# Geospatial and Data-Driven Risk Management Methods for Tracking Anomalies in Transportation and Environmental Systems

---

A Dissertation

Submitted to the faculty of the School of Engineering and Applied Science

of the University of Virginia

---

By

Rayshaun L. Wheeler

In Partial Fulfillment of the requirements for the degree

Doctor of Philosophy in Civil Engineering

May 2025

# Approval Sheet

This Dissertation

is submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Author: Rayshaun L. Wheeler

This Dissertation has been read and approved by the examining committee:

Advisor: …….James H. Lambert
Co-Advisor: ……...Cody A. Pennetti
Committee Member:……...Lisa M. Colosi Peterson
Committee Member:……...John S. Miller
Committee Member:……...Venkataraman Lakshmi
Committee Member:……...Garrick E. Louis

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science
May 2025

# Abstract

This dissertation presents a unified framework for proactive risk analysis and systems management across critical lifeline infrastructures, specifically for the domains of transportation safety and environmental resilience. With advances in connected vehicle (CV) technologies, machine learning, and geospatial analytics, this research introduces scalable, data-driven methodologies to identify, predict, and mitigate risks in both human mobility and agricultural systems. There are three component methods as follows. (i) The first component of this research addresses pedestrian safety in high-risk school zones by applying clustering techniques (DBSCAN) and hotspot analysis (Getis-Ord Gi*) to CV data capturing harsh braking and acceleration events. These techniques enable proactive identification of hazardous driving patterns before crashes occur, supporting Vision Zero initiatives and equitable infrastructure planning. (ii) The second component explores the information management systems required to use such insights at scale. Through the development and implementation of a geospatial, model-based application, this study demonstrates how digital twins and centralized knowledge repositories can support data governance, multi-criteria decision-making, and lifecycle infrastructure planning. (iii) The third component expands the geospatial risk modeling framework to environmental systems, focusing on the impact of extreme heat on agricultural productivity. Using machine learning models (XGBoost, SVR, Random Forest) alongside spatial interpolation methods (Kriging and IDW), the study generates heat-based risk indices to assess vulnerability in food systems and human health under changing climate conditions. With the above synthesis of geospatial anomalies, or *hot spots*, methodologies across transportation and environmental sectors, this dissertation thus contributes a comprehensive, systems approach to risk management with real-time data, spatial

intelligence, and predictive analytics. The results support scalable, cross-sector solutions that enhance resilience and safety in the face of evolving threats to infrastructure and public well-being.

Keywords: Hotspots, resilience, civil engineering, machine learning, predictive analytics, risk ranking

## Acknowledgements

The motivation for the work presented in this dissertation stems from my desire to explore how principles learned in my Risk Analysis course, taught by Professor Dr. James H. Lambert, can be integrated into the rapidly evolving domains of Model-Based Systems Engineering (MBSE) and Data Science. I am especially grateful for the foundational

framework established by the Center for Risk Management of Engineering Systems (CRMES), which provided a platform for me to extend and adapt these principles to modern applications in transportation engineering and data-driven analysis. I would like to express my deepest gratitude to Dr. Joi Williams and Ms. Lessie Oliver-Clark, who first introduced me to the possibility of pursuing a doctoral degree during my undergraduate studies at Virginia State University in Information Logistics Technology. Their early encouragement set in motion an academic journey that would ultimately lead me to the University of Virginia.

To my colleagues within CRMES, thank you for your continued collaboration, encouragement, and intellectual exchange throughout this doctoral journey. The camaraderie and shared sense of purpose have made this experience both professionally enriching and personally meaningful. I am thankful to the University of Virginia for its unwavering support throughout my graduate studies in Civil Engineering and for providing numerous opportunities to travel, present, and engage with professionals at conferences across the country and internationally. These experiences have been critical to my academic and professional development.

I would also like to acknowledge the support of several organizations whose resources and collaboration significantly enriched this work: The Commonwealth Center for Advanced Logistics Systems (CCALS), the Virginia Department of Transportation

(VDOT), the National Science Foundation Center for Hardware and Embedded Systems Security and Trust, and the Port of Virginia. I extend sincere thanks to the National GEM Consortium for its commitment to empowering underrepresented minorities in STEM and for fostering a vibrant community of scholars and professionals across academia, government, and industry.

I am especially grateful to my GEM sponsor, the National Renewable Energy Laboratory, and to the National Aeronautics and Space Administration (NASA) in Titusville, Florida, for providing me with the opportunity to contribute to research at the Swamp Works Laboratory an experience that reinvigorated my passion for scientific discovery and reaffirmed my decision to pursue a Ph.D. The many IEEE and SEIDS conferences I had the privilege to attend offered invaluable insight and academic inspiration, equipping me with the knowledge and perspective necessary to build a strong foundation for this research.

Finally, I thank my family for their unwavering love and support throughout this journey. To my siblings, my fiancé, my mother, and all those who stood by me, this achievement would not have been possible without your encouragement, patience, and belief in me. Above all, I dedicate the entirety of my work throughout this Ph.D. journey to achieving safer and more efficient transportation risk analysis systems and improved walkability for all pedestrians and children to the memory of my little brother, Jaylen Hix. Jaylen's life was tragically cut short at the age of eleven due to a traffic crash, and his loss left an unfillable void in my heart. Yet from that tragedy came a renewed sense of purpose to ensure that no other family experiences the pain that mine has endured. His memory fuels my commitment to research and to making our transportation systems safer and more just. This work is for him.

# Contents



# List of Figures

## List of Tables

# List of Abbreviations

AADT: Annual Average Daily Traffic

AI: Artificial Intelligence

BSM: Basic Safety Message

CV: Connected Vehicle

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

DT: Digital Twin

Gi*: Getis-Ord Gi Statistic

GIS: Geographic Information Systems

IDEF: Integrated DEFinition

IDW: Inverse Distance Weighting

ITS: Intelligent Transportation Systems

LRS: Linear Referencing System

ML: Machine Learning

P4P: Pathways for Planning

RF: Random Forest

SUMO: Simulation of Urban MObility

SVR: Support Vector Regression

V2I: Vehicle-to-Infrastructure

VDOT: Virginia Department of Transportation

VISSIM: VerkehrsInfrastrukturSimulationsModell (Traffic Simulation Software)

XGBoost: eXtreme Gradient Boosting

# Chapter 1 | Introduction

## 1.1. Introduction

This chapter provides the foundational context for the dissertation. Section 1.2 outlines the motivation for this study, rooted in the growing need for proactive, data-informed approaches to transportation safety and environmental resilience. Section 1.3 presents the problem statement, highlighting the philosophical underpinnings of the research and introducing the integration of data-driven geospatial analysis and machine learning applications to identify and assess risk and safety factors. This section also discusses the use of IDEF modeling to structure and formalize the research process. Section 1.4 defines the purpose and scope of the dissertation, detailing the objectives and boundaries of the study. Finally, Section 1.5 provides an overview of the dissertation structure, guiding the reader through the organization and progression of the subsequent chapters.

## 1.2. Motivation

The ongoing and latest emergence of intelligent transportation systems, coupled with advancements in in-vehicle networking, sensors, and communication technologies, has enabled the collection of high-volume, near-real-time data on both vehicle and driver behavior (1). Analyzing this data presents significant opportunities to gain deeper insights into driving behavior analysis (DBA). Understanding driver behavior is essential in numerous transportation research domains, including traffic safety, connected vehicle development, energy efficiency, fuel consumption, risk evaluation, and driver profiling (2). Progress in DBA techniques, such as detecting driver distraction or impaired driving. Holds promise to reduce severe and fatal crashes. Moreover, identifying driving styles, whether eco-conscious or aggressive, can support fuel optimization strategies and risk management (3). As a result, there is a growing research focus on addressing the complexities and opportunities within DBA. Current research methodologies within traffic engineering often prioritize infrastructure investments and safety improvements based on locations with a high volume of recorded road traffic crashes, rather than utilizing intelligent transportation systems to take a proactive approach to predicting and preventing future incidents (4). The growing complexity of transportation systems demands innovative approaches to improving safety, efficiency, and resilience. In the face of rapid urbanization, climate change, and evolving mobility patterns, traditional engineering methods often fail to capture the dynamic nature of transportation networks. As a result, data-driven decisionmaking has emerged as a critical tool in modern transportation engineering,

offering the ability to analyze vast amounts of real-time data to identify risks, optimize infrastructure, and inform proactive strategies (5).

This dissertation is motivated by the potential of integrating IDEF (Integrated Definition) models, machine learning algorithms, and geospatial analysis to address pressing challenges in transportation safety and system resilience. IDEF models offer a structured way to represent complex processes, making them ideal for understanding and improving transportation workflows. Machine learning enables predictive insights into crash likelihood, traffic patterns, and driver behavior, while geospatial analysis enhances our ability to detect high-risk locations and visualize the impact of interventions.

Together, these methodologies provide a powerful framework for advancing transportation research and practice. By combining these tools, this dissertation aims to contribute to a more intelligent, adaptive, and data-informed approach to transportation planning and risk management ultimately supporting the goals of Vision Zero and sustainable infrastructure resilience. Table 1 describes a comprehensive breakdown of how emerging technologies and methodologies intersect with key components of transportation risk and safety analysis. Categorizing each research function across three strategic dimensions risk analysis, safety analysis, and data-driven decision-making, highlights potential of integrating intelligent systems into transportation planning and operations. For example, IDEF modeling enhances traditional risk assessment by identifying potential failure points and mapping critical safety components in transportation workflows while supporting data pipeline structuring for connected vehicle (CV) and environmental data.

Connected vehicle data provides real-time insights into hazardous driving behaviors and supports predictive modeling that informs timely interventions. Geospatial analysis complements these approaches by visualizing spatial risk exposure, such as in school zones, and enhancing the granularity of safety dashboards. Environmental factors, like extreme heat, add another layer of analysis by informing both human vulnerability and infrastructure resilience. Machine learning and predictive analytics amplify this framework by modeling crash likelihood and driver behavior trends, contributing to a proactive safety posture. Lastly, information management systems centralize diverse data inputs, integrate historical and real-time crash information, and support digital decision-making tools such as digital twins. Collectively, the matrix illustrates how the convergence of these technologies and methodologies can transform transportation systems into adaptive, datainformed networks capable of anticipating and mitigating safety and risk challenges.

**Table 1** - Framework for Data-Driven Decision Making in Safety and Risk Analysis. This table outlines the integration of various research components, from data modeling to predictive analytics, to enhance safety analysis and inform decision-making processes.

| **Research Function/Component** | **Risk Analysis** | **Safety Analysis** | **Data Driven Decision Making** |
|---|---|---|---|
| IDEF Modeling | Identifies failure points in process flows. | Maps safetycritical systems in workflows. | Structures data pipelines for CV/environmental data. |
| Connected Vehicle Data | Supports risk detection via event-based data. | Detects hazardous behavior in realtime. | Enables predictive modeling for interventions. |

| | | | |
|---|---|---|---|
| Geospatial Information Systems (GIS) Analysis | Assesses spatial risk exposure. | Visualizes highrisk zones (school zones etc.) | Enhances location-specific insights and dashboards. |
| Environmental Factors (Climate/Heat) | Reveals highimpact weather scenarios. | Informs pedestrian and worker safety measures. | Predicts agricultural stress using ML and Kriging. |
| Machine Learning/ Predictive Analytics | Models future risk probability. | Identifies behavioral trends before crashes. | Drives proactive planning and decision support. |
| Information Management Systems | Centralizes risk data inputs. | Integrates crash reports and studies. | Powers tools like VDOT P4P and digital twins. |

## 1.3. Problem Statement

One of the most critical gaps in the current civil engineering body of knowledge is the integration of data-driven decision-making into planning, design, and operations. Recent studies have shown leveraging big data and predictive analytics can improve infrastructure performance by up to 20%, enabling engineers to make more informed, proactive decisions that reduce costs and increase resilience (6). With the rapid pace of global urbanization, urban traffic volumes have experienced a dramatic surge, leading to increasingly severe congestion challenges. Persistent traffic disruptions significantly reduce transportation efficiency and contribute to the overall decline in system performance.

In addition, congestion is closely linked to a rise in traffic accidents and increased environmental pollution. As a result, addressing urban traffic congestion has become a pressing issue that demands effective and innovative solutions in the evolving landscape of transportation systems, managing complexity and scale demands structured methodologies

that align data, processes, and decision-making tools. The use of IDEF (Integrated DEFinition) modeling and robust information management systems, as demonstrated in this paper, provides a vital bridge between conventional traffic engineering practices and emerging Intelligent Transportation Systems (ITS).

IDEF modeling contributes to the structured analysis of business processes within transportation agencies by explicitly identifying inputs, controls, mechanisms, and outputs (7). In the VDOT case, these models were enhanced to include risk identification, which is crucial for transportation systems where performance can be influenced by dynamic and uncertain variables (e.g., weather, policy changes, infrastructure failures). This enriched IDEF framework allows traffic engineers and planners to visualize how project components interrelate, assess vulnerabilities in workflows, and embed safety and risk considerations directly into the planning and operational lifecycle.

However, the information management system, specifically VDOT's *Pathways for Planning* platform, operationalizes these models by integrating geospatial analytics, temporal data, and multiple data sources (such as crash data, weather, and traffic volume) into a single, interactive platform. This system acts as a digital twin for Virginia's transportation infrastructure, enabling real-time situational awareness, performance monitoring, and predictive modeling. The integration of more than 130 geospatial data sources and advanced filtering tools (such as the Linear Referencing System) exemplifies how traffic operations data can be curated, visualized, and transformed into actionable insights.

Together, these tools enable data-driven decision making that aligns with the principles of ITS. Such as system interconnectivity, automation, and adaptive response. Instead of relying solely on historical crash data or post-event analysis, transportation stakeholders can now proactively identify at-risk corridors, optimize resource allocation, and simulate the impact of potential interventions before implementation.

Ultimately, by embedding IDEF modeling into the broader framework of information management, this approach enhances system resilience, safety analysis, and operational efficiency. It allows agencies to move beyond siloed, reactive methodologies toward integrated, proactive strategies that are the hallmark of intelligent transportation systems.

## 1.4. Purpose and Scope

With the above motivation, the purpose of this dissertation is to develop and evaluate an integrated framework that leverages data-driven methodologies, geospatial analytics, and system modeling to enhance risk management and safety strategies within transportation and environmental systems. Specifically, this research investigates how connected vehicle

(CV) data, when combined with predictive analytics and intelligent transportation systems (ITS), can be utilized to proactively identify risk-prone areas and inform decision-making processes at both the operational and strategic levels.

This approach addresses critical gaps in current traffic engineering practices, which have relied on retrospective crash data and reactive infrastructure investments. By incorporating Integrated DEFinition (IDEF) modeling, machine learning algorithms, and

geospatial analysis, the research provides a forward-looking approach to understanding and mitigating transportation risks particularly in urban areas experiencing rapid development and mobility challenges. The integration of environmental data further expands the scope, offering insights into how external stressors such as climate and extreme heat influence both infrastructure performance and user safety.

The scope of this study encompasses the analysis of large-scale connected vehicle datasets, including event-based records (e.g., harsh braking, acceleration, and speeding), and their spatial-temporal relationship to crash occurrences, pedestrian safety concerns, and roadway conditions. The study also includes the development and demonstration of a digital information management system that synthesizes these data sources to support planning, risk assessment, and resilience strategies. By focusing on the intersection of emerging technologies and traditional infrastructure management, this dissertation contributes to the growing body of knowledge on proactive transportation safety and environmental resilience. It offers a replicable framework that can inform state and federal agencies, metropolitan planning organizations, and policymakers in advancing Vision Zero and climate adaptation goals.

## 1.5. Structure of Dissertation

Table 2 provides an overview of the dissertation structure, outlining the organization and purpose of each chapter. The table serves as a roadmap for the reader, illustrating how the research progresses from foundational concepts to applied case studies and synthesis. It begins with the introduction and literature review, which establish the research problem,

theoretical grounding, and gaps in existing practice. The methodology chapter presents the conceptual framework, including the integration of IDEF modeling, geospatial analysis, and machine learning. Subsequent chapters detail three distinct case studies that demonstrate the application of these methods across real-world transportation and environmental challenges. The final chapters synthesize key findings, identify best practices, and compare the proposed framework against traditional approaches to transportation safety and risk analysis. Together, these chapters provide a comprehensive and logical progression of inquiry, culminating in a set of actionable insights and contributions to the field.

**Table 2.** Overview of dissertation structure

| CHAPTER | TITLE | DESCRIPTION |
|---|---|---|
| 1 | Introduction | Introduces the research problem, motivation, significance, purpose, scope, and organization of the dissertation. |
| 2 | Problem Definition | Definition current practices in transportation safety, risk analysis, intelligent transportation systems, connected vehicle data, geospatial analytics, and machine learning. Identifies knowledge gaps and research opportunities |
| 3 | Conceptual Framework and Methodology | Describes the theoretical basis for integrating risk analysis, geospatial methods, and datadriven decision-making into transportation engineering. Introduces the use of IDEF modeling, digital twins, and information management systems. |

| | | |
|---|---|---|
| 4 | Chapter 4 Case Study: Risk Mapping with Geospatial and Environmental Data | Applies geospatial analysis, kriging, and environmental datasets to identify and visualize transportation vulnerabilities, particularly in relation to extreme climate conditions. |
| 5 | Chapter 5 Case Study: Classification Model for Connected Vehicle Event Data | Demonstrates the development of a machine learning classification model to analyze harsh events (e.g., braking, acceleration, turning) using connected vehicle data to predict and classify risky driving behavior. |
| 6 | Chapter 6 Case Study: Information Management and IDEF Modeling for VDOT Planning | Applies IDEF modeling and information systems to streamline risk-informed transportation project planning, using the VDOT Pathways for Planning system as an example. |
| 7 | Synthesis of Methodologies and Best Practices | Summarizes key takeaways across the case studies. Proposes a generalized framework that integrates machine learning, geospatial analysis, and information management in transportation safety and resilience planning. |
| 8 | Comparative Evaluation and Future Applications | Benchmarks the proposed methods against traditional approaches. Evaluates performance, scalability, and potential for broader deployment across agencies and regions. |
| 9 | Conclusion | Summarizes findings, discusses the implications for practice and research, and outlines contributions, limitations, and future work. |

# Chapter 2 | Problem Definition

## 2.1. Introduction

This chapter outlines the foundations in literature for the dissertation by reviewing key research areas relevant to geospatial and data-driven risk management strategies and provides a guiding philosophy for the subsequent chapters. Section 2.1 introduces the chapter and outlines its structure. Section 2.2 examines the application of geospatial analysis techniques in transportation and environmental risk, highlighting spatial modeling

approaches, including hotspot analysis and interpolation methods. Section 2.3 reviews the use of connected vehicle data and surrogate safety measures to detect risky driving behaviors and enhance proactive safety assessments. Section 2.4 discusses machine learning and predictive modeling techniques used to analyze crash patterns and climaterelated risks. Section 2.5 explores the role of information management systems and digital twins in supporting integrated risk-informed decision-making. Finally, Section 2.6 identifies gaps in the existing literature and presents opportunities for advancing interdisciplinary research at the intersection of transportation safety, environmental resilience, and predictive analytics.

## 2.2. Applications of Geospatial Analysis in Transportation and Environmental Risk

Traffic safety remains a critical challenge in rural areas across the United States. Although only 19% of the U.S. population resides in rural regions, these areas account for a disproportionate 43% of all roadway fatalities (8, 9). This alarming disparity underscores the need for advanced, data-driven approaches to identify and mitigate transportation risks in vulnerable communities. Road traffic crashes impose significant global health and economic burdens, resulting in more than 1.3 million fatalities each year (10). In addition to the loss of life, millions sustain serious injuries that contribute to long-term disability and reduced quality of life. Economically, these incidents are estimated to result in a 3%

loss of gross domestic product for many countries, with low- and middle-income nations bearing the most significant impact (11).

Spatial analysis methods are increasingly recognized as vital for examining the spatial dimensions of traffic crashes and understanding underlying patterns (12). These techniques facilitate the synthesis and mapping of multiple data sources, enabling the identification of location-based risk factors and clusters. By doing so, geospatial approaches enhance the precision of road safety strategies and support more effective, evidence-based decision-making (13).

The National Oceanic and Atmospheric Administration (NOAA) and others have highlighted record-breaking climate extremes over the past two years (14). An increase in both the frequency and severity of extreme weather events, such as heat waves, which are becoming more prevalent as a result of ongoing climate change (15). Heat waves are particularly dangerous due to their capacity to significantly impact public health, often placing the greatest burden on vulnerable groups, including the elderly, individuals with limited financial resources, and those with underlying health conditions (16). Numerous studies have shown a direct relationship between prolonged exposure to extreme heat and elevated rates of heat-related illnesses and fatalities (17).

If there are accelerating features of climate change, it is increasingly critical to evaluate population vulnerability to heat stress and to identify high-risk geographic areas. Such assessments are essential for informing the development of effective mitigation strategies and public health interventions. The Intergovernmental Panel on Climate Change characterizes vulnerability as the degree to which individuals or systems are likely to

experience harm, encompassing both their sensitivity to climate impacts and their capacity to adapt or respond (18).

Figure 1 illustrates the conceptual integration of transportation, environmental, and geospatial data to support predictive analysis and decision-making within the domains of transportation safety and environmental risk. The diagram highlights three core data sources: Connected Vehicle Data, Environmental Data, and Geospatial Data. These sources are increasingly available through modern sensing technologies, spatial databases, and transportation networks, and offer unique insights into dynamic conditions on roadways and across ecosystems.

At the center of the diagram is the Integrated Geospatial and Environmental Data node, which represents the integration of these datasets through advanced analytical techniques. This integration allows for spatially and temporally resolved insights that are essential for understanding context-sensitive risk factors. For example, combining connected vehicle data with environmental variables such as extreme temperatures or precipitation can help reveal patterns of hazardous driving behavior during adverse conditions. Similarly, geospatial mapping of environmental exposure and population vulnerability allows for the creation of spatial risk indices.

The two output nodes reflect practical and actionable outcomes of this integrated data framework. Transportation Safety Clustering Analysis enables researchers and policymakers to identify high-risk zones for crashes or near-misses using methods like DBSCAN or hotspot analysis. This supports proactive infrastructure and policy responses, particularly in sensitive areas such as school zones or rural corridors. On the other side, the

Environmental Risk Index for Policy Decisions provides a spatially informed metric that highlights regions most susceptible to climate-related health risks, such as extreme heat. This index can guide resource allocation, emergency response planning, and long-term adaptation strategies. Overall, the diagram encapsulates a systems-thinking approach to data-driven decision-making. Demonstrating how disparate datasets can be synthesized through geospatial and machine learning techniques underscores a key theme of the dissertation: Integrating data across disciplines will be key to understanding the anomalies and hotspots of disparate systems, focused here on cases of safety and risk in transportation and environmental domains.

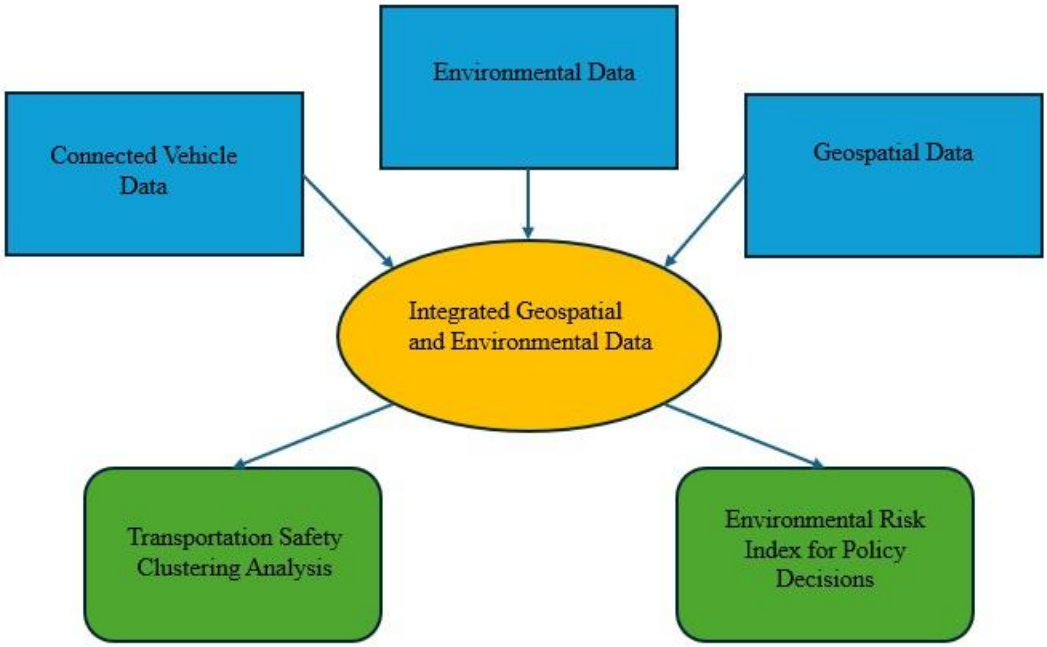


Figure 1. Integrated framework for data-driven transportation safety and environmental risk analysis

## 2.3. Connected Vehicle Data and Safety Measures

A significant proportion of vehicle crashes, nearly half, occur at roadway intersections and access points, with studies showing elevated crash rates along highways that experience high traffic volumes and dense access spacing (19). Traditional aggregated transportation performance metrics, however, often mask critical localized variations in system behavior, making it difficult to identify specific areas of concern (20). Limitations in data availability and granularity further complicate performance assessment, particularly during timesensitive operations. For instance, school arrival and dismissal periods involve brief but critical windows of high activity, sometimes accompanied by temporary speed limit changes, that are not captured in daily or hourly averages. Aggregated datasets from sources such as GPS or probe vehicles may fail to reflect these localized and temporal variations in travel behavior.

The growing availability of connected vehicle (CV) data introduces new possibilities for gaining detailed insights into transportation system performance. With its high-resolution spatial and temporal attributes, CV data allows for more precise analysis but it also demands thoughtful data management strategies to ensure efficient processing and resource allocation. As explored in this research, CV data enables enhanced risk detection and performance evaluation along corridors by providing a more nuanced understanding than traditional datasets. Vision Zero, a globally adopted traffic safety strategy first introduced in Sweden in 1997, aspires to eliminate all roadway fatalities and serious injuries (21). The initiative is grounded in the belief that while human errors in traffic are unavoidable, the transportation system should be designed in a way that prevents

these mistakes from resulting in severe harm. This is typically accomplished through infrastructure improvements, vehicle speed regulation, and technological innovations. The present research contributes to this proactive safety paradigm by applying predictive analytics to identify hazardous locations based on patterns of harsh braking and acceleration surrogate indicators of elevated crash risk. Leveraging high-resolution connected vehicle (CV) data and machine learning techniques, this study identifies potential risk zones before crashes occur, allowing agencies to deploy targeted interventions such as traffic signal timing adjustments, signage enhancements, or roadway redesigns. This proactive, data-driven approach aligns directly with the goals of *Vision Zero* by focusing on injury prevention rather than reactive solutions.

Despite efforts to improve roadway safety, national data shows that the goal of zero fatalities remains distant. While there has been modest progress, with overall traffic deaths declining, pedestrian fatalities have surged rising by 14.1% compared to 2019 and by 77% since 2010 (22). These troubling statistics highlight the persistent vulnerability of nonmotorized road users and underscore the need for renewed attention to equitable safety strategies. Systemic challenges in roadway design, enforcement, and accessibility remain barriers to true progress, and addressing these requires better data, smarter analysis, and informed investment in high-risk areas.

Traditional traffic data collection methods, such as fixed-location sensors, have served an important role in understanding roadway conditions but come with notable limitations. Typically, these systems monitor traffic volumes or speeds during predefined windows, such as peak or off-peak periods. As a result, they may miss key behavioral trends

occurring during non-peak hours or in less predictable conditions like school zone arrival and dismissal periods (23). Additionally, the cost and complexity of deploying and maintaining these sensors restricts their scalability, especially on extensive highway networks or in rural regions. Many widely used vehicle trajectory datasets, such as the FHWA's NGSIM program, only capture short roadway segments and limited durations, which constrains their usefulness for comprehensive behavior modeling (24). Similarly, emerging technologies like drones and LiDAR offer high-resolution data but are constrained by spatial coverage and context, often lacking critical information like weather, lighting, or roadway characteristics (25).

Connected vehicle data offers a compelling alternative to these legacy approaches. With the growing adoption of vehicle-to-infrastructure and vehicle-to-vehicle communication technologies, CV data provides real-time, wide-area traffic insights that are both spatially and temporally rich. This allows researchers to move beyond isolated snapshots of roadway performance and instead monitor dynamic traffic behavior continuously across diverse operating conditions. Additionally, CV data supports the analysis of surrogate safety measures like rapid deceleration events that can signal risk even in the absence of crash reports. As such, CV data enables a shift toward predictive, preventive safety planning that is more aligned with modern transportation system management goals and Vision Zero's emphasis on eliminating preventable harm.

Road traffic crashes remain a major public health and economic concern worldwide, contributing to a significant number of fatalities, injuries, and financial losses each year. In response, countries across the globe have adopted various strategies and

interventions aimed at enhancing traffic and roadway safety (26). The economic burden of road traffic injuries is considerable, with estimates indicating that such incidents can cost nations between 1% and 5% of their Gross Domestic Product (GDP) annually. According to the World Health Organization (WHO), road traffic incidents are the foremost cause of death among children and young adults globally (27).

Recognizing the gravity of this issue, the United Nations (UN) incorporated road safety into its Sustainable Development Goals. These include two primary targets: Reducing global road traffic fatalities and injuries by 50% by the year 2030, and ensuring equitable access to safe, affordable, and sustainable transportation systems. To support these goals, the Second Decade of Action for Road Safety (2021–2030) promotes the implementation of the Safe Systems approach an initiative rooted in the Vision Zero philosophy. This framework acknowledges the inevitability of human error while emphasizing a system design that minimizes the likelihood of fatal or serious outcomes through layered safety protections (28).

## 2.4. Machine Learning and Predictive Modeling in Safety and Environmental Systems

Climate change is a widely acknowledged global challenge with profound implications for both ecological systems and human society. In recent years, its impacts have become increasingly visible, marked by the rising intensity and frequency of extreme weather events. These phenomena, including hurricanes, floods, wildfires, and droughts,

have caused extensive damage to critical infrastructure, disrupted economic activity, and imposed significant financial and environmental costs (29).

If there is acceleration of these risks, it is essential that the private sector proactively evaluates its exposure to climate-related disruptions and implements strategies to enhance operational resilience. Businesses, which are deeply embedded in social, environmental, and economic systems, face multifaceted threats from climate change. These range from rising input costs and supply chain interruptions to shifts in consumer demand and increasing regulatory pressures (30). Empirical evidence suggests that global economic output may decline significantly as temperatures rise; for example, each 1°C increase in average global temperature is associated with an estimated 1.2% reduction in global GDP (31). Additionally, agricultural productivity, especially in regions like the United States, is expected to suffer due to changing climatic conditions, jeopardizing food security and rural economies.

To respond effectively, businesses must integrate climate science into their decision-making frameworks. This integration is crucial for assessing exposure, forecasting potential disruptions, and crafting strategies to build long-term resilience. However, leveraging climate data in business contexts is inherently complex. The spatial and temporal variability of climate models, along with the uncertainty of future scenarios, often makes this data difficult to interpret and apply directly to corporate planning.

To bridge this gap, advanced technologies such as artificial intelligence (AI) and deep learning have emerged as valuable tools. AI enables machines to replicate human reasoning by learning from data and making autonomous decisions (32). Within this

domain, deep learning uses multilayered neural networks to identify intricate relationships in large datasets and has already demonstrated success in areas such as image processing, language modeling, and increasingly, climate science (33).

The application of AI in climate risk management presents a promising avenue for enhancing business resilience. By harnessing large-scale climate datasets, AI algorithms can identify vulnerabilities, simulate potential impacts, and forecast adverse events with improved accuracy (34). These insights allow businesses to make informed decisions in areas such as supply chain logistics, capital investment, and risk mitigation planning. AIpowered decision support systems also play a critical role in optimizing operations amid uncertain environmental conditions. Notably, deep learning techniques are increasingly used to predict specific climate hazards, such as floods, droughts, and tropical cyclones, based on historical and real-time data, enabling earlier warnings and more precise responses (35).

## 2.5. Information Management Systems and Digital Twins

A digital twin provides an interactive virtual representation using multidimensional, multiscalar, multidisciplinary, and multiphysical parameters, effectively simulating the attributes, behaviors, and operational rules of real-world entities, making it especially valuable for traffic engineering applications (36). By thoroughly incorporating lifecycle data, digital twins can precisely mirror the performance, operations, environmental interactions, geometry, and current state of transportation systems (37). Nevertheless, realizing these digital twins for enhanced traffic management and planning is complicated

by significant challenges such as ensuring reliable real-time communication, accurate data analysis, and managing extensive data storage requirements (38).

In recent years, the exponential growth in data generation and sharing across businesses, governmental institutions, industrial sectors, nonprofit entities, and academic research has significantly impacted the development of information management systems and digital twins (39). Although advancements have improved data availability, many sectors now face the critical issue of data saturation characterized by vast amounts of data coupled with insufficient analytical resources and limited time for proper assessment. This proliferation complicates data-driven decision-making, as organizations are increasingly burdened by the necessity to allocate additional resources for extracting relevant and actionable insights (40). Essential operations such as data management, integration, and processing have become increasingly complex, with over 65% of organizations reporting an inability to efficiently analyze their accumulated data (41).

Several challenges necessitate the careful management and curation of data, particularly within the context of information management systems and digital twins. Issues related to high dimensionality and large-scale datasets result in heightened computational costs and increased risk of algorithmic instability. Moreover, large datasets in the realm of Big Data frequently originate from diverse sources spanning multiple temporal domains and utilize varied technologies, often incompatible and prone to obsolescence (42).

As detailed by the Data Management Body of Knowledge, information management encompasses the collection, management, and dissemination of data across multiple sources to various stakeholders (43). The intrinsic value of data is contingent upon

structured processing, necessitating substantial resource investments to pinpoint meaningful information. A notable threat to effective information management is entropy, characterized by increasing disorder or unpredictability within data sets. This threat is amplified when spatial and temporal dimensions are integral parts of the data schema (44, 45).

The design process of a Digital Twin (DT) within the transportation domain closely mirrors that of developing a physical transportation component, typically involving stages such as concept exploration, preliminary design, detailed design, implementation, testing and evaluation, as well as ongoing operations and maintenance (46). However, the DT design diverges notably from traditional physical designs by emphasizing comprehensive data collection and sophisticated analytical processes.

A standardized methodology for DT design, outlined by previous research, highlights essential decision-making processes relevant to transportation applications. This methodology explicitly addresses key high-level considerations, particularly the choice of input and output parameters and the technologies necessary for robust DT functionality (Fig. 2). Decision-making aspects, including the selection of suitable technologies, identifying essential inputs, and defining critical outputs, precede and directly inform the practical implementation of the DT. For example, decisions about DT usage frequency (whether continuously monitored or intermittently activated) and operational dynamics (static versus dynamic responsiveness to changing conditions) significantly influence design specifications.

Relevant DT technologies in transportation include machine learning and artificial intelligence (AI) for predictive analytics, experimental design methodologies, and knowledge-based systems, such as ontologies, for enhanced decision-making support. Precise characterization of inputs and outputs encompassing unit measures, data types (historical or real-time), and update frequencies is critical. The explicit documentation of these characteristics ensures that transportation DTs remain transparent, easily comprehensible, and reusable across different transportation-related scenarios.

This dissertation recognizes a gap in existing DT standardization methods by exploring the decision-making processes in greater depth within transportation applications. Specifically, it subdivides broader technology identification steps into finer, more detailed considerations such as geometric modeling decisions, the selection of appropriate simulation solvers, and the determination of suitable physics parameters.

Figure 2 describes a systematic framework developed to guide the design and deployment of Digital Twins (DTs) for transportation systems, specifically focused on applications that support proactive risk management and decision-making through integrated data-driven methods. This framework aligns with the core objective of the dissertation by incorporating key elements of geospatial analytics, connected vehicle data, and predictive modeling to create an adaptive and resilient transportation management tool. At the outset, the framework begins with the definition of a transportation objective. This objective sets the foundation for all subsequent steps and must be clearly aligned with measurable goals such as improving roadway safety, reducing travel time variability, or enhancing system efficiency. These objectives are context-sensitive and depend on the

unique challenges faced by transportation networks under study, whether urban arterials, rural corridors, or school zones vulnerable to pedestrian-vehicle conflicts. In the context of this dissertation, safety and resilience are emphasized, especially through the lens of minimizing crash risks and mitigating environmental disruptions like extreme weather events.

Once the objective has been established, the DT design process is initiated. The left-hand section of the framework outlines the hierarchical steps necessary to develop a Digital Twin that is both representative of real-world conditions and scalable for broader application. The process begins with identifying the infrastructure or system to emulate. This could include intersections, corridors, roundabouts, or entire regional networks depending on the scope of the analysis. Notably, in this research, particular emphasis is placed on intersections and access points, which have been identified as frequent locations for crash occurrences and erratic vehicle maneuvers based on connected vehicle event data.

The DT design then proceeds with the definition of both input and output parameters. Input parameters refer to observable variables that can be captured through sensor technologies or derived from connected vehicle telemetry. Examples include vehicle speed, location, acceleration or deceleration values, and lateral movement, all of which serve as proxies for driver behavior and potential risk. Output parameters represent the modeled or predicted outcomes derived from the DT, such as estimated crash risk, travel time, delay, or exposure to heat stress in vulnerable environments. These parameters form the basis for predictive analytics and surrogate safety assessment models developed in later sections of this dissertation.

To ensure consistency and data integrity across simulations, it is critical to define the characteristics of each parameter. This includes setting the appropriate units of measurement, determining sampling frequency, and specifying temporal or spatial resolution. The fidelity of the Digital Twin is directly tied to how well these characteristics mirror real-world dynamics, and inaccuracies or inconsistencies at this stage can propagate through the model, leading to flawed conclusions.

Once the digital model design is complete, the right-hand side of the framework guides the implementation and evaluation of the DT. Using modeling environments such as SUMO or VISSIM, the DT is developed with the capability to simulate the defined infrastructure and input-output dynamics. These tools provide the computational backbone for evaluating how changes in behavior, infrastructure, or policy affect the modeled transportation system.

The next phase involves performance measurement, where the effectiveness of the DT is assessed based on its ability to replicate real-world phenomena or meet the predefined transportation objectives. If the Digital Twin does not achieve performance agreement defined here as the alignment between modeled outputs and empirical data or policy goals the design loop is reinitiated. This iterative process ensures that the model evolves and adapts based on new data or revised objectives, an essential feature when dealing with complex and dynamic systems influenced by environmental, behavioral, and infrastructural variability.

If performance agreement is achieved, the DT is deployed in real-time or batch applications for continuous monitoring, scenario analysis, or strategic decision-making.

The deployment stage signifies the transition of the Digital Twin from a conceptual tool to a practical asset for transportation planners and policymakers. In the context of this research, DT deployment can support a variety of applications, including the real-time identification of crash-prone segments, the proactive adjustment of traffic signal timing to prevent vehicle-pedestrian conflicts in school zones and the spatial overlay of transportation risk with environmental hazards such as extreme heat.

By offering a structured and repeatable methodology for creating and operationalizing Digital Twins, the framework in Figure 2 supports the broader goals of this dissertation. It operationalizes the integration of diverse data sources from connected vehicles, environmental variables, and geospatial information within a modeling architecture that is capable of supporting predictive and preventive decision-making. This structured approach underscores the interdisciplinary nature of the work, bridging transportation engineering, data science, environmental resilience, and risk analysis into a cohesive and practical modeling strategy.

In summary, the framework provides a blueprint for translating theoretical concepts into actionable digital systems that enhance the safety, efficiency, and resilience of transportation networks. It demonstrates how the Digital Twin concept can be tailored to specific risk management objectives using a systematic and data-informed approach, making it a critical component of the proposed solution in this dissertation for modern transportation challenges.

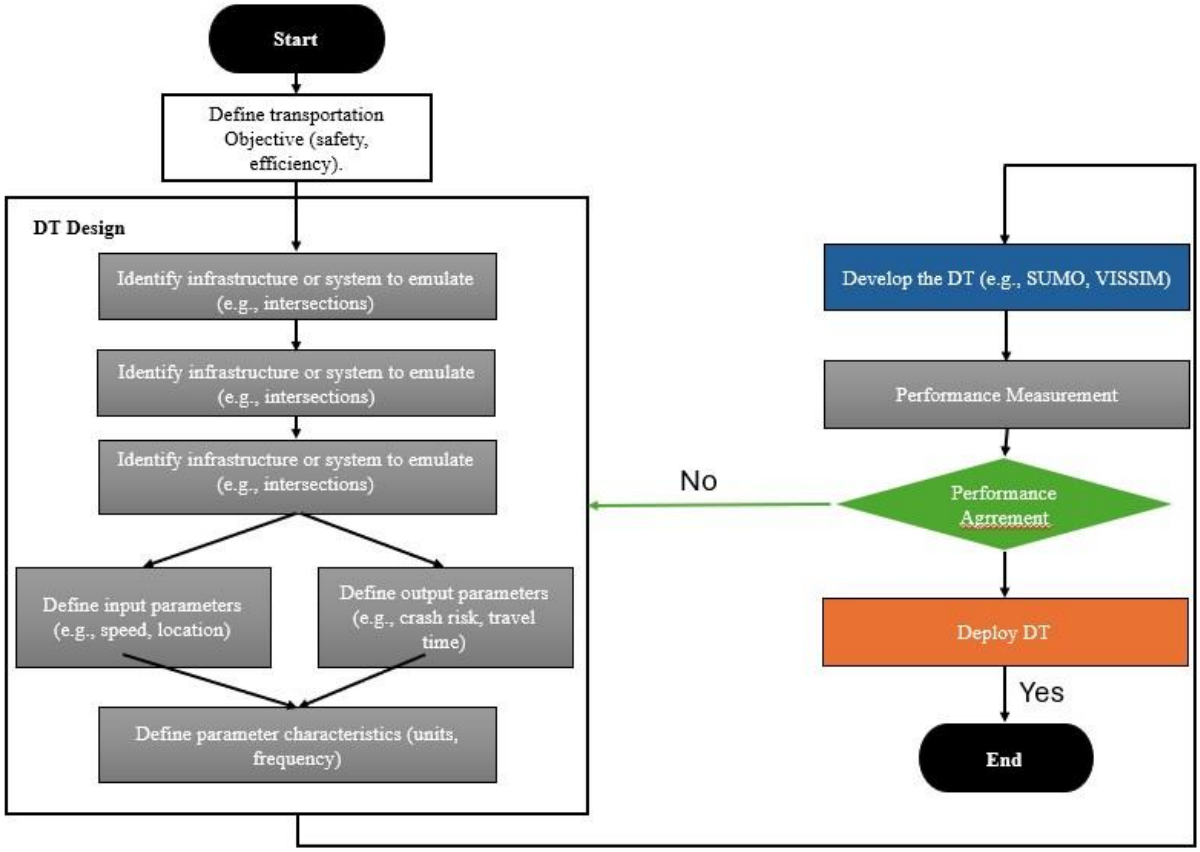


Figure 2. Digital Twin design and deployment framework for transportation systems

## 2.6. Research Gaps and Opportunities

A persistent challenge in developing an effective and reliable Digital Twin for transportation systems lies in the physical dimension of the infrastructure itself. The accuracy and dependability of digital representations are deeply influenced by the capabilities and limitations of the underlying sensor technologies used to capture realworld data. In the context of connected vehicle systems and environmental monitoring, current sensor technologies face multiple constraints that directly hinder the precision and

effectiveness of data collection. These limitations often stem from the susceptibility of sensors to environmental variability, hardware degradation, and the complex dynamics of transportation environments. For example, sensors are frequently affected by varying levels of noise, which can originate from fluctuations in ambient conditions such as temperature, humidity, or the presence of particulate matter. These forms of noise degrade signal quality and impair the measurement accuracy of critical environmental and operational variables. Cameras used in transportation networks are susceptible to changes in lighting throughout the day and across seasons, which can result in inconsistent image quality and a reduced ability to detect and classify objects or events reliably.

In addition to visual sensors, technologies such as LIDAR and RADAR, which are integral for three-dimensional spatial awareness, often encounter problems due to occlusion. When objects obstruct the direct line of sight, these sensors may either fail to detect entities or produce incomplete spatial models. Over time, sensor calibration can drift, leading to further inaccuracies and necessitating periodic maintenance to maintain measurement integrity. This need for ongoing recalibration introduces another layer of logistical and financial complexity. A significant operational challenge in deploying digital twin systems for transportation is the inherently dynamic nature of mobile agents such as vehicles, cyclists, and pedestrians. These entities exhibit unpredictable and non-linear movement patterns, making them particularly difficult to track accurately in real-time. Although advancements in real-time tracking algorithms and object detection models have shown promise, the current state of technology still leaves room for substantial improvement, especially in congested or unpredictable urban settings.

Moreover, the financial cost associated with widespread sensor deployment remains a formidable barrier to full-scale implementation. Comprehensive coverage of transportation networks demands the installation of a large number of fixed and mobile sensors, and the associated expenses for hardware acquisition, installation, calibration, and ongoing maintenance can escalate rapidly. These deployments also generate vast quantities of raw data, often in real-time, which places a heavy demand on data transmission infrastructure. High bandwidth requirements pose challenges for both wired and wireless communication networks, particularly in remote or underserved regions. To mitigate this issue, strategies such as Compressed Sensing have been proposed as effective methods for reducing data transmission burdens. Compressed Sensing enables the reconstruction of full signals from fewer samples by leveraging the inherent sparsity present in many types of transportation and environmental data. However, for such approaches to be viable, sensors must be supported by adequate edge computing capabilities that can perform real-time filtering and compression before transmitting data to centralized repositories or cloud platforms.

Beyond the limitations of hardware and communication systems, integrating heterogeneous data sources into a coherent digital twin framework introduces additional complexities. Transportation-focused digital twin systems must increasingly incorporate data from external and non-transportation sources such as weather forecasts, emergency response notifications, and public health records. These data streams often exist in disparate formats and with varying levels of resolution and temporal granularity. Ensuring interoperability between these sources and transportation-specific datasets is essential to

support predictive analytics and risk modeling. Furthermore, external data streams often contain uncertainties or inaccuracies that can propagate through analytical models, resulting in flawed predictions or assessments. For example, weather reports may vary significantly in accuracy depending on the source or the forecasting model used, and traffic incident reports may contain inconsistencies in spatial tagging or time of occurrence. These uncertainties present a significant challenge to transportation risk management efforts that rely on high-confidence data to inform decision-making and proactive intervention strategies.

Advancing sensor technology and data integration methods is therefore crucial to overcoming current limitations in digital twin development. Innovations such as multispectral cameras offer a promising avenue by enabling the capture of a wider range of electromagnetic wavelengths within a single device. Unlike traditional RGB or infrared cameras, multispectral sensors can detect variations in surface materials, thermal patterns, and environmental conditions, allowing for more robust and resilient sensing across different lighting or weather conditions. These sensors can help fill existing gaps by compensating for the weaknesses of individual sensor types. As research into transportation resilience and predictive analytics continues to evolve, the development and deployment of smarter, more adaptive sensing and computing architectures will be essential to enable real-time, high-fidelity modeling of complex transportation systems. Such advancements will enhance the ability of transportation agencies and researchers to conduct proactive risk assessments, optimize mobility strategies, and ultimately improve the safety of infrastructure and resilience of environmental systems that support daily life.

## 2.7. Summary

Chapter 2 provides a comprehensive review of the interdisciplinary research that forms the foundation of this dissertation, emphasizing the integration of geospatial analysis, connected vehicle data, machine learning, and digital twin technologies to support datadriven risk management in transportation and environmental contexts. The chapter begins by exploring spatial modeling techniques, including hotspot analysis and interpolation, to identify geographic clusters of transportation risk and environmental vulnerability. It then discusses the potential of connected vehicle data to improve proactive safety assessments through the detection of surrogate indicators like harsh braking and acceleration. The review extends to predictive modeling methods that leverage machine learning to anticipate crash patterns and climate-related disruptions. In examining digital twins and information management systems, the chapter highlights the technical and operational challenges of integrating high-resolution data across dynamic and complex systems. Finally, it identifies key research gaps related to sensor limitations, data interoperability, and the high cost of implementation, while also presenting opportunities for sensors, hotspots/anomalies analytics, and associated decision-making frameworks.

# Chapter 3 | Methods

## 3.1. Introduction

Chapter 3 presents the foundational methodology used to support the integrated risk management framework developed in this dissertation. This chapter describes the theoretical underpinnings, analytical techniques, and modeling tools applied throughout the research to analyze transportation and environmental risks using geospatial and connected vehicle data. The goal is to provide a general methodological foundation that informs the more context-specific analyses found in the subsequent case study chapters.

Section 3.1 introduces the chapter and situates the methodology within the broader research objectives of the dissertation. This section emphasizes the interdisciplinary nature of the framework, which integrates transportation engineering, geospatial science, data analytics, and environmental modeling.

Section 3.2 presents the overall research design and data sources, describing how connected vehicle data, environmental datasets, and spatial information were selected and prepared to support the study. This section also outlines the criteria used for selecting tools and techniques for modeling risk across diverse scenarios.

Section 3.3 introduces the IDEF (Integrated Definition) modeling methodology, which is used to represent transportation processes and identify system-level risk points. IDEF modeling provides a structured way to break down transportation workflows, stakeholder roles, data flows, and points of potential failure. It plays a critical role in helping design and evaluate Digital Twin systems later discussed in this research.

Section 3.4 details the spatial clustering and hotspot detection techniques used for identifying concentrations of risky behavior in transportation systems. Specifically, this section describes the application of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to detect clusters of harsh driving events, as well as the Getis-Ord Gi* statistic to identify statistically significant safety hotspots. These tools support proactive safety planning by highlighting areas where interventions may be needed even in the absence of crash data.

Section 3.5 focuses on spatial interpolation methods used to estimate environmental and transportation-related variables across space. This section compares and applies

Inverse Distance Weighting (IDW) and Kriging interpolation techniques to analyze spatial patterns in heat exposure, environmental risk, and transportation data gaps. These methods are particularly useful for estimating conditions in rural or underserved areas where sensor coverage may be limited.

Section 3.6 outlines the machine learning models used for predictive analytics and feature interpretation. It highlights the use of eXtreme Gradient Boosting (XGBoost) to build robust prediction model outputs and rank the influence of different features on predicted outcomes. These tools are critical for understanding the driving factors behind crash risk and climate vulnerability.

Section 3.7 concludes the chapter with a summary of the methodological components and explains how they come together to support the overall goals of the dissertation. It also offers a transition into the case study chapters by explaining how each methodological tool will be applied in specific contexts whether analyzing school zone safety, agricultural vulnerability, or digital twin design for infrastructure resilience.

Together, these methodological components provide a comprehensive and adaptable framework that supports data-driven decision-making and risk mitigation across multiple domains within transportation and environmental systems.

## 3.2. Research Design and Theoretical Approach

This dissertation employs a mixed-methods research design that integrates geospatial analysis, data science, and systems engineering to develop a comprehensive framework for transportation and environmental risk management. The research is grounded in a

systemsthinking perspective, acknowledging the interconnectedness of infrastructure performance, human behavior, and environmental conditions. This approach supports the development of predictive and preventive strategies that can be applied across diverse use cases, from identifying risky driving patterns in school zones to assessing heat-based vulnerability in agricultural systems and designing digital twins for critical infrastructure. The methodological framework developed in this research is operationalized through three independent but thematically linked case studies, each of which applies the broader framework to a distinct domain of interest.

The foundation of this work lies in the integration of multiple high-resolution datasets that capture the spatial, temporal, behavioral, and environmental dimensions of risk. The primary data source for transportation risk analysis is connected vehicle (CV) data, which includes detailed Basic Safety Messages (BSMs) capturing vehicle speed, acceleration, location, heading, and time at sub-second intervals. These data allow for the detection of surrogate safety events such as harsh braking and rapid acceleration, which serve as proxies for crash risk without actual collision reports.

Connected vehicle (CV) data represents a transformative advancement in assessing transportation system performance and safety. Unlike traditional data sources that rely on aggregated or static measurements, CV data is derived from individual vehicles equipped with embedded communication systems that collect and transmit real-time information about vehicle speed, location, acceleration, braking, and other critical performance metrics. These vehicles operate as mobile sensors, generating rich streams of data at high temporal

and spatial resolution, often every 1 to 3 seconds, thereby enabling granular insights into mobility patterns, safety behavior, and environmental conditions (48, 49, 50).

Historically, transportation agencies have relied on metrics such as Annual Average Daily Traffic (AADT) or fixed-location detectors, which provide coarse, time-insensitive information that obscures important variations in driver behavior and roadway performance. For example, aggregated probe vehicle data or GPS-based averages often fail to capture critical conditions near school zones during specific times of the day, such as student arrival or dismissal. CV data addresses this limitation by capturing dynamic driving behaviors including harsh braking, sharp acceleration, and speed fluctuations at specific locations and times. These observations provide surrogate safety indicators that can be used to infer crash risk in the absence of direct collision data.

The Federal Highway Administration (FHWA) distinguishes CV data from connected and automated vehicle systems, clarifying that CVs refer specifically to vehicles equipped with communication technologies such as cellular modems or vehicle-toinfrastructure (V2I) modules. These systems have evolved significantly since the introduction of OnStar in 1996, and their adoption continues to expand. Although current penetration rates are still modest, averaging around 5% across U.S. highway systems, they are comparable to traditional probe vehicle datasets and sufficient for meaningful pattern recognition and risk modeling when supported by appropriate sampling strategies and infrastructure (51).

One of the primary advantages of CV data lies in its spatial and temporal granularity. In contrast to segment-level reporting over kilometers, CV data can identify critical

roadway features such as access points, school entrances, intersections, and midblock crossings. This level of detail is particularly valuable for conducting hotspot analysis, which aggregates vehicle behaviors across fine-grained geographies and temporal windows to prioritize areas of concern.

Moreover, CV event data, including harsh braking and harsh acceleration, has been increasingly validated as a meaningful surrogate for crash prediction. In this dissertation, CV event data will be used to inform the detection of risk-prone segments and to support predictive modeling strategies using machine learning.

To fully harness the potential of CV data, this research also considers the data architecture and processing requirements needed to work with datasets of this magnitude. For instance, the sample size used to demonstrate this architecture comprises approximately 55 billion CV observations, amounting to more than two terabytes of data provided by the Virginia Department of Transportation (VDOT). This dissertation adopts similar practices, integrating Extract-Transform-Load (ETL) pipelines, spatial indexing, and columnar storage solutions to ensure that CV data can be analyzed efficiently and reproducibly.

The incorporation of CV data into this research framework is a shift in how transportation risk is measured and modeled. By enabling proactive detection of safety hazards, continuous performance monitoring, and context-sensitive analysis, CV data allows transportation agencies and researchers to go beyond static, reactive approaches and toward predictive, adaptive, and data-informed decision-making.

In parallel, environmental data such as ambient temperature, solar radiation, and relative humidity are sourced from weather stations, remote sensing platforms, and publicly

available climate repositories to assess exposure to heat and other hazards. Geospatial datasets provide the spatial reference framework and include roadway networks, land use layers, school zone boundaries, agricultural regions, and other physical infrastructure layers.

To ensure interoperability and temporal alignment, all datasets are standardized to a common spatial reference system and synchronized based on a timestamp or observational period. Noise filtering, projection transformation, and unit normalization are applied during preprocessing. Event detection algorithms are used to extract relevant behavioral signals from raw CV data, while environmental records are geocoded and interpolated to estimate exposure at unsampled locations. These preprocessing steps enable a unified and scalable dataset that serves as the foundation for the modeling and analysis stages described in later chapters.

A technical strategy applied in the cases that follow is built upon a modular analytical framework, where each method contributes a specific dimension of insight. The IDEF (Integrated Definition) modeling approach is first used to represent key processes, actors, data flows, and risk points across the transportation system. This structured modeling tool supports the formalization of process-based risks and is particularly useful for developing digital twin systems in infrastructure management. For safety-related spatial analysis, clustering techniques such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) are applied to detect dense groupings of harsh events, while the Getis-Ord Gi* statistic is used to perform statistically significant hotspot analysis of spatial

risk concentrations. These methods are employed to identify hazardous areas that may not be apparent through traditional crash data alone.

To analyze environmental conditions across space, spatial interpolation methods are applied, including both Inverse Distance Weighting (IDW) and Kriging. These geostatistical methods are used to generate continuous surfaces of environmental risk indicators, such as heat index values, across the study area. These interpolated surfaces support the identification of high-risk regions that may suffer from under-monitoring or sparse sensor coverage. In conjunction with geospatial techniques, predictive modeling is conducted using machine learning algorithms, focusing on eXtreme Gradient Boosting (XGBoost) due to its high performance on tabular data and ability to capture nonlinear relationships.

The implementation of these analytical methods is supported by a comprehensive suite of software tools and computational resources designed to facilitate large-scale data processing, spatial analysis, and machine learning applications. Python is used for most machine learning, data preprocessing, and statistical analysis, leveraging packages such as scikit-learn, XGBoost, and Pandas. Geospatial analysis and visualization are conducted using ArcGIS Pro, while spatial data storage and querying are managed through PostgreSQL with the PostGIS extension. In cases requiring large-scale simulation and modeling, environments such as SUMO and VISSIM are used to support the development of digital twins. These tools provide the computational infrastructure necessary for processing large datasets and running scenario-based risk assessments.

In addition to the technical methods, ethical considerations are addressed with respect to data privacy and responsible analysis. Connected vehicle data used in this research is anonymized and aggregated to prevent individual identification, and no personally identifiable information (PII) is used. The analysis framework emphasizes transparency, reproducibility, and policy relevance, ensuring that findings can be used to support equitable and evidence-based decision-making.

To ensure alignment between research goals and methodology, Table 3 provides a mapping of the dissertation's key research questions to the corresponding analytical techniques:

**Table 3.** Alignment of research questions with analytical methods used in the dissertation

| Research Question | Analytical Methods |
|---|---|
| How can connected vehicle data be used to identify and mitigate risky driving zones? | DBSCAN, Gi* Hotspot Analysis |
| How can environmental risk exposure be estimated across agricultural regions? | Kriging, IDW, XGBoost, |
| What process models can support the design of digital twins for infrastructure safety? | IDEF Modeling, System Mapping, Simulation Integration |

This methodological framework provides a flexible and comprehensive platform for addressing complex transportation and environmental risk questions. It supports the application of advanced analytics to real-world problems while maintaining relevance to policy, planning, and engineering practice.

### 3.3. IDEF Process Modeling for System and Risk Representation

Integrated Definition (IDEF) modeling is a systematic, graphical method designed to represent complex systems and processes clearly and comprehensively (52). Initially

developed for defense applications, IDEF modeling has been widely adopted across multiple engineering domains due to its robust ability to depict detailed processes, system relationships, and the various components involved. Within the context of this dissertation, IDEF modeling is employed to provide a structured approach for visualizing and analyzing transportation and environmental resilience systems, specifically emphasizing the representation of system dynamics, data flows, and potential risk points.

IDEF models systematically illustrate system activities by decomposing processes into their fundamental components: inputs, outputs, controls, mechanisms, and sources of risk. Inputs represent data or resources that initiate or are required for system activities (53). Outputs indicate the products, results, or information generated from these activities. Controls describe constraints, policies, or conditions governing the activities, whereas mechanisms specify tools, technologies, or procedures enabling their execution. Importantly, sources of risk highlight areas or conditions within the system where uncertainty or potential failure could occur, emphasizing the critical need for proactive management strategies.

IDEF modeling offers a structured means to pinpoint where and how data-driven analytics can be most effectively implemented in transportation and environmental risk management systems, such as those investigated in this dissertation. By mapping complex interactions between connected vehicle data streams, environmental variables, and geospatial information, IDEF diagrams enable researchers and practitioners to identify critical junctures at which risks emerge or intensify. These risk points are typically linked

to uncertainties associated with data accuracy, sensor reliability, dynamic environmental conditions, or variations in human behavior.

As illustrated conceptually, the generalized IDEF modeling framework includes an "activity" or system process central to the analysis. Inputs for such models can generically include raw data from sensors, vehicle telematics, or environmental measurements. Outputs are often presented as predictive risk scores, safety evaluations, or other decisionsupport indicators. Transportation scenarios' control commonly includes regulatory frameworks, safety protocols, or established performance thresholds. Mechanisms encompass computational methods, analytical algorithms, geospatial tools, and machine learning techniques utilized to analyze data and produce outputs. Finally, the explicit identification of risk sources within this model supports targeted interventions and refined system designs to enhance resilience and performance.

Methodologically, developing an IDEF model for system and risk representation involves several iterative steps: step 1 clearly defines system boundaries and activities of interest, step 2 identifies relevant inputs, outputs, controls, and mechanisms, step 3 systematically maps data and process flows, and step 4 explicitly marks potential sources of uncertainty or system failure points. These steps facilitate a comprehensive understanding of complex interactions, ensuring transparency, reproducibility, and ease of communication among interdisciplinary stakeholders.

IDEF modeling's systematic structure ensures consistency in process representation, making it an ideal foundational tool for digital twin development, predictive analytics integration, and geospatial risk assessments (54). By providing clear

visualizations of system processes and associated risks, IDEF models serve as effective decision-making tools, supporting risk-informed, proactive management strategies in transportation safety, environmental resilience, and infrastructure planning.

## 3.3 Clustering and Hotspot Detection Techniques (DBSCAN, Gi*)

Identifying areas of elevated risk in transportation systems necessitates spatial analytical techniques capable of isolating clusters and significant concentrations of high-risk events, such as harsh braking incidents or environmental stressors. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and the Getis-Ord Gi* hotspot analysis are powerful and complementary spatial techniques used to identify and visualize locations of significant risk within geographic data (55, 56, 57). These methods offer robust, datadriven mechanisms for detecting and quantifying risk patterns that may not be evident through traditional data aggregation approaches.

DBSCAN is a density-based clustering algorithm particularly suited to identifying clusters within large datasets containing noise and spatially irregular shapes. Unlike traditional clustering methods, DBSCAN does not require the specification of a predefined number of clusters. Instead, the algorithm operates based on two key parameters: the minimum number of points required to form a cluster (MinPts) and the radius (ε) within which points are considered neighbors (58). DBSCAN classifies data points into three categories: core points, border points, and noise, based on density thresholds, making it highly effective in identifying clusters of traffic events, such as harsh braking or rapid acceleration, often occurring in localized and irregularly shaped zones (59). The

densitybased nature of DBSCAN ensures that clusters reflect accurate spatial patterns rather than imposed geometric assumptions.

The Getis-Ord Gi* statistic, often referred to as hotspot analysis, is employed to statistically determine areas where the spatial concentration of events is significantly higher or lower than would be expected by random chance. The Gi* method calculates z-scores for each spatial feature (such as roadway segments or geographic grids), indicating areas of statistically significant spatial clusters of high (hotspot) or low (coldspot) values (60). This technique considers both the location and attribute value of each feature in relation to its neighbors, making it particularly suitable for pinpointing precise hotspots of risk activity such as crash-prone zones, environmentally vulnerable regions, or localized operational issues within transportation networks.

Recent studies underscore the efficacy of combining spatial clustering and hotspot techniques for transportation risk assessment (61). For instance, DBSCAN has been effectively applied to aggregate harsh braking events to detect clusters correlated strongly with crash occurrences in highway construction zones. This approach demonstrated that clusters of abrupt braking events often correspond directly to locations of increased collision risk. Moreover, these analyses confirmed that high-density braking zones frequently overlapped with areas of known infrastructural issues, such as narrow lanes or uneven pavement surfaces, highlighting DBSCAN's practical utility in proactive risk identification and management.

Similarly, the Gi* statistic offers an additional layer of statistical validation, quantifying the significance of identified clusters to ensure that interventions target areas

of statistically validated risk. Its application enables transportation planners and decisionmakers to prioritize areas most urgently requiring attention, efficiently guiding resource allocation and infrastructure improvements. By statistically validating risk clusters, the hotspot analysis facilitates data-driven decision-making beyond anecdotal evidence or raw incident counts.

This dissertation integrates DBSCAN and hotspot analysis within the methodological framework to systematically identify and validate risk hotspots across diverse transportation contexts. Utilizing connected vehicle data, environmental measurements, and geospatial attributes, these techniques are leveraged to proactively manage and mitigate risks. The combined application of DBSCAN for exploratory spatial pattern detection and Gi* for statistical confirmation ensures robust and actionable insights. Thus, the deployment of these clustering and hotspot methodologies significantly advances the capacity for predictive and preventive interventions, aligning with broader transportation safety goals and resilience strategies.

## 3.4 Spatial Interpolation (IDW, Kriging) and Predictive Modeling (XGBoost)

Spatial interpolation and predictive modeling are fundamental analytical components within the methodological framework of geospatial and data-driven risk management. These methods enable researchers and decision-makers to estimate risk-related variables across geographic spaces where empirical observations are incomplete or sparse, thereby

enhancing the understanding of spatial variability in transportation and environmental systems.

Spatial interpolation techniques, such as Inverse Distance Weighting (IDW) and Kriging, address gaps inherent in geographic datasets by estimating unknown values based on measured data points (62). IDW is a deterministic interpolation method that calculates unknown values as a weighted average of neighboring points, where the weighting factor decreases with increasing distance from the interpolated location (63). IDW is straightforward to apply and computationally efficient, making it particularly suitable for datasets with a relatively uniform distribution of measurement points and minimal spatial complexity.

In contrast, Kriging is a geostatistical interpolation technique that considers both the distance and the spatial correlation among data points. Unlike IDW, Kriging incorporates statistical models of spatial autocorrelation, making it more flexible and typically more accurate for complex spatial patterns, especially where underlying spatial processes are non-uniform or anisotropic. Kriging techniques, such as Ordinary Kriging or Universal Kriging, can produce spatially explicit uncertainty estimates in addition to predictions, providing valuable insights into confidence intervals and the reliability of interpolated surfaces (64). These qualities make Kriging particularly useful in environmental applications, such as estimating exposure levels to extreme heat or other climate-related risks, as well as transportation analyses where risk factors like vehicle event density may vary spatially in complex ways.

Predictive modeling is another essential methodological tool employed in this dissertation, with eXtreme Gradient Boosting (XGBoost) serving as a primary predictive modeling algorithm. XGBoost is a robust, decision-tree-based ensemble algorithm recognized for its high performance and computational efficiency, particularly suited for handling large and diverse datasets characteristic of connected vehicle data and environmental observations (65). This method iteratively builds a sequence of models, where each successive model corrects errors from previous iterations, effectively capturing complex nonlinear relationships and interactions among predictive features. As such, XGBoost significantly enhances predictive accuracy and model reliability when forecasting events like traffic incidents, environmental vulnerabilities, or infrastructure risks.

Combining spatial interpolation and predictive modeling techniques enhances the methodological robustness and applicability of this dissertation. Spatial interpolation addresses data gaps and spatial uncertainties, while predictive modeling quantifies relationships and forecasts risk conditions. XGBoost, ensures that risk models are both highly accurate and transparent, thus facilitating informed decision-making and proactive management strategies. Collectively, these advanced analytical tools form the backbone of a robust, data-driven framework aimed at enhancing transportation safety, environmental resilience, and overall infrastructure reliability.

## 3.5 Model Validation and Evaluation Metrics

Ensuring the reliability and accuracy of analytical models is crucial when applying geospatial and predictive techniques for transportation and environmental risk

management. Validation procedures and clear evaluation metrics provide confidence in the models' predictive capability, helping to ensure their effectiveness in real-world decisionmaking contexts. This dissertation systematically employs robust model validation and evaluation methods to assess performance and guarantee the validity of insights derived from connected vehicle data, environmental indicators, and spatial analyses.

Model validation involves assessing how well a predictive or spatial model performs on previously unseen data, providing a realistic measure of its generalizability and applicability. A commonly employed validation strategy is k-fold cross-validation, a resampling procedure that partitions data into k subsets or "folds." In this approach, models are iteratively trained on k-1 subsets and tested on the remaining subset, averaging results across multiple iterations to ensure unbiased performance estimates. Cross-validation ensures that predictive models, such as XGBoost models applied in risk predictions, avoid issues related to overfitting and can generalize effectively to new or unseen data scenarios.

Evaluation metrics used in this dissertation vary depending on the modeling context. For predictive modeling applications involving continuous outcomes such as estimating the magnitude or probability of risk metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination ($R^2$) are employed. RMSE provides a measure of the average prediction error magnitude, giving greater weight to larger errors, while MAE offers a straightforward measure of average absolute error without such weighting. The $R^2$ metric quantifies the proportion of variance explained by the model, indicating how well the model predictions fit observed data.

Collectively, these metrics offer a comprehensive assessment of predictive accuracy.

For classification-based risk analyses, such as identifying locations with high or low risk, metrics like accuracy, precision, recall, and the F1 score are more appropriate. Precision measures the proportion of correctly predicted positive observations among all positive predictions, whereas recall assesses the proportion of correctly identified positive cases among actual positive observations. The F1 score balances precision and recall, providing an integrated measure particularly useful when classes are imbalanced, as often encountered in transportation and environmental risk scenarios.

Additionally, spatial models, including DBSCAN clustering and Gi* hotspot analyses, require spatially explicit validation methods. Spatial validation methods include evaluating the consistency of clustering patterns through silhouette scores, assessing spatial randomness using methods such as Moran's I, or leveraging known historical risk patterns as benchmarks. These spatial metrics confirm whether detected clusters and hotspots reflect significant spatial patterns rather than random chance.

In summary, the model validation and evaluation metrics employed in this research are selected for their ability to provide robust, reliable, and transparent assessments of model performance across diverse contexts. These validation procedures and metrics collectively ensure that the integrated geospatial and predictive analytical framework developed in this dissertation effectively supports evidence-based decision-making, enhancing transportation safety, environmental resilience, and overall infrastructure management.

## 3.6 Summary

Chapter 3 has established the methodological foundation underpinning this dissertation, providing a structured, interdisciplinary framework for analyzing transportation safety and environmental resilience using geospatial and connected vehicle data. The chapter outlined the theoretical rationale, practical applications, and systematic approaches adopted to manage risk proactively in diverse and complex scenarios. By integrating methods from transportation engineering, geospatial analytics, data science, and environmental modeling, the research design offers robust and comprehensive tools for effective risk analysis.

Beginning with the overall research design, the chapter emphasized the importance of high-resolution datasets, including connected vehicle data and environmental observations. The transformative capabilities of connected vehicle data, characterized by high spatial and temporal resolution, were highlighted for their unique ability to capture detailed driving behaviors and surrogate safety events, critical for precise risk detection and predictive modeling.

Section 3.3 introduced IDEF process modeling, demonstrating its utility for representing complex transportation and environmental systems. This structured approach facilitates clear visualization and understanding of inputs, outputs, controls, mechanisms, and potential risk sources within these systems, essential for systematic risk assessment and proactive decision-making. Spatial clustering and hotspot detection techniques, detailed in Section 3.4, further strengthened the analytical framework by enabling the identification and statistical validation of high-risk locations through methods such as DBSCAN and Getis-Ord Gi*. These techniques provide a scientifically rigorous

foundation for prioritizing interventions and resource allocation.

Section 3.5 presented spatial interpolation techniques (IDW, Kriging) alongside predictive modeling methodologies, highlighting their importance in accurately estimating unobserved values across spatial domains and forecasting risk scenarios. The integration of machine learning models, ensures transparency and practical usability of predictions, crucial for informed decision-making.

Finally, the chapter described rigorous model validation and evaluation strategies in Section 3.6. Techniques such as k-fold cross-validation and specific evaluation metrics like RMSE, MAE, precision, recall, and F1 score were identified as critical components for ensuring the reliability, accuracy, and robustness of analytical outcomes. Collectively, the methodologies described in Chapter 3 provide a comprehensive, transparent, and scalable analytical foundation, forming the basis for detailed explorations within subsequent case studies. These methodological tools collectively support proactive risk management, aligning closely with the overarching goals of improving transportation safety and enhancing environmental resilience.

# Chapter 4 | Case Study: Information Management for Lifeline Infrastructure

## 4.1. Introduction

Chapter 4 presents a detailed case study on Information Management for Lifeline Infrastructure, focusing on the systematic integration and management of data to enhance

infrastructure resilience. Section 4.2 introduces the application of a Modified IDEF Modeling approach specifically tailored for comprehensive project risk assessment, providing structured insights into potential failure points and interactions within infrastructure systems. Section 4.3 discusses the role and importance of geospatial features and Linear Referencing Systems (LRS) in accurately locating, analyzing, and managing infrastructure assets. Section 4.4 explores the integration of crash data, previous safety studies, and analytical dashboards, emphasizing their combined use for improving decision-making and operational safety outcomes. The benefits of maintaining a centralized knowledge repository to streamline information sharing, enhance data accessibility, and support efficient risk management processes are outlined in Section 4.5. Section 4.6 discusses practical applications of these methodologies for the development and operationalization of digital twins, highlighting their potential to proactively manage and optimize lifeline infrastructure. Finally, Section 4.7 provides a summary, synthesizing key insights and contributions from this case study.

## 4.2. Modified IDEF Modeling for Project Risk Assessment

Effective information management is a critical component within transportation agencies, mainly when supporting operations and strategic planning through geospatial model-based frameworks. This section explores a modified version of Integrated Definition (IDEF) modeling, specifically adapted to enhance traditional business process models (BPM) utilized during various phases of transportation projects such as initiation, scoping, preliminary and detailed design, procurement, construction, and operational maintenance.

Originally, IDEF modeling is employed as a structured graphical language to systematically visualize business processes, clearly defining inputs, outputs, controls, and mechanisms involved in various activities (66). Building upon this foundational model, recent research introduced the integration of risk factors as a distinct category within the IDEF framework. This addition significantly enhances the capacity to proactively identify, evaluate, and manage potential risks associated with project processes, as depicted conceptually in Figure 3.

Figure 3 presents the enhanced IDEF modeling structure, adapted from prior research, illustrating the explicit incorporation of risk identification elements into traditional IDEF models (67). The inputs to project development processes typically encompass diverse reports, engineering calculations, design models, and a wide variety of documentation drawn from multiple data sources and stakeholders. Due to the complexity and dynamic nature of transportation projects, these inputs are inherently subject to various risks, including data inaccuracies, inconsistencies arising from multiple contributors, policy fluctuations, funding volatility, and ongoing updates from operational feedback.

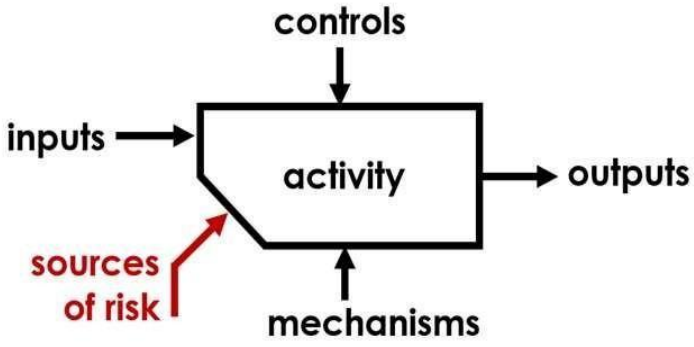


Figure 3. Integrating risk identification into IDEF

The inclusion of risk sources within the modified IDEF model explicitly acknowledges these uncertainties and threats. By doing so, the model enables stakeholders to systematically evaluate the quality, reliability, and criticality of information used in each process activity. Such evaluation is essential for advancing effective risk management strategies that rely on accurate and timely information.

To mitigate identified risks, robust information management systems are employed, providing stakeholders with high-quality, current, and trusted data. These systems facilitate disciplined data curation and proactive risk control, thus supporting the continuous monitoring and management processes necessary for optimal operational performance and informed decision-making.

Conceptually extending beyond traditional project boundaries, transportation agencies increasingly pursue the establishment of comprehensive digital twins for largescale, statewide infrastructure. For instance, a transportation agency might aim to digitally replicate a vast network comprising thousands of kilometers of roadway, extensive personnel resources, multiple concurrent projects, and substantial operating budgets. The concept of a digital twin originally developed by NASA in the context of space missions entails continuously updating digital representations of physical assets to maintain accurate and timely reflections of current conditions, thus ensuring ongoing relevance and operational effectiveness.

Recently, digital twin applications have scaled beyond isolated assets to encompass extensive urban and regional infrastructure systems. Such city-scale or region-scale digital

twins integrate multiple interconnected sectors, including transportation, buildings, energy systems, communication networks, and environmental monitoring frameworks. Although these expansive digital twins hold substantial potential for enhancing decision-making capabilities, they also present significant challenges concerning technological infrastructure, data quality and confidence, privacy concerns, security vulnerabilities, and substantial data processing demands.

Nevertheless, the strategic adoption of modified IDEF modeling within robust information management frameworks significantly contributes to overcoming these challenges. By ensuring data integrity, effective risk identification, and proactive management capabilities, transportation agencies can fully realize the benefits of digital twins, resulting in improved operational resilience, strategic planning efficacy, and informed decision-making across extensive infrastructural and organizational domains.

Transportation agencies consistently encounter challenges related to the management, integration, and reliability of data throughout the various phases of project development. State-level departments of transportation manage expansive infrastructure networks and complex operational frameworks requiring precise and reliable information for informed decision-making. In response to these challenges, a modified Integrated Definition (IDEF) modeling methodology has been developed and applied to systematically represent, assess, and mitigate risks inherent in transportation project planning and operational activities.

Figure 4 exemplifies the use of this modified IDEF approach during the scoping phase of transportation project development. This phase is particularly critical, as

inaccurate or inaccessible data can significantly compromise subsequent project outcomes, including cost overruns, schedule delays, and deviations from the defined scope. By explicitly incorporating incorrect or inaccessible data as a source of risk within the IDEF framework, transportation agencies can proactively recognize and mitigate potential threats that may impact the scoping process and subsequent project phases.

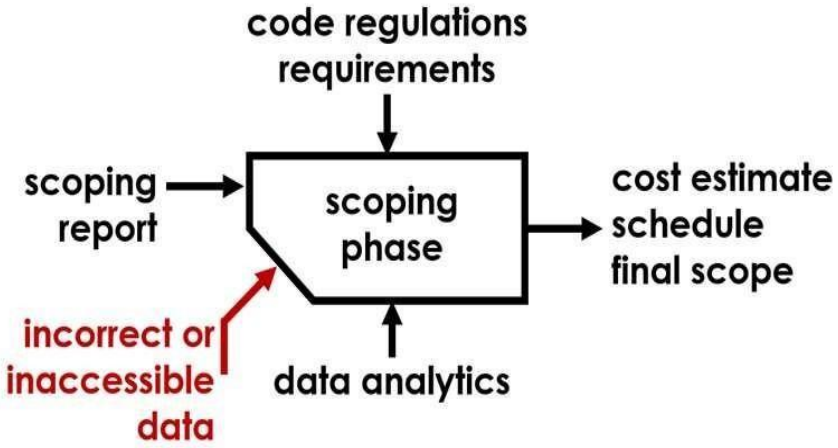


Figure 4. Modified IDEF representation highlighting data risk in project scoping

In practice, transportation agencies like VDOT leverage information management systems to address these challenges. These systems act as centralized repositories that aggregate, validate, and disseminate essential transportation-related data to stakeholders. For instance, VDOT employs a comprehensive geospatial analytics architecture within their information management platform, facilitating detailed spatial and temporal analyses. The system integrates diverse geospatial data sets such as crash records with extensive attributes, infrastructure inventories, and performance metrics ensuring that transportation stakeholders have reliable, current, and actionable information.

Analytical functions within these information management systems further enhance their value. Techniques such as data fusion allow agencies to integrate multiple datasets concurrently, enabling a comprehensive understanding of transportation conditions and performance. Geospatial visualization tools play an essential role by offering intuitive, map-based interfaces that stakeholders can use to interpret complex data, communicate findings effectively, and inform decision-making processes.

Moreover, modules designed within these systems enable precise referencing to specific transportation network elements, in-depth analysis of spatial and temporal asset characteristics, and the integration of operational reports and studies linked explicitly to geographic regions. These capabilities are essential for identifying areas requiring intervention, planning infrastructure enhancements, and optimizing operational efficiency.

The methodological framework provided by the modified IDEF model thus serves as a robust basis for information and risk management, offering transportation agencies a structured approach to systematically address data-related uncertainties. By embedding risk assessment directly into the process models used for transportation planning and management, agencies like VDOT ensure greater data integrity, improved decision confidence, and enhanced operational effectiveness, ultimately contributing to safer, more efficient transportation systems.

## 4.3 Geospatial Features and Linear Referencing System

Effective transportation system management necessitates precise and efficient management of extensive and dynamic geospatial data (68). Departments of transportation (DOTs) regularly encounter the challenge of managing vast quantities of geographic data associated

with transportation infrastructure, including roadway alignments, intersection points, mile markers, and related linear features. To systematically address these challenges, agencies commonly employ Linear Referencing Systems (LRS), which enable the precise spatial referencing of transportation assets and events along linear network structures.

A Linear Referencing System is an established method of spatial referencing designed specifically for transportation networks, allowing agencies to identify locations using linear measures relative to defined reference points rather than absolute coordinates (69). The principal advantage of an LRS is its ability to simplify the representation and referencing of linear infrastructure, such as highways and railways, into manageable, logically segmented entities. By capturing location information relative to defined reference markers, such as mileposts or roadway segments, LRS streamlines the spatial management and analysis of linear assets, simplifying data integration and exchange across diverse datasets and improving overall data consistency and quality (70).

In combination with Integrated Definition (IDEF) models, which explicitly detail business processes, activities, data flows, and points of risk, LRS significantly enhances the capabilities of information management systems utilized by transportation agencies. IDEF modeling methods establish a structured representation of workflow processes, clearly defining the types of data inputs, the processes required to transform inputs into outputs, and the inherent risks or uncertainties involved in transportation infrastructure projects. When paired with geospatial referencing provided by an LRS, IDEF models gain additional context and depth, as the spatial dimension of data flows and risk sources can be precisely visualized and analyzed within a geographical context. Such integration improves

transparency, facilitates clear communication among stakeholders, and supports informed decision-making regarding infrastructure planning, operation, and risk mitigation.

From a practical standpoint, transportation agencies employ these integrated methods to ensure data management efficiency and accuracy. For example, roadway assets such as signage, pavement conditions, guardrails, and incident occurrences can all be dynamically located through linear referencing. When data from numerous independent sources such as crash reports, maintenance records, and environmental conditions are integrated within an LRS, transportation managers and planners gain comprehensive insights about the status, condition, and performance of transportation assets. Moreover, this geospatial data integration allows for sophisticated analytical capabilities, such as hotspot identification, proximity analysis, and corridor-level assessments, thus supporting targeted intervention strategies.

Visualization is another essential benefit derived from the integration of LRS and IDEF methodologies. By representing transportation infrastructure data visually, agencies can intuitively communicate complex relationships and critical risks to diverse stakeholder groups. Such visualizations support intuitive risk assessment, spatial trend identification, and the tracking of asset conditions and performance metrics. Geographic Information Systems (GIS) software platforms often facilitate the visualization and analytics processes inherent in an LRS-based framework, thereby enhancing the effectiveness of communication among agency personnel, planners, public stakeholders, and policy decision-makers.

Furthermore, the incorporation of an LRS within an overall information management framework ensures scalability and adaptability. Transportation networks regularly undergo expansion, modification, and improvement. Linear Referencing Systems inherently support dynamic updates to asset inventories and spatial relationships, thereby accommodating continuous modifications in the transportation infrastructure. This adaptability is especially critical for departments of transportation, whose project portfolios are expansive and constantly evolving.

In summary, integrating Linear Referencing Systems and IDEF modeling methodologies within a transportation agency's information management framework offers robust capabilities for geospatial data management, spatial analysis, and visualization. Such integrated frameworks streamline data integration, improve consistency and accuracy of infrastructure information, enable effective identification and mitigation of project-level risks, and ultimately enhance the agency's decision-making capacity. By adopting such methodologies, transportation agencies significantly strengthen their operational effectiveness, infrastructure planning processes, and overall risk-informed strategic decision-making.

## 4.4. Integration of Crash Data, Studies, and Analytical Dashboards

Effective transportation management increasingly relies upon comprehensive and timely data integration to inform operational decisions and mitigate potential risks. Transportation agencies, regularly aggregate diverse data sources such as crash records, infrastructure studies, traffic performance metrics, and safety evaluations into centralized analytical dashboards. These dashboards play a critical role in translating raw data into actionable

insights, enhancing decision-making capabilities across different project development phases.

The process of integrating transportation data poses inherent risks related to accuracy, completeness, timeliness, and interoperability. Particularly in the context of crash data, incomplete or inaccurate information can lead to flawed analyses, ineffective safety strategies, and suboptimal resource allocation. This dissertation emphasizes the importance of explicitly identifying and managing these risks through structured methodologies. Leveraging the modified IDEF framework described previously (see Figure 4), this section systematically identifies and categorizes risk sources throughout data integration and management lifecycle stages, from initial data collection to visualization and ongoing maintenance.

Table 4 outlines critical risk sources that transportation agencies like VDOT encounter in the integration of crash data, related studies, and analytical dashboard systems. Each listed source of risk is systematically associated with relevant lifecycle stages, highlighting key points at which errors or vulnerabilities might emerge. For instance, the risk of "Inaccurate Data Entry" (R1) is notably critical during the initial data collection stage and continues to pose challenges during data visualization and maintenance. Similarly, the risk of "Data Incompatibility" (R3) emerges prominently during integration and dashboard development stages, potentially compromising analytical modeling accuracy and consistency across systems.

Furthermore, effective analytical dashboards require reliable integration of spatial and temporal data attributes to visualize critical patterns, trends, and hotspot locations effectively. Spatial referencing errors particularly those associated with Linear Referencing Systems (LRS) represent a significant and ongoing concern. Misalignments or incorrect referencing of crash incidents, asset locations, or project boundaries can severely limit the effectiveness of analytical dashboards, potentially leading to incorrect risk assessments or ineffective management strategies.

VDOT's analytical dashboard infrastructure exemplifies a successful integration strategy, incorporating over one million crash records with frequent data updates and approximately 150 unique attributes, including incident location, severity, contributing factors, roadway characteristics, and environmental conditions. The system employs robust data fusion methodologies to reconcile and analyze diverse data streams, ensuring consistent and accurate representations of transportation network performance. Advanced visualization tools embedded in these dashboards support the clear communication of transportation conditions, risk hotspots, and infrastructure performance across jurisdictional boundaries and administrative divisions. Figure 5 illustrates the spatial visualization capabilities of the Virginia Department of

Transportation’s (VDOT) Pathways for Planning platform, which serves as a comprehensive digital interface for monitoring both physical infrastructure assets and realtime performance metrics displayed in figure 5. The interface allows users to explore a variety of interactive map layers, such as crash data categorized by severity (e.g., fatal, severe, visible, minor, and property damage only), STARS study areas, and traffic volume

represented by Annual Average Daily Traffic (AADT). Each dataset is visually encoded using distinct color schemes and symbols, enhancing the user's ability to identify patterns and assess infrastructure conditions at both macro and corridor levels. This geospatial functionality plays a critical role in supporting data-driven transportation planning, enabling engineers, planners, and policymakers to visualize risks, prioritize improvements, and align projects with safety and congestion mitigation goals.


Figure 5. VDOT Pathways for Planning, geospatial viewer of multiple information layers, stylized by feature attributes.

Ultimately, the structured identification and management of data integration risks, as illustrated in Table 4, strengthen transportation agencies' decision-making processes. Explicitly recognizing potential risk points and addressing them through disciplined data governance, quality control procedures, and robust information management systems

enhance agencies' abilities to proactively mitigate risks. This structured approach ensures that analytical dashboards and associated studies deliver actionable and reliable insights, fundamentally contributing to the overarching goals of transportation safety, operational efficiency, and risk-informed decision-making.

**Table 4.** Sources of risk across data integration and management lifecycle stages for transportation agencies

| Sources of Risk | Data Collection | Data Integration | Dashboard Development | Analytical Modeling | Visualization & Reporting | Maintenance & Updating |
|---|---|---|---|---|---|---|
| R1 - Inaccurate Data Entry | + | | | | + | + |
| R2 - Data Loss or Corruption | + | + | + | + | + | + |
| R3 - Data Incompatibility (format/systems) | | + | + | + | | |
| R4 - Unauthorized Data Access | | + | + | | + | + |
| R5 - Network or Infrastructure Failures | | + | + | + | | + |
| R6 - Incorrect Spatial Referencing (LRS errors) | + | + | | + | + | + |
| R7 - Software Bugs or Defects | | | + | + | + | |
| R8 - Data Misinterpretation (analytical errors) | | | | + | + | |
| R9 - Incomplete or Missing Attributes | + | + | | + | + | + |
| R10 - Timeliness and Latency Issues | + | + | | + | + | + |
| R11 - Changes in Reporting Standards or Regulations | | | | + | + | + |
| R12 - Visualization Errors (misleading maps/charts) | | | + | | + | |

## 4.5. Benefits of a Centralized Knowledge Repository

A centralized knowledge repository provides significant advantages for transportation agencies and organizations responsible for large-scale transportation management, project implementation, and operational risk mitigation (71). In the context of transportation

agencies, the volume, complexity, and dynamic nature of data generated by projects and operations necessitate an organized and integrated approach to knowledge management.

Centralized knowledge repositories offer structured and streamlined access to comprehensive data, which enhances the ability of transportation agencies to make timely and informed decisions. Unlike fragmented or isolated data storage methods, centralized systems provide stakeholders across multiple jurisdictions and project phases the ability to easily retrieve and utilize critical information. This integrated data structure is particularly valuable for supporting decisions related to project scoping, resource allocation, safety management, and operational improvements.

A centralized repository supports consistent and standardized data management, reducing redundancies and minimizing inconsistencies that commonly occur when multiple stakeholders manage their own independent data systems (72). The standardized data structures improve interoperability and information sharing, enabling smoother collaboration among different divisions or jurisdictions within a transportation agency. Such uniformity also significantly reduces the likelihood of misinterpretations or errors, which could otherwise lead to costly project delays or ineffective operational responses.

Moreover, transportation projects involve numerous stakeholders across diverse teams including planners, engineers, emergency response personnel, and policymakers. A centralized knowledge repository allows these diverse groups to interact more efficiently by providing a common reference point and ensuring all parties have access to the same accurate, timely, and relevant information. As a result, these systems facilitate improved

communication, coordination, and cooperation, fostering a collaborative environment that enhances overall organizational effectiveness.

One of the crucial benefits of centralized knowledge management systems is their capability to support proactive risk identification and mitigation. These systems can integrate and analyze vast amounts of historical and real-time data, helping transportation agencies detect emerging risks or patterns early, often before significant impacts manifest. With centralized repositories, transportation agencies can leverage historical data to identify trends, predict future conditions, and proactively address potential safety hazards or operational inefficiencies.

Centralized repositories also facilitate continuity in knowledge retention and transfer, significantly enhancing institutional memory (73). Projects in transportation agencies are typically executed over extended periods and often involve shifting project teams. Without effective knowledge management, valuable insights and experiences are frequently lost during project transitions. Centralized repositories ensure knowledge gained throughout project lifecycles is systematically captured and retained, enabling the seamless transfer of knowledge across different project phases and to future projects. This approach mitigates the risk associated with turnover of key personnel and helps maintain consistent project performance.

Finally, analytical dashboards and geospatial visualization integrated within centralized knowledge repositories enhance decision-making by presenting complex information clearly and intuitively. Visual analytics capabilities assist stakeholders in interpreting intricate datasets and support quicker, data-driven decision-making processes.

By centralizing geospatial and operational data, transportation agencies can employ advanced visualization techniques to illustrate spatial and temporal patterns, effectively highlighting critical issues or priority areas for intervention.

In summary, centralized knowledge repositories are foundational for modern transportation agencies managing extensive networks, complex projects, and multifaceted operational responsibilities. Through standardized data management, proactive risk detection, effective communication, and continuity of knowledge transfer, such repositories substantially enhance organizational efficiency, safety outcomes, and strategic decision-making capabilities.

## 4.6 Applications for Digital Twin Development

Digital twin technology presents a transformative capability for transportation agencies, enabling the detailed, dynamic representation of real-world transportation assets and operational conditions within a digital environment (74). Within the context of transportation risk management, digital twins offer transportation agencies the ability to proactively simulate, monitor, and optimize their extensive networks of infrastructure, spanning thousands of kilometers of roadway, numerous bridge structures, tunnels, and associated infrastructure assets. The deployment of digital twins allows agencies to forecast and evaluate various operational scenarios, improving planning effectiveness, enhancing safety, and mitigating risks associated with infrastructure management and maintenance.

The development and effective implementation of digital twin technology fundamentally rely upon structured modeling frameworks, such as the Integrated Definition (IDEF) methodology. IDEF models facilitate systematic decomposition of complex

transportation processes, capturing the intricate interplay of data flows, operational constraints, decision points, and the critical identification of risk sources. By establishing clearly defined relationships between various system components including inputs, outputs, mechanisms, controls, and associated risks IDEF models provide essential foundational structures necessary for digital twin construction. Specifically, these models clarify the logical architecture and data requirements essential for representing physical infrastructure digitally, enabling accurate and reliable digital twins that reflect real-world operational contexts and risks.

Transportation agencies leverage digital twins in numerous practical applications, each significantly benefiting from integration with IDEF modeling frameworks. One primary application includes real-time monitoring and predictive analysis of asset conditions and operational performance. By continuously integrating sensor data, environmental measurements, and asset conditions into a digital twin, transportation agencies can proactively detect anomalies or potential risks, such as infrastructure deterioration, environmental vulnerabilities, or operational inefficiencies. Through the structured processes outlined by IDEF models, transportation agencies ensure that digital twins are systematically updated with accurate data streams, thereby enhancing their reliability as decision-support tools.

Furthermore, digital twins provide powerful capabilities for scenario analysis and planning. Transportation agencies frequently encounter uncertainties related to extreme weather events, changes in traffic patterns, or infrastructure expansions. A digital twin, grounded in the systematic process modeling of IDEF frameworks, supports sophisticated

simulations and "what-if" analyses to understand potential outcomes under varying operational conditions or environmental scenarios. These analyses allow transportation agencies to identify and evaluate risks proactively, enabling informed strategic planning and optimized allocation of resources.

Digital twin applications are also notably effective in emergency response planning and crisis management. In situations such as severe weather events or large-scale incidents, digital twins, integrated with geospatial data and real-time sensor information, facilitate rapid assessment of infrastructure conditions, traffic congestion, and potential safety hazards. Leveraging the clarity and structured representation provided by IDEF process models, agencies can more effectively coordinate resources, prioritize interventions, and mitigate risks in real-time during emergencies.

In addition to operational and emergency management applications, digital twins significantly enhance transportation infrastructure lifecycle management. Infrastructure projects involve complex interactions between multiple stakeholders, data streams, and evolving operational conditions. Digital twins developed through systematic IDEF modeling practices facilitate effective management of infrastructure throughout its lifecycle, from initial design and construction to long-term maintenance and eventual decommissioning or replacement. These digital representations ensure consistency in project information, improve communication among stakeholders, and support comprehensive risk management at every lifecycle phase.

Ultimately, the integration of digital twin technology and structured IDEF modeling frameworks enables transportation agencies to manage complex infrastructure networks

with unprecedented accuracy, efficiency, and foresight. The digital twin serves not merely as a static digital replica but as a dynamically evolving representation of infrastructure assets, continuously updated through structured data integration processes. As demonstrated throughout this dissertation, the strategic deployment of digital twins, informed by robust IDEF process models, fundamentally enhances transportation system resilience, safety, operational efficiency, and overall risk-informed decision-making capacity.

## 4.7 Summary

Chapter 4 of the dissertation provides a comprehensive case study on Information Management for Lifeline Infrastructure, emphasizing how structured data integration can significantly enhance infrastructure resilience. The chapter introduces a modified Integrated Definition (IDEF) modeling approach tailored for project risk assessment within transportation projects. This enhanced model explicitly incorporates risk identification elements, helping agencies systematically identify, evaluate, and manage potential threats such as data inaccuracies, inconsistencies, policy fluctuations, and funding volatility. By integrating these risk factors, the model improves decision-making, supports proactive risk mitigation, and fosters effective project management throughout various phases, including scoping, design, procurement, and maintenance.

The chapter further explores the significance of Geospatial Features and Linear Referencing Systems (LRS), detailing their role in managing spatially extensive and dynamic transportation data. It discusses how LRS simplifies asset management by enabling precise referencing of assets and incidents along linear networks, significantly

enhancing data consistency, integration, and analytical capability. Additionally, the integration of crash data, historical safety studies, and analytical dashboards within centralized knowledge repositories demonstrates substantial improvements in operational safety, data reliability, and decision-making efficiency. Practical applications of digital twin technology are also discussed, highlighting their potential for real-time monitoring, predictive analytics, emergency response planning, and lifecycle management of transportation infrastructure, all systematically guided by the structured processes outlined in the modified IDEF modeling framework.

# Chapter 5 | Case Study: School Zone Safety Using Connected Vehicle Data

## 5.1. Introduction

Chapter 5 presents a case study focused on school zone safety by leveraging connected vehicle data to identify high-risk driving behaviors near educational institutions. This chapter illustrates how event-based data, specifically harsh braking and acceleration incidents, can be used to understand traffic conditions, assess risk patterns, and inform proactive safety improvements in alignment with Vision Zero goals. Through a geospatial and data-driven methodology, the study applies clustering and hotspot analysis techniques to uncover spatial trends and safety-critical zones. The results are used to generate actionable insights for improving infrastructure and traffic operations around school zones.

The chapter begins by introducing the relevance of connected vehicle data for school zone safety (Section 5.1) and describes the selected study area and datasets used (Section 5.2). It then details the process for detecting harsh braking and acceleration events (Section 5.3), followed by the application of the DBSCAN clustering algorithm to identify spatial clusters of these events (Section 5.4). Section 5.5 outlines the hotspot analysis methodology, which enhances spatial understanding of where the most critical events occur. Section 5.6 demonstrates the combined approach in action, showcasing key findings. Based

on the analysis, Section 5.7 provides targeted safety recommendations and discusses the broader implications for Vision Zero initiatives. The chapter concludes with a summary of key insights and contributions (Section 5.8).

## 5.2. Study Area and Data Description

The case study presented in this chapter focuses on a school zone in Northern Virginia, strategically situated at the midway point between two of the region's most significant transportation hubs: Reagan National Airport and Dulles International Airport. This location is not only representative of a typical high-traffic urban-suburban interface but also presents a unique mix of transportation dynamics influenced by local commuter traffic, school operations, and transient flows associated with airport-bound travel. The integration of connected vehicle (CV) data in such a complex environment offers unprecedented opportunities to study safety conditions in real time, providing granular insights into risk patterns that traditional traffic monitoring systems often overlook.

To support this analysis, Table 4 presents a curated selection of connected vehicle data attributes that are particularly relevant to this project's objectives. These variables were extracted from the broader CV data architecture provided by the Virginia Department of Transportation (VDOT) and selected based on their utility in geospatial modeling, risk detection, and event classification. Each attribute supports a unique analytical lens through which behavioral and environmental factors impacting school zone safety can be evaluated. Collectively, these variables facilitate a robust framework for risk analysis, cluster detection, and decision support.

The spatial components latitude, longitude, and heading are essential to geolocating and interpreting vehicle behaviors within the school zone's defined boundaries. These parameters allow for the visualization of vehicle trajectories and the identification of high density conflict zones, particularly around key infrastructure such as school entrances, pedestrian crossings, and signalized intersections. In the context of Northern Virginia's dense and often congested roadways, accurate spatial positioning becomes even more critical for capturing the dynamic nature of school zone interactions.

Recent advancements in transportation safety research have increasingly emphasized the use of high-resolution, data-driven methodologies, with connected vehicle (CV) data emerging as a critical resource for evaluating driving behavior and system performance. This dissertation builds upon that momentum by utilizing multiple clustering analysis techniques and machine learning approaches to examine harsh braking and acceleration events. By analyzing objective sensor-based measurements collected from CV systems, this methodology enables spatial visualization of high-risk driving behavior and supports the identification of hazardous zones based on empirical evidence rather than subjective reporting.

These techniques facilitate the detection of meaningful geospatial patterns in vehicle dynamics, allowing researchers and practitioners to uncover areas where dangerous driving behavior is more likely to occur. The ability to map and interpret these patterns provides a valuable foundation for developing targeted safety strategies and interventions aimed at reducing crash risk and improving roadway performance.

To further support the analysis of driving behavior, Figure 6 presents a conceptual illustration of the two primary acceleration axes captured in the dataset. Longitudinal acceleration (AccelerationX) corresponds to forward and backward motion, such as rapid acceleration or harsh braking, while lateral acceleration (AccelerationY) captures side-toside movement, often associated with abrupt lane changes or sharp turns. Together, these variables offer a multidimensional view of vehicle behavior, contributing to a more comprehensive understanding of risk and system performance in real-world driving conditions.

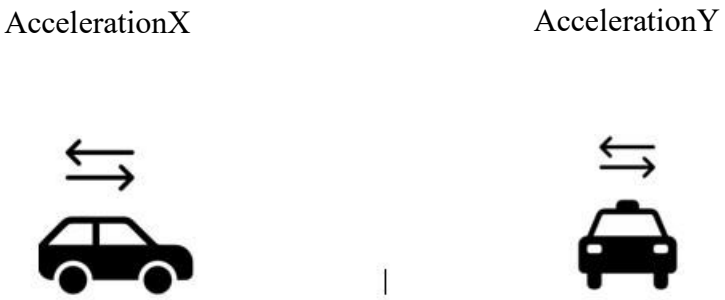


Figure 6. Illustration of AccelerationX (Forward/Backward) and AccelerationY (Lateral) movement in vehicle dynamics

The speed attribute is also critical in evaluating compliance with reduced school zone speed limits. Combined with acceleration patterns, speed can differentiate between typical deceleration due to traffic congestion and deliberate harsh braking in response to unexpected conditions such as pedestrian movement or erratic driving behavior. Moreover, captured date/time attributes enable fine-grained temporal segmentation of events, facilitating analysis across morning arrival, mid-day lull, and afternoon dismissal windows. This time-sensitive analysis is essential in school zone studies, where traffic conditions

fluctuate significantly over short time frames and are often misrepresented in aggregate data.

Additionally, environmental and driver state variables, such as exterior temperature, windshield wiper speed, seat belt status, and ABS activation, provide supplementary context for interpreting risky events. For example, poor weather conditions inferred from active wipers and low temperatures could increase the likelihood of brakingrelated incidents. Similarly, activation of anti-lock braking systems (ABS) serves as an automated proxy for abrupt deceleration on potentially slick or uneven surfaces an important factor when assessing safety near school entrances, where pavement conditions and stop-and-go behavior are more pronounced.

The journey ID and ignition state are used to filter out non-movement data and analyze complete trip patterns that pass through or originate near the school zone. This is particularly helpful for longitudinal studies aimed at tracking patterns in recurring behaviors, such as habitual speeding or frequent harsh braking by the same vehicle across multiple days. These identifiers also support the segmentation of trip stages, such as entry into and exit from the zone, allowing for localized analysis of behavior change in response to posted signage or traffic calming infrastructure.

Finally, turn signal and seat occupancy data add interpretability to vehicle behavior patterns, especially in the context of parent drop-offs, school bus movements, and pedestrian conflict areas. For example, inconsistent use of turn signals at critical access points may suggest confusion or poor visibility, while high seat occupancy may infer

student transport trips, aligning with the overall objectives of school zone safety evaluations.

Taken together, the data elements outlined in Table 4 provide the methodological foundation for the spatial analysis, machine learning, and hotspot detection techniques discussed in this chapter. They enable a multi-dimensional exploration of school zone operations, blending vehicle behavior, environmental context, and spatial distribution of risk into a cohesive analytical approach. This level of granularity is crucial for proactive transportation planning, particularly in vulnerable contexts such as school zones where the margin for error is small and the consequences of unsafe behavior are severe.

In summary, Table 4 serves as more than a list of connected vehicle attributes it is the scaffolding upon which the risk assessment framework is constructed. By leveraging this diverse and high-resolution dataset, the case study is able to identify, interpret, and address safety risks in a complex transportation environment where children, caregivers, and commuters converge. The use of CV data in this Northern Virginia school zone exemplifies the future of data-driven transportation safety management, showcasing the value of integrated digital systems in enhancing public safety and operational resilience.

**Table 5.** Relevant connected vehicle data attributes for transportation risk analysis

| Attribute Name | Description | Source | Application in Dissertation |
|---|---|---|---|
| Longitude | Geographic coordinate for vehicle location (X-axis) | Movement & Event Data | Spatial mapping of vehicle location. |
| Latitude | Geographic coordinate for vehicle location (Y-axis) | Movement & Event Data | Supports hotspot detection and spatial clustering. |

| | | | |
|---|---|---|---|
| Speed | Vehicle speed at time of record (mph) | Movement Data | Detection of risky behaviors (e.g., speeding, sudden deceleration). |
| Heading | Direction of vehicle movement in degrees (0–360) | Movement Data | Analyzing trajectory and turning behavior. |
| AccelerationX | Longitudinal acceleration, negative for braking, positive for acceleration | Event Data | Used to identify harsh braking and acceleration events. |
| AccelerationY | Lateral acceleration, useful for detecting sharp turns or swerving | Event Data | Detection of lateral risk behaviors (lane changes, turning risk). |
| Captured Date/Time | Timestamp of the observation | Movement & Event Data | Enables temporal analysis and time-of-day risk patterns. |
| Journey ID | Unique identifier for each vehicle journey | Event Data | Supports analysis of vehicle behavior across trip segments. |
| Ignition State | Indicates whether the vehicle is turned on or off | Event Data | Filters out stationary vehicles; ensures active driving observations. |
| Exterior Temperature | Measured ambient temperature from vehicle sensors | Event Data | Useful for correlating weather conditions with driving behavior. |
| Seat Belt Status | Indicates whether driver seat belt is engaged | Event Data | Contributes to safety compliance analysis. |
| ABS Activation | Indicates if anti-lock braking system was triggered | Event Data | Correlation with harsh braking and loss of traction. |
| Windshield Wiper Speed | Reflects whether the vehicle is operating under precipitation | Event Data | Environmental condition inference (e.g., rain during events). |
| Turn Signal | Indicates if a turn signal was active during the event | Event Data | Important for determining intentional vs. unintentional lane maneuvers. |
| Seat Occupancy | Detects number of seat positions occupied | Event Data | Contextualizes risk depending on vehicle load. |

By analyzing key vehicle dynamics such as acceleration, this study offers a detailed view into patterns of risky driving behavior within the vicinity of a designated school zone. These patterns support the identification of localized safety concerns and inform the development of targeted interventions aimed at reducing risk in these sensitive areas. Figure 6 displays a histogram of longitudinal acceleration (AccelerationX) values recorded within a 1.609 km (1-mile) radius surrounding the selected school zone. The data reveal a

distribution that skews slightly toward deceleration, with a mean value near -0.70 m/s². The median of -0.16 m/s² suggests that many of the recorded events involve minor braking or low-level acceleration, while a standard deviation of 2.23 indicates a broad range of values, encompassing both aggressive acceleration and severe braking.

Two notable peaks emerge in the histogram. One, centered around -3.9 m/s², reflects a concentration of harsh braking events; another, near 3.7 m/s², corresponds to episodes of rapid acceleration. These extreme values may be indicative of sudden driver reactions, possibly triggered by unexpected obstacles, congested traffic flow, or poor visibility and signage. Previous research has associated such behaviors with aggressive or inattentive driving, driver misjudgment, or infrastructural challenges that may contribute to abrupt vehicle maneuvers (33). These results underscore the importance of investigating the root causes of such behaviors in school zones, where vulnerable road users like children and pedestrians are present. Interventions guided by this type of analysis whether infrastructure redesign, signage enhancement, or speed regulation can help mitigate these risks and improve overall roadway safety.

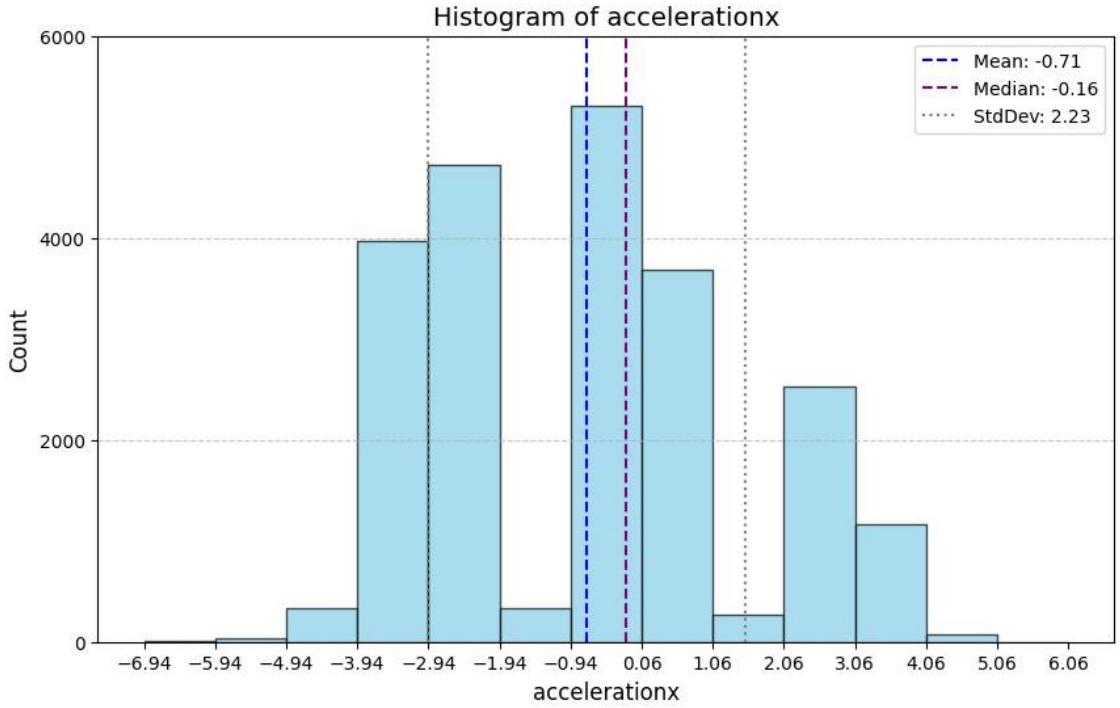


**Figure 7.** Histogram of AccelerationX displays the distribution of acceleration values in the dataset, highlighting the count of events at different acceleration levels with the mean, median, and standard deviation

## 5.3. Detection of Harsh Braking and Acceleration Events

The identification of harsh braking and acceleration events is a critical component of modern transportation safety analysis, particularly in sensitive areas such as school zones. These events serve as surrogate indicators of elevated crash risk, reflecting abrupt driver responses to traffic conflicts, pedestrian activity, or insufficient infrastructure design. In this study, connected vehicle (CV) data is leveraged to detect such behaviors in real time and with a level of detail that far exceeds the capabilities of traditional traffic monitoring systems.

Conventional data collection methods such as loop detectors, fixed-location cameras, or manual traffic counts are inherently limited in spatial and temporal resolution. These systems typically record aggregated traffic volumes or average speeds over fixed intervals, offering only a high-level perspective of roadway conditions. Moreover, they are constrained to specific locations and often fail to capture the nuanced behaviors that occur between monitoring points. As a result, traditional data sources may obscure the variability in driving behavior, particularly during short-duration, high-risk events such as sudden braking or aggressive acceleration.

In contrast, CV data offers a highly granular and continuous stream of vehicle performance information, capturing attributes such as longitudinal and lateral acceleration, speed, heading, and position at sub-second intervals. This level of precision allows for the identification of individual driving maneuvers and the exact locations where they occur. For the purposes of this case study, harsh events were defined using thresholds derived from acceleration in the X-axis (AccelerationX), where significantly negative values represent rapid braking and high positive values reflect sudden acceleration. These thresholds were applied to a dataset of Basic Safety Messages (BSMs) collected within a one-mile radius of a Northern Virginia school zone.

Unlike crash data, which is often retrospective and reliant on police reports or selfreporting, CV data allows for proactive safety analysis. It enables the detection of conflictprone areas before crashes occur, offering transportation agencies a powerful tool for early intervention. By applying spatial clustering techniques such as DBSCAN or other densitybased algorithms clusters of harsh braking and acceleration events were identified in

close proximity to school entrances, intersections, and mid-block crossings. These clusters represent zones of elevated driving volatility and potential danger to pedestrians, particularly during school arrival and dismissal periods when traffic volumes and behavioral unpredictability are highest.

The integration of harsh event detection with geospatial analysis provides a more complete understanding of safety conditions than has been traditionally possible. Not only can transportation professionals identify specific locations where risky behavior is concentrated, but they can also evaluate the likely contributing factors, such as signal timing, signage placement, or crosswalk visibility. This data-driven approach supports the development of targeted safety countermeasures, including traffic calming infrastructure, improved signage, or enhanced crossing treatments designed specifically for the school environment.

In summary, the use of CV data in detecting harsh braking and acceleration events offers a transformative advantage over traditional traffic monitoring methods. It enables a continuous, high-resolution, and behaviorally informed view of roadway risk, supporting proactive safety planning and helping to protect vulnerable users in school zones. This methodology sets the foundation for further predictive modeling and intervention prioritization in the chapters that follow.

## 5.4. DBSCAN-Based Clustering Methodology

To identify spatial patterns of harsh vehicle maneuvers, we employed the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, a widely accepted un-supervised learning method well-suited for discovering clusters of varying shapes and densities, particularly in spatial datasets. The algorithm was applied to filtered connected vehicle datasets that recorded acceleration data along with corresponding latitude and longitude coordinates.

### *Data Preparation and Filtering*

The input dataset consisted of geolocated vehicle events, including time-stamped acceleration values on the x-axis (longitudinal acceleration), recorded in meters per second squared ($m/s^2$). Two subsets of the dataset were created based on thresholds designed to identify critical safety events:

- **Harsh Braking Events (HBE):** Events where longitudinal acceleration $a_x \leq -4$ $m/s^2$, reflecting significant deceleration.
- **Harsh Acceleration Events (HAE):** Events where longitudinal acceleration $a_x \geq 4$ $m/s^2$, indicating rapid forward motion.

These threshold values were selected based on established safety literature and prior empirical studies of connected vehicle telemetry, which suggest that such accelerations are indicative of potentially dangerous driving behaviors.

### *DBSCAN Model Application*

DBSCAN was chosen for its ability to classify spatially dense regions without requiring a predefined number of clusters. The algorithm groups closely packed points together while labeling isolated points as noise.

Let $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ be the set of input data points, each represented as a vector $x_i = [\text{longitude}_i, \text{latitude}_i]$. DBSCAN uses two key parameters:

- $\varepsilon$: The maximum distance between two points for them to be considered neighbors.
- MinPts: The minimum number of points required to form a dense region.

The $\varepsilon$-neighborhood of each point $x_i$ is defined as:

$$N_\varepsilon(x_i) = \{x_j \in \mathbf{X} \mid \text{dist}(x_i, x_j) \leq \varepsilon\} \quad (1)$$

where the distance function is typically Euclidean:

$$\text{dist}(x_i, x_j) = \sqrt{(x_i^{\text{lat}} - x_j^{\text{lat}})^2 + (x_i^{\text{lon}} - x_j^{\text{lon}})^2} \quad (2)$$

A point $x_i$ is classified as:

- A *core point* if $|N_\varepsilon(x_i)| \geq \text{MinPts}$
- A *border point* if it is reachable from a core point but has fewer than MinPts neighbors
- *Noise* if it does not meet either condition

In this study, the DBSCAN parameters were tuned based on domain knowledge and visual inspection of spatial distributions. Specifically, $\varepsilon$ was set to 0.0002 degrees (approximately 20 meters), and MinPts was set to 10 to avoid overfitting to transient noise or isolated points.

Output and Post-Processing The DBSCAN model's output was a labeled dataset in which each data point was assigned a cluster ID or marked as noise (-1). These cluster assignments were appended to the original dataset for further spatial analysis and visualization.

Let $C_i$ denote the set of points in cluster $i$, and $\hat{y}_i$ be the predicted cluster label. The final output dataset can be described as:

$$\mathbf{Y} = \{(x_i, \hat{y}_i) \mid x_i \in \mathbf{X}\} \quad (3)$$

Points labeled $\hat{y}i = -1$ were excluded from further cluster-based spatial hotspot analyses, as they represent non-dense or isolated events. The processed data were exported in CSV format, preserving the geographic coordinates and cluster labels for integration into

ArcGIS Pro for mapping and visual pattern detection.

## 5.5. Hotspot Analysis Methodology

Transportation improvements differ from policy or procedural changes in that they are inherently spatially rooted in physical location. This spatial characteristic provides a unique opportunity to examine system resilience through geospatial analytical techniques. One such technique is spatial autocorrelation, which assesses whether patterns of a variable, such as risky driving behaviors, occur in clusters or are randomly distributed across a geographic area.

To analyze where concentrations of harsh braking and acceleration events occur, the Getis-Ord Gi* statistic is commonly used. This spatial statistical method evaluates the intensity of attribute values within a spatial context to identify statistically significant clusters, known as hot spots (areas with high values) and cold spots (areas with low values). In this study, attribute values represent the frequency or severity of harsh vehicle maneuvers, and locations are defined by their associated GPS coordinates.

The Gi* statistic is expressed mathematically as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \quad (4)$$

Here, $x_j$ refers to the attribute value (count of harsh braking or acceleration events) for location *j, wi, j,* is the spatial weight between features *i* and *j*, and *n* is the total number of spatial features (all observed events). The terms *X* and *S* represent the global mean and standard deviation of the attribute values, respectively, calculated as:

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (5)$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2} \quad (6)$$

This methodology allows for the detection of spatial clusters in the harsh event data. A significant hot spot indicates an area where a high frequency of harsh braking or

acceleration occurs in proximity to other high-frequency locations. Conversely, cold spots reflect areas with consistently low event intensity, while locations without statistical significance suggest spatial randomness.

The use of Gi* analysis in this context can highlight zones of elevated driving risk, where harsh maneuvering behaviors are geographically concentrated. Identifying these zones can inform targeted interventions, such as traffic calming measures or infrastructure upgrades.

It is important to note that Gi* analysis assumes a minimum number of spatial observations (generally 30 or more) to ensure statistical robustness. Additionally, meaningful spatial relationships depend on the assumption that nearby features are more likely to share similar characteristics in this case, aggressive driving behaviors. When such assumptions are met, the method can be used not only to detect clusters, but also to guide investment decisions by highlighting areas in need of safety improvements.

In broader research, similar spatial techniques have been successfully applied to detect clustering in crash data, crime incidents, and other urban phenomena. In this study, these principles are extended to connected vehicle data, emphasizing geospatial patterns in driving behavior as a proxy for roadway risk.

By interpreting spatial association or disassociation within this dataset, planners and policymakers can make more resilient decisions. If harsh driving behaviors are geographically concentrated, then interventions can be localized for greater effectiveness. If, on the other hand, such behaviors are widespread and unclustered, a broader systemic

approach may be necessary. Either outcome contributes valuable insight into how driver behavior and infrastructure interact spatially.

## 5.6. Demonstration of Approach

Figure 8 illustrates the spatial distribution and clustering outcomes of harsh acceleration events (HAE) based on the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. The visualized data represents connected vehicle events mapped across the School area in Fairfax County, Virginia, and surrounding arterials. Each point in the figure corresponds to an individual event characterized by a longitudinal acceleration value of at least 4 $m/s^2$, which meets the threshold for harsh acceleration defined in this study.

The symbology in Figure 7 reflects the DBSCAN-assigned cluster designations. Events marked in green correspond to cluster = -1, representing noise or outliers according to DBSCAN's classification criteria. These are points that do not belong to any highdensity cluster and typically exist in spatial isolation or in areas with low event density. In contrast, events highlighted in orange correspond to cluster = 0, which DBSCAN identified as a distinct spatial cluster of harsh acceleration behaviors. The presence of this cluster along Lake Braddock Drive adjacent to the school campus and key intersections suggests a repeated pattern of aggressive acceleration maneuvers that may be tied to congestion, poor

traffic signal timing, or driver behavior influenced by school-related activity such as pick-up/drop-off operations.

This visualization confirms the applicability of DBSCAN in differentiating between significant spatial concentrations of harsh events and more isolated occurrences.

The presence of a well-defined cluster (cluster 0) supports the notion that certain roadway segments demonstrate consistent high-risk driving behaviors. Such insights are foundational to transportation resilience planning and can inform targeted interventions such as speed management strategies, signage improvements, or signal adjustments. Furthermore, the use of ArcGIS for spatial representation provides an intuitive and interpretable medium for communicating complex data mining outputs to planning agencies and stakeholders.

The results of this analysis not only demonstrate the feasibility of the proposed clustering approach but also underscore the importance of geospatial methods in understanding behavior-based risk patterns in connected vehicle environments.

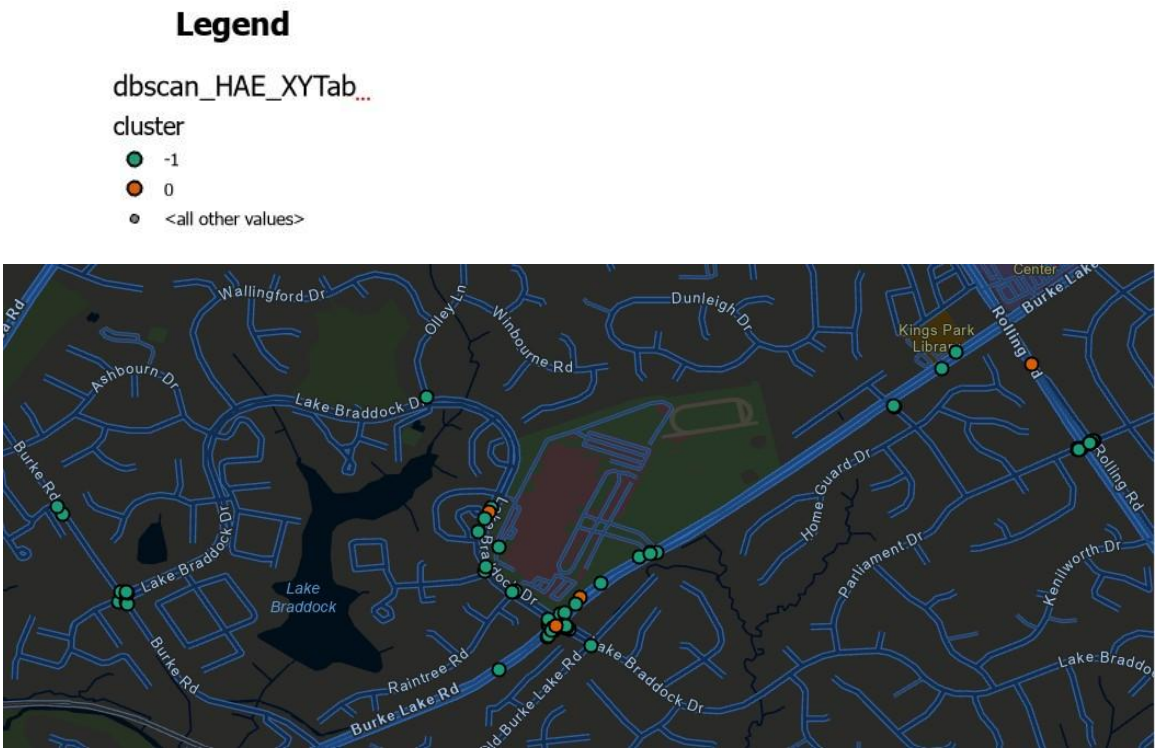


Figure 8. : DBSCAN of Harsh Acceleration Events in a Northern Virginia school zone.

Figure 9 provides a detailed spatial representation of clustering results for harsh braking events (HBE) within the same study area as previously shown centered around Lake Braddock Secondary School in Fairfax County, Virginia. These events were extracted from connected vehicle data based on the threshold $a_x \leq -4\,\mathrm{m/s^2}$ , signifying abrupt deceleration behaviors that may reflect driver overreactions, signal changes, or inadequate roadway design.

The figure displays the results of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, applied to the filtered harsh braking dataset. The resulting cluster labels range from -1 (noise) to 12, each depicted with a distinct color. Unlike Figure 7, which only yielded a single significant cluster for harsh acceleration,

the DBSCAN analysis of harsh braking events produced a much more complex spatial distribution, revealing thirteen distinct cluster groups as well as noise points.

Clusters are prominently concentrated along Lake Braddock Drive, Burke Lake Road, and Rolling Road particularly at key intersections and entrance/exit points around the school. Notably, clusters labeled 3, 4, 6, and 10 appear along corridors where vehicular congestion and school-related activities such as drop-offs or traffic signal transitions likely contribute to abrupt stopping patterns. The presence of numerous smaller clusters suggests that harsh braking events are not isolated to a single risk zone, but rather dispersed throughout the broader school zone area.

The -1 cluster (depicted in dark cyan) again represents noise, capturing spatially isolated braking events that were not part of any dense grouping. However, these may still be relevant when considered individually, particularly if they occur near pedestrian zones or unsignalized crossings.

The spatial variability in clustering outcomes for HBE, as shown in Figure 8, demonstrates the utility of DBSCAN in capturing both macro-level and micro-level patterns in driving behavior. The ability to detect multiple distinct zones of high braking activity provides transportation planners and engineers with actionable insights into areas where traffic calming interventions, updated signage, or signal timing adjustments may be needed.

Furthermore, this figure highlights how localized behavior patterns can vary between braking and acceleration events, reinforcing the need for tailored countermeasures for each behavior type. The map-based visualization serves not only as an analytical output

but also as a communication tool to engage stakeholders, support risk assessment, and prioritize infrastructure investments that address aggressive or unsafe driving behavior near sensitive land uses like schools.

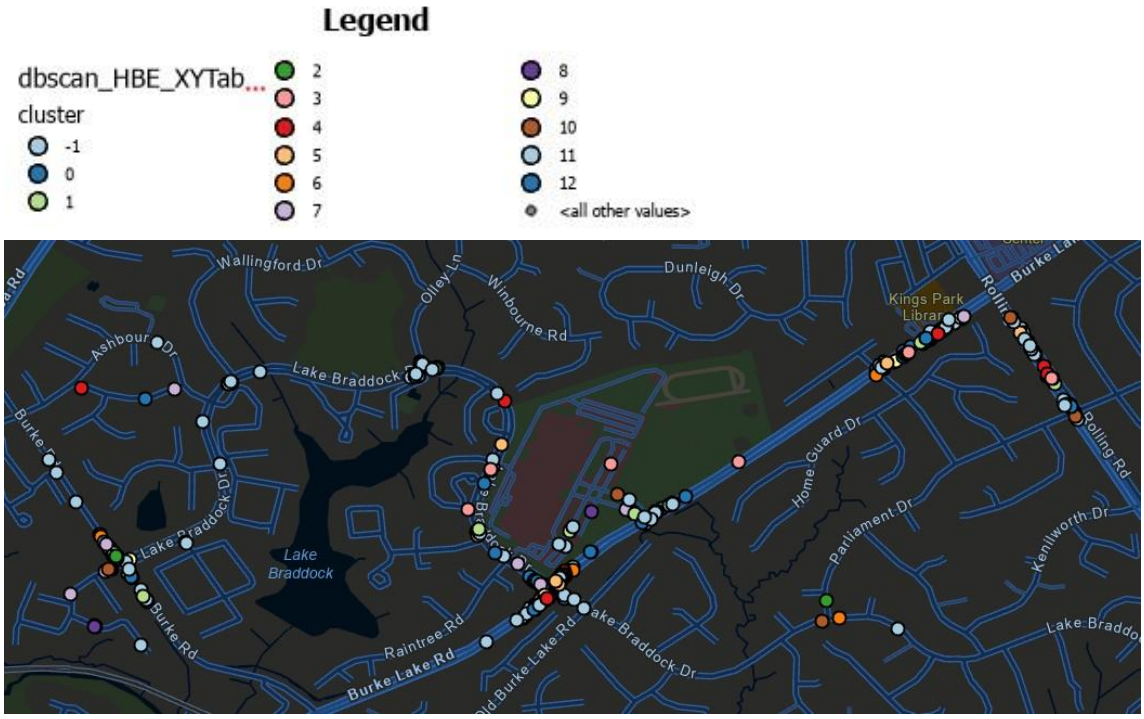

Figure 9. DBSCAN of Harsh Braking events in a northern Virginia school zone.

Figure 9 presents the results of a Getis-Ord Gi* hot spot analysis applied to connected vehicle events recorded near a Northern Virginia school zone. This method was used to identify statistically significant clusters of harsh braking and acceleration behaviors that could pose elevated safety risks to both motorists and pedestrians, particularly in a school-centric environment. The spatial distribution of events shown in the figure represents the combined outcome of previous preprocessing and DBSCAN filtering steps,

which ensured only qualifying events (e.g., ax≤−4 m/s2a_x \leq -4 \, \text{m/s}^2ax ≤−4m/s2 for braking and ax≥4 m/s2a_x \geq 4 \, \text{m/s}^2ax≥4m/s2 for acceleration) were included.

Each point in the figure is symbolized based on its assigned Gi_Bin value, representing confidence levels for statistical significance:

- **Red (Bin = 3)**: 99% confidence hot spot
- **Orange (Bin = 2)**: 95% confidence hot spot
- **Yellow (Bin = 1)**: 90% confidence hot spot
- **Green (Bin = 0)**: Not statistically significant
- **Light to dark blue (Bins = -1 to -3)**: Cold spots at varying levels of confidence • **Purple (Bin = -3)**: 99% confidence cold spot

The results show a prominent concentration of hot spots along Burke Lake Road, Lake Braddock Drive, and Rolling Road, which are key arterials providing access to Lake Braddock Secondary School. These areas correspond closely to previously identified DBSCAN clusters and provide additional inferential evidence that aggressive driving behaviors are not randomly distributed but spatially associated with specific road segments and intersections.

The aggregation of high Gi_Bin values around school access points such as the school driveway, adjacent traffic signals, and pedestrian crosswalks points to operational concerns that may be influenced by high vehicle volumes during student arrival and dismissal times. The presence of cold spots along secondary neighborhood roads suggests that traffic calming or lower speed limits in residential areas may be contributing to safer driving behaviors in those zones.

This analysis underscores the value of spatial inferential methods such as the Gi* statistic in transportation safety analytics. Unlike clustering methods that group based solely on proximity and density, the Gi* hot spot method incorporates both attribute intensity and spatial relationships, allowing for a statistically robust identification of critical safety zones. These results can directly inform targeted engineering and policy responses such as signal re-timing, crossing guard placement, signage improvements, or reconfiguration of pick-up/drop-off zones.

Incorporating spatial autocorrelation metrics into connected vehicle analysis enhances situational awareness for planners and helps ensure that resource allocation is both data-driven and equity-conscious. As seen in Figure 9, visualizing statistically significant concentrations of harsh events offers a practical framework for prioritizing interventions in high-risk school environments.

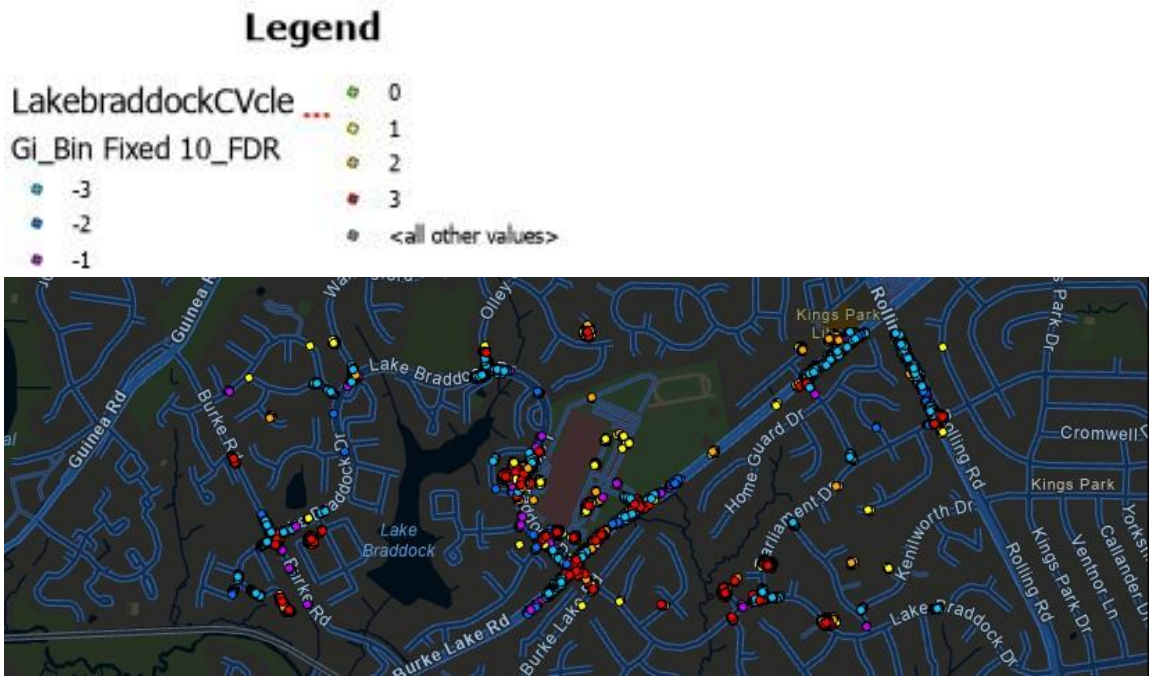

Figure 10. Hot spot analysis of Northern Virginia school zone.

## 5.7. Safety Recommendations and Vision Zero Implications

The results from the spatial clustering and hotspot analysis conducted in Figures 7, 8, and 9 highlight critical areas of concern related to driver behavior near school zones. Both the DBSCAN and Getis-Ord Gi* methodologies revealed consistent geographic patterns of harsh acceleration and braking events, with significant clustering occurring along high-volume corridors such as Lake Braddock Drive, Rolling Road, and Burke Lake Road. These findings have direct implications for advancing Vision Zero goals, which aim to eliminate traffic-related fatalities and serious injuries through data-driven, system-level interventions.

The DBSCAN clustering of harsh acceleration events (Figure 7) indicated a concentration of rapid vehicle movements near school access points, suggesting that drivers may be speeding up aggressively after delays caused by congestion or signal timing. These behaviors increase the likelihood of collisions, particularly in environments with high pedestrian activity such as student drop-off and pick-up zones. Conversely, the clustering of harsh braking events (Figure 8) was more spatially dispersed but also exhibited clear concentrations along the same corridors, pointing to a combination of unpredictable traffic patterns, limited sight distances, or poorly timed signals that compel drivers to stop abruptly.

The hot spot analysis (Figure 9), which integrates both attribute intensity and spatial association, confirmed that several of these clustered areas are statistically significant at the 90%, 95%, and 99% confidence levels. This further supports the assertion that

aggressive or erratic driving behavior is not random, but rather geographically linked to infrastructural or operational deficiencies.

Based on these findings, the following safety recommendations are proposed:

(1) Traffic Calming Measures: Implement speed tables, raised crosswalks, and bulbouts along key corridors where clusters of harsh acceleration and braking events overlap. These treatments are known to reduce vehicle speeds and improve pedestrian safety in school zones.

(2) Signal Timing Optimization: Adjust signal phase and timing plans, particularly along Burke Lake Road and Rolling Road, to reduce start-stop conditions and mitigate the need for sudden acceleration or braking.

(3) School Zone Redesign: Reconfigure access points to the school to reduce conflict points. Consider implementing a dedicated student pick-up/drop-off loop that minimizes the mixing of through-traffic with school-related traffic.

(4) Enhanced Signage and Enforcement: Deploy flashing school zone signage with dynamic speed displays and expand automated enforcement (e.g., speed and redlight cameras) in areas identified as high-risk through hotspot analysis.

(5) Public Awareness Campaigns: Develop targeted driver education campaigns aimed at modifying driver behavior near schools. Emphasize the risk associated with aggressive driving and promote awareness of pedestrian presence during peak hours.

The integration of connected vehicle data with geospatial analytics offers a scalable and proactive approach to traffic safety that aligns with Vision Zero principles. By identifying specific locations where driver behavior increases risk, agencies can move beyond reactive crash-based analysis and toward preventative design and operational

strategies. Moreover, these insights support equitable investment by directing resources to infrastructure that directly impacts vulnerable road users, such as children walking or biking to school.

This case study demonstrates how transportation agencies can leverage real-time vehicle data and spatial statistics to better understand behavioral risk patterns and develop localized, evidence-based interventions. In doing so, they take meaningful steps toward realizing a transportation system where safety is prioritized for all users.

## 5.8. Summary

Chapter 5 presents a comprehensive case study focused on enhancing school zone safety through the application of connected vehicle (CV) data analytics. The chapter demonstrates how high-resolution, sensor-based vehicle data can be used to detect and map risky driving behaviors specifically harsh braking and acceleration within the vicinity of a Northern Virginia school zone. By leveraging data attributes such as longitudinal acceleration, speed, heading, and environmental conditions, the study builds a multidimensional understanding of vehicle behavior under real-world traffic conditions. Using clustering methods like DBSCAN, the analysis identifies concentrated zones of aggressive maneuvers, while Getis-Ord Gi* hot spot analysis validates the statistical significance of these clusters. Together, these methods provide an evidence-based framework for detecting spatial patterns of concern and assessing the role of contextual variables such as congestion, weather, and roadway design.

The findings from this chapter offer critical insights for proactive safety interventions in school environments. Consistent clustering of harsh events along arterial roads and near

school entrances underscores the need for targeted engineering, enforcement, and educational strategies. Recommendations such as traffic calming measures, signal timing optimization, and enhanced signage are presented in alignment with Vision Zero goals to eliminate fatalities and serious injuries. This case study exemplifies how connected vehicle data, when combined with spatial analytics, can support a shift from reactive crash-based planning to preventive, data-driven decision-making. As such, Chapter 5 not only advances the methodological contributions of this dissertation but also offers practical implications for agencies seeking to protect vulnerable road users particularly children in complex, high-conflict urban-suburban environments.

# Chapter 6 | Case Study: Environmental Risk in Agricultural Systems

## 6.1. Introduction

Chapter 6 explores environmental risk within agricultural systems, with a particular focus on the role of extreme temperature events in shaping vulnerability and resilience. The case study presented in this chapter integrates geospatial analysis, statistical interpolation, and machine learning techniques to develop a comprehensive risk assessment framework. This chapter supports broader goals of enhancing food security and environmental systems under changing climate conditions by identifying heat-based environmental stressors and modeling their potential impact on agricultural productivity.

The chapter begins by introducing the motivation for assessing agricultural risk in the context of environmental change (Section 6.1) and providing an overview of observed temperature trends and their implications for crop viability and rural economies (Section 6.2). Section 6.3 outlines the development of a heat-based risk index, while Section 6.4 compares spatial interpolation methods Inverse Distance Weighting (IDW) and Kriging to model temperature variations across agricultural regions. Section 6.5 applies machine learning models, including XGBoost, Random Forest (RF), and Support Vector Regression (SVR), to predict risk levels based on spatial and climatic variables. Section 6.6 focuses on

the evaluation of these predictive models and the visualization of risk layers using GIS. The chapter concludes with a discussion on the implications of the findings for improving agricultural resilience and informing food security strategies in the face of environmental stress (Section 6.7).

## 6.2. Overview of Agricultural Risk and Temperature Trends

Global agriculture is increasingly vulnerable to the intensifying impacts of extreme heat, driven by persistent shifts in climate patterns and record-breaking temperature events. The rise in global mean temperatures has led to more frequent and prolonged heatwaves that directly affect crop growth, soil moisture, and overall farm productivity. In temperate climates, optimal plant development typically occurs between 20°C and 30°C, with nighttime temperatures playing a critical role in maintaining plant immunity. When temperatures exceed these optimal ranges, crops experience physiological stress, reduced yields, and heightened susceptibility to disease. For example, mango cultivars require a safe range of 10°C to 12°C to avoid irreversible damage. The heatwave of July 2023, which recorded the hottest days in Earth's history, highlights a growing trend that threatens not only agricultural productivity but also the resilience of food systems around the world.

Agricultural risk is further amplified by variations in regional climate conditions and population exposure. Densely populated agricultural zones face greater challenges in mitigating and responding to extreme temperatures, as more people and infrastructure are affected. To address these risks, this study employs a heat-based risk index that integrates

temperature thresholds and population density, enabling targeted spatial analysis of vulnerability. With climate volatility expected to increase, traditional farming strategies are no longer sufficient. Advanced tools such as Geographic Information Systems (GIS), supported by spatial interpolation and machine learning, offer the precision needed to identify and respond to risk hotspots. These innovations provide essential support for policymakers, farmers, and planners seeking to adapt agricultural practices to the realities of a warming world.

This begins the development of the Heat-Based Risk Index. To construct the heat-based risk index, population density was first calculated using the Calculate Field tool in ArcGIS. This involved dividing population values by the area of a circle with a 100-kilometer radius, yielding an estimate of individuals per square kilometer for each location in the dataset. These population density figures were then integrated with temperature data using the expression: (Temperature – 30) × Population
Density. This index with Table 6 enabled the identification of areas where both high temperatures and high population exposure intersect, highlighting regions most vulnerable to heat-related impacts.

Following the creation of the risk index, geostatistical interpolation techniques namely Kriging and Inverse Distance Weighting (IDW) were applied to generate continuous surface layers representing spatial variations in heat-related risk across the study area. IDW offered a computationally efficient approach based on proximity weighting, while Kriging provided a more advanced method that incorporates spatial autocorrelation and quantifies prediction uncertainty. These resulting geospatial surfaces

offered valuable insight into regional risk patterns and enabled the identification of highpriority zones for mitigation.

These geostatistical outputs are instrumental for agricultural and climate resilience planning. By visualizing areas where extreme temperatures overlap with dense populations, decision-makers can better allocate resources, implement targeted adaptation strategies, and prioritize interventions. Including demographic exposure in the risk index also increases its relevance for public health and policy applications. Engaging stakeholders such as farmers, urban planners, and health officials in interpreting these spatial layers ensures the usability and relevance of the outputs. Ultimately, this integrated approach strengthens efforts to mitigate the impacts of extreme heat and improve resilience in vulnerable communities, particularly as heat continues to be a leading cause of climate related mortality.

**Table 6.** Field calculator for heat-based risk index

| | Lat | Long | Temp (C°) | Pop | Heat Risk | Pop Density |
|---|---|---|---|---|---|---|
| Count | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 | 15.00 |
| Mean | -0.62 | 11.57 | 29.67 | 61774.93 | -3.23 | 1.97 |
| Std | 1.57 | 1.66 | 3.13 | 145325.49 | 14.50 | 4.63 |
| Min | -3.42 | 8.75 | 23.00 | 1200.00 | -55.21 | 0.04 |
| 25% | -1.64 | 10.44 | 27.50 | 9318.50 | -1.20 | 0.30 |
| 50% | -0.72 | 11.48 | 31.00 | 20714.00 | 0.48 | 0.66 |
| 75% | 0.68 | 13.07 | 32.00 | 36559.00 | 1.62 | 1.16 |
| Max | 2.08 | 13.93 | 33.00 | 578156.00 | 4.10 | 18.40 |

## 6.3. Interpolation Techniques: IDW vs Kriging

### *Inverse Distance Weighing*

Inverse Distance Weighting (IDW) is a commonly used spatial interpolation method that estimates unknown values by leveraging the values of surrounding, known data points. The core assumption of IDW is that data points exert greater influence the closer they are to the location being estimated. In other words, nearby values carry more weight than those farther away. To generate predictions, the method calculates a weighted average of known values, assigning weights based on the inverse of their distances from the target location.

In the context of environmental risk modeling, IDW can be applied to create a continuous surface that visualizes spatial patterns of heat exposure. For instance, when using a heat-based risk index formulated as (Temperature - 30) multiplied by population density IDW can interpolate the values across a region to identify areas where populations are more vulnerable to elevated temperatures. While IDW is praised for its ease of use and computational speed, it does have limitations. The method assumes a consistent relationship between proximity and similarity, which may not hold in areas with sparse or uneven data distribution, potentially leading to distorted outputs. Nonetheless, its straightforward implementation and repeatability make it a valuable tool for preliminary spatial analyses and GIS-based assessments.

The value at the unknown location x0 is calculated as a weighted sum of the known values: Where $\hat{z}(x_0)$ is the estimated value are the known values at $x_0, z(x_i)$ locations and λi are the weights calculated from the distances.

$$\hat{z}(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i) \quad (7)$$

### *Kriging*

Kriging is a geostatistical interpolation method that enhances spatial analysis by accounting for both the distances between observed data points and the spatial relationships, or autocorrelation, among them. Unlike deterministic techniques such as Inverse Distance Weighting, Kriging not only estimates unknown values at unsampled locations but also provides a measure of the associated prediction uncertainty. This dual output prediction and confidence makes Kriging especially valuable for applications that require both spatial accuracy and statistical reliability.

The process begins by constructing a variogram, which characterizes how the similarity between data points diminishes over increasing spatial separation. This spatial model is then used to solve a set of equations that assign optimal weights to known data points, enabling the estimation of values at unknown locations. In the context of a heatbased risk index, Kriging generates a continuous surface that spatially represents varying levels of heat-related agricultural risk. This results in a risk map that captures thermal stress's intensity and distribution across a given region. Its strength lies in producing highly detailed environmental maps, particularly useful for variables such as temperature, allowing agricultural stakeholders to better anticipate areas of concern and implement proactive resilience strategies. However, it should be noted that Kriging demands more computational resources and specialized knowledge in spatial statistics due to the complexity of variogram modeling and its sensitivity to data quality.

The semivariogram is a key component of Kriging, describing the spatial autocorrelation of the data. It is defined as:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2 \quad (8)$$

Where γ(h) is the semivariance at distance h, N(h) is the number of pairs of observations separated by h, and z(xi) is the observed value at location xi. Once the semivariogram is modeled, Kriging uses it to estimate the value at an unknown location $x_0$ as a weighted sum of the known values:

$$\hat{z}(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i) \quad (9)$$

$$\sigma^2(x_0) = \gamma(0) - \sum_{i=1}^{N} \lambda_i \gamma(x_i - x_0) \quad (10)$$

### *Techniques for Layer Development*

Figure 10 presents the heat-based risk index visualized using the Kriging interpolation method, a probabilistic geostatistical approach that incorporates spatial autocorrelation to generate a smooth, continuous surface. This technique effectively captures spatial dependencies between known data points, allowing for the estimation of risk levels in areas without direct measurements. The use of an exponential variogram model supports the accurate representation of localized heat variations, making Kriging particularly suitable for identifying subtle gradients and patterns in regions prone to thermal stress.

In contrast, Figure 11 displays the same risk index derived through Inverse Distance Weighting (IDW), a deterministic technique that places greater emphasis on nearby data

points. This results in a less refined, more segmented surface when compared to Kriging. While IDW offers a faster and more straightforward approach, it may lack the detail necessary to reveal nuanced spatial trends. Nonetheless, both mapping techniques offer valuable perspectives: the Kriging output provides in-depth insights into microclimatic differences and their implications for agricultural productivity, while the IDW map clearly pinpoints zones that may require urgent intervention. Together, these visualizations support strategic decision-making in environmental risk management, aiding efforts to enhance working conditions and bolster the resilience of agricultural systems under increasing climate stress.

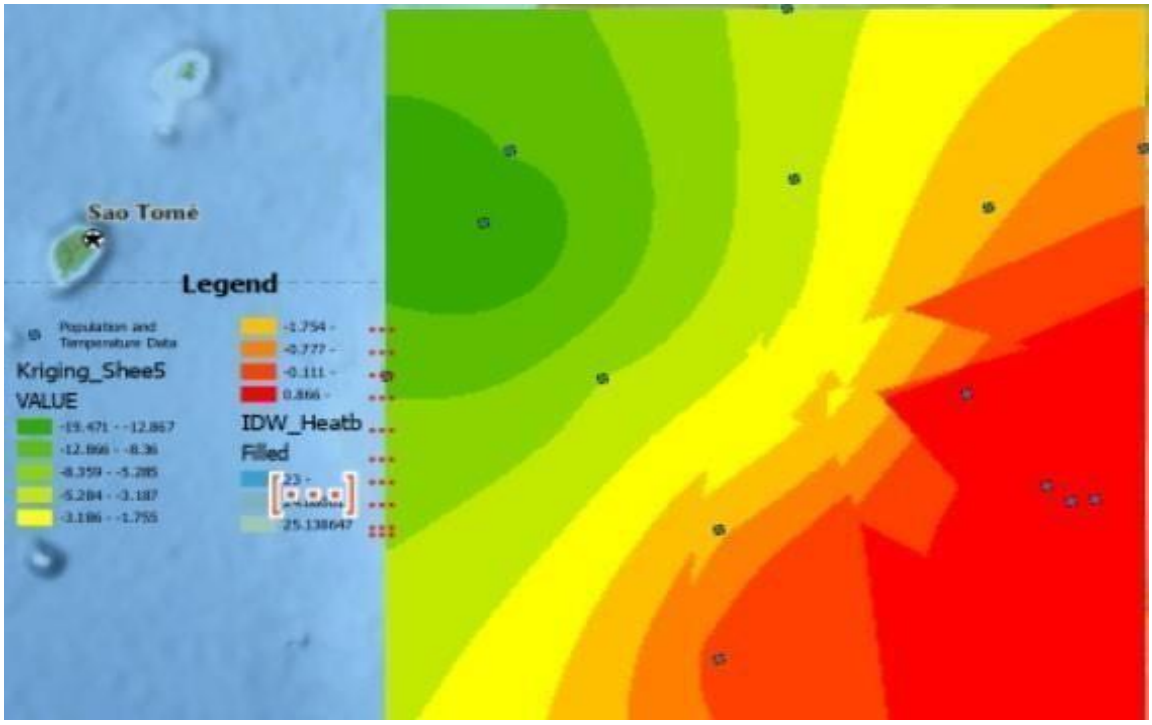


Figure 11. Illustrates a heat-based risk index map for multiple countries in West Africa, generated using Kriging, a geostatistical technique that utilizes spatial autocorrelation and an exponential variogram to provide a continuous and detailed representation of Crop Yield.

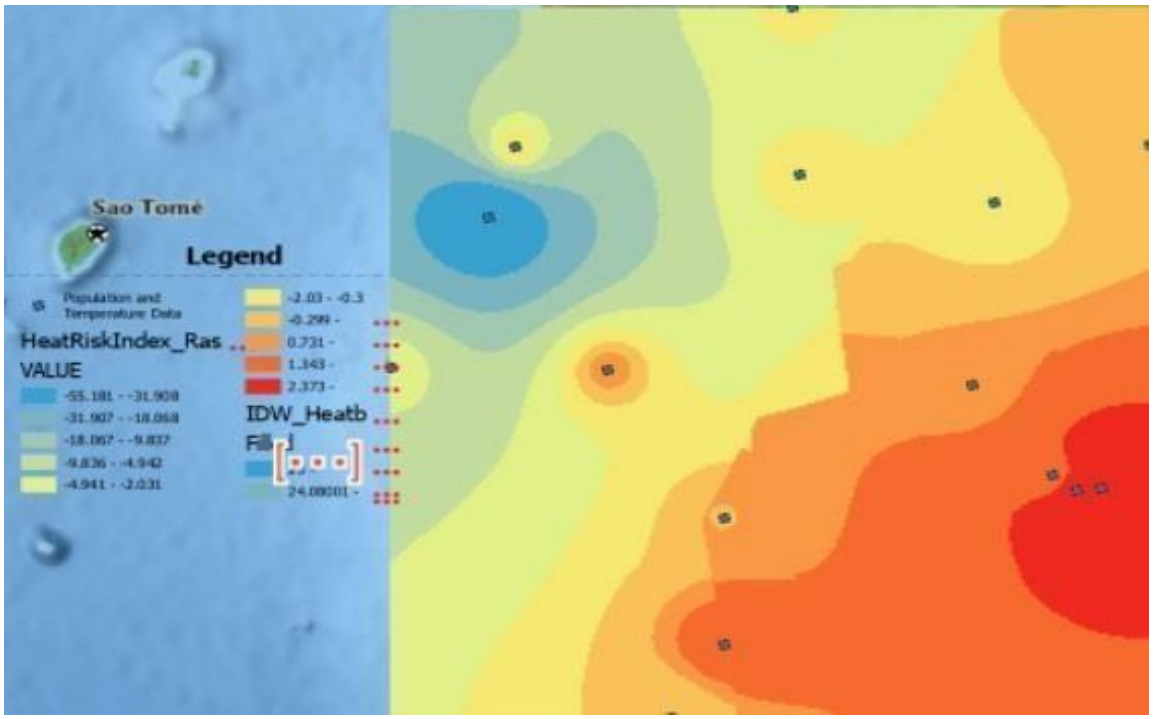


Figure 12. A heat risk index map for São Tomé, Africa, created using Inverse Distance Weighting (IDW), a deterministic method that emphasizes the influence of nearby data points, resulting in a less smooth and more segmented representation of heat risk variations

## 6.4. Evaluation and Visualization of Risk Layers

This section outlines a structured methodology for evaluating multiple regression models used to predict a heat-based risk index. As illustrated in Table 7, the workflow encompasses data preparation, model training, performance evaluation, and comparison of several machine learning algorithms. The models tested include Linear Regression, Random Forest Regression, Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost), each assessed using key performance metrics Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination ($R^2$). These metrics provide a comprehensive view of each model's predictive accuracy and reliability.

To enhance robustness and minimize overfitting, cross-validation techniques are incorporated, ensuring model performance is consistent across various data subsets. This entire modeling process is designed with automation in mind, enabling seamless execution of data cleaning, model fitting, and validation without manual intervention. Automating these steps not only improves efficiency and reduces the potential for human error, but also supports regular updates to the heat-based risk index as new data becomes available. Furthermore, the ability to systematically compare model outputs accelerates the selection of the most effective model, streamlining decision-making for heat risk assessment and management.

The comparative analysis of regression models revealed that Extreme Gradient Boosting (XGBoost) outperformed the other methods in predicting the heat-based risk index. XGBoost achieved the most favorable evaluation metrics, including the lowest Mean Squared Error (MSE) of 0.0516, Root Mean Squared Error (RMSE) of 0.2271, and Mean Absolute Error (MAE) of 0.0749. Additionally, it produced a positive $R^2$ score of 0.076, indicating a modest yet meaningful ability to explain variance within the dataset. In contrast, more traditional approaches such as Linear Regression and Support Vector Regression (SVR) underperformed, as evidenced by higher error values and negative $R^2$ scores, highlighting their limited suitability for this particular application.

Despite being more computationally intensive, Kriging also showed competitive performance, with an MSE of 0.053, RMSE of 0.230, and MAE of 0.090, further validating its value in geospatial modeling. The results are visually summarized in Figure 12, which illustrates the predicted spatial distribution of heat-related risk using the XGBoost model.

Its ability to capture localized variations with a high degree of accuracy underscores its potential for improving predictive capacity in environmental risk assessments. Implementing XGBoost in this context enhances both the precision and reliability of risk estimation, thereby informing more effective strategies for mitigating the impacts of extreme heat on agriculture and vulnerable populations.

**Table 7.** Evaluation of various machine learning models (Linear Regression, Random Forest Regression, Support Vector Regression, XGBoost, and Kriging).

| Model | MSE | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| Linear Regression | 6.4783 | 2.5452 | 0.6877 | -115.07 |
| Random Forest | 0.0586 | 0.2422 | 0.0955 | -0.0510 |
| Support Vector | 0.0672 | 0.2592 | 0.1468 | -0.2040 |
| XGBoost | 0.0515 | 0.2270 | 0.0749 | 0.0759 |
| Kriging | 0.0530 | 0.2300 | 0.0900 | 0.0500 |
| Inverse Density Weighting | N/A | 2.4196 | 2.4196 | N/A |

## 6.5. Machine Learning for Risk Prediction

The implemented code in Table 8 outlines a systematic approach for evaluating multiple regression models aimed at predicting a heat-based risk index. As detailed in Table 7, the workflow includes data preprocessing, model training, and performance assessment across a range of machine learning algorithms namely Linear Regression, Random Forest, Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost). The models are compared using standard evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination ($R^2$), which collectively offer insights into prediction accuracy and model robustness.

To enhance the reliability of results, cross-validation is applied to ensure each model is tested across multiple data partitions, reducing the risk of overfitting and increasing generalizability. The entire pipeline from data input to model evaluation can be automated, allowing for streamlined updates and reducing human error. This automation ensures that as new data becomes available, the risk index can be efficiently recalculated using the most accurate model. Ultimately, this framework accelerates model selection and supports data-driven decision-making for identifying and managing regions at risk from extreme heat.

A high level of automation can be achieved when implementing Kriging and Inverse Distance Weighting (IDW) techniques, streamlining the entire workflow from data preprocessing to final output generation. Tasks such as data cleaning, normalization, and transformation can be efficiently handled using Python libraries like pandas and scikitlearn. For Kriging, key geostatistical operations such as analyzing spatial autocorrelation and fitting variogram models can be automated using tools like pykrige, ensuring consistency and repeatability across datasets. Similarly, IDW computations can be scripted within ArcGIS using the arcpy library, allowing for uniform application of proximity-based weighting methods.

Once the preprocessing stage is completed, both Kriging and IDW can be executed programmatically with predefined parameters, removing the need for manual intervention. The automated workflow can also handle the generation of geospatial layers, including raster surfaces and risk maps. Additional scripting components can be implemented for error-checking, logging, and exception handling, thereby increasing the robustness and reliability of the process. Visualization tasks such as plotting spatial data and creating maps

can also be automated using Python libraries like matplotlib and seaborn, or integrated within GIS platforms like ArcGIS. These workflows can be scheduled to run periodically, ensuring that heat-based risk assessments remain current and reflective of incoming data.

**Table 8.** Pseudocode for model evaluation and risk index prediction

**Algorithm 1** Model Evaluation and Risk Index Prediction

```
Input: DataFrame df
Output: Evaluation metrics for different models and Risk Index Map

X ← df[['Latitude',' Longitude']]
y ← df['RiskIndex']

function EVALUATE_MODEL(model, model_name)
    predictions ← cross_val_predict(model, X, y, cv = len(X))
    mse ← mean_squared_error(y, predictions)
    rmse ← √mse
    mae ← mean_absolute_error(y, predictions)
    r2 ← r2_score(y, predictions)
    print("model_name")
    print("MSE : {mse}")
    print("RMSE : {rmse}")
    print("MAE : {mae}")
    print("R² : {r2}")
end function

lr ← LinearRegression()
EVALUATE_MODEL(lr, "Linear Regression")
rf ← RandomForestRegressor(n_estimators = 100, random_state = 42)
EVALUATE_MODEL(rf, "Random Forest Regression")
svr ← SVR(kernel =' rbf')
EVALUATE_MODEL(svr, "Support Vector Regression")
X_dmat ← xgb.DMatrix(data = X, label = y)
kf ← KFold(n_splits = len(X))
predictions ← np.zeros(len(X))
for train_index, test_index ∈ kf.split(X) do
    X_train, X_test ← X.iloc[train_index], X.iloc[test_index]
    y_train, y_test ← y.iloc[train_index], y.iloc[test_index]
    dtrain ← xgb.DMatrix(data = X_train, label = y_train)
    dtest ← xgb.DMatrix(data = X_test, label = y_test)

    params ← {objective : reg:squarederror,
              max_depth : 3,
              learning_rate : 0.1,
              n_estimators : 100,
              seed : 42}

    model ← xgb.train(params, dtrain)
    predictions[test_index] ← model.predict(dtest)
end for

mse ← mean_squared_error(y, predictions)
rmse ← √mse
mae ← mean_absolute_error(y, predictions)
r2 ← r2_score(y, predictions)

print("XGBoostRegression")
print("MSE : {mse}")
print("RMSE : {rmse}")
print("MAE : {mae}")
print("R² : {r2}")
```

Following the identification of the most suitable regression model for predicting the heat-based risk index, the final step involves visualizing the predicted values on a geographic map. This visualization is crucial for conveying spatial variations in risk and supporting interpretation for decision-makers. Table 9 outlines the methodology used to generate the risk index map, which begins by extracting latitude and longitude coordinates from the dataset. Risk index values are then used to color-code each point on a scatter plot, with longitude and latitude serving as the horizontal and vertical axes. The 'coolwarm' color scale is applied to visually distinguish levels of risk, and a color bar is included to improve readability. Titles and axis labels are added to provide clear context, including the identification of the predictive model used, such as XGBoost. The resulting visualization highlights spatial hotspots and supports more targeted and informed mitigation strategies.

**Table 9.** Pseudocode for Plotting Predicted Risk Index Map

**Algorithm 2** Plotting Predicted Risk Index Map

```
Input: DataFrame df, predictions
Output: Risk Index Map

X ← df[['Latitude', 'Longitude']]
y ← df['RiskIndex']

plt.scatter(X['Longitude'], X['Latitude'], c=predictions, cmap='coolwarm')
plt.colorbar(label='Predicted Risk Index')
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.title('Predicted Risk Index (XGBoost)')
plt.show()
```

Figure 13 visually represents the Predicted Risk Index using XGBoost, highlighting the model's superior performance. Overall, XGBoost was the most effective model for predicting the heat-based risk index, demonstrating superior accuracy and better variance

explanation than Linear Regression, Random Forest Regression, and Support Vector Regression.

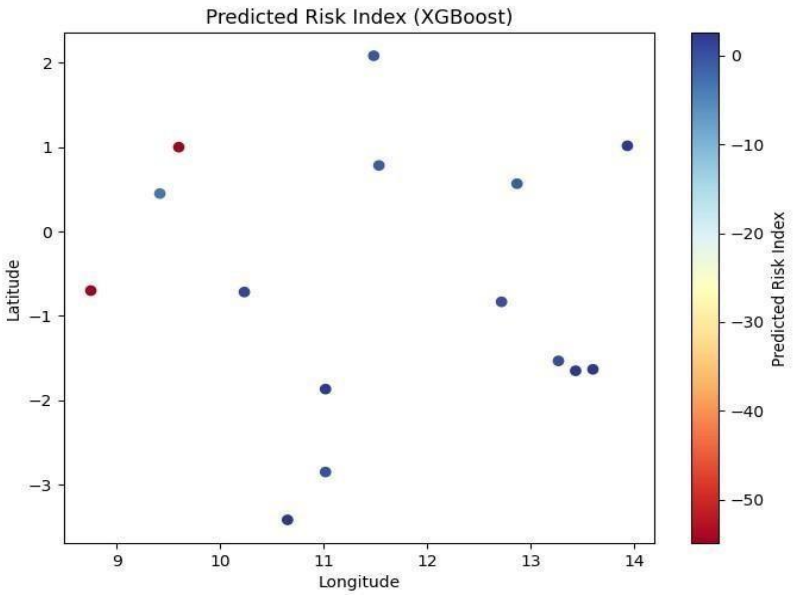


**Figure 13.** Demonstrates the spatial distribution of the Predicted Risk Index

## 6.6. Implications for Agricultural Resilience and Food Security

The findings from this study underscore the urgent need to integrate data-driven approaches into agricultural planning and climate-resilience strategies. As global temperatures continue to rise, heat-related stress is expected to increase in frequency and severity, posing a direct threat to crop yields, soil quality, and water availability. Developing a heat-based risk index combined with geostatistical techniques such as Kriging and IDW provides a robust framework for identifying spatial patterns of vulnerability. These tools allow for detecting high-risk zones where crops and farming communities are most exposed

to extreme heat. Such spatial insights enable stakeholders to prioritize adaptation efforts, such as implementing shade structures, altering planting schedules, or deploying droughtresistant crop varieties in critical areas. By forecasting where agricultural systems are most at risk, this methodology promotes proactive rather than reactive responses to climateinduced threats.

Furthermore, the integration of machine learning models into environmental risk assessment enhances the precision and scalability of decision-making tools for agricultural resilience. Models like XGBoost demonstrated strong predictive performance in identifying areas where heat exposure and population density intersect, a key factor in assessing not just environmental impact but also socio-economic vulnerability. These insights are critical in regions where food insecurity is already prevalent and where smallholder farmers lack the resources to adapt effectively. By automating the analysis and visualization processes, this framework allows for real-time updates as new data becomes available, ensuring dynamic monitoring and early warning capabilities. Ultimately, this research contributes to the broader goal of strengthening food systems against climate variability by equipping policymakers, farmers, and development agencies with the spatial intelligence needed to allocate resources efficiently and equitably.

## 6.7. Summary

Chapter 6 presented a comprehensive case study on assessing environmental risk in agricultural systems, with a particular emphasis on the influence of extreme heat events. The chapter introduced the development of a heat-based risk index that integrates

temperature thresholds with population density to identify regions most vulnerable to thermal stress. Using geospatial analysis techniques namely Inverse Distance Weighting (IDW) and Kriging the study generated continuous surface layers to visualize risk patterns. Kriging's incorporation of spatial autocorrelation offered greater precision in identifying microclimatic variations, while IDW provided a simpler, more computationally efficient alternative. The comparison of these interpolation methods provided valuable insights into how spatial modeling can inform climate adaptation strategies in agriculture.

To enhance predictive capabilities, machine learning models, including XGBoost, Random Forest, Support Vector Regression (SVR), and Linear Regression, were evaluated using key performance metrics. XGBoost emerged as the most accurate and reliable model, effectively predicting risk based on climatic and spatial variables. The process was further strengthened through automation of data preprocessing, model training, and visualization tasks, supporting continuous monitoring of agricultural risk. These combined approaches geostatistical and machine learning demonstrated the potential of data-driven tools to inform targeted adaptation, improve resource allocation, and support long-term agricultural resilience. The findings highlight the importance of integrating advanced spatial analytics with predictive modeling to address food security challenges in the face of ongoing climate variability.

# Chapter 7 | Conclusions & Future Work

## 7.1. Introduction

Chapter 7 provides a comprehensive conclusion to this dissertation, synthesizing the research findings and highlighting their broader implications. The chapter begins with Section 7.2, which summarizes the key findings from each case study and outlines how they collectively address the research objectives. In Section 7.3, cross-case insights and thematic connections are explored, revealing patterns and commonalities that emerged across different scenarios and geographic contexts. Section 7.4 presents a critical reflection on the dissertation's contributions to theory, methodology, and application, offering a philosophical perspective on how this work advances the field of risk analysis in transportation and environmental resilience.

Section 7.5 outlines the value of this research to scholars, practitioners, and policymakers, demonstrating the practical, academic, and policy-relevant benefits of

leveraging connected vehicle data for proactive safety strategies. This is followed by Section 7.6, which highlights methodological contributions made through novel geospatial techniques, predictive modeling, and the integration of big data sources for risk assessment. Section 7.7 discusses the limitations of the research, addressing constraints related to data availability, generalizability, and model assumptions. Building on this, Section 7.8 offers clear and actionable recommendations for future research, identifying opportunities to expand the scope, test new use cases, and deepen the theoretical foundation.

Finally, Section 7.10 details the practical and policy implications of the research, emphasizing how these findings can inform safer infrastructure planning and targeted policy interventions. The chapter concludes with Section 7.10, a set of final reflections that reaffirm the importance of this work and envision a continued interdisciplinary effort to optimize transportation systems through data-driven approaches.

## 7.2. Summary of Key Findings

This dissertation developed and demonstrated a comprehensive, geospatial, and datadriven framework for proactive risk management in transportation and environmental systems. The findings across the three major case studies reflect how the integration of connected vehicle (CV) data, machine learning, geospatial analytics, and information management can significantly enhance the ability to detect, model, and mitigate risks before adverse events occur.

In chapter 4 the case study focused on the research focused on the implementation of an information management system, applying IDEF modeling to represent and streamline risk-informed transportation planning. This study showed how modified IDEF

diagrams, when integrated with a digital twin approach, enable transportation planners to identify critical decision points and enhance the visualization of systemic risk. The development of centralized data repositories and geospatial dashboards further demonstrated the value of structuring largescale data for multi-criteria decision-making and long-term infrastructure resilience.

The Chapter 5 Case Study explored the application of clustering techniques (DBSCAN) and hotspot analysis (Getis-Ord Gi*) to analyze harsh braking and acceleration events in school zones. The results highlighted that surrogate safety measures derived from connected vehicle data can effectively identify high-risk zones in the absence of crash records. These findings support proactive interventions aligned with Vision Zero goals by revealing risky driving behavior patterns that would otherwise go undetected. The case study also produced actionable safety recommendations, such as adjustments to signage, speed controls, and enforcement strategies to protect vulnerable users like children and pedestrians.

In the Case Study of Chapter 6, the framework was extended to assess environmental risks, particularly the impacts of extreme heat on agricultural systems in West Africa. This study used Kriging and IDW interpolation methods along with machine learning models

(XGBoost, SVR, and Random Forest) to generate spatial risk indices that quantified heat vulnerability across regions. The findings demonstrated that geospatial risk modeling can provide critical insights for food security and climate adaptation planning. The ability to predict and visualize environmental stressors with high resolution confirmed the utility of the

proposed methods for decision-makers operating in resource-constrained or climatesensitive areas.

Across all three case studies, the research confirmed that integrating real-time data, spatial intelligence, and predictive modeling leads to more effective, proactive, and scalable risk mitigation strategies. These results validate the interdisciplinary framework proposed in this dissertation as a practical, replicable approach for advancing transportation safety and environmental resilience.

## 7.3. Cross-Case Insights and Thematic Connections

The three case studies presented in this dissertation—spanning transportation planning, school zone safety, and environmental resilience—share several unifying themes and insights that reinforce the value of a geospatial and data-driven approach to risk management. While each study applied distinct methodologies tailored to its context, together they demonstrate a cohesive framework rooted in systems thinking, predictive analytics, and the strategic use of emerging data sources such as connected vehicle (CV) data.

A primary cross-case insight is the importance of proactive risk detection. In contrast to traditional, reactive approaches that rely heavily on historical crash data or disaster response, each case study leverages real-time or high-frequency data to anticipate future risk conditions. In the school zone safety study, CV-based surrogate safety measures such as harsh braking and acceleration successfully identified risky driver behavior before crashes occurred. Similarly, the environmental risk analysis in West Africa provided early indicators of vulnerability due to extreme heat, using spatial models that forecast exposure

even in data-sparse regions. Across all cases, this predictive orientation facilitated more timely and targeted decision-making.

Another key theme is the integration of diverse data sources and analytic techniques into unified, actionable frameworks. The digital twin-based information management system developed in partnership with VDOT incorporated over 130 geospatial datasets, enabling a layered understanding of risks, infrastructure, and performance. This integrated data environment mirrors the environmental study's combination of remote sensing, weather data, and machine learning, as well as the school zone study's fusion of CV data and spatial statistics. These cases illustrate how interdisciplinary data fusion when managed systematically can enhance insight, reduce uncertainty, and support datainformed planning.

A third thematic connection lies in the use of geospatial intelligence and visualization to communicate complex risk patterns. Whether through DBSCAN clustering in school zones, Kriging-based heat maps in agricultural regions, or geospatial dashboards in transportation planning, each case study underscores the utility of spatial tools in enhancing transparency and accessibility. Visualization was not only instrumental in technical analysis but also served as a bridge between research outputs and practitioner needs, reinforcing the relevance of spatial analytics in both research and policy contexts.

Moreover, the transferability and scalability of methods used in each case affirm the framework's applicability across sectors. Techniques such as hotspot analysis, machine learning, and IDEF modeling proved effective in both transportation and environmental domains, suggesting that this integrated risk management framework can be adapted to other infrastructure systems, including energy, public health, and emergency management.

Lastly, each case study emphasized equity and resilience as core pillars. The school zone analysis supported safer infrastructure for vulnerable road users, particularly children in underserved communities. The environmental study provided insight into food insecurity risks in resource-constrained areas, and the digital twin development at VDOT enabled more inclusive planning through data democratization. Together, these efforts contribute to a broader vision of resilient, just, and sustainable systems.

In summary, the thematic connections across the case studies illustrate the strength of a unified framework that combines advanced analytics, real-time data, and geospatial intelligence. This approach offers a replicable model for anticipatory risk management and provides a foundation for cross-sector innovation in transportation and environmental systems.

## 7.4. Contributions to Theory, Methodology, and Application

This dissertation contributes meaningfully to the theoretical, methodological, and applied understanding of risk analysis in transportation and environmental systems. By integrating connected vehicle (CV) data, geospatial analytics, and predictive modeling into a systemsengineering framework, this work advances the field across three critical dimensions: theory, method, and practice.

### *Theoretical Contributions*

At the theoretical level, this research reframes risk analysis as a proactive, interdisciplinary process that leverages surrogate safety indicators and real-time data sources to anticipate and mitigate hazards before they manifest in the form of crashes or

system failures. Traditional transportation safety theory has historically been reactive relying on crash histories and aggregated data—but this work positions CV-based behavioral indicators (e.g., harsh braking, acceleration) as valid proxies for latent risk. This shift represents a philosophical evolution in how safety and risk are conceptualized and operationalized within engineering systems.

Additionally, the dissertation builds a conceptual bridge between transportation and environmental resilience theory, recognizing both as components of a broader infrastructure risk landscape. Through its application to agricultural heat vulnerability in West Africa, the research demonstrates how risk management strategies developed in transportation contexts can be extended to environmental domains expanding theoretical discourse into global development, and climate adaptation frameworks.

### *Methodological Contributions*

Methodologically, this dissertation offers a novel, integrated framework that combines IDEF modeling, digital twins, clustering algorithms (e.g., DBSCAN), spatial statistics (Getis-Ord Gi*), interpolation techniques (Kriging, IDW), and machine learning (e.g., XGBoost, SVR, Random Forest) within a unified risk assessment pipeline. This multi-method approach addresses key limitations in current practice, including the siloed nature of data sources and the lack of scalable, replicable analytical workflows.

The modification and application of IDEF models to include data governance, risk identification, and system-level performance analysis represent a unique extension of process modeling in transportation planning. The framework's adaptability to VDOT's

Pathways for Planning (P4P) platform illustrates how structured modeling can support enterprise-wide information management and decision-making.

Equally innovative is the use of CV event data to detect behavioral risk in school zones, demonstrating that high-resolution mobility data can uncover safety issues in contexts where crash data is sparse or unavailable. The methodological integration of hotspot analysis and surrogate indicators contributes to more granular, equity-focused transportation analysis and supports Vision Zero goals.

The inclusion of machine learning for environmental heat risk modeling contributes to climate analytics by offering scalable tools for estimating vulnerability in data-poor regions. The use of Kriging and IDW for creating continuous spatial risk surfaces and the evaluation of predictive models with explainable outputs further establish this work as methodologically rigorous and adaptable.

### *Applied Contributions*

Practically, the dissertation presents three robust, real-world applications of its framework—each with direct implications for practitioners and policymakers. The VDOT case study informs how agencies can structure large-scale information systems to support digital twin deployment and long-term infrastructure planning. The school zone safety analysis offers actionable insights for municipal planners and DOTs seeking to improve child pedestrian safety without waiting for crash data to accumulate. The environmental risk case study supports international development and agricultural policy planning, providing spatial tools to assess vulnerability to extreme heat.

Across all applications, the research emphasizes equity, early detection, and actionable insight. It empowers agencies to move from reactive to proactive management of transportation and environmental risks, thus aligning with national safety, resilience, and crop yield goals. The scalability and transferability of the framework demonstrate its potential for deployment across different jurisdictions and sectors.

## 7.5. Value to Scholars, Practitioners, and Policymakers

This dissertation provides distinct value to academic researchers, transportation practitioners, and policymakers by offering a replicable, data-driven framework that bridges theoretical innovation with real-world application. Through its interdisciplinary approach, the research delivers meaningful contributions to the fields of risk analysis, transportation safety, environmental resilience, and intelligent infrastructure planning.

### *Value to Scholars*

For scholars, this work contributes to academic discourse by advancing the theoretical foundations of proactive risk assessment through the use of surrogate safety measures, spatial intelligence, and machine learning. It broadens the existing literature on connected vehicle (CV) data by demonstrating its efficacy beyond conventional crash analysis, offering novel applications in predictive safety modeling, environmental risk assessment, and digital twin design. The integration of IDEF modeling with geospatial analytics and predictive algorithms creates a theoretical bridge between systems engineering and data science. Furthermore, the dissertation identifies key research gaps in

sensor integration, data fusion, and real-time modeling—establishing a strong foundation for future studies in resilient infrastructure and intelligent transportation systems (ITS).

### *Value to Practitioners*

Transportation engineers, planners, and infrastructure managers will find practical utility in the methodological tools and case study applications presented. The framework's ability to process and analyze billions of connected vehicle observations to detect high-risk driving behaviors provides a powerful approach for proactive safety interventions. Practitioners can apply the clustering and hotspot detection techniques demonstrated in school zone environments to other high-risk areas—enabling data-informed decisions without the long delays required for crash data collection.

Additionally, the digital twin and centralized knowledge repository developed for the State departments of Transportaiton, exemplifies how agencies can streamline project planning, enhance risk visibility, and improve coordination across departments. These tools can be readily adapted to support planning, asset management, and emergency response, especially in jurisdictions seeking to align with Vision Zero and environmental systems goals.

### *Value to Policymakers*

For policymakers, the dissertation offers a framework for evidence-based policy development rooted in real-time data and predictive insight. The ability to identify and visualize risk—both behavioral and environmental—enables the design of more equitable and effective transportation policies. For example, the school zone safety case demonstrates how connected vehicle data can be used to prioritize interventions in underserved

communities before crashes occur, supporting fair resource distribution and public safety initiatives.

The environmental risk modeling applied to agricultural regions in West Africa further supports international development policy, climate adaptation planning, and food security assessments. These findings can guide funding allocation, regulatory oversight, and cross-agency coordination on global resilience initiatives.

By connecting predictive analytics with actionable outcomes, this dissertation empowers policymakers to move from reactive, post-incident interventions to proactive strategies that address systemic risk across transportation and environmental sectors. In doing so, it supports broader societal goals, including reduced injury and fatality rates, increased infrastructure resilience, and improved quality of life.

## 7.6. Methodological Contributions to Risk Analysis

This dissertation introduces a novel, interdisciplinary methodological framework that significantly advances risk analysis practices within transportation and environmental systems.

Traditionally, risk analysis has relied on historical crash data, simplified statistical models, and siloed datasets. By contrast, the methodology presented in this research integrates connected vehicle (CV) data, machine learning, geospatial analytics, and systems modeling to support proactive, data-driven, and scalable risk assessment.

One of the core methodological contributions is the integration of IDEF modeling with risk-informed digital twin development. While IDEF (Integrated DEFinition) modeling is traditionally used for process analysis, this research extends its use to identify

system-level vulnerabilities and support the design of digital twins for transportation planning. By explicitly modeling inputs, outputs, controls, mechanisms, and sources of risk, the IDEF diagrams developed in this work allow practitioners to visualize not only transportation processes but also the points at which data-driven interventions can be embedded. This modeling approach has been operationalized in collaboration with the Virginia Department of Transportation (VDOT), where it guided the structuring of a centralized knowledge repository and geospatial dashboard to support lifecycle planning and multi-criteria decision-making.

Another significant contribution is the use of spatial clustering (DBSCAN) and hotspot analysis (Getis-Ord Gi*) to detect and quantify surrogate safety measures. These techniques were applied to billions of connected vehicle observations, allowing for the identification of high-risk driving patterns—such as harsh braking and rapid acceleration—in school zones. The use of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) allowed for the identification of naturally occurring clusters of risky behavior without needing to predefine the number of clusters or impose geometric constraints. This was further enhanced by Getis-Ord Gi*, which statistically validated the significance of detected hotspots. Together, these tools offer an advanced, replicable methodology for identifying risk-prone areas even in the absence of crash data, thereby improving on traditional reactive approaches.

The dissertation also contributes to environmental resilience modeling through the application of spatial interpolation methods—Kriging and Inverse Distance Weighting

(IDW)—to develop continuous surfaces of heat-based risk. These methods allowed for the spatial estimation of environmental vulnerability across West Africa, especially in rural and data-sparse regions. The comparison of Kriging and IDW demonstrated how choice of interpolation method can influence spatial outputs, and validated Kriging's superior ability to capture spatial autocorrelation and provide uncertainty measures. These interpolation techniques represent a methodological enhancement over coarse, region-level risk models used in conventional environmental studies.

Further methodological innovation lies in the use of machine learning algorithms such as XGBoost, Random Forest, and Support Vector Regression (SVR) for risk prediction. These models were trained on integrated datasets combining environmental variables, CV data, and spatial indicators. The use of XGBoost, in particular, enabled the identification of key predictive variables and allowed for scalable, high-performance modeling of risk indices. The models supported both transportation safety and environmental vulnerability use cases, highlighting their versatility and generalizability.

Lastly, the research presents a unified, modular framework that connects all these methodologies within a broader decision-support system. This system is capable of ingesting real-time data, visualizing spatial patterns of risk, and generating predictive outputs to guide proactive interventions. By embedding these analytical capabilities into a digital twin environment, the research transitions from traditional, descriptive risk analysis to an adaptive, intelligent approach that supports ongoing monitoring, evaluation, and improvement of transportation and environmental systems.

In sum, this dissertation contributes a flexible, replicable, and interdisciplinary methodology that enhances both the rigor and the relevance of risk analysis. It equips scholars, practitioners, and policymakers with tools to better understand complex systems, forecast emerging risks, and deploy timely, equitable, and effective interventions.

## 7.7. Limitations

While this dissertation offers valuable insights and methodological advancements in transportation safety and environmental risk management, several limitations must be acknowledged. These limitations primarily stem from data availability, model assumptions, and practical constraints related to implementation.

One of the primary limitations lies in the availability and coverage of connected vehicle data. Although the datasets used were extensive and included billions of observations, they represent only a subset of all vehicles on the road. Vehicle participation in connected data networks is still limited by technology adoption rates and coverage varies geographically. This can lead to incomplete representations of traffic conditions, especially in less urbanized areas or during periods of low vehicle activity. Furthermore, not all vehicle events are recorded with the same frequency or quality, which may introduce sampling bias or underreporting in certain locations or time frames.

Another limitation is associated with the use of surrogate safety measures. While indicators such as harsh braking and rapid acceleration provide meaningful proxies for potential crash risk, they do not guarantee that a crash will occur. These events may reflect unsafe behavior, but they can also result from necessary defensive driving in unpredictable traffic environments. As a result, caution must be exercised when interpreting clusters of

surrogate events as definitive indicators of high risk without supporting observational or contextual data.

From a methodological perspective, the interpretation of spatial patterns and predictive models is influenced by assumptions inherent to the tools used. For instance, clustering results from DBSCAN are sensitive to parameter selection, which can affect how many clusters are detected and their size. Similarly, interpolation techniques such as Kriging and IDW rely on the assumption that nearby values are spatially correlated, which may not always hold true in areas with sharp environmental or behavioral transitions. Machine learning models, though powerful, are influenced by the quality and balance of the training data and may not generalize well in areas or time periods with different characteristics than those used during model development.

Another limitation is the challenge of real-time integration and operational deployment. While the research demonstrates the feasibility of integrating diverse datasets into a digital twin and centralized information management system, deploying such systems at scale requires significant coordination across agencies, investment in digital infrastructure, and staff capacity to interpret and act on model outputs. This can be particularly challenging in regions where digital transformation in transportation agencies is still emerging.

Additionally, environmental data used in the analysis of heat-based agricultural risk was limited in spatial and temporal resolution. Despite the use of interpolation to estimate continuous surfaces, actual ground truth data in many of the rural regions studied remains

scarce. This restricts the ability to validate predictive models with high confidence and may limit the applicability of findings outside the studied regions.

Lastly, while the dissertation presents a unified risk analysis framework, its implementation was focused on selected case studies, including a school zone in Northern Virginia and agricultural regions in West Africa. These case studies were chosen to demonstrate the flexibility of the framework across domains, but they do not capture all possible contexts or challenges that might emerge in other regions or infrastructure systems.

In summary, the research is constrained by data availability, modeling assumptions, and implementation feasibility. These limitations do not undermine the contributions of the work but instead highlight important areas for future refinement and continued development.

## 7.8. Recommendations for Future Research

This dissertation has laid the foundation for a data-driven and geospatially informed approach to risk analysis in both transportation and environmental contexts. While the results demonstrate promising outcomes across multiple domains, there remain several opportunities for further exploration and enhancement of the methods and applications introduced.

One key area for future research is the expansion of connected vehicle data sources. As participation in connected vehicle networks increases and data quality improves, researchers will be able to conduct more comprehensive and granular analyses. Future studies should explore how the integration of additional vehicle metrics, such as steering

angles, lane positioning, and pedestrian detection events, can improve the understanding of complex driving behaviors. These enhancements could lead to more refined indicators of near misses and unsafe maneuvers, especially in urban corridors and pedestrian-heavy areas.

Another recommendation involves the application of the framework across a wider variety of locations and conditions. While this research focused on specific case studies in Virginia and West Africa, future work should test the transferability of the models in different geographic, economic, and infrastructure settings. Expanding the framework to rural, suburban, and high-density urban environments across various regions can validate the adaptability and robustness of the risk indicators and predictive models. Doing so would help identify how infrastructure layout, cultural differences in driving behavior, and climate conditions influence risk patterns.

The integration of additional data streams and sensors presents another valuable opportunity. Incorporating data from mobile devices, roadside units, weather monitoring systems, and crowdsourced reports could offer a more complete picture of the transportation ecosystem. Multimodal datasets, including those capturing bicycle and pedestrian movements, could also be incorporated into future risk assessments to support more inclusive mobility planning.

There is also an opportunity to further advance predictive modeling techniques. While this research applied established machine learning models such as XGBoost and Random Forest, future research could explore the use of deep learning architectures that can capture temporal and spatial dependencies more effectively. Recurrent neural networks,

graph neural networks, and hybrid models could be tested to model the evolving nature of risk in both transportation systems and environmental conditions. Additionally, explainable artificial intelligence methods should be further developed to ensure that model outputs remain transparent and actionable for planners and decision-makers.

Future research should also continue to refine digital twin capabilities. This includes building more dynamic and automated linkages between real-time data feeds and decision support dashboards. Greater automation in updating geospatial layers, integrating sensor inputs, and triggering alerts would allow digital twins to become more operational and responsive tools. This would be particularly beneficial in time-sensitive contexts such as school zone monitoring, evacuation planning, or weather-related disruptions.

Lastly, future research should place increased emphasis on evaluating the impact of interventions informed by this framework. While the current work focuses on detection and prediction, additional studies should explore how data-driven interventions affect realworld outcomes. Pilot programs that implement safety countermeasures based on hotspot analysis, for example, could be evaluated to measure reductions in risk-related behaviors or incidents. Similarly, testing the effectiveness of agricultural adaptation recommendations in areas predicted to face extreme heat can help validate model utility and guide future improvements. These recommendations aim to expand the scope, precision, and impact of the research introduced in this dissertation. By pursuing these directions, future studies can continue to bridge the gap between advanced analytics and real-world risk mitigation across diverse sectors.

## 7.9 Schedule and Timeline


Figure 14. Timeline of conference presentations and publications. Annotations above the timeline represent conference presentations, and annotations show journal and conference publications.

## References

[1] H. Dui, S. Zhang, M. Liu, X. Dong and G. Bai, "IoT-Enabled Real-Time Traffic Monitoring and Control Management for Intelligent Transportation Systems," IEEE Internet of Things Journal, vol. 11, no. 9, pp. 15842–15854, May 1, 2024, doi: 10.1109/JIOT.2024.3351908.

[2] M. Nasr Azadani and A. Boukerche, "Driving Behavior Analysis Guidelines for Intelligent Transportation Systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6027–6045, July 2022, doi: 10.1109/TITS.2021.3076140.

[3] P. Arthurs, L. Gillam, P. Krause, N. Wang, K. Halder and A. Mouzakitis, "A Taxonomy and Survey of Edge Cloud Computing for Intelligent Transportation Systems and Connected Vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6206–6221, Jul. 2022, doi: 10.1109/TITS.2021.3084396.

[4] K. Aati, M. Houda, S. Alotaibi, A. M. Khan, N. Alselami, and O. Benjeddou, "Analysis of Road Traffic Accidents in Dense Cities: Geotech Transport and ArcGIS," *Transportation Engineering*, vol. 16, p. 100256, 2024, doi: 10.1016/j.treng.2024.100256.

[5] K. Tsolaki, T. Vafeiadis, A. Nizamis, D. Ioannidis, and D. Tzovaras, "Utilizing machine learning on freight transportation and logistics applications: A review," ICT Express, vol. 9, no. 3, pp. 284–295, 2023, doi: 10.1016/j.icte.2022.02.001.

[6] A. C. Ikegwu, H. F. Nweke, C. V. Anikwe, et al., "Big data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions," Cluster Computing, vol. 25, pp. 3343–3387, 2022, doi: 10.1007/s10586-022-03568-5..

[7] R. L. Wheeler, C. A. Pennetti, T. L. Polmateer, G. Jones, Y. Dhanpal, and J. H. Lambert, "Information Management of a Lifeline Infrastructure for Mobility of People and Goods," in Proc. 2022 IEEE International Systems Conference (SysCon), Montreal, QC, Canada, 2022, pp. 1–4, doi: 10.1109/SysCon53536.2022.9773882.

[8] National Highway Traffic Safety Administration, “Rural Road Safety,” NHTSA, [Online]. Available: https://www.nhtsa.gov/rural. [Accessed: Mar. 25, 2025].

[9] Bureau of Transportation Statistics, “Rural Transportation Statistics,” U.S. Department of Transportation, 2022. [Online]. Available: https://www.bts.gov/rural. [Accessed: Mar. 25, 2025].

[10] World Health Organization, “Road traffic injuries,” WHO, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. [Accessed: Mar. 25, 2025].

[11] W. Wijnen, W. Weijermars, A. Schoeters, W. van den Berghe, R. Bauer, L. Carnis, R. Elvik, and H. Martensen, “An analysis of official road crash cost estimates in European countries,” Safety Science, vol. 113, pp. 318–327, 2019, doi: 10.1016/j.ssci.2018.12.004.

[12] A. Qalb, H. S. H. Arshad, M. S. Nawaz, and A. Hafeez, “Risk reduction via spatial and temporal visualization of road accidents: a way forward for emergency response optimization in developing countries,” International Journal of Injury Control and Safety Promotion, vol. 30, no. 2, pp. 310–320, 2023, doi: 10.1080/17457300.2022.2164312.

[13] M. S. Hussain, A. K. Goswami, and A. Gupta, “Predicting pedestrian crash locations in urban India: An integrated GIS-based spatiotemporal HSID technique,” Journal of Transportation Safety & Security, vol. 15, no. 2, pp. 103–136, 2022, doi: 10.1080/19439962.2022.2048759.

[14] National Oceanic and Atmospheric Administration, “World just sweltered through its hottest August on record,” NOAA, Sep. 14, 2023. [Online]. Available: https://www.noaa.gov/news/world-just-sweltered-through-its-hottest-august-on-record.

[15] A. Marsha, S. R. Sain, M. J. Heaton, et al., “Influences of climatic and population changes on heat-related mortality in Houston, Texas, USA,” Climatic Change, vol. 146, pp. 471–485, 2018, doi: 10.1007/s10584-016-1775-1.

[16] A. Bratu, K. G. Card, K. Closson, N. Aran, C. Marshall, S. Clayton, M. K. Gislason, H. Samji, G. Martin, M. Lem, C. H. Logie, T. K. Takaro, and R. S. Hogg, “The 2021 Western North American heat dome increased climate change anxiety among British Columbians: Results from a natural experiment,” The Journal of Climate Change and Health, vol. 6, p. 100116, 2022, doi: 10.1016/j.joclim.2022.100116.

[17] C. Faurie, B. M. Varghese, J. Liu, and P. Bi, “Association between high temperature and heatwaves with heat-related illnesses: A systematic review and meta-analysis,” Science of The Total Environment, vol. 852, p. 158332, 2022, doi: 10.1016/j.scitotenv.2022.158332.

[18] IPCC, Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, H.-O. Pörtner, D. C. Roberts, M. Tignor, E. S. Poloczanska, K. Mintenbeck, A. Alegría, et al., Eds. Cambridge, UK and New York, NY, USA: Cambridge University Press, 2022, doi: 10.1017/9781009325844.

[19] Federal Highway Administration, Impact of Intersection Angle on Highway Safety, FHWA-HRT-20-067, McLean, VA, USA, 2021.

[20] C. A. Pennetti, J. Jun, G. S. Jones, and J. H. Lambert, “Temporal disaggregation of performance measures to manage uncertainty in transportation logistics and scheduling,” ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, vol. 7, no. 1, p. 04020047, 2021, doi: 10.1061/ajrua6.0001096.

[21] W. Kumfer et al., “Systemic predictive safety analysis of pedestrian crashes for Montgomery County’s Vision Zero program,” Transportation Research Record, vol. 2678, no. 11, pp. 2165–2180, 2024. [Online]. Available: https://doi.org/10.1177/03611981241247178.

Connelly, E. B., Colosi, L. M., Clarens, A. F., & Lambert, J. H. (2015). Risk analysis of biofuels industry for aviation with scenario-based expert elicitation. Systems Engineering, 18(2), 178-191.

[22] J. Wang, T. Fu, and Q. Shangguan, "Wide-area vehicle trajectory data based on advanced tracking and trajectory splicing technologies: Potentials in transportation research," Accident Analysis & Prevention, vol. 186, Art. no. 107044, 2023. [Online]. Available: https://doi.org/10.1016/j.aap.2023.107044.

[23] Y. Gu, H. Zhang, L. D. Han, and A. Khattak, "Modeling spatiotemporal heterogeneity in interval-censored traffic incident time to normal flow by leveraging crowdsourced data: A geographically and temporally weighted proportional hazard analysis," Accident Analysis & Prevention, vol. 195, p. 107406, 2024. [Online]. Available: https://doi.org/10.1016/j.aap.2023.107406.

[24] T. Syum Gebre, L. Beni, E. Tsehaye Wasehun, and F. Elikem Dorbu, "AI-integrated traffic information system: A synergistic approach of physics informed neural network and GPT-4 for traffic estimation and real-time assistance," IEEE Access, vol. 12, pp. 65869–65882, 2024, doi: 10.1109/ACCESS.2024.3399094.

[25] Z. Meng et al., "Traffic object detection for autonomous driving fusing LiDAR and pseudo 4D-radar under bird's-eye-view," IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 11, pp. 18185–18195, Nov. 2024. [Online]. Available: https://doi.org/10.1109/TITS.2024.3417826.

[26] A. Abulibdeh, "Planning for congestion pricing policies in the Middle East: Public acceptability and revenue distribution," Transportation Letters, vol. 14, no. 3, pp. 282– 297, 2022. [Online]. Available: https://doi.org/10.1080/19427867.2020.1857908.

[27] S. Chen, M. Kuhn, K. Prettner, and D. E. Bloom, "The global macroeconomic burden of road injuries: estimates and projections for 166 countries," The Lancet Planetary Health, vol. 3, no. 9, pp. e390–e398, 2019, doi: 10.1016/S2542-5196(19)30170-6.Hadj-Mabrouk, H. (2017). Preliminary Hazard Analysis (PHA): New hybrid approach to railway risk analysis. *International Refereed Journal of Engineering and Science*, *6*(2), 51-58.

[28] I. A. Rossi, D. Vienneau, M. S. Ragettli, B. Flückiger, and M. Röösli, "Estimating the health benefits associated with a speed limit reduction to thirty kilometres per hour: A health impact assessment of noise and road traffic crashes for the Swiss city of Lausanne," Environment International, vol. 145, p. 106126, 2020, doi: 10.1016/j.envint.2020.106126.Haskins, C. INCOSE Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities; Wiley: New York, NY, USA, 2006

[29] C. Izaguirre, I. J. Losada, P. Camus, et al., "Climate change risk to global port operations," Nature Climate Change, vol. 11, pp. 14–20, Jan. 2021, doi: 10.1038/s41558020-00937-z.

[30] V. Manzione Filho, "Climate Change and Its Impacts on Businesses," in Climate Change, S. A. Bandh, Ed. Cham: Springer, 2022. doi: 10.1007/978-3-030-86290-9_6.

[31] M. Burke, S. Hsiang, and E. Miguel, "Global non-linear effect of temperature on economic production," Nature, vol. 527, pp. 235–239, Nov. 2015, doi: 10.1038/nature15725.

[xx] Iacobescu, C., Oltean, G., Florea, C., & Burtea, B. (2021). Unified interplanetary smart parking network for maximum end-user flexibility. *Sensors*, *22*(1), 221.

[32] L. Alzubaidi, J. Zhang, A. J. Humaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," J. Big Data, vol. 8, article no. 53, 2021, doi: 10.1186/s40537-021-00444-8.

[33] X. Hao, G. Zhang, and S. Ma, "International Journal of Semantic Computing," Int. J. Semant. Comput., vol. 10, no. 03, pp. 417–439, 2016.

[34] A. T. D. Perera, V. M. Nik, D. Chen et al., "Quantifying the impacts of climate change and extreme climate events on energy systems," Nat. Energy, vol. 5, pp. 150–159, Feb. 2020, doi: 10.1038/s41560-020-0558-0.

[35] S. Scher and G. Messori, "Predicting weather forecast uncertainty with machine learning," Quart. J. Roy. Meteorol. Soc., vol. 144, pp. 2830–2841, 2018, doi: 10.1002/qj.3410.

[xx] ISO/IEC/IEEE. 2010. Systems and Software Engineering - System and Software Engineering Vocabulary (SEVocab). Geneva, Switzerland: International Organization for Standardization

[36] C. Wu, Y. Zhou, M. V. P. Pessôa, Q. Peng, and R. Tan, "Conceptual digital twin modeling based on an integrated five-dimensional framework and TRIZ function model," J. Manuf. Syst., vol. 58, pt. B, pp. 79–93, 2021, doi: 10.1016/j.jmsy.2020.07.006.

[37] Z. Zhu, C. Liu, and X. Xu, "Visualisation of the Digital Twin data in manufacturing by using Augmented Reality," Procedia CIRP, vol. 81, pp. 898–903, 2019, doi: 10.1016/j.procir.2019.03.223. Johnson, D. A., Melo, V., & Lambert, J. H. (2022, October). Risk Identification with Entity Attributes Diagrams in Business Process Modeling. In *2022 IEEE International Symposium on Systems Engineering (ISSE)* (pp. 18). IEEE.

[38] H. M. Carlin, P. A. Goodall, R. I. M. Young, and A. A. West, "An interactive framework to support decision-making for Digital Twin design," J. Ind. Inf. Integr., vol. 41, 2024, Art. no. 100639, doi: 10.1016/j.jii.2024.100639.

[39] S. Madumidha, P. S. Ranjani, S. S. Varsinee and P. S. Sundari, "Transparency and traceability: In food supply chain system using blockchain technology with internet of things", Proc. Int. Conf. Trends Electron. Informatics ICOEI 2019, pp. 983-987.

[40] S. V. Buer, G. I. Fragapane and J. O. Strandhagen, "The Data-Driven Process Improvement Cycle: Using Digitalization for Continuous Improvement", IFACPapersOnLine, vol. 51, no. 11, pp. 1035-1040.

[41] R. Bean and T. H. Davenport, "Companies Are Failing in Their Efforts to Become Data-Driven", Harvard Business Review, 2019, [online] Available: http://hbr.org/2019/02/companies-are-failing-in-their-efforts-tobecome-data-driven.
Kaufmann, U., & Schuler, R. (2016). Systems Re-Engineering-ein Beitrag zur Integration von MBSE und PLM. *Schulze, S.-O.; Tschirner, C.; Kaffenberger, R*, 343-353.

[42] Z. A. Collier and J. H. Lambert, "Managing obsolescence of embedded hardware and software in secure and trusted systems", Front. Eng. Manag, vol. 7, pp. 172-181, [online] Available: https://doi.org/10.1007/s42524-019-0032-5.

[43] DAMA International Data Management Body of Knowledge. Technics Publications, 2017.
[xx] Krishnan, R., & Bhada, S. V. (2022). Integrated system design and safety framework for modelbased safety assessment. IEEE Access, 10, 79311-79334.

[44] C. A. Pennetti, J. Jun, G. S. Jones and J. H. Lambert, "Temporal Disaggregation of Performance Measures to Manage Uncertainty in Transportation Logistics and Scheduling", ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng, vol. 7, no. 1, pp. 04020047.

[45] C. A. Pennetti, J. Jun, G. S. Jones and J. H. Lambert, "Temporal Disaggregation of Performance Measures to Manage Uncertainty in Transportation Logistics and Scheduling", ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng, vol. 7, no. 1, pp. 04020047.

[46] J. Bickford, D. L. Van Bossuyt, P. Beery, and A. Pollman, "Operationalizing digital twins through model-based systems engineering methods," Systems Engineering, vol. 23, pp. 724–750, 2020, doi: 10.1002/sys.21559.

[47] J. Bickford, D. L. Van Bossuyt, P. Beery, and A. Pollman, "Operationalizing digital twins through model-based systems engineering methods," Systems Engineering, vol. 23, pp. 724–750, 2020, doi: 10.1002/sys.21559.

[48] Y. Feng, K. L. Head, S. Khoshmagham, and M. A. Zamanipour, "A real-time adaptive signal control in a connected vehicle environment," Transp. Res. Part C Emerg. Technol., vol. 55, pp. 460–473, 2015.

[49] N. Goodall, B. L. Smith, and B. B. Park, "Traffic signal control with connected vehicles," Transp. Res. Rec. J. Transp. Res. Board, vol. 2381, pp. 65–72, 2013.

[50] J. Zheng and H. X. Liu, "Estimating traffic volumes for signalized intersections using connected vehicle data," Transp. Res. Part C Emerg. Technol., vol. 79, pp. 347– 362, 2017.

[51] M. Hunter, J. K. Mathew, E. Cox, M. Blackwell, and D. M. Bullock, "Estimation of connected vehicle penetration rate on Indiana roadways," JTRP Affiliated Reports, no. 37, pp. 1–6, 2021, doi: 10.5703/1288284317343.

MITRE Corporation (2014) MITRE Systems Engineering Guide

[52] J. Leng, J. Guo, J. Xie, X. Zhou, A. Liu, X. Gu, D. Mourtzis, Q. Qi, Q. Liu, W. Shen, and L. Wang, "Review of manufacturing system design in the interplay of Industry 4.0 and Industry 5.0 (Part I): Design thinking and modeling methods," J. Manuf. Syst., vol. 76, pp. 158–187, 2024, doi: 10.1016/j.jmsy.2024.07.012.

[53] R. G. Alsarraj, A. M. Altaie, and E. Z. Majeed, "Developing an automated modelbased software testing tool from the design phase," IEEE Access, doi: 10.1109/ACCESS.2025.3553967.

[54] L. Luo, C.-S. Cheng, W. Mobley, A. Novoselac, K. Lieberknecht, and F. Leite, "Scenario generation for built environment decision support under uncertainty: Case studies of airflow modeling and climate-resilient infrastructure system design," J. Comput. Civ. Eng., vol. 39, no. 4, 2025, doi: 10.1061/JCCEE5.CPENG-6253.

[55] A. M. Amiri, N. Nadimi, V. Khalifeh, and M. Shams, "GIS-based crash hotspot identification: A comparison among mapping clusters and spatial analysis techniques," Int. J. Inj. Control Saf. Promot., pp. 325–338, May 2021, doi: 10.1080/17457300.2021.1925924.

[56] W. Zhang, "An improved DBSCAN algorithm for hazard recognition of obstacles in unmanned scenes," Soft Comput., vol. 27, pp. 18585–18604, Dec. 2023, doi: 10.1007/s00500-023-09319-x. OMG, 2018. What is SysML? http://www.omgsysml.org/what-is-sysml.html

[57] Y. Tu, Z. Tang, and B. Lev, "Regional flood risk grading assessment considering indicator interactions among hazard, exposure, and vulnerability: A novel FlowSort with DBSCAN," J. Hydrol., vol. 639, p. 131587, 2024, doi: 10.1016/j.jhydrol.2024.131587. [58] A. Starczewski, "A novel approach to determining the radius of the neighborhood required for the DBSCAN algorithm," in Artificial Intelligence and Soft Computing. ICAISC 2021. Lecture Notes in Computer Science, vol. 12854, L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, and J. M. Zurada, Eds. Cham, Switzerland: Springer, 2021, pp. 373–382, doi: 10.1007/978-3-030-87986-0_32.

[59] P. Bhattacharjee and P. Mitra, "A survey of density based clustering algorithms," Front. Comput. Sci., vol. 15, art. no. 151308, 2021, doi: 10.1007/s11704-019-9059-3.

[60] M. G. R. Fahad, W. C. Zech, R. Nazari, and M. Karimi, "Developing a geospatial framework for severe occupational injuries using Moran's I and Getis-Ord Gi* statistics for Southeastern United States," Nat. Hazards Rev., vol. 23, no. 3, 2022, doi: 10.1061/(ASCE)NH.1527-6996.0000566.

[61] D. D. Desai et al., "Optimal ambulance positioning for road accidents with deep embedded clustering," IEEE Access, vol. 11, pp. 59917–59934, 2023, doi: 10.1109/ACCESS.2023.3284993.

[62] K. H. Yasin, T. B. Gelete, A. D. Iguala, and E. Kebede, "Optimal interpolation approach for groundwater depth estimation," MethodsX, vol. 13, p. 102916, 2024, doi: 10.1016/j.mex.2024.102916.

[63] C. Munyati and N. I. Sinthumule, "Comparative suitability of ordinary kriging and inverse distance weighted interpolation for indicating intactness gradients on threatened savannah woodland and forest stands," Environ. Sustain. Indicators, vol. 12, p. 100151, 2021, doi: 10.1016/j.indic.2021.100151.

[64] R. Abdulmanov, I. Miftakhov, M. Ishbulatov, E. Galeev, and E. Shafeeva, "Comparison of the effectiveness of GIS-based interpolation methods for estimating the spatial distribution of agrochemical soil properties," Environ. Technol. Innov., vol. 24, p. 101970, 2021, doi: 10.1016/j.eti.2021.101970.

[65] T. Kavzoglu and A. Teke, "Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost)," Arab. J. Sci. Eng., vol. 47, pp. 7367–7385, Jun. 2022, doi: 10.1007/s13369-022-06560-8.

[66] Z. A. Collier, A. Gaskins, and J. H. Lambert, "Business Process Modeling for Semiconductor Production Risk Analysis Using IDEF0," IEEE Engineering Management Review, vol. 51, no. 1, pp. 183–188, Firstquarter, Mar. 2023, doi: 10.1109/EMR.2022.3230374.

[67] D. A. Johnson, R. Wheeler, and J. H. Lambert, "Risk Description Augmenting a Business Process Model," in Proc. IEEE Int. Syst. Conf. (SysCon), Vancouver, BC, Canada, 2023, pp. 1–7, doi: 10.1109/SysCon53073.2023.10131196.

[68] Z. Karami and R. Kashef, "Smart transportation planning: Data, models, and algorithms," Transportation Engineering, vol. 2, Art. no. 100013, 2020, doi: 10.1016/j.treng.2020.100013.

[69] M. J. Heaton, A. Datta, A. O. Finley, et al., “A Case Study Competition Among Methods for Analyzing Large Spatial Data,” JABES, vol. 24, pp. 398–425, Sep. 2019, doi: 10.1007/s13253-018-00348-w.

[70] L. Liu, X. Wang, X. Yang, H. Liu, J. Li, and P. Wang, “Path planning techniques for mobile robots: Review and prospect,” Expert Syst. Appl., vol. 227, Art. no. 120254, 2023, doi: 10.1016/j.eswa.2023.120254..

[71] M. Zichichi, S. Ferretti, and G. D’angelo, “A Framework Based on Distributed Ledger Technologies for Data Management and Services in Intelligent Transportation Systems,” IEEE Access, vol. 8, pp. 100384–100402, 2020, doi: 10.1109/ACCESS.2020.2998012.

[72] A. I. Shahrul, N. M. Nik Mustapha, M. S. Ahmad, et al., “Development of software for collecting cleft-specific data in Malaysia,” BMC Oral Health, vol. 25, Art. no. 333, Mar. 2025, doi: 10.1186/s12903-025-05583-5.

[73] M. C. Annosi, A. Martini, F. Brunetta, and L. Marchegiani, “Learning in an agile setting: A multilevel research study on the evolution of organizational routines,” J. Bus. Res., vol. 110, pp. 554–566, 2020, doi: 10.1016/j.jbusres.2018.05.011.

[74] X. Chang, R. Zhang, J. Mao, and Y. Fu, “Digital Twins in Transportation Infrastructure: An Investigation of the Key Enabling Technologies, Applications, and Challenges,” IEEE Trans. Intell. Transp. Syst., vol. 25, no. 7, pp. 6449–6471, Jul. 2024, doi: 10.1109/TITS.2024.3401716.