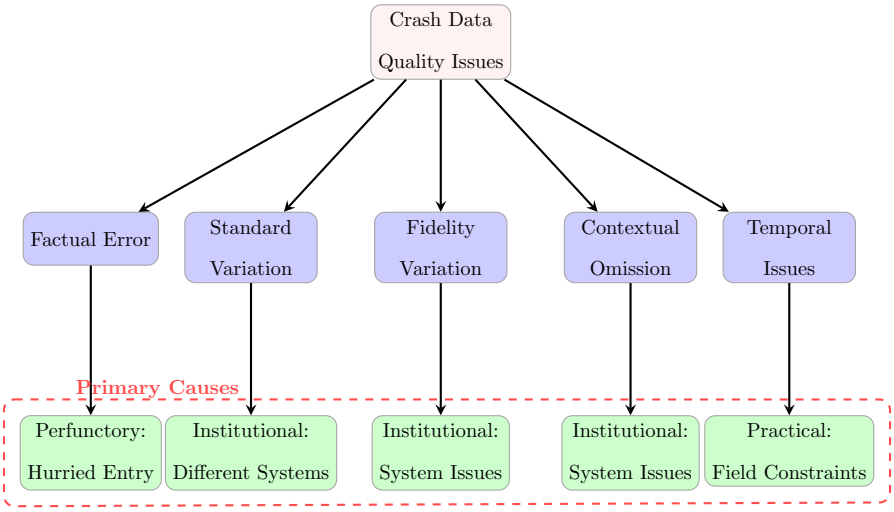


Figure 4.1: Hierarchy of crash data quality issues with their primary sources

Data quality issues in crash reporting stem from three fundamental sources: perfunctory data entry (reflecting carelessness or insufficient attention), practical constraints of field data collection, and systematic institutional differences in documentation approaches. These sources manifest differently across five distinct categories of data quality issues, as explained below and visualized in Figure 4.1.

- **Factual Error**: Incorrect recording of directly observable crash details.
  - *Primary source*: Perfunctory - Misrecorded facts due to carelessness or inattention during data entry.
  - *Secondary source*: Practical - Errors resulting from adverse environmental conditions or time pressure at crash scenes.
- **Standard Variation**: Differences in how crash information is classified and recorded.
  - *Primary source*: Institutional - Fundamental differences between classification systems.

  - *Secondary source*: Practical - Variations arising from limitations in available technology or local practices.

- **Fidelity Variation**: Differences in the level of detail captured in crash reports.
  - *Primary source*: Institutional - Variations in granularity requirements across agencies (e.g., "at intersection" versus specific approach arm).
  - *Secondary source*: Practical - Detail limitations due to resource constraints or time pressure.

- **Contextual Omission**: Missing information that could affect crash analysis and contribute to endogeneity concerns.
  - *Primary source*: Institutional - Systematic exclusion of factors not aligned with institutional priorities or protocols.
  - *Secondary source*: Perfunctory - Failure to document readily available contextual information.
  - *Tertiary source*: Practical - Inability to capture temporary conditions or environmental factors.

- **Temporal Issues**: Problems related to timing and sequencing aspects of crash data.
  - *Primary source*: Practical - Difficulties in establishing exact crash times or event sequences.
  - *Secondary source*: Perfunctory - Imprecise recording of timing information despite availability.

4.0.3 Autonomous Vehicle Classification Challenges

Issues of data quality affect several components of crash reports, not just crash location. A particularly clear example of how Law Enforcement Officers are left to interpret

details without sufficient instruction and education is the classification of autonomous vehicles in crash reports. The classifications present at least two challenges:

1. Lack of clear definitions: While forms may include up to "Full Automation" (Level 5), even robotaxis are at level 4, and only certain cities allow them today. Among vehicles currently available for purchase, SAE Level 2 is the highest level of autonomy allowed on public roads in most jurisdictions. Without expert guidance, officers may struggle to identify the correct level in the field.

2. Rapidly evolving technology: As autonomous vehicle technology advances, the definitions and capabilities associated with each level may change faster than crash report forms can be updated. Only certain states allow Level 3 automation, and even then only in specific areas, but more states may follow soon.

This situation exemplifies how even well-intentioned efforts to capture detailed data can lead to quality issues, particularly when dealing with complex and rapidly evolving technologies. The potential for misclassification or inconsistent application of these categories across different incidents or jurisdictions poses significant challenges for data aggregation, analysis, and the development of evidence-based safety policies.

4.1 HOW CRASH LOCATION IS RECORDED BY LAW ENFORCEMENT OFFICERS

Crash location recording procedures vary across jurisdictions. Based on our review of several states, we find several common approaches to documenting crash locations, though specific requirements differ among them.

In Florida, crash location reporting offers four options to identify the crash site:

1. **Street, Road, or Highway**: Record the name of the highest classification of the trafficway where the crash occurred. For parking lot crashes, include the address; for private property, specify *private property* and the address.

2. **Street Address**: Provide the street address number if applicable. This field is not required if other location data such as latitude/longitude, intersection, or milepost is used.

3. **Latitude and Longitude**: Enter the coordinates of the crash location in float format (e.g., -85.869586). Latitude and longitude values are optional and can substitute for other location fields.

4. **Intersection or Milepost**: Specify the distance and direction from the nearest intersection or milepost. Measurements can be in feet or miles, and the direction should indicate N, S, E, or W.

Ohio maintains stricter requirements, mandating latitude and longitude coordinates for all crashes, unlike Florida's optional approach. Texas is similar to Florida, allowing but not requiring GPS coordinates. Both Ohio and Texas provide fields for route numbers and road names, with clear rules for prioritizing route systems and using secondary references in intersections.

The approaches across Florida, Ohio, and Texas reflect different priorities in balancing flexibility, precision, and redundancy:

- **Latitude and Longitude**: While optional in Florida and Texas, these coordinates are mandatory in Ohio, showing varying approaches to geospatial data collection.

- **Road System and Street Names**: All three states require the use of street names or roadway systems when available, with Texas implementing detailed hierarchical rules.

- **Street Address Requirements**: Texas specifically emphasizes that GPS coordinates do not replace the need for street address information, which must always be provided. Florida offers more flexibility by allowing latitude/longitude to substitute for other fields.

4.1.1 International Perspectives: The Swedish Model

International standards for crash reporting can deviate significantly from those in the United States. Sweden, for instance, is recognized as a leader in traffic safety and is the birthplace of Vision Zero—a strategy aimed at eliminating all traffic fatalities and severe injuries.

According to Swedish guidelines for Law Enforcement Officers, the location of a crash must be captured in a manner that leaves no doubt as to where the incident occurred. This is achieved by:

- **Precise Location Identification**: Documenting the accident site using the road number and/or street name, along with the distance to the nearest intersecting street or road.
- **Supplementary Locality Data**: Including the name of the district, municipality, or locality when possible.
- **Enhanced Precision via GPS**: When available, employing GPS coordinates to pinpoint the exact location of the incident.

The method employed in Sweden is similar to practices observed in some U.S. states—emphasizing the need for detailed and unambiguous location information. However, regarding technical precision, Ohio's crash reporting system is the most stringent among the examples discussed, through the mandatory use of GPS coordinates for all crash reports. While this approach is intended to promote precise geolocation by capturing latitude and longitude data, practical challenges such as data entry errors can still result in inaccuracies. Thus, the benefit of the mandated precision offered by GPS data may sometimes obscure underlying data quality issues such as those discussed in the preceding section.

4.2 LITERATURE REVIEW

The impact of inaccurate crash data on road safety analyses is likely significant. When Imprialou and Quddus [25] reviewed the literature, they found that data quality problems vary in severity and extent across different attributes and are especially severe for crash location and timing, challenges in linking databases due to inconsistencies, misclassification of crash severity, incomplete or inaccurate demographic information of those involved, and incorrect identification of factors contributing to crashes.

The validation of traffic accident locations from police reports has challenged researchers for three decades, with various methodological approaches proposed. Work by Levine and Kim in the late 1990s pointed to the need for "efforts to enhance data quality [involving] better training and standardization of location reporting throughout the entire data management process."

The error rate in crash locations varies across studies. Miler et al. [24] found that 33.5% of crashes in a database of 8,550 observations had inaccurate location attributes. Their innovative approach employed fuzzy string matching using the Jaro-Winkler distance, achieving a 15% improvement over classical methods. This work was particularly notable for its use of OpenStreetMap data, which provided access to local variants of street names.

Looking toward the future, Imprialou and Quddus [25] suggest that emerging intelligent crash reporting systems, incorporating GPS-based applications and automated data collection, could significantly reduce location errors. These systems are being implemented in several countries, including the US, UK and Italy, though their adoption faces challenges related to cost, training requirements, and potential technological vulnerabilities.

4.2.1 Research Gap

While previous work has established various methodologies for location validation, from probabilistic matching to sophisticated map-matching algorithms, these approaches have generally relied on either exact string matching or predetermined similarity metrics. The potential of leveraging modern language models' semantic understanding capabilities remains unexplored in this domain. Furthermore, while advanced algorithms using artificial intelligence concepts have achieved high accuracy rates in correcting crash locations - with some approaches reaching 98.9% accuracy - these methods are rarely implemented in practice. This limited adoption may be due to the complexity of developing and validating such algorithms, which often require hundreds of manual location verifications.

Our work aims to address these gaps by developing and evaluating an LLM-based approach to location validation, potentially offering a more robust and adaptable solution that could see wider practical adoption.

4.3 PROPOSED SOLUTION

Our proposed approach represents a significant methodological departure from previous work by utilizing both structured data validation and visual-textual analysis for location validation. This hybrid approach offers several potential advantages over traditional methods:

1. *Multi-source Verification*: By comparing the coordinates given by the Department of Transportation (ODOT) and the Police (ODPS), we increase confidence in location accuracy.

2. *Contextual Understanding*: Our approach leverages a multi-modal LLMs to extract and compare information from crash diagrams and written narratives.

3. *Spatial Validation*: Geospatial database queries verify consistency with administrative boundaries and known road networks.

4. *Progressive Confidence Building*: Rather than binary validation, we implement a credibility scoring system that accumulates evidence across multiple dimensions.

Our primary data source consists of individual crash reports from the Ohio Department of Transportation, an example of which is shown in Figure 4.2. These reports contain rich structured and unstructured information including both ODOT and ODPS (local police department) coordinates, reference location information, narrative description, and a crash diagram.

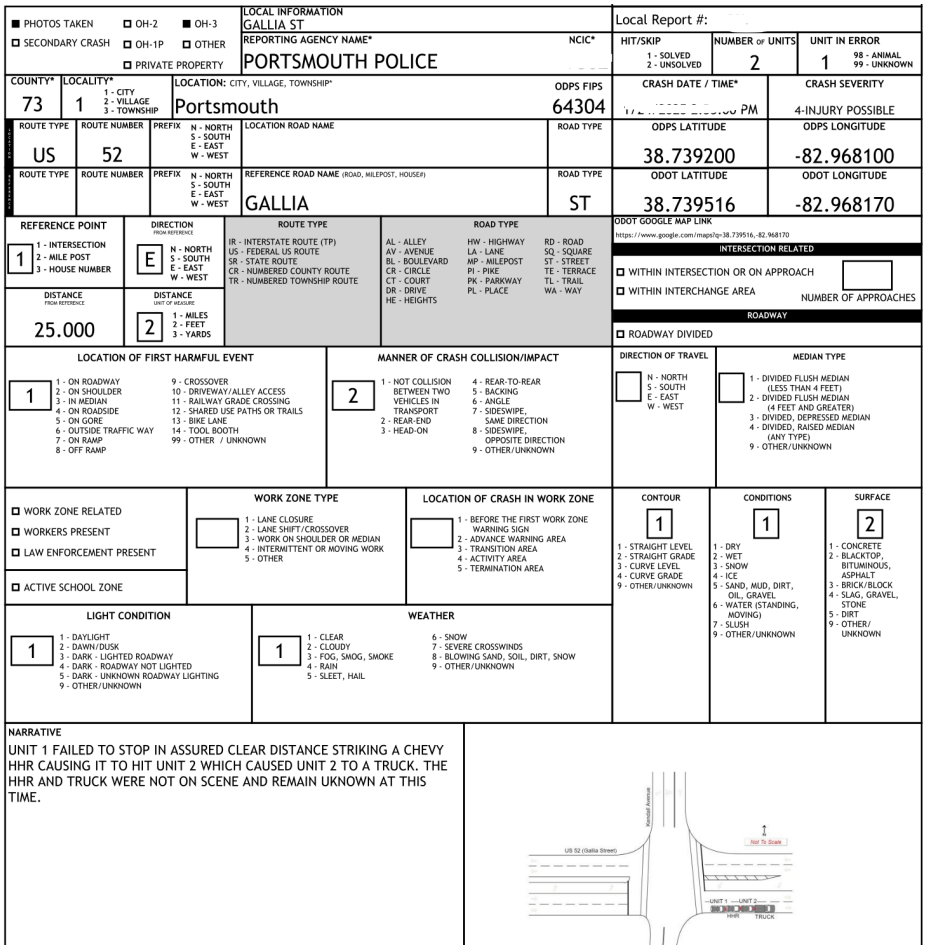


Figure 4.2: Example of an Ohio crash report showing the key data elements used in our validation process: ODOT and ODPS coordinates, reference point information (intersection of US 52 and Gallia St), and the crash diagram that visually represents the location and circumstances.

The validation process employs a robust, object-oriented Python framework designed for maximum efficiency and accuracy. Our system processes crash reports in batches with comprehensive error handling, enabling smooth processing of thousands of documents simultaneously. When location discrepancies are detected, our methodology applies a systematic validation approach that dynamically selects the most appropriate strategy based on available data:

For intersection references, the system executes precise geospatial queries against comprehensive road network data. For milepost references, it interfaces with a specialized milepost database for accurate linear referencing. For house number references, it leverages external geocoding services to pinpoint specific addresses.

Each successful validation incrementally builds the report's overall credibility score. The framework implements rigorous validation thresholds, rejecting any potential corrections that exceed 800 meters from official coordinates. All validated reports are seamlessly integrated into a spatial database, facilitating efficient aggregation, visualization, and analysis across the entire dataset.

Figure 4.4 shows the flowchart of our credibility-based validation system for determining whether the geolocation provided in a crash report is accurate. The validation process builds a cumulative credibility score through a series of checks:

- **Coordinate Consistency** (0.7 points): Compares Ohio Department of Transportation (ODOT) and Police Department (ODPS) coordinates, awarding points if they match within 50 meters. For example, in the report shown in Figure 4.2, ODOT coordinates (38.739516, -82.968170) and ODPS coordinates (38.739200, -82.968100) are within the 50-meter threshold.

- **Administrative Boundary Verification** (0.2 points): Verifies that coordinates fall within the expected FIPS county boundaries. The crash report shows FIPS code 64304 for Portsmouth, which corresponds to the county code 73 (Scioto County).

- **Crash Diagram & Narrative Consistency** (0.5 points): Uses a Large Language Model to extract road information from the crash diagram and verify it against the written narrative. In our example, the diagram shows the intersection of US 52 and Gallia Street, which matches the location data in the form fields.

- **Reference Point Validation** (0.5 points): Validates location against specific reference points (intersections, mileposts, or house numbers) through spatial database queries and geocoding. The example report indicates reference point type 1 (intersection) with a distance of 25 feet in the East direction.

We recommend a credibility threshold of 1.0 for most applications, which typically requires successful validation across at least two independent checks. This threshold ensures both geographic consistency and corroboration with other report elements. The system also implements a minimum credibility score of 0.5 before attempting more computationally expensive validation steps, improving efficiency.

4.3.1 Real-Time Application

Our system can be implemented for real-time validation during incident entry. This process ensures accurate geographical information through a two-stage validation approach: county-level validation and road-level verification.

The process begins with automatic GPS coordinate acquisition from the officer's device. These coordinates undergo immediate validation against the expected county boundaries. When coordinates fall within the expected county, the system proceeds directly to road entry. However, if coordinates indicate an unexpected county, the system alerts the officer and requires confirmation. This geographic validation step prevents inadvertent out-of-jurisdiction entries while maintaining flexibility for legitimate cross-boundary cases.

Following county validation, the system progresses to road-level verification. The entered road/location is checked against an authorized database. For roads not

immediately found, the system provides nearby suggestions to account for potential spelling variations or unofficial road names. Officers can select from these suggestions or, if necessary, proceed with manual road entry. This multi-tiered approach balances automation with officer discretion, ensuring both accuracy and operational flexibility.

The process concludes by transitioning to reference point validation only after both county and road information have been properly verified. This structured approach maintains data integrity while accommodating the various scenarios officers encounter in the field.

4.4 EXPERIMENTAL VALIDATION AND RESULTS

We ran our algorithm on a sample of 5,000 crashes in Ohio, approximately 1,000 random samples for each severity level recorded in Ohio (fatal, serious injury suspected, minor injury suspected, injury possible, and property damage only). Our analysis revealed that approximately 20% of reports required location corrections, indicating significant geospatial discrepancies in official crash data.

The correction rates showed minimal variation across severity levels: 17.7% for fatal crashes, 20.3% for serious injuries, 21.5% for minor injuries, 22.8% for possible injuries, and 20.1% for property damage only cases. This relatively consistent pattern across severity categories suggests that location validation challenges are fundamental to crash reporting methodology rather than being influenced by crash severity.

The nature of these corrections varied systematically based on reference point types. Intersection references (type 1) dominated the validation process across all severity levels, accounting for 931 of 1027 successful corrections (90.7%). Milepost references (type 2) were less common but still significant, particularly for fatal crashes where they represented 19.8% of corrections compared to 7.9% for property damage only crashes. House number references (type 3) proved extremely rare, appearing in only 2 cases, both for property damage only crashes.

These findings reveal important patterns in crash location reporting. The predominance of intersection-based corrections (90.7%) likely reflects two key factors: (1) the higher frequency of crashes at intersections, which are known conflict points in the roadway network, and (2) the relative ease of validating locations where two named roads meet, providing clear reference points for both manual and automated correction systems. Despite this intersection bias, the presence of successful corrections using milepost and house number references, particularly in fatal crashes, demonstrates the value of maintaining multiple reference systems in crash location validation. The consistent correction rates across severity categories suggests that location reporting challenges represent a systematic issue in crash reporting infrastructure rather than being influenced by the specific circumstances or severity of individual crashes.

4.5 EMERGING TECHNOLOGIES AND THEIR IMPACT

4.5.1 Event Data Recorders in Europe

The European Union's mandate requiring all new vehicles sold from July 2024 to include an Event Data Recorder (EDR) represents a significant step toward improving road safety through data-driven analysis. EDRs are designed to capture critical crash-related data, such as vehicle speed, braking activity, seatbelt usage, and airbag deployment, providing valuable insights into crash dynamics. However, to address privacy concerns, the EU regulations explicitly exclude the recording of GPS location, audio, video, or any data that could identify the driver or passengers. This restriction ensures that the EDR focuses solely on technical vehicle performance rather than personal or behavioral monitoring, thereby balancing the need for safety improvements with the protection of individual privacy. Consequently, EDR data is anonymized and only accessible under specific legal or investigative circumstances, maintaining a clear boundary between public safety objectives and personal data protection.

4.5.2 Automated Reporting by Manufacturers

Correctly identifying the level of autonomy is one thing, but it may be even more challenging for an officer to accurately determine which level of autonomy was engaged at the time of the crash, especially if the system has disengaged or been manually overridden in response to the incident.

Recognizing challenges and opportunities with autonomous vehicles, California has introduced Assembly Bill 3061, which addresses the need for more comprehensive and accurate reporting of autonomous vehicle incidents. Key provisions of this bill include:

- Requiring manufacturers of autonomous vehicles to report to the Department of Motor Vehicles (DMV) on any vehicle collision, traffic violation, disengagement, or barrier to access or incident of discrimination for a passenger with a disability that involves a manufacturer's vehicle in California, starting July 31, 2025.
- Mandating specific criteria to be reported for various types of incidents, including collisions, traffic violations, disengagements, and accessibility issues.
- Establishing a system for public reporting of incidents involving autonomous vehicles, with a process for the DMV to verify and investigate these reports.
- Implementing penalties for non-compliance, including fines and potential suspension or revocation of a manufacturer's permit.

4.5.3 Connected Vehicle Data Opportunities

Connected Vehicles are mainly associated with Autonomy but even with semi-autonomous vehicles, it would be possible for vehicles to share data. Work is underway on protocols for such data sharing and although privacy will need to be maintained, formalizing data collection and sharing hold promise for a better future when it comes to data quality.

4.6 INTEGRATED APPROACHES TO CRASH REPORTING

4.6.1 Multi-Source Data Integration

The value of comprehensive data integration is not limited to academic studies but is also recognized in national reporting systems. For instance, Sweden's national traffic injury reporting system, STRADA, categorizes the degree of completeness in injury reporting based on the integration of various data sources [26]. Figure 4.3 illustrates this categorization, demonstrating how different combinations of data sources contribute to a more complete picture of road accidents. Note that even when including both police reports and hospital reports, a subset (grey in figure 4.3) of crashes are not reported anywhere. In fact, a 2017 study in Sweden found that only about 63% of traffic-related injuries were captured in STRADA, while the Patient Registry (PAR) captured approximately 65% of cases. The overlap between these systems was surprisingly low, with only about 30% of road traffic injuries being recorded in both systems.

In Sweden, hospitals are legally mandated to report all patients injured on public roads to the national injury database. While similar data consolidation efforts exist in the United Kingdom and the Netherlands, these practices remain voluntary rather than legally required. Beyond official reporting systems, some academic researchers have successfully incorporated insurance data to complement police and hospital records [17]. However, such insurance data are typically only available for specific time-bound analyses and rarely accessible at regional or national scales.

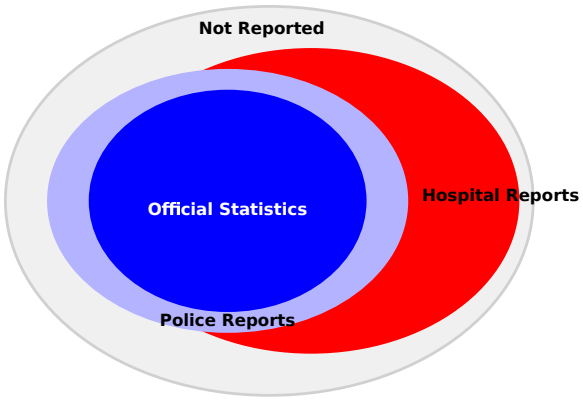


Figure 4.3: Injury Reporting Sources in Sweden's STRADA System

4.7 DISCUSSION

Our findings reveal significant opportunities for improvement in crash location data quality. This section examines the broader implications of these findings for transportation safety practice and research.

4.7.1 Implications for Practice

The implementation of our validation and correction methodology has several important implications for transportation safety practitioners:

1. **Data-Driven Decision Making**: More accurate crash locations enable more precise identification of high-risk roadway segments and intersections. This improved spatial precision can substantially enhance hotspot analysis and countermeasure selection, potentially leading to more effective safety interventions.

2. **Resource Allocation Optimization**: Transportation agencies operate with limited resources. By improving location accuracy, our methodology enables more targeted allocation of safety improvements where they are most needed, maximizing return on investment for safety infrastructure.

3. **Real-time Reporting Improvements**: One opportunity to improve data quality presents itself in the initial data collection process. Implementing our validation methodology as part of electronic report filing could provide immediate feedback to officers, potentially eliminating most location errors at the source.

4. **Cross-agency Coordination**: The observed discrepancies between ODOT and ODPS coordinates highlight the need for better standardization and coordination between agencies. Our framework provides a mechanism for identifying and resolving these inconsistencies systematically.

5. **Reference Point Prioritization**: Our results indicate that the most effective reference point type varies based on initial coordinate agreement. Agencies should prioritize different validation methods based on this observed pattern:

    - When coordinates show high consistency (within 50 meters), intersection references should be prioritized
    - When coordinates disagree significantly, milepost references become more reliable validation sources

6. **Integration with Existing Systems**: The modular nature of our validation framework allows for integration with existing crash reporting systems with minimal disruption to current workflows. This facilitates gradual adoption and improvement of location data quality.

The practical impact of these improvements extends beyond mere data accuracy. More precise crash locations enable better targeting of engineering countermeasures, more effective enforcement strategies, and more focused educational campaigns—the three pillars of comprehensive transportation safety programs.

4.8 CONCLUSIONS AND FUTURE RESEARCH

This study introduces a novel methodology for crash location validation and correction that leverages both structured data validation and visual-textual analysis through multi-modal LLMs. Our empirical evaluation using 5,000 crash reports from Ohio demonstrates the effectiveness of this approach, with approximately 20% of reports requiring and receiving successful location corrections.

The key contributions of this work include:

1. Development of a credibility-based scoring system that systematically integrates multiple sources of location information

2. Demonstration that reference point types have varying reliability depending on initial coordinate agreement

3. Implementation of an efficient, scalable framework that balances computational requirements with validation thoroughness

4. Integration of multi-modal LLMs to extract and validate location information from crash diagrams and narratives

These contributions collectively represent a significant advancement in crash data quality methodology and provide practical tools for transportation safety practitioners.

Future research opportunities include analyzing accident causes and injury severity using larger datasets, particularly for pedestrian-vehicle accidents, and exploring how LLMs might complement emerging intelligent crash reporting systems. Additionally, several promising avenues for future research emerge from this work:

1. **Real-time Implementation and Evaluation**: Implementing our validation framework as part of electronic crash reporting systems and evaluating its impact on data quality in real-time represents a natural extension of this work.

2. **Transfer Learning for Diagram Interpretation**: While our current approach uses general-purpose multi-modal LLMs, future research could explore specialized models trained specifically for crash diagram interpretation, potentially improving accuracy and computational efficiency.

3. **Temporal Analysis of Location Accuracy**: Investigating how location accuracy has evolved over time as reporting technologies have advanced could provide valuable insights into the effectiveness of technological interventions.

4. **Multi-jurisdictional Comparison**: Expanding this analysis to include crash data from multiple states or countries could reveal jurisdiction-specific patterns and best practices in location reporting.

5. **Integration with Connected Vehicle Data**: As connected and autonomous vehicle technologies proliferate, exploring how vehicle-generated location data could complement or replace manually reported locations represents an important future direction.

4.8.1 Limitations

Several limitations of our current study should be acknowledged:

1. **Geographic Scope**: Our analysis is limited to crash data from Ohio. Different states and countries may have different reporting systems and location accuracy challenges that are not addressed by our current methodology.

2. **Technology Dependency**: Our approach relies on the availability of digital crash reports with specific data elements. Implementation in jurisdictions with paper-based reporting or limited data fields would require adaptation.

3. **LLM Interpretation Variability**: The performance of multi-modal LLMs in interpreting crash diagrams can vary based on diagram quality and complexity.

More standardized diagram formats would likely improve performance.

4. **Reference Database Quality**: The effectiveness of our validation methods depends on the quality and completeness of the reference databases used for spatial queries. Outdated or incomplete road network data would limit validation accuracy.

5. **Manual Ground Truth Verification**: While our methodology provides significant improvements, we lack a comprehensive manual verification of "ground truth" locations for all crashes in our dataset, which would require extensive field verification.

4.8.2 Recommendations

Based on our findings, we offer the following recommendations for improving crash location accuracy:

1. **Agency-level Implementation**:

    - Integrate location validation into electronic crash reporting systems to provide real-time feedback to officers
    - Establish cross-agency standardization of coordinate systems and location references
    - Implement regular data quality audits using automated validation frameworks similar to ours

2. **Policy Recommendations**:

    - Develop standards for minimum location accuracy in crash reporting
    - Establish protocols for resolving discrepancies between different location information sources

- Create incentives for improved data quality in reporting systems

3. **Technology Enhancements**:
    - Leverage GPS-enabled devices to automatically capture crash coordinates
    - Develop user-friendly interfaces that facilitate accurate location reporting
    - Implement augmented reality tools to improve on-scene location documentation
4. **Training and Education**:
    - Provide targeted training for law enforcement on accurate location reporting
    - Educate transportation safety analysts on location validation methodologies
    - Develop best practices guides for crash location documentation

By addressing these recommendations, transportation agencies can significantly improve crash location accuracy, leading to more effective safety interventions and, ultimately, lives saved on our roadways.

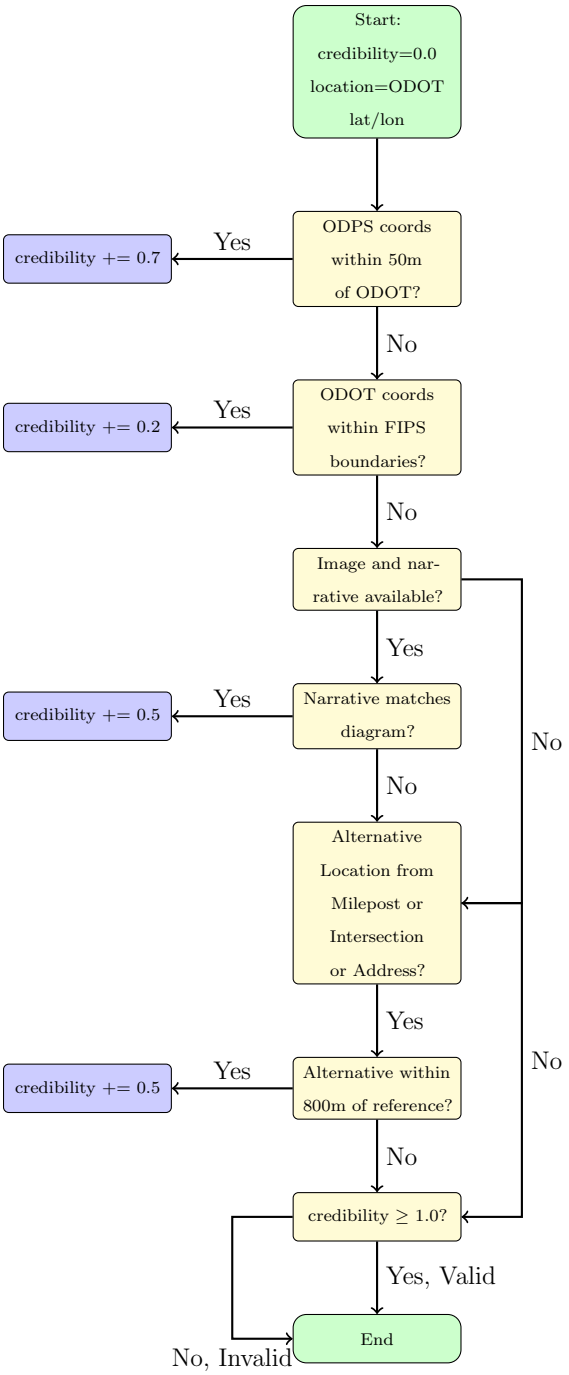


Figure 4.4: Crash Report Data Validation Process

CHAPTER 5

RISK-AWARE NAVIGATION FRAMEWORK: INTEGRATING CRASH PROBABILITY DATA FOR SAFER MOBILITY

Building on the location validation methodology established in Chapter 4, this chapter presents a comprehensive framework for incorporating crash risk data into navigation systems. The methodology developed here directly addresses the implementation gap identified in our systematic review (Chapter 3), where sophisticated analytical methods fail to translate into practical safety improvements. By creating a risk-aware navigation system that operates with commonly available data sources, we bridge the divide between research capabilities and real-world deployment.

We use state-of-the-art GIS and Data Engineering solutions (GDAL, Spark, Sedona, Overture Maps) to incorporate crash risk assessments directly into route planning algorithms for both autonomous and human-driven vehicles, with the explicit goal of global applicability. The framework provides safety intelligence for autonomous vehicles while simultaneously offering enhanced navigation options for human drivers. The core algorithm for integrating safety metrics into routing decisions is the subject of a U.S. patent application, representing a novel approach to navigation optimization.

Our implementation requires only two widely available data sources: (1) crash reports from law enforcement officers (LEOs) that include geolocation or clear location descriptions, and (2) traffic counts for a subset of roads. These data requirements are met across large parts of the world, enabling safety-aware navigation beyond the constraints of previous infrastructure-specific approaches. To demonstrate our solution, we processed 3.3 million crashes spanning from January 2013 to February 2025 from the U.S. state of Ohio, to quantify relative crash risk for individual road

segments based on historical patterns and traffic exposure metrics.

Our solution is a natural fit for autonomous vehicles, as a navigation system is an integral part of their operation. For people driving, it allows for a balance between travel time and risk exposure, creating a new dimension in route optimization. This dual-purpose solution is not only appropriate for individual travelers but also for connected vehicle fleets, autonomous mobility services, and transportation agencies working to mitigate systemic safety challenges. By providing a standardized methodology for risk assessment that functions across diverse operational environments, our framework contributes to both immediate safety improvements and long-term CAV development objectives.

5.1 METHODOLOGY

The proposed methodology incorporates crash risk assessment into route optimization algorithms through a systematic approach combining data preparation, risk quantification, and integration into routing cost functions. To clarify the different uses of the term "cost" in this chapter, we use the following distinct notation:

- C_e = Economic cost of crashes on segment e (dollars)
- $cost_e$ = Routing cost function for optimization (dimensionless)
- r_e^c = Cost-weighted crash rate (dollars per DVMT)

5.1.1 Risk Estimation

Let $G = (V, E)$ represent the road network, where V is the set of vertices (intersections) and E is the set of edges (road segments). For each segment $e \in E$, we define the following key attributes:

$$l_e = \text{length of segment } e \text{ (miles)} \tag{5.1}$$

$$\text{ADT}_e = \text{Average Daily Traffic on segment } e \text{ (vehicles/day)} \tag{5.2}$$

$$c_e = \text{number of crashes on segment } e \text{ during study period} \tag{5.3}$$

$$T = \text{duration of study period (days)} \tag{5.4}$$

To normalize for differential exposure across segments, we compute the Daily Vehicle Miles Traveled (DVMT) as:

$$\text{DVMT}_e = \text{ADT}_e \times l_e \tag{5.5}$$

The exposure-normalized crash rate r_e for each segment is then calculated as:

$$r_e = \frac{c_e}{\text{DVMT}_e \times T} \tag{5.6}$$

This provides the raw crash rate. While we use this raw rate in our calculations, it's worth noting that for reporting and comparison with standard literature, this rate can be expressed as:

$$r_e^{\text{reported}} = r_e \times 10^6 \tag{5.7}$$

This scaling expresses the rate in crashes per million vehicle miles traveled, consistent with standard practice used by FHWA and transportation safety literature.

5.1.2 Crash Cost by Severity

The severity of crashes matters greatly to the actual as well as perceived risk. Similar to analyses by the Federal Highway Administration (FHWA), we account for the severity of crashes in our calculations by assigning economic costs to crashes based on their severity level using the KABCO scale:

Table 5.1: Crash costs by KABCO severity classification (2016 dollars)

KABCO	Severity Code	Cost per Crash
K	1	$11,295,400
A	2	$655,000
B	3	$198,500
C	4	$125,600
O	5	$11,900

For each segment e, we compute the total crash cost C_e as:

$$C_e = \sum_{k \in K} n_{e,k} \times cost_k \tag{5.8}$$

where K is the set of severity levels, $n_{e,k}$ is the number of crashes of severity k on segment e, and $cost_k$ is the economic cost associated with severity level k.

This allows us to calculate a cost-weighted crash rate r_e^c for each segment:

$$r_e^c = \frac{C_e}{\text{DVMT}_e \times T} \tag{5.9}$$

Weighting the crash risk by severity biases our algorithm to avoid severe crashes while allowing more fender benders. This approach is consistent with standard safety evaluation practices and reflects typical drivers' risk perception and preferences, as research indicates drivers are willing to accept longer travel times to avoid routes with higher likelihood of severe crashes. However, the practical implementation of such severity-weighted risk metrics in navigation systems depends critically on whether drivers will actually modify their routing behavior in response to safety information—a behavioral assumption that merits empirical examination.

5.1.3 Risk Percentile Transformation

In order for risk values to be easier to compare across different road types and locations, we transform the raw crash rates into percentiles. For each road segment e with type t (e.g., primary, secondary, residential):

$$p_e = \text{percentile}(r_e, \{r_{e'} | e' \in E_t\}) \tag{5.10}$$

where E_t represents all segments of type t. This transformation normalizes risk values to a scale between 0 and 1, making them easier to use in routing algorithms while still recognizing that different types of roads have different underlying risks. Such a percentile is also easier to interpret for analysts reviewing the data.

This ensures route selection accounts for relative safety differences within each road type rather than being dominated by baseline differences between road classifications. Once the crash risk has been calculated for each segment, we needed to make this new feature available for navigation. To this end, we traversed the OSM data structure to append a risk tag containing the appropriate risk value according to the following decision framework:

$$\text{risk}(id_i) = \begin{cases} r_i, & \text{if } (id_i, r_i) \in R \\ \mu_h, & \text{if highway}(id_i) = h \text{ and } \mu_h \text{ exists} \\ \mu_{\text{global}}, & \text{otherwise} \end{cases} \tag{5.11}$$

where μ_{global} represents the global median risk across all segments with empirically derived values.

This hierarchical approach ensures comprehensive coverage of the road network while maintaining the highest possible fidelity to observed risk patterns. The categorization by highway type provides valuable risk discrimination even in the absence of direct empirical measurements. Our analysis revealed substantive variations in median

risk across different road classifications that would be obscured by a simplistic global median approach. Unclassified roads demonstrated a markedly elevated risk profile compared to service roads, despite both categories often being considered secondary in conventional road hierarchies.

5.2 IMPLEMENTATION

We selected GraphHopper to demonstrate our solution for three key reasons: (1) it is open source, providing full access to modify and extend its capabilities; (2) it utilizes OpenStreetMap (OSM) data, which is also open source and forms the basis of our road network data; and (3) it offers extensive customization options for routing algorithms, enabling the integration of our crash risk metrics directly into route calculations.

For this research, we created a fork of GraphHopper to implement risk-aware routing modifications. In our implementation, each OSM way in the road network is augmented with a risk metric encoded as a tag with value $r \in [0, 1]$ with resolution $\delta = 10^{-3}$. The risk values are stored using a 10-bit representation, providing sufficient granularity for risk differentiation while maintaining computational efficiency.

Our GraphHopper implementation directly corresponds to the multiplicative cost adjustment approach established in the methodology section. The routing engine applies the theoretical cost function:

$$cost_e = \frac{cost_e^{base}}{1 - k \cdot p_e} \tag{5.12}$$

through GraphHopper's priority-based weight system, where $cost_e^{base}$ represents the standard routing cost and $k \cdot p_e$ is stored as the road_risk tag for each segment.

The model uses a priority function that adjusts the weight of each road segment based on its risk value:

```
{
```

```
  "priority": [
    {
      "if": "true",
      "multiply_by": "1-road_risk"
    }
  ]
}
```

This configuration implements the theoretical framework by setting the priority multiplier to $(1 - \text{road\_risk}) = (1 - k \cdot p_e)$ from Equation 5.12, ensuring consistency between the methodology and implementation. For a road segment with risk value r, the priority multiplier becomes $(1 - r)$, meaning that segments with higher crash risk receive proportionally higher costs and are less likely to be selected in the optimal route. By disabling GraphHopper's optimization algorithms (contraction hierarchies and landmarks) through API parameters, we ensure that the custom model is fully applied during route calculation.

The resulting routes balance safety considerations with practical routing constraints, demonstrating how crash risk data can be integrated into real-world navigation systems. Our approach introduces no computational overhead compared to standard routing algorithms. Risk values are pre-computed and embedded as static attributes in the road network data, enabling CAV routing systems to achieve identical performance to conventional navigation while incorporating safety optimization.

Although our implementation requires disabling GraphHopper's Contraction Hierarchies and Landmarks optimizations to enable custom priority models, performance benchmarking validates that this does not impact routing performance: our risk-aware configuration achieved identical response times to standard GraphHopper configuration with optimizations enabled—both averaging 3.5ms per request across 1,120 route calculations. For the typical commute patterns between adjacent counties tested in

this study, disabling these optimizations to enable safety-aware routing imposes no performance penalty compared to conventional navigation systems.

The direct compatibility with GraphHopper demonstrates its usefulness in an established, high-performance routing engine capable of multi-criteria path optimization. Finally, the normalized risk expression also facilitates interpretation by users reviewing the data.

5.2.1 Data Preparation

The integration of crash risk metrics with a navigation system requires rigorous data preparation to ensure spatial accuracy, compatibility with routing software, and computational efficiency. For different navigation systems, there will be different considerations to make, but the key step is to project geolocated crash counts onto official road inventory data with Average Daily Traffic counts (ADT). The result is then projected onto mapping data that can be used by routing software. The road inventory would be maintained by regional or national transportation authorities—Departments of Transportation (DOT) in the USA. Crash counts are collected by local Law Enforcement Officers (police). In our case, we were working with crash reports that already had been geolocated. In many localities, the description of the location of the crash would have to be geolocated first.

5.2.2 Risk-Weighted Route Evaluation

To evaluate the effectiveness of risk-aware routing, we developed a method to calculate the weighted average risk along a complete route. This metric provides a quantitative measure of overall risk exposure:

$$\text{Route Risk} = \frac{\sum_i (\text{distance}_i \times \text{risk}_i)}{\sum_i \text{distance}_i} \tag{5.13}$$

where distance_i represents the length of a road segment and risk_i its associated

risk value. The implementation leverages GraphHopper's path details, which provide run-length-encoded information about various attributes along the route, including our custom risk metric and segment distances.

5.2.3 Performance Analysis

Our implementation enables systematic comparative analysis between traditional shortest-path routes and risk-optimized alternatives. The framework supports quantitative evaluation of the trade-offs between route efficiency and safety through multiple metrics captured simultaneously during route calculation.

The GraphHopper API configuration allows for the collection of comprehensive route characteristics through the details parameter:

```
"details": ["road_risk", "road_class", "distance"]
```

The three listed features are the normalized road risk, the highway type, and the distance for each segment traversed in a route. With these details, we can calculate the weighted risk exposure for an entire route.

By comparing risk-optimized routes with traditional shortest paths, the system provides a foundation for empirical assessment of the relationship between route efficiency and safety. We use this to assess the effectiveness with which our solution allows a user to prioritize safety versus travel time.

5.3 CASE STUDY RESULTS

The substantial variation in absolute crash rates across road classes illustrates the complexity of risk patterns that could benefit from more sophisticated statistical modeling approaches. Table 5.2 presents absolute crash risk metrics across Ohio's major road classifications.

Table 5.2: Absolute crash risk metrics by road class

Road Class	Segment Count	Mean	Median	P25	P75
residential	84,299	11,638.43	562.01	171.62	2,683.02
tertiary	26,174	25,691.63	1,917.10	313.11	12,321.43
secondary	17,738	18,465.81	750.95	152.78	4,797.85
primary	13,512	6,305.76	530.71	124.15	2,386.09
footway	12,214	1,074.40	332.84	142.29	879.46
motorway	9,831	5,286.28	884.15	197.24	3,903.90
service	9,663	1,577.30	248.09	91.39	786.54
unclassified	6,771	31,260.41	6,461.54	915.01	27,269.66
motorway_link	6,277	661.12	278.64	110.19	691.07
trunk	3,959	7,496.57	963.96	211.64	4,657.65

Crash rates per million vehicle miles traveled (VMT)

Only road classes with > 100 segments shown

The 75th percentile values (45% risk reduction, 536m distance increase, 319s time increase) indicate that substantial safety improvements are achievable for most commute patterns, with the majority of routes requiring less than 9 additional minutes of travel time. These findings convincingly demonstrate that risk-informed routing can achieve significant safety improvements while imposing minimal penalties on travel efficiency. While this demonstration using Ohio commuting data establishes the feasibility of integrating crash risk into navigation systems, the methodology is designed for broader applicability and the precision of risk assessments can be enhanced through the improvements discussed in Section 5.6.

Figure A.1 demonstrates the practical implementation of risk-aware routing using our GraphHopper modification. The route pair shows a representative trade-off: 35% risk reduction (0.433 to 0.281 weighted risk score) for an additional 1.3 km distance

and 89-second time penalty.

5.3.1 Implications for Transportation Planning

Our empirical analysis of commuting data in Ohio suggests that substantial crash risk reductions can be achieved with acceptable travel time penalties. Transportation planners can leverage such analysis to strategically prioritize infrastructure investments. Specifically, high-risk segments along frequently traversed commuter corridors between major employment centers (such as those identified in our MSA county-pair analysis) represent prime targets for safety interventions. The spatial distribution of these high-leverage segments varies significantly across the state's urban and rural contexts.

Notably, we identified several corridors where safer routing alternatives coincide with shorter distances. Such instances may be worth a review by local road engineers. We identified significant regional heterogeneity in risk-reduction potential, which suggests that reviews need to be localized. By quantitatively integrating crash risk into analyses, regional and national transportation engineers can target interventions where they are most needed.

5.4 APPLICATIONS FOR CONNECTED AND AUTONOMOUS VEHICLES

The methodology presents several key contributions to the field of connected and automated vehicle technology and transportation safety:

First, our percentile-based risk ranking approach provides a standardized, interpretable measure of relative safety across heterogeneous road networks that can be seamlessly integrated into CAV decision-making systems. This transformation mitigates the challenges posed by outliers and data sparsity while maintaining an intuitive scale that can be effectively communicated to both autonomous systems and human users through appropriate interfaces. For CAVs specifically, this standardized

risk metric can serve as a validation parameter for operational design domain (ODD) assessment and safety assurance.

Second, the routing cost function (Equation 5.12) offers a flexible mechanism for integrating safety considerations into existing CAV path planning solutions without sacrificing computational efficiency. The parameterization through adjustable coefficients enables customization based on operational needs, safety requirements, and vehicle-specific performance characteristics. The adaptability is useful for CAV fleet operators seeking to optimize routing systematically but also for drivers with different risk profiles and preferences.

Third, our case study using Ohio's road network and crash data demonstrates a real-world implementation of the solution. The methodology is generalizable to any region with available crash and traffic data, requiring only minimal adaptation to local data structures and road classification systems. The scalability of our solution is a key feature that sets it apart from other studies constrained to small areas or particular types of infrastructure.

The implementation using the open-source GraphHopper routing library with OpenStreetMap data illustrates how the theoretical framework can be translated into a functioning navigation system compatible with existing CAV software architectures. This implementation pathway could be adopted by both commercial CAV developers and transportation agencies seeking to enhance both autonomous and human-driven navigation safety. By providing risk-aware navigation capabilities, the system supports functional safety requirements for SAE Level 3+ automation while simultaneously offering benefits for conventional vehicles.

As autonomous vehicles transition from testing to widespread deployment and connected vehicle technologies become standard across the transportation ecosystem, the incorporation of safety metrics into route planning represents a significant opportunity for crash reduction and enhanced CAV reliability. For autonomous vehicles

specifically, risk-aware navigation can serve as an additional safety layer beyond immediate perception-based obstacle avoidance, potentially addressing edge cases and long-tail risks that challenge current validation approaches. Future refinements to this approach—particularly those addressing the spatiotemporal dynamics of risk, V2X information integration, and system-specific risk profiles—could further enhance its impact on autonomous vehicle safety verification and validation.

5.5 METHODOLOGICAL CONSIDERATIONS

The risk assessment framework presented in the previous section provides a mathematical foundation to integrate safety considerations into route optimization. Here, we discuss ways to improve some limitations of our approach.

5.5.1 Data Quality and Modeling Assumptions

While our methodology normalizes crash counts by exposure through the r_e calculation, it treats crashes as uniformly distributed along segments. In reality, risk often varies within segments, particularly near intersections or geometric features. The percentile transformation $p_e = F(r_e)$ mitigates some effects of outliers, but does not account entirely for heterogeneity within segments.

Our study period (January 2013 to February 2025) includes the COVID-19 pandemic years (2020-2021), during which traffic patterns and volumes differed substantially from normal conditions. For example, the crash risk of an interchange can affect multiple connected segments, violating the independence assumption underlying our percentile-based approach.

An Empirical Bayes (EB) approach could enhance the stability of our risk estimates, particularly for segments with small sample sizes. Using this method, the estimated crash rate would become:

$$r_e^{EB} = w_e \cdot r_e + (1 - w_e) \cdot r_e^{SPF} \tag{5.14}$$

where r_e^{SPF} is the expected crash rate from a calibrated Safety Performance Function based on segment characteristics, and $w_e \in [0, 1]$ is a weight determined by the reliability of the observed data. This approach would be particularly valuable for refining the median substitution we currently employ for segments with missing data.

5.5.2 Implementation Challenges

The risk assignment framework defined in Equation 5.12 modifies standard routing costs by incorporating crash risk through a multiplicative penalty. A key implementation challenge involves calibrating the risk parameter k in this formulation. A high k value could divert traffic to paths that, while statistically safer, might introduce practical issues such as significantly longer travel times or routing through inappropriate corridors. A low k value, conversely, might render the safety adjustment ineffective.

The current formulation applies a linear relationship between the risk percentile p_e and the cost penalty. Alternative formulations could consider non-linear transformations that apply greater penalties to the highest-risk segments:

$$cost_e = \frac{cost_e^{base}}{1 - k \cdot f(p_e)} \tag{5.15}$$

where $f(p_e)$ could be p_e^2 or $e^{p_e} - 1$ to create steeper penalties for high-risk segments.

A further challenge is that traffic counts (ADT), a necessary component of our system, are not available for all segments. To account for this, we apply the median crash rate by road type:

$$r_e = \text{median}(\{r_i | i \in E_t, \text{ data available for } i\}) \quad \text{if data unavailable for } e \tag{5.16}$$

where E_t represents all segments of highway type t.

Assigning the median risk to segments with missing data provides a practical solution which could be improved by imputation methods that consider spatial and network context. A Bayesian hierarchical model, for instance, could potentially provide more nuanced estimates for these segments by leveraging information from similar roads within the network.

5.5.3 Commercialization Challenges

The core risk-aware navigation process operates without requiring personally identifiable information (PII) or real-time user tracking. Risk assessments are computed from anonymized historical crash data and traffic counts, with routing decisions made locally within navigation software. This design minimizes privacy concerns for the fundamental solution.

However, real-time enhancements incorporating live traffic and weather data would introduce additional data requirements and cybersecurity considerations requiring secure APIs. Such enhancements would face distinct regulatory challenges across global markets. In the European Union, these enhanced versions would require compliance with GDPR for location tracking, while the EU Data Act mandates fair access to transportation authority datasets. The EU Digital Markets Act may impose interoperability requirements for large-scale deployments, potentially affecting proprietary risk algorithms.

Deployment in emerging markets across the Global South, ASEAN, and Africa faces different challenges primarily related to data availability and reliability. The core requirement for comprehensive crash reporting systems and regular traffic counting may not exist or may lack consistent update intervals necessary for reliable risk assessment. Our framework was primarily designed with the robust data infrastructure of the United States in mind, where crash reports and traffic inventories are systematically maintained and regularly updated.

Across all markets, automotive functional safety standards (ISO 26262) require systematic validation of safety claims, while product liability frameworks remain undefined for navigation services making explicit safety recommendations.

5.6 FUTURE WORK

The risk assessment and routing optimization solution presented in this chapter establishes a foundation for data-driven safety-aware navigation. In this section, we will discuss possible future developments that would increase the accuracy and precision of our solution.

A limitation of our solution is that it treats crash risk as a static function of travel volumes. An improvement to this approach would be to run iterative models to determine the equilibrium state of risk distribution, similar to established practices in regional travel demand modeling. Recent research by Li et al. has demonstrated the feasibility of incorporating safety considerations into network equilibrium frameworks.

5.6.1 Autonomous Vehicle-Specific Risk Modeling

The current framework's reliance on historical crash data from human-driven vehicles represents a fundamental limitation that future research must address. As CAV market penetration increases and crash data specific to autonomous systems becomes available, the development of differentiated risk models will become both feasible and necessary. Such models would account for:

- Systematic behavioral differences between algorithmic and human decision-making
- CAV-specific failure modes distinct from human error patterns
- Interaction effects between CAVs and human drivers in mixed traffic
- Sensor-dependent risk factors under various environmental conditions

The transition from human-derived to CAV-specific risk metrics will likely require probabilistic frameworks that can accommodate the deterministic nature of autonomous systems while capturing edge-case behaviors and system limitations.

5.6.2 Network-Level Effects and Traffic Distribution

Our framework demonstrates a complete, implementable solution for integrating crash risk into navigation decisions. While our primary objective was to establish the feasibility and effectiveness of risk-aware routing, we recognize that large-scale deployment would benefit from additional constraints to manage traffic distribution across the network.

In developing this proof-of-concept, we focused on enabling individual route optimization based on empirically-derived risk metrics. This approach successfully demonstrates significant safety improvements while maintaining computational efficiency and compatibility with existing navigation systems. Notably, our empirical analysis reveals that safety-optimized routes differ only modestly from standard routes—averaging just 369 meters longer—suggesting that the algorithm naturally avoids dramatic diversions through residential neighborhoods or other inappropriate corridors. However, system-wide implementation would naturally evolve to incorporate explicit network-level considerations.

Specifically, future deployments could extend our framework to include:

- Time-varying restrictions for sensitive areas (e.g., school zones during arrival/dismissal times)
- Volume thresholds based on functional classification and design capacity
- Land use buffers to preserve neighborhood character

These enhancements would build upon our routing cost function through additional

constraint terms:

$$\text{Cost}_e = \alpha \cdot t_e + \beta \cdot p_e \cdot l_e + \gamma \cdot \max(0, v_e - C_e) \tag{5.17}$$

where the third term penalizes routes exceeding capacity thresholds C_e.

Such extensions align with the equilibrium modeling approaches discussed in this section, where iterative optimization would naturally balance individual route safety with system-wide traffic distribution. Transportation agencies implementing our solution have the flexibility to incorporate local priorities and constraints while leveraging the core risk assessment methodology we have established.

5.6.3 Severity Classification Validation and Bias Correction

The documented inflation of KABCO severity classifications—with "A" designations overstated in up to 65% of cases according to Burdett et al.—introduces systematic bias into cost-weighted risk calculations that could propagate through route selection algorithms. The economic cost differential between severity levels spans two orders of magnitude (from $11,900 for "O" crashes to $11.3 million for "K" crashes), meaning that misclassification directly affects the relative weighting of road segments in routing decisions.

Future research should investigate the sensitivity of route selection to severity misclassification through controlled simulation studies. By systematically adjusting severity distributions according to documented bias patterns—such as downgrading a percentage of "A" classifications to "B" or "C" levels—researchers can quantify how misclassification affects both segment-level risk rankings and actual route recommendations. This analysis would determine whether the percentile-based transformation provides sufficient robustness or whether bias correction mechanisms are necessary for reliable implementation.

Where jurisdictions maintain linked crash-medical records, comparative studies could validate field severity assessments against clinical injury documentation. Such

validation would enable development of region-specific correction factors that account for local reporting practices. The research would also inform whether KABCO bias exhibits systematic patterns by road type, crash circumstances, or geographic factors that could be incorporated into improved risk assessment models.

5.7 CONCLUSION

This chapter introduced a novel way to account for crash risk in navigation software. We showed how to use road inventory data combined with crash reports by Law Enforcement Officers to develop risk assessments that we then used to minimize crash risk when selecting a route for travel. By formalizing the relationship between exposure-normalized crash rates and routing decisions, we have demonstrated the feasibility of navigation that balances traditional metrics of efficiency with quantifiable safety considerations—a critical advancement for Connected and Automated Vehicles as well as manually driven vehicles.

The goal of this study is to contribute to the broader goal of creating safer automated and connected transportation systems while developing solutions that can be used broadly including by human drivers. By bridging current and future mobility paradigms, we provide a practical pathway toward the incremental safety improvements necessary for public acceptance and regulatory compliance of autonomous vehicle technology.

The framework establishes not just a technical solution, but a comprehensive approach to data-driven safety management that can evolve with advancing transportation technologies. The work presented in this chapter represents the integration of the research approach established throughout this dissertation, combining methodological sophistication with practical implementation, enhanced data quality with real-world deployment, and current applications with future technologies.

CHAPTER 6
CONCLUSIONS AND FUTURE DIRECTIONS

Examining the three research components presented in this dissertation together reveals relationships not apparent when considering each separately. The systematic review documented advances in crash analysis methodologies over recent decades but also revealed that a lot of effort goes into correcting for biases, unobserved heterogeneity, and entry errors in source data. With our location validation work, we uncovered that it's possible to improve the quality of existing data. Although the effort to identify correct locations is not new (Loo [147], Chung et al. [148]), our work also provides a solution to improve quality at the point of data entry.

Data quality limitations may partially explain the persistent gap between methodological sophistication and practical implementation. While multiple factors contribute, such as computational resources, training requirements, organizational capacity, data quality represents a fundamental constraint. Advanced analytical methods depend on accurate input data; when crash locations are systematically misreported, even well-designed algorithms produce less reliable insights for safety interventions.

Traditional approaches to bridging the research-practice gap emphasize simplifying methods or improving training. While valuable, these efforts address only part of the challenge. Without adequate data quality, simplified methods may still produce unreliable outputs. The LLM-based validation offers a complementary approach: automated data quality improvement that enables sophisticated analytical methods.

The final contribution presented here introduces road safety as a parameter in route planning. To the best of our knowledge, navigating according to road safety is a novel idea. The only related research we have found is focused on hazardous materials

transportation, e.g., Hu et al. [149].

6.1 A NOTE ON COMPREHENSIVE LOCATION CORRECTIONS

While the LLM-based validation algorithm could theoretically correct crash locations across the entire Ohio dataset, practical constraints limited this application. The validation approach requires access to original crash report PDFs, which were not available for all crashes in the timeframe needed for route optimization analysis. Additionally, the computational resources and time required to process and validate the full historical crash database exceeded our resources and the scope of our research. Consequently, the routing algorithm relies on crash locations as recorded in the official database, accepting the inherent data quality limitations documented in our validation study.

However, the framework's design mitigates the impact of these limitations through min-max normalization: by expressing risk as relative rather than absolute values, the system prioritizes comparative safety assessment across routes rather than precise risk quantification for individual segments. Given that location errors appear randomly distributed and affect approximately 20% of reports, aggregate risk patterns across the network remain sufficiently robust for route comparison purposes.

The solution remains functional with standard crash databases and can incorporate improved location data when available. This flexibility aligns with our broader objective of developing methods that operate within real-world constraints rather than requiring idealized conditions.

Together, our contributions provide a pathway for advancing transportation safety: developing analytical methods, ensuring sufficient data quality, and designing approaches practitioners can adopt within existing constraints.

6.2 IMPLICATIONS FOR CONNECTED AND AUTOMATED VEHICLES

CAV deployment requires understanding real-world risk distributions. Our integrated framework provides solutions to create validated risk maps that could support CAV development and testing. New data could potentially be incorporated (such as CAV sensor data) to update risk assessments as deployment scales, though this capability requires further validation.

A navigation paradigm that invites drivers to make informed decisions about safe paths raises questions about transition strategies for autonomous vehicles. If users become accustomed to understanding and evaluating routing trade-offs between safety and efficiency, could this familiarity ease acceptance of how semi-autonomous and autonomous systems make similar decisions? The risk communication approaches developed here might inform how CAVs explain their navigation choices to passengers, though this connection requires empirical investigation.

6.3 METHODOLOGICAL ADVANCES

The credibility scoring system provides a framework to combine multiple imperfect information sources to achieve higher confidence than any single source alone. By treating different data sources (coordinates, narratives, diagrams) as components that collectively provide stronger validation than individual sources, the system adapts ensemble learning principles to data validation.

The routing algorithm treats risk as varying with infrastructure, exposure, and behavioral factors rather than as an inherent segment characteristic. This approach enabled the development of utility functions that balance safety and efficiency objectives.

6.4 LIMITATIONS

The empirical validation focuses on Ohio. While methods are designed for broader applicability, performance in other contexts requires validation. Different jurisdictions have varying crash reporting standards, road network characteristics, and data quality patterns that may affect the effectiveness of an implementation.

The LLM-based validation depends on narrative descriptions and diagrams in crash reports. Where these elements are routinely omitted or completed inadequately, effectiveness would be reduced. The risk assessment approach uses historical data to predict future risk and cannot account for sudden changes such as construction, special events, or weather conditions.

6.5 FUTURE RESEARCH DIRECTIONS

Immediate priorities include expanding validation to other geographic contexts and adapting methods for jurisdictions with more limited data. LLM applications could extend to identifying inconsistencies in crash narratives, detecting underreporting patterns, and predicting missing data fields.

Real-time risk assessment integrating streaming data from connected vehicles, weather services, and traffic sensors could enable dynamic risk maps. Behavioral research should examine how different demographic groups respond to risk information and what presentation formats effectively communicate risk without causing alarm.

As vehicles become increasingly connected, they could share near-miss incidents and hazard observations, creating crowd-sourced risk maps that complement historical crash data.

6.6 CLOSING PERSPECTIVE

This dissertation addresses transportation safety by developing implementable solutions that apply contemporary technology to longstanding problems. The artificial intelligence techniques that enable natural language processing can validate crash locations. The routing algorithms that minimize travel time can incorporate crash risk. The data documenting past incidents can inform prevention efforts.

The solutions developed provide components for integrated safety approaches. Effectiveness will ultimately be measured by safety improvements achieved through implementation. As transportation systems evolve, ensuring that safety remains a fundamental design principle becomes increasingly important. The theoretical advances and practical tools provided establish a foundation for continued research and implementation efforts.

BIBLIOGRAPHY

[1] Parichat Curry, Ramesh Ramaiah, and Monica S. Vavilala. Current trends and update on injury prevention. *International Journal of Critical Illness and Injury Science*, 1(1):57–65, 2011.

[2] Ashar Ahmed, Ahmad Farhan Mohd Sadullah, and Ahmad Shukri Yahya. Errors in accident data, its types, causes and methods of rectification-analysis of the literature. *Accident Analysis & Prevention*, 130:3–21, September 2019.

[3] Lars Skaug, Mehrdad Nojoumian, Nolan Dang, and Amy Yap. Road crash analysis and modeling: A systematic review of methods, data, and emerging technologies. *Applied Sciences*, 15(13), 2025.

[4] Lars Skaug and Mehrdad Nojoumian. A multimodal artificial intelligence framework for intelligent geospatial data validation and correction. *Inventions*, 10(4), 2025.

[5] Lars Skaug and Mehrdad Nojoumian. Risk-aware navigation framework for autonomous and human-driven vehicles: Integrating crash probability data for safer mobility. *SAE International Journal of Connected and Automated Vehicles*, 2025. Manuscript under review.

[6] Fangming Qu, Nolan Dang, Borko Furht, and Mehrdad Nojoumian. Comprehensive study of driver behavior monitoring systems using computer vision and machine learning techniques. *Journal of Big Data*, 11(32):44, 2024.

[7] Mehrdad Nojoumian. Active occupant status and vehicle operational status warning system and methods, 2025. US Patent 17/542,807.

[8] Kavi Bhalla and Kevin Gleason. Effects of vehicle safety design on road traffic deaths, injuries, and public health burden in the latin american region: a modelling study. *The Lancet Global Health*, 8(6):e819–e828, 2020.

[9] Ishtiaque Ahmed et al. Road infrastructure and road safety. *Transport and Communications Bulletin for Asia and the Pacific*, 83(13):19–25, 2013.

[10] Lars Åberg. Traffic rules and traffic safety. *Safety science*, 29(3):205–215, 1998.

[11] Mohammad Mahdi Rezapour Mashhadi, Promothes Saha, and Khaled Ksaibati. Impact of traffic enforcement on traffic safety. *International Journal of Police Science & Management*, 19(4):238–246, 2017.

[12] Adam Fletcher, Kirsty McCulloch, Stuart D Baulk, and Drew Dawson. Countermeasures to driver fatigue: a review of public awareness campaigns and legal approaches. *Australian and New Zealand Journal of Public Health*, 29(5):471–476, 2005.

[13] Amanda Delaney, Bella Lough, Michelle Whelan, Max Cameron, et al. A review of mass media campaigns in road safety. *Monash University Accident Research Centre Reports*, 220:85, 2004.

[14] Mehrdad Nojoumian and Lars Skaug. Road-risk awareness system (RAS) in semi or fully autonomous vehicles, 2025. US Patent 19/016,485.

[15] Mehrdad Nojoumian and Lars Skaug. Sun glare avoidance system (SAS) in semi or fully autonomous vehicles, 2025. US Patent 19/016,240.

[16] Fred L. Mannering and Chandra R. Bhat. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, 1:1–22, 2014.

[17] J. Short and B. Caulfield. Record linkage for road traffic injuries in ireland using police hospital and injury claims data. *Journal of Safety Research*, 58:1–14, 2016.

[18] Vishal Mahajan, Christos Katrakazas, and Constantinos Antoniou. Crash risk estimation due to lane changing: A data-driven approach using naturalistic data. *IEEE Transactions on Intelligent Transportation Systems*, 23(4):3756–3765, 2022.

[19] Xuesong Wang, Qian Liu, Feng Guo, Shou'en Fang, Xiaoyan Xu, and Xiaohong Chen. Causation analysis of crashes and near crashes using naturalistic driving data. *Accident Analysis & Prevention*, 177:106821, 2022.

[20] Yasir Ali, Md. Mazharul Haque, and Fred Mannering. A bayesian generalised extreme value model to estimate real-time pedestrian crash risks at signalised intersections using artificial intelligence-based video analytics. *Analytic Methods in Accident Research*, 38:100264, 2023.

[21] Lisa Pinals, Alex Kerin, Craig Van Alsten, Richard Sharp, and Sam Madden. Telematics-enabled usage-based insurance (ubi) and its impact on driving behavior, 2023. Accessed: 2023-10-11.

[22] R. Fix, C. Wilkinson, and G. Siegmund. Comparing event data recorder data (edr) in front/rear collisions from the crash investigation sampling system (ciss) database. Technical Report 2024-01-2892, SAE Technical Paper, 2024.

[23] A. Watson et al. The under-reporting of road crash injuries to police. *Australian Journal of Road Safety*, 26(2):12–22, 2015.

[24] Mario Miler, Filip Todić, and Marko Ševrović. Extracting accurate location information from a highly inaccurate traffic accident dataset: A methodology based on a string matching technique. *Transportation Research Part C: Emerging Technologies*, 68:185–193, 2016.

[25] Marianna Imprialou and Mohammed Quddus. Crash data quality for road safety research: Current state and future directions. *Accident Analysis & Prevention*, 130:84–90, 2019.

[26] Transportstyrelsen. Om strada, 2023.

[27] D. A. Lombardi et al. Improving crash injury identification through integration of hospital discharge and crash report data. *Injury Prevention*, 28:167–172, 2022.

[28] K. H. Janstrup et al. Understanding traffic crash under-reporting: Linking police and medical records to individual and crash characteristics. *Traffic Injury Prevention*, 17(6):580–584, 2016.

[29] Bucko Burdett, Andrea Bill, and David Noyce. Evaluation of law enforcement agency injury severity assessments. *Transportation Research Record*, 2676(9):246–255, 2022.

[30] Jia Li, Chengqian Li, and Xiaohua Zhao. Optimizing crash risk models for freeway segments: A focus on the heterogeneous effects of road geometric design features, traffic operation status, and crash units. *Accident Analysis & Prevention*, 205:107665, 2024.

[31] Fred L. Mannering, Venky Shankar, and Chandra R. Bhat. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research*, 11:1–16, 2016.

[32] Kazi Redwan Shabab, Tanmoy Bhowmik, Mohamed H. Zaki, and Naveen Eluru. A systematic unified approach for addressing temporal instability in road safety analysis. *Analytic Methods in Accident Research*, 43:100335, 2024.

[33] Li-Yen Chang and Fred L. Mannering. Predicting vehicle occupancies from accident data: An accident severity approach. *Transportation Research Record*, 1635(1):93–104, 1998.

[34] Shamsunnahar Yasmin, Naveen Eluru, and Md. Mazharul Haque. Addressing endogeneity in modeling speed enforcement, crash risk and crash severity simultaneously. *Analytic Methods in Accident Research*, 36:100242, 2022.

[35] William Jr. Haddon. A logical framework for categorizing highway safety phenomena and activity. *The Journal of Trauma: Injury, Infection, and Critical Care*, 12(3):193–207, March 1972. President, Insurance Institute for Highway Safety, Washington, D. C.

[36] Marjan P. Hagenzieker, Jacques J.F. Commandeur, and Frits D. Bijleveld. The history of road safety research: A quantitative approach. *Transportation Research Part F: Traffic Psychology and Behaviour*, 25:150–162, 2014. Special Issue: The history of road safety research and the role of traffic psychology.

[37] Salvatore Cafiso, Alessandro Di Graziano, Giacomo Di Silvestro, Grazia La Cava, and Bhagwant Persaud. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. *Accident Analysis & Prevention*, 42:1072–1079, 2010.

[38] Jonathan Aguero-Valverde and Paul P. Jovanis. Spatial analysis of fatal and injury crashes in pennsylvania. *Accident Analysis & Prevention*, 38(3):618–625, 2006.

[39] Jonathan Aguero-Valverde and Paul P. Jovanis. Analysis of road crash frequency with spatial models. *Transportation Research Record*, 2061(1):55–63, 2008.

[40] Jonathan Aguero-Valverde and Paul P. Jovanis. Bayesian multivariate poisson lognormal models for crash severity modeling and site ranking. *Transportation Research Record*, 2136(1):82–91, 2009.

[41] Yu-Chiun Chiou and Chiang Fu. Modeling crash frequency and severity using multinomial-generalized poisson model with error components. *Accident Analysis & Prevention*, 50:73–82, 2013.

[42] Yu-Chiun Chiou, Chiang Fu, and Hsieh Chih-Wei. Incorporating spatial dependence in simultaneously modeling crash frequency and severity. *Analytic Methods in Accident Research*, 2:1–11, 2014.

[43] James A. Bonneson and Michael P. Pratt. Procedure for developing accident modification factors from cross-sectional data. *Transportation Research Record*, 2083(1):40–48, 2008.

[44] Amirfarrokh Iranitalab and Aemal Khattak. Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, 108:27–36, 2017.

[45] Pengpeng Xu and Helai Huang. Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. *Accident Analysis & Prevention*, 75:16–25, 2015.

[46] Xuesong Wang, Xueyu Zhang, and Yingying Pei. A systematic approach to macro-level safety assessment and contributing factors analysis considering traffic crashes and violations. *Accident Analysis & Prevention*, 194:107323, 2024.

[47] Andrew P. Jones and Stig H. Jørgensen. The use of multilevel models for the prediction of road accident outcomes. *Accident Analysis & Prevention*, 35(1):59–69, 2003.

[48] Panagiotis Ch. Anastasopoulos and Fred L. Mannering. A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis & Prevention*, 41(1):153–159, 2009.

[49] S.M. Sohel Mahmud, Luis Ferreira, Md. Shamsul Hoque, and Ahmad Tavassoli. Using a surrogate safety approach to prioritize hazardous segments in a rural highway in a developing country. *IATSS Research*, 44(2):132–141, 2020.

[50] Changjian Zhang, Jie He, Mark King, Ziyang Liu, Yikai Chen, Xintong Yan, Lu Xing, and Hao Zhang. A crash risk identification method for freeway segments with horizontal curvature based on real-time vehicle kinetic response. *Accident Analysis & Prevention*, 150:105911, 2021.

[51] Mohamed Abdel-Aty and Joanne Keller. Exploring the overall and specific crash severity levels at signalized intersections. *Accident Analysis & Prevention*, 37(3):417–425, 2005.

[52] Kibrom A. Abay. Examining pedestrian-injury severity using alternative disaggregate models. *Research in Transportation Economics*, 43(1):123–136, 2013. The Economics of Transportation Safety.

[53] Marisol Castro, Rajesh Paleti, and Chandra R. Bhat. A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation Research Part B: Methodological*, 46:253–272, 2012.

[54] Mohamed Ahmed, Helai Huang, Mohamed Abdel-Aty, and Bernardo Guevara. Exploring a bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. *Accident Analysis & Prevention*, 43(4):1581–1589, 2011.

[55] Xunjia Zheng, Di Zhang, Hongbo Gao, Zhiguo Zhao, Heye Huang, and Jianqiang Wang. A novel framework for road traffic risk assessment with hmm-based prediction model. *Sensors*, 18(12):4313, 2018.

[56] Dominique Lord and Fred Mannering. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5):291–305, 2010.

[57] Ali S Al-Ghamdi. Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, 34(6):729–741, 2002.

[58] Tariq Usman Saeed, Thomas Hall, Hiba Baroud, and Matthew J. Volovski. Analyzing road crash frequencies with uncorrelated and correlated random-parameters count models: An empirical assessment of multilane highways. *Analytic Methods in Accident Research*, 23:100101, 2019.

[59] Panagiotis Ch. Anastasopoulos, Fred L. Mannering, Venky N. Shankar, and John E. Haddock. A study of factors affecting highway accident rates using the random-parameters tobit model. *Accident Analysis & Prevention*, 45:628–633, 2012.

[60] H.M. Abdul Aziz, Satish V. Ukkusuri, and Samiul Hasan. Exploring the determinants of pedestrian–vehicle crash severity in new york city. *Accident Analysis & Prevention*, 50:1298–1309, 2013.

[61] Panagiotis Ch. Anastasopoulos, Andrew P. Tarko, and Fred L. Mannering. Tobit analysis of vehicle accident rates on interstate highways. *Accident Analysis & Prevention*, 40(2):768–775, 2008.

[62] Marisol Castro, Rajesh Paleti, and Chandra R. Bhat. A spatial generalized ordered response model to examine highway crash injury severity. *Accident Analysis & Prevention*, 52(28), 2013.

[63] Wan-Hui Chen and Paul P. Jovanis. Method for identifying factors contributing to driver-injury severity in traffic crashes. *Transportation Research Record*, 1717(1):1–9, 2000.

[64] F.D. Bijleveld. The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. *Accident Analysis & Prevention*, 37(4):591–600, 2005.

[65] Chandra R. Bhat and Subodh K. Dubey. A new estimation approach to integrate latent psychological constructs in choice modeling. *Transportation Research Part B: Methodological*, 67:68–85, 2014.

[66] Chandra R. Bhat, Kathryn Born, Raghuprasad Sidharthan, and Prerna C. Bhat. A count data model with endogenous covariates: Formulation and application to roadway crash frequency at intersections. *Analytic Methods in Accident Research*, 1:53–71, 2014.

[67] Hassan T. Abdelwahab and Mohamed A. Abdel-Aty. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transportation Research Record*, 1746(1):6–13, 2001.

[68] Yu-Chiun Chiou, Lawrence W. Lan, and Wen-Pin Chen. A two-stage mining framework to explore key risk conditions on one-vehicle crash severity. *Accident Analysis & Prevention*, 50:405–415, 2013.

[69] Weitiao Wu, Shuyan Jiang, Ronghui Liu, Wenzhou Jin, and Changxi Ma and. Economic development, demographic characteristics, road network and traffic accidents in zhongshan, china: gradient boosting decision tree model. *Transportmetrica A: Transport Science*, 16(3):359–387, 2020.

[70] Pei Li, Mohamed Abdel-Aty, and Jinghui Yuan. Real-time crash risk prediction on arterials based on lstm-cnn. *Accident Analysis & Prevention*, 135:105371, 2020.

[71] Kingsley Adjenughwure, Gerdien Klunder, Jeroen Hogema, and Richard Van der Horst. Monte carlo-based microsimulation approach for estimating the collision probability of real traffic conflicts. *Transportation Research Record*, 2677(9):314–326, 2023.

[72] Kibrom A. Abay, Rajesh Paleti, and Chandra R. Bhat. The joint analysis of injury severity of drivers in two-vehicle crashes accommodating seat belt use endogeneity. *Transportation Research Part B: Methodological*, 50:74–89, 2013.

[73] Saleh Altwaijri, Mohammed Quddus, and Abigail Bristow. Analysing the severity and frequency of traffic crashes in riyadh city using statistical models. *International Journal of Transportation Science and Technology*, 1(4):351–364, 2012.

[74] Khaled A Abbas. Traffic safety assessment and development of predictive models for accidents on rural roads in egypt. *Accident Analysis & Prevention*, 36(2):149–163, 2004.

[75] Md Istiak Jahan, Tanmoy Bhowmik, and Naveen Eluru. Enhanced aggregate framework to model crash frequency by accommodating zero crashes by crash type. *Transportation Research Record*, 2678(2):506–519, 2024.

[76] E. Papadimitriou et al. Comparative analysis of road safety performance indicators. *Accident Analysis & Prevention*, 129:317–327, 2019.

[77] Y. Wu et al. Influence of multiple road environment conditions on traffic accidents. *Transportation Research Part D: Transport and Environment*, 93:102766, 2021.

[78] S. Moslem et al. An integrated model of ahp-bwm for evaluating driver behavior factors. *Transportation Research Part F: Traffic Psychology and Behaviour*, 71:46–57, 2020.

[79] D. Farooq et al. Assessing factors affecting frequent lane changing using ahp-bwm with dbq data. *Journal of Transportation Safety & Security*, 13(8):791–810, 2021.

[80] Richard A. Retting, Helen B. Weinstein, and Mark G. Solomon. Analysis of motor-vehicle crashes at stop signs in four u.s. cities. *Journal of Safety Research*, 34(5):485–489, 2003.

[81] Amin Mirza Boroujerdian, Mahmoud Saffarzadeh, Hassan Yousefi, and Hassan Ghassemian. A model to identify high crash road segments with the dynamic segmentation method. *Accident Analysis & Prevention*, 73:274–287, 2014.

[82] E. Amoros, J.L. Martin, and B. Laumon. Comparison of road crashes incidence and severity between some french counties. *Accident Analysis & Prevention*, 35(4):537–547, 2003.

[83] James A. Bonneson and Patrick T. Mccoy. Estimation of safety at two-way stop-controlled intersections on rural highways. *Transportation Research Record*, 1993.

[84] Muamer Abuzwidah and Mohamed Abdel-Aty. Crash risk analysis of different designs of toll plazas. *Safety Science*, 107:77–84, 2018.

[85] Nardos Feknssa, Narayan Venkataraman, Venky Shankar, and Tewodros Ghebrab. Unobserved heterogeneity in ramp crashes due to alignment, interchange geometry and truck volume: Insights from a random parameter model. *Analytic Methods in Accident Research*, 37:100254, 2023.

[86] Jodi Carson and Fred Mannering. The effect of ice warning signs on ice-accident frequencies and severities. *Accident Analysis & Prevention*, 33(1):99–109, 2001.

[87] Erdong Chen and Andrew P. Tarko. Modeling safety of highway work zones with random parameters and random effects models. *Analytic Methods in Accident Research*, 1:86–95, 2014.

[88] J.W.H. (Jan Hendrik) van Petegem and Fred Wegman. Analyzing road design risk factors for run-off-road crashes in the netherlands with crash prediction models. *Journal of Safety Research*, 49:121.e1–127, 2014.

[89] Jaimyoung Kwon and Pravin Varaiya. Effectiveness of california's high occupancy vehicle (hov) system. *Transportation Research Part C: Emerging Technologies*, 16(1):98–115, 2008.

[90] Rory A. Austin and Barbara M. Faigin. Effect of vehicle and crash factors on older occupants. *Journal of Safety Research*, 34(4):441–452, 2003. Senior Transportation Safety and Mobility.

[91] Ulf Brüde and Jörgen Larsson. Models for predicting accidents at junctions where pedestrians and cyclists are involved. how well do they fit? *Accident Analysis & Prevention*, 25(5):499–509, 1993.

[92] Mohamed Abdel-Aty and Hassan Abdelwahab. Modeling rear-end collisions including the role of driver's visibility and light truck vehicles using a nested logit structure. *Accident Analysis & Prevention*, 36(3):447–456, 2004.

[93] Michael F Ballesteros, Patricia C Dischinger, and Patricia Langenberg. Pedestrian injuries and vehicle type in maryland, 1995–1999. *Accident Analysis & Prevention*, 36(1):73–81, 2004.

[94] Hsin-Li Chang and Tsu-Hurng Yeh. Risk factors to driver fatalities in single-vehicle crashes: Comparisons between non-motorcycle drivers and motorcyclists. *Journal of Transportation Engineering*, 132, 2006.

[95] Michel Bédard, Gordon H. Guyatt, Michael J. Stones, and John P. Hirdes. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accident Analysis & Prevention*, 34(6):717–727, 2002.

[96] Jacob A. Benfield, William J. Szlemko, and Paul A. Bell. Driver personality and anthropomorphic attributions of vehicle personality relate to reported aggressive driving tendencies. *Personality and Individual Differences*, 42(2):247–258, 2007.

[97] Chandra R. Bhat and Naveen Eluru. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological*, 43(7):749–765, 2009.

[98] Ahmed Sajid Hasan, Mohammad Jalayer, Eric Heitmann, and Joseph Weiss. Distracted driving crashes: A review on data collection, analysis, and crash prevention methods. *Transportation Research Record*, 2676(8):423–434, 2022.

[99] Amenah S.M. Thabit, Chaker Abdelaziz Kerrache, and Carlos T. Calafate. A survey on monitoring and management techniques for road traffic congestion in vehicular networks. *ICT Express*, 2024.

[100] Allan M de Souza, Celso ARL Brennand, Roberto S Yokoyama, Erick A Donato, Edmundo RM Madeira, and Leandro A Villas. Traffic management systems: A classification, review, challenges, and future perspectives. *International Journal of Distributed Sensor Networks*, 13(4):1550147716683612, 2017.

[101] Vishal Mandal, Abdul Rashid Mussah, Peng Jin, and Yaw Adu-Gyamfi. Artificial intelligence-enabled traffic monitoring system. *Sustainability*, 12(21), 2020.

[102] Vicente Milanes, Jorge Villagra, Jorge Godoy, Javier Simo, Joshué Perez, and Enrique Onieva. An intelligent v2i-based traffic management system. *IEEE Transactions on Intelligent Transportation Systems*, 13(1):49–58, 2012.

[103] Alena Høye. Are airbags a dangerous safety measure? a meta-analysis of the effects of frontal airbags on driver fatalities. *Accident Analysis & Prevention*, 42(6):2030–2040, 2010.

[104] Kristofer D. Kusano and Hampton C. Gabler. Safety benefits of forward collision warning, brake assist, and autonomous braking systems in rear-end collisions. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1546–1555, 2012.

[105] Shengxuan Ding, Mohamed Abdel-Aty, Natalia Barbour, Dongdong Wang, Zijin Wang, and Ou Zheng. Exploratory analysis of injury severity under different levels of driving automation (sae levels 2 and 4) using multi-source data. *Accident Analysis & Prevention*, 206:107692, 2024.

[106] Fanny Malin, Ilkka Norros, and Satu Innamaa. Accident risk of road and weather conditions on different road types. *Accident Analysis & Prevention*, 122:181–188, 2019.

[107] John D. Bullough, Eric T. Donnell, and Mark S. Rea. To illuminate or not to illuminate: Roadway lighting as it affects traffic safety at intersections. *Accident Analysis & Prevention*, 53:65–77, 2013.

[108] Xuan Zhang, Huiying Wen, Toshiyuki Yamamoto, and Qiang Zeng. Investigating hazardous factors affecting freeway crash injury severity incorporating real-time weather data: Using a bayesian multinomial logit model with conditional autoregressive priors. *Journal of Safety Research*, 76:248–255, 2021.

[109] Alma Cohen and Liran Einav. The Effects of Mandatory Seat Belt Laws on Driving Behavior and Traffic Fatalities. *The Review of Economics and Statistics*, 85(4):828–843, 11 2003.

[110] Lai Zheng and Tarek Sayed. Application of extreme value theory for before-after road safety analysis. *Transportation Research Record*, 2673(4):1001–1010, 2019.

[111] Ozlem Yanmaz-Tuzel and Kaan Ozbay. A comparative full bayesian before-and-after analysis and application to urban road safety countermeasures in new jersey. *Accident Analysis & Prevention*, 42(6):2099–2107, 2010.

[112] A. Ziakopoulos and G. Yannis. A review of spatial approaches in road safety. *Accident Analysis & Prevention*, 135:105323, 2020.

[113] J. Ryan et al. A risk-aware path planning algorithm using kernel density estimation and self-organizing maps. *Transportation Research Part C: Emerging Technologies*, 116:102659, 2020.

[114] Amir Pooyan Afghari, Md Mazharul Haque, and Simon Washington. Applying a joint model of crash count and crash severity to identify road segments with high risk of fatal and serious injury crashes. *Accident Analysis & Prevention*, 144:105615, 2020.

[115] American Association of State Highway and Transportation Officials (AASHTO). Highway safety manual, n.d. Accessed December 2024.

[116] Ezra Hauer, Douglas W. Harwood, Forrest M. Council, and Michael S. Griffith. Estimating safety by the empirical bayes method: A tutorial. *Transportation Research Record*, 1784(1):126–131, 2002.

[117] Meagan Powers and Jodi Carson. Before-after crash analysis: A primer for using the empirical bayes method. Final Report Z3419, Montana State University, Department of Civil Engineering, Bozeman, MT, May 2004. Research performed in cooperation with the Montana Department of Transportation and the U.S. Department of Transportation, Federal Highway Administration.

[118] Bhagwant Persaud and Craig Lyon. Empirical bayes before–after safety studies: Lessons learned from two decades of experience and future directions. *Accident Analysis & Prevention*, 39(3):546–555, 2007.

[119] Ezra Hauer. Overdispersion in modelling accidents on road sections and in empirical bayes estimation. *Accident Analysis & Prevention*, 33(6):799–808, 2001.

[120] Rune Elvik. The predictive validity of empirical bayes estimates of road safety. *Accident Analysis & Prevention*, 40(6):1964–1969, 2008.

[121] Juneyoung Park, Mohamed Abdel-Aty, and Jaeyoung Lee. Use of empirical and full bayes before–after approaches to estimate the safety effects of roadside barriers with different crash conditions. *Journal of Safety Research*, 58:31–40, 2016.

[122] Théophile Bougna, Gursmeep Hundal, and Peter Taniform. Quantitative analysis of the social costs of road traffic crashes literature. *Accident Analysis & Prevention*, 165:106282, 2022.

[123] Wim Wijnen, Wendy Weijermars, Annelies Schoeters, Ward van den Berghe, Robert Bauer, Laurent Carnis, Rune Elvik, and Heike Martensen. An analysis of official road crash cost estimates in european countries. *Safety Science*, 113:318–327, 2019.

[124] Eduard Zaloshnja, Ted Miller, Forrest Council, and Bhagwant Persaud. Crash costs in the united states by crash geometry. *Accident Analysis & Prevention*, 38(4):644–651, 2006.

[125] Ali Pirdavani, Tom Brijs, Tom Bellemans, Bruno Kochan, and Geert Wets. Evaluating the road safety effects of a fuel cost increase measure by means of zonal crash prediction modeling. *Accident Analysis & Prevention*, 50:186–195, 2013.

[126] Sheikh Shahriar Ahmed, Sarvani Sonduru Pantangi, Ugur Eker, Grigorios Fountas, Stephen E. Still, and Panagiotis Ch. Anastasopoulos. Analysis of safety benefits and security concerns from the use of autonomous vehicles: A grouped random parameters bivariate probit approach with heterogeneity in means. *Analytic Methods in Accident Research*, 28:100134, 2020.

[127] Alexandra M. Boggs, Behram Wali, and Asad J. Khattak. Exploratory analysis of automated vehicle crashes in california: A text analytics & hierarchical bayesian heterogeneity-based approach. *Accident Analysis & Prevention*, 135:105354, 2020.

[128] Xin Chang, Haijian Li, Jian Rong, Xiaohua Zhao, and An'ran Li. Analysis on traffic stability and capacity for mixed traffic flow with platoons of intelligent connected vehicles. *Physica A: Statistical Mechanics and its Applications*, 557:124829, 2020.

[129] OECD. Road accidents (indicator). https://doi.org/10.1787/2fe1b899-en, 2023. Accessed on 06 July 2023.

[130] Tingting Huang, Shuo Wang, and Anuj Sharma. Highway crash detection and risk estimation using deep learning. *Accident Analysis & Prevention*, 135:105392, 2020.

[131] Zihe Zhang, Qifan Nie, Jun Liu, Alex Hainen, Naima Islam, and Chenxuan Yang. Machine learning based real-time prediction of freeway crash risk using crowdsourced probe vehicle data. *Journal of Intelligent Transportation Systems*, 28(1):84–102, 2024.

[132] Emre Esenturk, Albert G. Wallace, Siddartha Khastgir, and Paul Jennings. Identification of traffic accident patterns via cluster analysis and test scenario development for autonomous vehicles. *IEEE Access*, 10:6660–6675, 2022.

[133] Cambridge Mobile Telematics. Distracted driving report. Technical report, Cambridge Mobile Telematics, 2023.

[134] Arity. Distracted driving trends report. Technical report, Arity, 2023.

[135] National Highway Traffic Safety Administration. Distracted driving 2022. Research Note DOT HS 813 382, National Highway Traffic Safety Administration, 2022.

[136] J. Yuan, M. Abdel-Aty, Y. Gong, and Q. Cai. Real-time crash risk prediction using long short-term memory recurrent neural network. *Transportation Research Record*, 2673(4):314–326, 2019.

[137] Bryan Lim, Sercan O. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

[138] Lei Han, Mohamed Abdel-Aty, Rongjie Yu, and Chenzhu Wang. Lstm + transformer real-time crash risk evaluation using traffic flow and risky driving behavior data. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):18383–18395, 2024.

[139] Corey Park and Mehrdad Nojoumian. Social acceptability of autonomous vehicles: Unveiling correlation of passenger trust and emotional response. In *4th International Conference on HCI in Mobility, Transport and Automotive Systems (MobiTAS)*, LNCS 13335, pages 402–415. Springer, 2022.

[140] Jamie Craig and Mehrdad Nojoumian. Should self-driving cars mimic human driving behaviors? In *3rd International Conference on HCI in Mobility, Transport and Automotive Systems (MobiTAS)*, LNCS 12791, pages 213–225. Springer, 2021.

[141] Shervin Shahrdar, Corey Park, and Mehrdad Nojoumian. Human trust measurement using an immersive virtual reality autonomous vehicle simulator. In *2nd AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 515–520, 2019.

[142] Shervin Shahrdar, Luiza Menezes, and Mehrdad Nojoumian. A survey on trust in autonomous systems. In *Computing Conference (CC)*, pages 368–386. Springer, 2018.

[143] Mehrdad Nojoumian. Adaptive speed-limit measurement (ASM) based on the traffic flow in semi or fully autonomous vehicles, 2024. US Patent 63/631,090.

[144] Mehrdad Nojoumian. Safety self talks (SST) by large language models in semi or fully autonomous vehicles, 2025. US Patent 63/747,463.

[145] D. McCarty and H. W. Kim. Risky behaviors and road safety: An exploration of age and gender influences on road accident rates. *PLoS ONE*, 19(1):e0296663, 2024.

[146] Zongni Gu, Binbin Peng, and Yu Xin. Higher traffic crash risk in extreme hot days? a spatiotemporal examination of risk factors and influencing features. *International Journal of Disaster Risk Reduction*, 116:105045, 2025.

[147] Becky P.Y. Loo. Validating crash locations for quantitative spatial analysis: A gis-based approach. *Accident Analysis & Prevention*, 38(5):879–886, 2006.

[148] Younshik Chung and IlJoon Chang. How accurate is accident data in road safety research? an application of vehicle black box data regarding pedestrian-to-taxi accidents in korea. *Accident Analysis & Prevention*, 84:1–8, 2015.

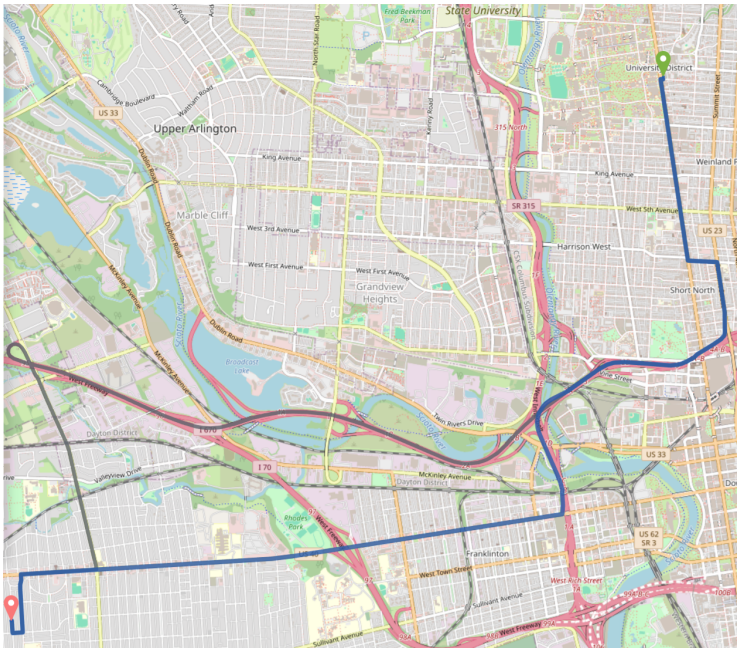
[149] Q. Hu, A. Mehdizadeh, A. Vinel, M. Cai, S. E. Rigdon, W. Zhang, and F. M. Megahed. Shortest path problems with a crash risk objective. *Transportation Research Record*, 2678(6):284–300, 2023. Original work published 2024.

APPENDICES

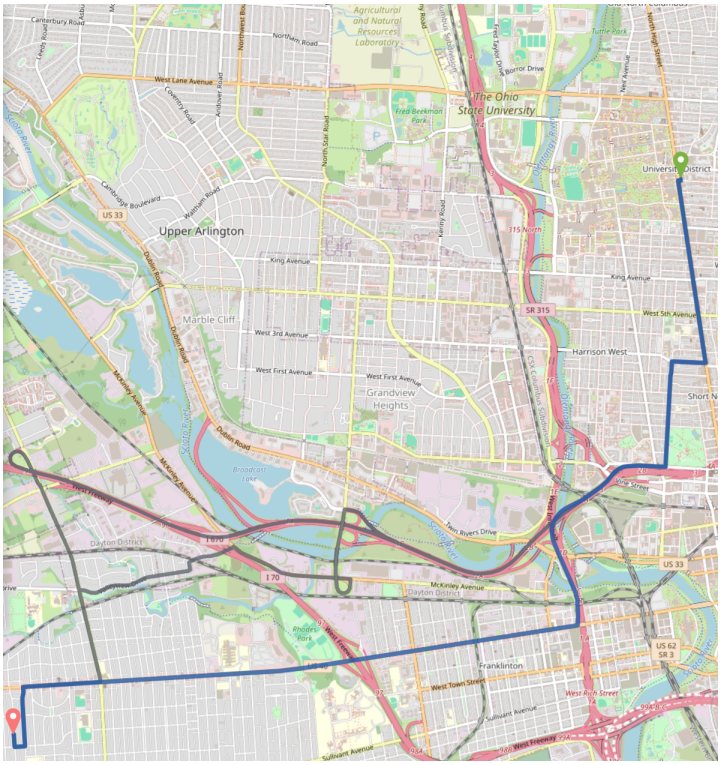
APPENDIX A

VISUAL ROUTE COMPARISON

Figure A.1 demonstrates the practical implementation of risk-aware routing using our GraphHopper modification.



(a) Risk-aware route (13.1 km, 950s, risk=0.281)



(b) Default fastest route (11.8 km, 861s, risk=0.433)

Figure A.1: Route comparison showing different path selections between risk-aware and default routing algorithms for the same origin-destination pair.