Investigation of Factors Contributing to Secondary Crashes

Prepared by: Mark Franz, Sara Zahedian

Investigation of Factors Contributing to Secondary Crashes

6/11/25

Acknowledgements: This research was funded by The Eastern Transportation Coalition (ETC). Many thanks to Sheryl Bradley from the ETC and Jason Dicembre from the Maryland Department of Transportation – State Highway Administration for their guidance and recommendations on this project.

The Eastern Transportation Coalition is a partnership of 19 states and the District of Columbia focused on connecting public agencies across modes of travel to increase safety and efficiency. Additional information on the Coalition, including other project reports, can be found on the Coalition's website: www.tetcoalition.org

Table of Contents

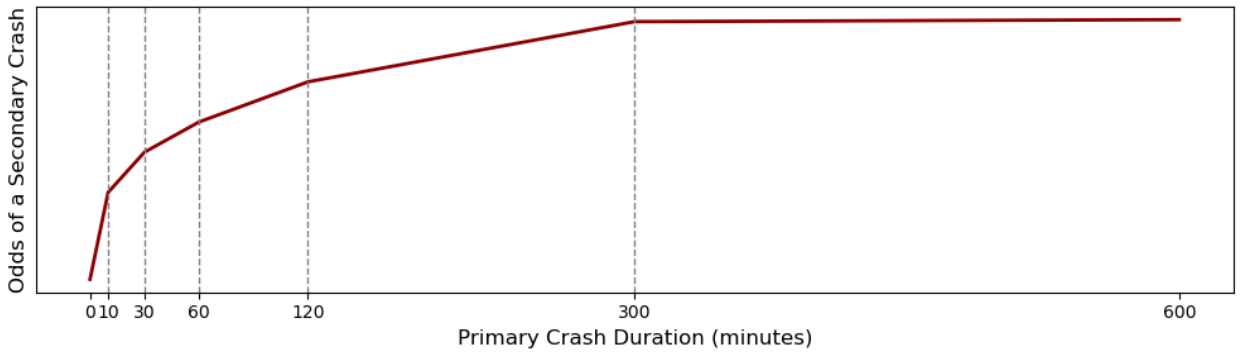
Executive Summary

A secondary crash is a collision that happens as a result of another initial or primary crash—often within the scene of the primary incident or in the resulting queue. These secondary crashes are essentially a chain reaction of crashes where one ultimately triggers another. There is an oft-quoted statistic that goes "the likelihood of a secondary crash increases by 2.8 percent for every minute that a primary incident remains a hazard." [1] Variations of this statistic have been presented in FHWA reports on the benefits of traffic incident management, the International Association of Chiefs of Police, numerous presentations, and related traffic safety studies ([2] and [3]) The transportation industry has adopted this statistic as fact despite the origin being based on an limited dataset covering a single roadway in Indiana from over 25 years ago. Updating this statistical talking point with broader datasets is essential to maintaining the credibility of our industry and our push for safety reforms.

The purpose of this research was to defensibly analyze and document the impact of incident duration on the probability of a secondary crash occurrence. To accomplish this goal, we conducted rigorous statistical analysis and probability modeling during a three-year, multi-state analysis of secondary crashes. This study evaluated nearly 3-million crashes during 2022, 2023, and 2024 in the states of Florida, Maryland, Tennessee, and Virginia. Our research determined that the percentage of crashes that could be classified as secondary varied between each state as shown in table below:

State	Percent of crashes that are considered secondary crashes.
Florida	4.8%
Maryland	3.9%
Tennessee	7.5%
Virginia	3.1%

One of the key contributions of this study is showing that the impact of each additional minute of incident duration on the likelihood of a secondary crash is not constant but varies depending on the total duration of the incident. This challenges the prior practice of reporting a single value to represent the effect of duration. It is important to note that with a variable like incident duration, we can interpret the effect on odds—defined as the probability of a secondary crash occurring divided by the probability of it not occurring—but the actual probability depends on a combination of other factors beyond duration alone. The figure below provides a view illustrating the impact of each additional minute added to the primary incident duration on the odds of a secondary crash. This study found that the impact of each minute of duration varies by the total duration of the primary incident. In general, the impact on odds ratio is highest for short duration incidents (0-10 minutes), and gradually decreases with increased primary crash incident duration.



Our key findings concluded the following:

1. Overall, for every minute a lane is blocked, the odds of a secondary crash can range from **less than 1% to 21%, depending on various incident factors**.
2. For incidents lasting 0–10 minutes, approximately 0.5% to 1.6% of incidents led to a secondary crash depending on the state. In this bin, each additional minute of duration increased **the odds of a secondary crash by 13% in Maryland, Virginia, and Tennessee, but up to 21% in Florida.**
3. For incidents lasting 10-30 minutes, around 2% of cases in Florida and Virginia and around 4% in Maryland and Tennessee resulted in a secondary crash. In this bin, **The odds of a secondary crash occurring increased approximately 3% for each minute added to the incident duration** for all states.
4. For incidents lasting 30-60 minutes, around 4% of cases in Florida and Virginia, and 6-7% in Maryland and Tennessee resulted in a secondary crash. Within this bin, **the odds of a secondary crash increased by about 1.5% for each additional minute of incident duration**.
5. While the proportion of incidents that led to a secondary crash ranged from approximately 4% to 7% for 60–120 minute bins, to about 8% to 15% for the 120–300 minute bin, and up to over 21% for the 300–600 minute bin (varying by state), **the increase in odds for each additional minute of duration within these bins remained low— generally less than 1%**.
6. We confirmed existing literature that states secondary crashes tend to occur closer to the start time of the primary crash.

Our research also concluded that characteristics of each crash played an important role in increasing (or decreasing) the odds of a secondary crash occurring. For example:

- Severity: Crash severity was statistically non-significant in predicting the odds of a secondary crash in Florida; however, in Tennessee, increases in injury severity increased the odds of a secondary crash by 50%, relative to a minor crash.
- Capacity Reduction: In Maryland and Virginia, a 20-30 percent reduction in capacity doubled the odds of a secondary crash, relative to no capacity reduction.
- Day of Week: For the impact of day of week (weekday or weekend), the results were also mixed.
    - The day of the week in which the crash occurred had little impact on secondary incident odds in Virginia.
    - In Maryland, the odds of secondary crashes were **higher during the weekend**.

    - In Tennessee and Florida, the **odds of a secondary crash was 7-17 percent higher on weekdays**.
- Volumes: Higher traffic flows increased the odds of a secondary crash across all states.
- Weather: Inclement weather was found to increase secondary crash likelihood.
    - **Rain increased the odds by 200-250 percent relative to clear weather.**
    - Snow showed a **large impact with increased odds of secondary crash of 550 percent in Maryland**.

While the primary objective of this research was to investigate the impact of incident duration on the probability of a secondary crash, the team made the following additional contributions:

- Conducted an in-depth review of recent studies in secondary crashes, highlighting methods to identify secondary crashes and methods to model secondary crashes
- Established procedures to fuse disparate data sources into a master database for safety analysis.
- Developed a methodology to identify secondary crashes using real-world speed data.
- Created and evaluated several secondary crash probability estimation models using rigorous statistical methods to test the assumptions of each model. These models were used to make inferences on the impact of key variables such as incident duration, weather, capacity reduction, and flow rates on the probability of secondary crashes.
- Documented challenges related to best practices in traffic management system crash data collection
- Made recommendations on the critical variables to collect to support secondary crash inference modeling

The research team believes that the next logical step for this research is to take the methodologies developed for historic secondary crash data analysis and modify them to function in real-time as a secondary crash prediction tool that could be integrated into existing traffic incident management decision support platforms. These prediction algorithms could engage at the onset of an incident and provide traffic incident management decision makers with valuable insights on an incident's impact soon after detection and throughout the incident management process. This information will enable proactive operational decisions which could improve safety and reduce delays, fuel consumption, emissions, and property destruction.

Introduction

Background

Traffic incidents are a major contributor to roadway congestion and safety concerns across the United States. A 2022 study conducted by the Center for Advanced Transportation Technology Laboratory found that in 2019, 18% of US traffic delay involved incidents as a contributing factor [4]. The resulting dashboard of this study can be found at [5] . Among incidents, secondary crashes—defined by the Federal Highway Administration (FHWA) as "unplanned incidents (starting at the time of detection) for which a response or intervention is taken, where a collision occurs either a) within the incident scene or b) within the queue (which could include the opposite direction) resulting from the original incident" [3]—are of particular concern. Secondary crashes pose elevated safety risks for responders and travelers and are a key metric in evaluating the performance of Traffic Incident Management (TIM) programs. Accurately identifying secondary crashes and understanding their relationship to primary incident characteristics, especially the duration of the initial incident, is critical for improving incident response strategies and reducing risk on the road.

Objective

The primary objective of this study is to investigate how the duration of a primary traffic incident influences the likelihood of a secondary crash. In addition to incident duration, the study also considers other potential contributing factors such as time of day, weather conditions, roadway features, and prevailing traffic conditions. The project is structured around the following goals:

- Develop a comprehensive methodology for identifying secondary crashes from large-scale incident datasets.
- Build models to quantify how incident duration and other features affect secondary crash probability.
- Assess how the availability and completeness of primary incident data influence prediction accuracy.

Project Scope and Contributions

This study undertook a first-of-its-kind, multi-state analysis of secondary crash dynamics using incident data from Maryland (MD), Virginia (VA), Tennessee (TN), and Florida (FL). A key contribution of this work was the development of a scalable and efficient data processing framework for secondary crash identification, which can be applied to large historical datasets. The project integrates diverse datasets, including:

- Incident data from state-level traffic management systems
- Radar-based road network data (TMC network)
- Historical traffic volume profiles
- Probe vehicle speed data
- Weather records from reliable meteorological sources

In addition to data integration and processing, this study developed a comprehensive inference modeling framework to extract and quantify the impact of primary incident characteristics, environmental and temporal factors, and prevailing traffic conditions on the likelihood of secondary crashes. These models enable a deeper understanding of the conditions under which secondary crashes are more likely to occur and support more effective Traffic Incident Management (TIM) strategies.

The fusion and processing of these multiple data sources, combined with the robust modeling framework, set the groundwork for scalable, data-driven secondary crash analysis across diverse geographies and conditions.

Organization of the Report

The remainder of this report is organized as follows:

- **Literature Review**: This section summarizes previous work on secondary crash detection and analysis. This includes a review of detection methodologies, descriptive statistics, inferential models, and machine learning-based prediction approaches. It also discusses key challenges in identifying and validating secondary crashes and reviews the spatial and temporal scales considered in past studies.
- **Data Evaluation**: This section reviews incident data schemas from MD, VA, TN, and FL as stored in the CATT
- **Lab database**: This section highlights the availability of critical fields such as incident start and clearance times and documents schema-specific differences. It also outlines how additional data sources (weather, volume, speed) were identified and reviewed.
- **Secondary Crash Identification**: This section describes the methodology used to detect secondary crashes, including initial temporal and spatial filtering of incident data, integration of segment-level attributes, and use of speed data for identifying traffic queues.
- **Data Processing**: This section details the steps to clean and prepare datasets, including feature engineering for incident, volume, and weather data. A summary table of all processed features is included at the end of this section.
- **Model Development**: This section presents the rationale for selecting logistic regression for inference modeling. It outlines how model assumptions were checked, provides descriptive insights, and presents model results for all four states.
- **Recommendations for Data Collection**: This section synthesizes findings related to data availability and quality, highlighting gaps and inconsistencies that impacted model performance. It also suggests ways to improve future data collection efforts across agencies.

Literature Review

Studies on secondary crashes were extensively explored to understand the current research landscape in this domain. Prior research on secondary crashes generally focused on two main areas: (1) methods for secondary crash detection or identification, and (2) risk analysis of secondary crash occurrence. These studies were organized as follows:

For secondary crash detection, methods were categorized as [6]:

- Static methods, which used predefined spatiotemporal thresholds to associate incidents.
- Dynamic methods, which considered evolving traffic conditions—such as speed drops or queue formation—to detect secondary crashes.
- Database-tagged methods, which relied on incident reports or management system records where secondary crashes were explicitly labeled.

Risk analysis studies were classified into three groups:

- Likelihood analyses, which employed parametric models (e.g., logistic regression) to estimate the influence of various factors on the probability of secondary crashes.
- Predictive modeling studies, which used machine learning techniques to forecast the likelihood of secondary crashes based on incident, environmental, and traffic features.

Secondary Crash Identification

Static Identification

The static approach identified secondary crashes using predefined spatial and temporal thresholds relative to each primary incident. For example, a crash occurring within two miles upstream and two hours following a primary incident's start was typically considered a secondary crash. This method was relatively easy to implement and more consistent than manual identification procedures. However, it was generally less reliable than dynamic methods, as it did not account for actual traffic conditions such as queue formation or speed reduction.

Several studies employed static criteria to identify secondary crashes [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. In addition to identifying crashes on the same directional approach, some studies also considered secondary crashes that occurred in the opposite direction due to the onlooker effect, using predefined spatial-temporal thresholds [10], [13], [14], [15].

Dynamic Identification

The dynamic approach identified secondary crashes based on the actual traffic queue formed due to the primary incident, offering a more accurate and reliable estimation of the incident's impact area compared to static methods. Although this approach provided the most precise identification of secondary crashes, it was resource-intensive and heavily dependent on traffic data availability and quality.

Previous studies generally categorized dynamic methods into three groups: queuing model-based, shockwave-based, and traffic data-driven approaches.

Queuing-Based Methods

Queuing-based approaches offered a realistic representation of the spatial and temporal extent of incident impact areas by estimating the maximum queue length and dissipation time caused by the primary crash. These methods typically relied on roadway characteristics such as capacity, arrival rate, and service rate. A deterministic queuing model was often applied, where total system delay was used to define the temporal boundary of potential secondary crashes. However, different road segments were subject to varying queuing behaviors due to differences in traffic flow, roadway geometry, incident severity, and environmental conditions [6].

Shockwave-Based Methods

Shockwave models assumed that the incident impact area formed a triangular region in the time-space diagram. These models defined the spatiotemporal extent of the queue by estimating the speed of the backward-forming and forward-dissipating shockwaves associated with the onset and clearance of the primary incident. The backward shockwave represented the rate at which the queue expanded upstream. In contrast, the forward shockwave began at the time of incident clearance and continued until it intersected the backward shockwave, signaling queue dissipation. Despite their conceptual clarity, shockwave-based methods were limited by simplified assumptions about constant traffic arrival and discharge rates, which often failed to reflect real-world variability [6], [15].

Both queuing and shockwave methods required assumptions that often oversimplified complex traffic conditions. These included constant arrival and departure rates, fixed shockwave speeds, and uniform roadway behavior, which could result in inaccurate estimations of incident impact areas and secondary crash boundaries [18], [19].

Data-Driven Estimation of Incident Impact Area

Data-driven methods aimed to estimate the spatiotemporal extent of incident impact areas by leveraging real-world traffic data, enabling more dynamic and realistic identification of secondary crashes. One prominent approach was developed by [20], in which an incident queue was defined as the segment where speeds dropped by at least 30% relative to the historical average for that segment and time of day. A time-space diagram was then constructed for all candidate incidents occurring within 30 minutes of the primary incident's start time and within 0.5 miles upstream.

In this method, a crash was considered secondary if a straight line could be drawn from the primary to the secondary incident on the time-space diagram, passing only through segments classified as non-recurring congestion.

However, this method had two notable shortcomings: If the connecting line between the primary and secondary incidents passed through even a single segment without experiencing non-recurring congestion, the incidents were classified as unrelated. The method only used the primary incident's start time to define the connecting line, which could be problematic for incidents that caused congestion only later in their duration. Some studies have introduced modifications to address these limitations. In one variation, only 90% of the time-space intervals between the primary and secondary incidents were required to exhibit non-recurring congestion. Additionally, the connection was allowed to originate from any point along the timeline of the primary incident, not just its start [21].

Another data-driven technique that gained prominence in recent years involved the use of speed contour plots to estimate the impact area of a primary incident. In this approach, traffic speed data were collected for several hours before and after an incident, covering a spatial buffer both upstream and downstream. The observed speed data were compared against average speeds from incident-free days at the same time and location to differentiate incident-induced congestion from routine traffic delays. By subtracting these baseline values, researchers created differential contour plots that highlighted areas of non-recurring congestion. These were then used to identify whether other crashes fell within the influence zone of the primary incident [22].

This method has become increasingly dominant in recent secondary crash studies. For instance, Zhang et al. [23] extracted 5-minute speed data for the six hours before and after each labeled secondary crash, using traffic data sources from approximately two miles upstream and downstream of the crash location. The authors built a new contour plot by subtracting the average speed profiles from crash-free days, allowing them to isolate the effects of non-recurring congestion and more accurately identify the secondary nature of those crashes.

In a similar effort, Li et al. [24] collected speed data covering a spatial range of five miles upstream and two miles downstream of the primary crash location, with a temporal window extending from one hour before to three hours after the incident. This configuration enabled them to observe the evolution and dissipation of congestion around the incident site in high resolution.

Liu et al. [25] employed a symmetric observation window, using speed data spanning two hours before and two hours after each primary crash, and covering two miles upstream and downstream. This balanced spatial-temporal window allowed for a focused examination of the immediate impact area and improved the classification of nearby crashes as secondary.

Together, these studies demonstrated the effectiveness of contour plot methods in leveraging high-resolution traffic data to identify the influence zones of primary incidents. By incorporating baseline comparisons to filter out recurring congestion, this approach offered a data-driven and adaptable framework for secondary crash detection.

Database Tag

Another approach to identifying secondary crashes is through explicit tagging in traffic incident databases, where each crash is labeled at the time of reporting as being secondary to a prior

incident. This method bypasses the need for spatiotemporal inference or dynamic modeling by relying directly on incident records maintained by responding agencies or traffic management centers. While this approach offers the advantage of operational clarity, its effectiveness is highly dependent on consistent and comprehensive data entry practices.

A notable example is the FHWA case study on Kentucky's TIM program. Through enhanced data collection, this effort aimed to establish a baseline for key TIM performance measures, including secondary crashes, roadway clearance time, and incident clearance time. The Kentucky Transportation Cabinet, in collaboration with FHWA, Kentucky State Police, local agencies, and the Kentucky Transportation Center, formed a TIM task force to promote structured data reporting. By tagging secondary crashes directly in incident databases, the program sought to improve transparency, operational analysis, and long-term planning. This approach highlighted the potential of coordinated interagency efforts to support performance-driven TIM practices and more accurately quantify the safety and economic impacts of secondary crashes. The study found that errors in the classification of secondary crashes had decreased over time, with correctly identified secondary crashes increasing from 8.3% in 2015 to 13.3% in 2017. Simultaneously, the number of incorrectly coded secondary crashes declined, a trend attributed to improved training of first response personnel. However, the study also identified specific agencies with persistently high error rates and recommended targeted training programs to further enhance reporting accuracy [26].

Another study [27] evaluated the reliability and consistency of database-tagged secondary crashes across several U.S. states that explicitly recorded secondary crashes in their crash reports. A key finding was that data quality varied considerably, with more than two-thirds of crashes labeled as secondary lacking identifiable primary crash candidates within two hours and two kilometers. This raised concerns about inconsistent application of the secondary crash definition, potential geospatial inaccuracies, and underreporting of primary crashes. The study suggested that some secondary crashes may be triggered by non-crash incidents—such as disabled vehicles or debris—that are not always logged in crash databases. Additionally, verifying secondary crashes through crash narratives, while more accurate, was found to be resource-intensive. These limitations underscored the need for better training, clearer guidelines, and possibly the integration of spatiotemporal analysis to validate secondary crash classifications in state databases.

Despite these challenges, the study conducted a comprehensive descriptive analysis of the verified secondary crashes. Most secondary crashes occurred on Interstate highways or other major arterials in urban areas, during daylight hours, and under clear weather. The majority did not involve injuries, and roughly two-thirds were rear-end collisions. Smaller proportions were sideswipes or non-collision events. Spatially and temporally, about 84% of secondary crashes occurred within half a kilometer of the primary crash, and nearly half occurred within 20 minutes—though some timing patterns may reflect reporting biases due to time rounding. Contributing circumstances were often left blank, but where noted, common factors included stopped vehicles, failure to reduce speed, and following too closely. Aggregated across states, 41% of secondary crashes were linked to driver behavior, 28% to road hazards, and 24% to roadway or traffic conditions. Case studies from states like Florida reinforced these findings, with driver inattention and distraction consistently identified as key contributors.

Secondary Crash Risk Analysis

Parametric Models

Several studies employed parametric models to estimate the probability or timing of secondary crashes, offering interpretable relationships between explanatory variables and crash outcomes. One study that explicitly quantified the impact of incident duration on the likelihood of a secondary

crash analyzed a limited dataset of 741 incidents that occurred on the Norman Expressway in Indiana. Using a logistic regression model, the authors reported an odds ratio of 1.028 for the incident clearance time variable. This indicates that for each additional minute of clearance time, the odds of a secondary crash occurring increased by 2.8%. It is worth noting that the study described this as a 2.8% increase in "likelihood," but this terminology is not technically accurate: in logistic regression, the odds ratio represents the multiplicative change in the odds (i.e., the ratio of the probability of a secondary crash to the probability of no secondary crash), not the direct probability itself [1].

While the Indiana study focused specifically on the impact of incident duration, several other studies have employed parametric models to investigate a broader set of factors influencing secondary crash likelihood. One study developed a Bayesian random effect logit model using real-time traffic data to estimate the likelihood of secondary crashes. The inclusion of dynamic traffic features—such as lane-level speed and volume variations—significantly improved the model's accuracy, underscoring the importance of real-time data in crash risk estimation [28]. Another study used a Bayesian complementary log-log model to estimate secondary crash likelihood, with input features selected using Random Forest. Key variables included occupancy, lane closures, and incident clearance duration, allowing for inferences about hazard rates associated with varying traffic and incident conditions [29]. Structural Equation Modeling (SEM) was applied in a different study to examine the underlying relationships between driver behavior, vehicle condition, environmental factors, and secondary crash occurrence. In combination with a multinomial logit model and crash modification factor estimation via negative binomial regression, the study provided insight into the causal and contributory factors of rear-end secondary crashes [22]. Zhang et al. [23] employed a binary logit model to predict the occurrence of secondary crashes and a hierarchical ordered probit model to assess injury severity. These parametric models revealed that daylight, young drivers, and weather conditions such as snow significantly increased the likelihood of secondary crashes, while factors like alcohol use and vehicle type were associated with injury severity. Additionally, survival analysis models—including the Proportional Hazard (PH) and Accelerated Failure Time (AFT) models—were used to examine the duration between primary and secondary crashes. The models quantified how variables such as peak hour traffic, lane and shoulder closures, and traffic volume affected both the likelihood and timing of secondary crashes [30].

Non-Parametric Models

Several studies employed non-parametric or machine learning-based approaches to model the likelihood, timing, or location of secondary crashes, prioritizing predictive accuracy and handling of high-dimensional data over direct interpretability. One study used Random Forest models to predict the time and distance gaps between primary and secondary crashes, followed by SHAP (Shapley Additive Explanations) to interpret the influence of variables. Results showed that traffic volume, speed, lighting, and population density were stronger predictors than primary crash features, with Random Forest outperforming KNN and multilayer perceptron regression [25].

Another study proposed a hybrid machine learning framework that combined two XGBoost models—one for predicting whether a crash would lead to a secondary crash and another for estimating the likelihood of a secondary crash occurring. By integrating both outputs, the hybrid model achieved a high AUC (area under the curve) of 0.89 and maintained strong sensitivity with minimal false alarms, demonstrating its value for real-time applications [24].

A state-wide study applied association rule mining (ARM) alongside Random Forest and the Boruta algorithm to detect patterns and select relevant features related to secondary crash severity. The analysis showed that most secondary crashes occurred within 30 minutes of a primary crash and identified peak hour traffic and roadway type as important predictors [31]. Next, Chen et al. proposed a generative and predictive hybrid model, VarFusiGAN-Transformer, to

predict both the occurrence probability and spatiotemporal distribution of secondary crashes. The model used multilayer perceptrons and long short-term memories (LSTMs) to synthesize static and dynamic inputs, while the transformer architecture enabled powerful sequence modeling. This model achieved outstanding classification and regression performance, surpassing traditional generative adversarial network (GAN) variants in sensitivity and balance between false positives and false negatives [32].

Literature Review Summary

The body of research on secondary crash identification and modeling has evolved significantly over the past decade. Early work focused on static identification methods, relying on fixed time-distance thresholds, while more recent studies emphasized dynamic identification through traffic data such as speed contour plots. These data-driven approaches, particularly those using high-resolution speed profiles and spatiotemporal filtering, have become the dominant method for identifying secondary crashes due to their ability to reflect real-time traffic conditions.

Modeling efforts have also shifted over time. Earlier studies relied on parametric statistical models (e.g., logit, probit, survival analysis), which were suitable for inference and interpretation. In recent years, the focus has turned toward machine learning (ML) approaches, such as Random Forest, XGBoost, and deep learning models like LSTM and GAN-based frameworks, which offer superior predictive accuracy at the expense of interpretability. Explainability tools such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) have helped bridge this gap.

Almost all modeling studies were conducted on limited, region-specific datasets, typically ranging from 3,000 to 25,000 incidents, often from single corridors or metro areas. Only a few studies adopted a multi-state perspective, mainly concerned with assessing data quality and performing descriptive analyses, rather than inference and predictive modeling.

One key output of the literature review is the synthesis of variables used in statistical and ML models to predict secondary crash likelihood or severity. These variables span multiple dimensions. **Table 1** provides a summary of common variables used in secondary crash likelihood analysis.

Table 1. Summary of common variables used in secondary crash likelihood analysis

Category	Variables
Traffic Characteristics	Traffic volume, speed, lane-level speed differences
Incident Characteristics	Severity and type of primary crash, incident duration/incident clearance time
Roadway Geometry	Number of lanes, road width, curvature, intersections, road surface conditions, horizontal alignment, road geometry
Driver & Vehicle Characteristics	Vehicle type, vehicle condition, service year, defects
Environmental Conditions	Weather, lighting condition (daylight, strong light), visibility, snow depth, wind speed, temperature, humidity, surface condition (wet, slush, oiled)
Temporal Features	Time of day, day of week, month, sunrise/sunset status
Spatial Features	Proximity to intersection, proximity to billboards/trees, speed limit, intersection presence, traffic signal presence

Data Evaluation and Processing

Incident Data

CATT Lab receives a wide range of event data from different agencies, each using its own system for incident reporting and data management. For this project, four states—Maryland (MD), Virginia (VA), Tennessee (TN), and Florida (FL)—were selected for analysis. Each state maintains its own incident data collection process and shares data with the CATT Lab at different levels of completeness and granularity. As a result, data from each state were stored in separate schemas within the CATT Lab databases and evaluated individually for the availability of key features relevant to secondary crash analysis, such as incident start and end times, location accuracy, lane closure details, and responder actions.

The incident data used in this project covered 2022-2024, providing a recent and sufficiently large window for robust analysis across the four study states. To our knowledge, this is the largest secondary crash likelihood analysis conducted to date.

Once the CATT Lab receives incident records, they are spatially aligned to Traffic Message Channels (TMCs) based on their reported geolocation and direction of travel. This alignment is a critical feature that greatly facilitates data fusion, as many other traffic datasets—such as speed, volume, and probe-based measurements—are indexed at the TMC level. Snapping incidents to the TMC network allows these datasets to be consistently merged and analyzed within a unified spatial framework.

Among the different event types available in these schemas, this study focused on records classified as incidents lasting less than 10 hours. These represent unplanned events likely to cause traffic disruption and are most relevant to secondary crash detection.

It is important to note that the Maryland incident data used in this study is limited to records reported by the Coordinated Highways Action Response Team (CHART), which primarily covers incidents occurring on National Highway System (NHS) routes. While other states provided incident data across all roadway types, the broader Maryland State Police data—although more comprehensive in coverage—was excluded from analysis due to the absence of incident start and end times received by the CATT Lab. Since temporal information is essential for secondary

crash detection and timing analysis, only the CHART-sourced incidents were retained for Maryland in this project.

One major finding from the incident data evaluation was that the accuracy of the geolocation fields reported in the datasets was questionable. Across all four states, the number of unique geolocations was significantly lower than the number of incident records, suggesting that many incidents were repeatedly assigned to the same coordinates. These repeated locations often corresponded to highway access ramps, plazas, or other familiar reference points, indicating limitations in how incident locations are captured or reported. In some cases, dispatch procedures or data entry tools may have defaulted to common locations rather than pinpointing the actual incident site.

This lack of spatial precision presents challenges for analyses that rely on accurate location data—for example, in identifying secondary crashes or fusing with weather and roadway datasets. However, this limitation has less impact on features reported at the TMC level, since those are aligned to standardized roadway segments rather than incident-specific coordinates.

Table 2 presents the total number of incident records received from each state for 2022–2024, the subset with valid start and end times used in the analysis, and the percentage of unique geolocations relative to total incidents. This percentage indicates how frequently incident locations were repeated within each dataset.

Table 2. Incident Counts, Temporal Validity, and Geolocation Uniqueness by State.

State	Total Incident Records	Incidents With Valid Start/End Time	% Unique Geolocations (out of all incidents)
Maryland	218,143	214,008	32.88%
Virginia	411,086	397,964	17.89%
Tennessee	294,967	268,952	5.36%
Florida	1,943,489	1,849,575	2.92%

Incident Data Processing

Once the incident databases were explored and the availability of relevant features was assessed, the next step was to extract and process data in a format suitable for statistical analysis. Four main tables were used for this purpose: the event table, responder table, lane table, and vehicle table. Each table required tailored processing to consolidate and engineer features at the incident (*event_id*) level.

Event Table

The event table includes one row per incident, identified by a unique *event_id*, which serves as the primary key. The first step was to filter the table to include only records classified as incidents, excluding other event types such as work zones, weather-related events, and planned closures. Once the relevant *event_ids* were identified, key features were extracted directly from the table. These features included the start and end times of the incident, geolocation, road weather conditions, lighting, and general weather conditions, where available. Since each *event_id* appears only once in this table, extracting these attributes was straightforward.

Responder Table

The responder table contains one row for each responding unit associated with an incident, meaning multiple rows can exist for a single *event_ids*. This table was processed to aggregate responder-related attributes to the incident level. Specifically, the total number of responders per

incident was calculated. In addition, binary indicators were created to flag the presence of specific types of responders during the incident. These types included, but were not limited to: transportation response units, fire units, state police units, light tow units, local police units, freeway service patrols, emergency vehicle units, HAZMAT units, private contractor response units, and others.

Lane Table

Similar to the responder table, the lane table contains multiple rows per incident, with each row representing an update to the lane configuration during the event. The table includes information about lane types, enabling the distinction between travel lanes and shoulder lanes. To extract relevant features, the data was aggregated at the incident level.

The primary features derived from this table were the existence of shoulder lane closures and the average capacity reduction for both travel and shoulder lanes. For each incident, the total closure time was calculated separately for travel lanes and shoulder lanes. These closure times were then normalized by the total duration of the incident and the total number of available lanes of each type, resulting in consistent measures of proportional capacity reduction across incidents.

Vehicle Table

The vehicle table contains one row for each vehicle involved in an incident, meaning multiple rows can be associated with a single *event_ids*. This table was processed to extract vehicle-related features at the incident level. Specifically, the total number of vehicles involved in each incident was calculated. Additional counts were generated for specific vehicle types such as passenger cars, motorcycles, commercial vehicles, and buses, where available. It is important to note that vehicle-level data were only available in the CATT Lab databases for Maryland and Florida**.**

Incident Data Summary

To summarize the findings of the incident data evaluation and processing across the four states, **Table 3** lists the key features expected from incident data and indicates the availability of each feature in Maryland, Virginia, Tennessee, and Florida data sets that could be readily queried.

Table 3. Summary of Incident Data Availability by State

Feature	Maryland	Virginia	Tennessee	Florida
Start / End time of incident	✓	✓	✓	✓
Location	✓	✓	✓	✓
Responder data	✓	✓	✓	✓
Lane information	✓	✓	✓	✓
Road weather conditions			✓	✓
Lighting			✓	✓
Weather condition			✓	✓
Vehicles involved	✓			✓

Speed Data

The speed data used in this study consisted of probe-based traffic speed measurements at the TMC level, accessed through the CATT Lab platform. These data are sourced from the Regional Integrated Transportation Information System (RITIS), which provides access to raw probe speed data via the Massive Data Downloader (MDD) interface available in the Probe Data Analytics (PDA) suite.

Due to the nature of speed data—which are streamed in real time for all roadway segments at regular intervals—the MDD is optimized for querying large volumes of data across fixed time windows and predefined segment lists. However, incidents are distributed irregularly in space and time, making the default query structure inefficient for secondary crash analysis. This project implemented custom queries written directly to the probe data API to overcome this challenge. These queries were designed to extract 5-minute aggregated speed data specific to each incident over its impacted TMC segments and during a configurable time window surrounding the incident period.

This tailored querying approach allowed the project team to efficiently retrieve the precise spatiotemporal slices of speed data needed to assess traffic conditions before, during, and after each incident, supporting dynamic identification of secondary crashes.

The retrieved speed data were specifically processed for secondary crash identification. This processing workflow is described in detail in this report's Secondary Crash Identification section.

Volume Data

Unlike speed data, real-time volume data are only available at specific locations within a transportation network where fixed sensors have been installed to record vehicle counts. As a result, most volume-based analysis in transportation relies on historical data sources and annual averages such as the average annual daily traffic (AADT). To enable time-specific analysis, AADT can be disaggregated using a method known as profiling [33], which estimates typical vehicle counts in 15-minute intervals across a standard week for each segment.

This study used profiling volume data from the National Performance Management Research Data Set (NPMRDS) as the primary source of traffic volume. This dataset, which is updated annually, provides volume profiles for segments on the NHS. For incidents occurring on non-NHS routes, INRIX 2019 profiling volumes were used when available to supplement the NPMRDS data.

This combined approach ensured that a consistent and comprehensive volume estimate was available for all analyzed incidents, allowing traffic demand to be considered in conjunction with speed and incident characteristics during secondary crash identification and modeling.

Once the appropriate volume data source was determined for each incident and data availability was confirmed, 15-minute profiling volumes were extracted for the entire incident duration, from start to end. These values were then aggregated to calculate an average volume for the incident period. In addition to the average, volume values at the exact start and end times of the incident were also retained separately, allowing for a more granular representation of traffic conditions at incident boundaries.

Radar Weather Data

In addition to the weather-related fields available in the incident datasets for some states, this project incorporated radar-based weather data as a consistent and comprehensive source across all four study states. This external weather data was used to ensure uniform coverage of precipitation conditions, regardless of variations in incident reporting practices.

The CATT Lab has developed an API to ingest and serve weather data from the National Oceanic and Atmospheric Administration (NOAA). The data, originally provided in raster format (gridded pixels), is processed and mapped to road segments using the TMC network. This mapping allows seamless integration with other TMC-based datasets used in the study.

The API delivers data from NOAA's Multi-Radar Multi-Sensor (MRMS) feed at 2-minute intervals, including attributes such as precipitation type and precipitation rate. Only segments with precipitation are returned. If no data is returned for a segment at a specific timestamp, it is

assumed that no precipitation occurred. A value of -99 was returned in cases where data is missing from the NOAA archive.

Because the weather data was already provided at the segment level, it was easily fused with incident records and traffic data to support analysis of environmental factors associated with secondary crashes.

Weather data were processed to generate incident-specific features that capture precipitation conditions before and during each event. Two clusters of time windows were defined for querying the API: one for the period before the incident and another for the duration of the incident itself.

The before-event cluster aggregated precipitation records at several lookback intervals: 2 minutes, 30 minutes, 1 hour, 3 hours, 6 hours, 12 hours, and 24 hours before the incident start time. The during-event cluster covered the full time span between the incident's start and end timestamps.

For each time window, the following nine attributes were computed:

- Precipitation percentage: the percentage of the time window during which precipitation was recorded
- Maximum precipitation rate: the highest precipitation rate observed
- Minimum precipitation rate: the lowest precipitation rate observed
- Average precipitation rate: total precipitation divided by the duration of the time window
- Snow flag: true if snow was recorded as a precipitation type
- Hail flag: true if hail or mixed rain and hail was recorded
- Rain flag: true if any other precipitation type (e.g., rain) was recorded
- Data gap flag for rate: true if negative values (e.g., -99) were recorded for precipitation rate
- Data gap flag for type: true if negative values were recorded for precipitation type

Segment Data

Segment-level roadway characteristics used in this study were primarily derived from the HERE TMC map, which provides standardized segment definitions across the road network. The HERE dataset includes essential attributes such as segment geometry, direction of travel, and functional road class. OpenStreetMap (OSM) was used to supplement this base data to extract the number of lanes associated with each segment. An internal process was developed to map OSM road segments to TMC segments, allowing the number of lanes to be assigned to each TMC with high spatial accuracy.

Additionally, speed limit data from the MAP-21 dataset was incorporated where available. Although MAP-21 includes segment-level speed limits, its coverage is limited and not uniform across all regions. As a result, it was used selectively to supplement other segment characteristics when available.

Secondary Crash Identification

Considering the size and complexity of the incident dataset analyzed in this study, a hybrid method for secondary crash identification was developed to support large-scale processing across multiple states. **Figure 1** presents an overview of the identification method.

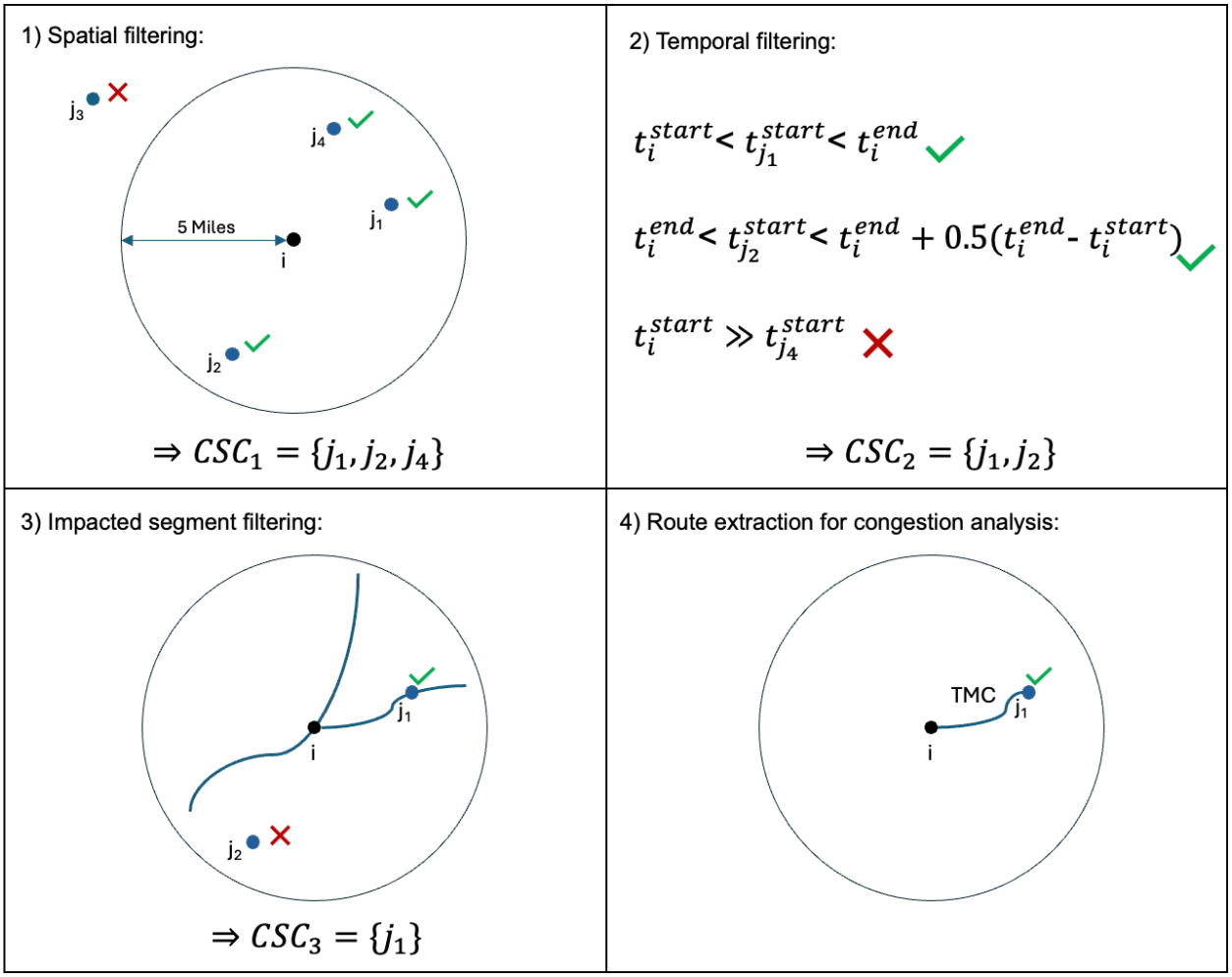


Figure 1. Secondary Crash Identification Method Overview.

Spatial and Temporal Filtering

As shown in **Figure 1**, the first step was to conduct a spatial search for each incident $i \in I$, where I is the set of all incidents analyzed. A 5-mile radius ($R = 5 - miles$) was used to search for nearby incidents. All incidents that fell within this radius were considered initial candidates. The set of these candidate incidents was labeled as CSC_1.

Next, a temporal filter was applied based on equation (1) to determine whether the start time of each candidate incident fell within a relevant window relative to incident i. Specifically, a candidate incident $j \in CSC_2$ was retained if its start time sc_j fell between the start time sc_i and end time ec_i of incident i plus 0.5 times its duration were retained for further consideration. The set of incidents that satisfied this condition was labeled CSC_2.

$$st_i \leq st_j \leq et_i + 0.5(st_i - et_i) \;\; \forall \, j \in CSC_2 \qquad (1)$$

In the third step, the network-level relationship between the primary incident i and each candidate $j \in CSC_2$ was evaluated. For this purpose, the list of TMC segments impacted by i was extracted in both the direction of travel and the opposite direction. The tracing algorithm developed for this study was capable of capturing complex network structures, including branches from ramps and parallel connectors, allowing for a realistic representation of how congestion may spread from the incident location.

From all candidate secondary crashes ($j \in CSC_2$), only those located on segments identified as potentially impacted by the primary incident were retained for further filtering. This ensured that spatial proximity considered the actual traffic flow, not just Euclidean distance. The set of these incidents was labeled as CSC_3. In this step, each primary incident could be associated with multiple secondary crashes; however, each secondary crash was assigned to only one primary. In cases where a secondary crash had more than one potential primary incident, the primary that was spatially closest, based on the calculated graph distance, was selected.

Speed-Based Filtering

Following the spatial, temporal, and network-level filters applied in earlier steps, an additional analysis stage was performed using speed data to evaluate the congestion impact of the primary incident. This step aimed to confirm that the candidate secondary crashes occurred in the presence of measurable disruption in traffic flow, as indicated by reduced travel speeds. The process involved extracting and analyzing probe speed data along the route between the candidate primary-secondary pairs.

Let P denote the set of all candidate primary-secondary crash pairs identified after spatial, temporal, and network-level filtering. For each $p \in P$ crash pair that passed the earlier filters, the following steps were performed:

Segment Identification and Speed Data Extraction

Let TMC_p represent the ordered set of TMC segments that form the route between the primary and secondary crash in pair p. For each segment $tmc \in TMC_p$, two types of speed data were extracted at 5-minute intervals from the start time of the primary crash to the start time of the secondary crash:

- S_{tmc}^t: the observed speed on segment tmc during interval t
- H_{tmc}^t: the historical speed on the same segment during the same time interval

In parallel, the total distance between the candidate primary and secondary crash locations was calculated as the sum of the lengths of the TMCs forming the route ($tmc \in TMC_p$), with adjustments for the offset positions of the crash locations within their respective segments.

Speed Aggregation by Time Intervals

For each time interval t, speeds were aggregated using the harmonic mean weighted by segment length:

- Observed speed averaged across the route was calculated based on equation (2):

$$Observed\ Speed_t^p = \frac{\sum_{tmc \in TMC_p} L_{tmc}}{\sum_{tmc \in TMC_p} \frac{L_{tmc}}{S_{tmc}^t}} \quad (2)$$

- Historical speed averaged across the route was calculated based on equation (3):

$$Historical\ Speed_t^p = \frac{\sum_{tmc \in TMC_p} L_{tmc}}{\sum_{tmc \in TMC_p} \frac{L_{tmc}}{H_{tmc}^t}} \quad (3)$$

Where:

- S_{tmc}^t: the observed speed on segment tmc during interval t
- H_{tmc}^t: the historical speed on the same segment during the same time interval
- L_{tmc}: the length of segment tmc

Speed Reduction Metrics

For each interval t in pair p, the following metrics were computed:

- Speed change:
- $Speed\ Change_t^p = Observed\ Speed_t^p - Historical\ Speed_t^p$ (4)
- Speed change percentage:
- $Speed\ Change\ Percentage_t^p = \frac{Speed\ Change_t^p}{Historical\ Speed_t^p} \times 100$ (5)

Temporally Aggregated Results

For each $p \in P$ cross all intervals between the primary and secondary crashes in pair p, the following were calculated:

- Overall average speed change:
- $Average\ Speed\ Change^p = \frac{1}{T^p}\sum_{t \in T^p} Speed\ Change_t^p$ (6)
- Overall average speed change percentage:
- $Overall\ Average\ Speed\ Change\ Percentage^p = \frac{1}{T^p}\sum_{t \in T^p} Speed\ Change\ Percentage_t^p$ (7)
- Speed change and speed change percentage for the specific time interval during which the secondary crash occurred, denoted as:
- $Speed\ Change_{ts}^p$ and $Speed\ Change\ Percentage\ _{ts}^p$

Where:

- T^p is the number of 5-minute time intervals between the primary and secondary crash p start times.
- ts is the time interval when the start time of the secondary crash.

Once the relevant speed metrics for the route between each candidate primary-secondary crash pair were calculated, these values were used to further filter the set P. This filtering aimed to retain only those pairs in which the secondary crash was more likely to have occurred due to the traffic disruption caused by the primary crash. Specifically, candidate pairs that did not exhibit reductions in speed were excluded from modeling and descriptive analysis.

The next part of this section presents summary statistics and distributions of the candidate primary-secondary crash pairs across the four states analyzed in this study.

Identification Results

Table 4 summarizes the candidate primary-secondary crash pairs identified through the hybrid filtering framework described earlier. For each state in the study, the table includes:

- The number of candidate primary crashes that were associated with at least one potential secondary crash on their potentially impacted segments (CSC_3)
- The number of candidate pairs that showed a speed drop along the route between the primary and secondary crash
    - $Overall\ Average\ Speed\ Change^p < 0$ and $Speed\ Change_{ts}^p < 0$

To further analyze the impact of incident timing, the counts of candidate pairs with observed speed drops are reported separately for two categories:

- Pairs where the secondary crash occurred during the clearance time of the primary crash
- Pairs where the secondary crash occurred after the primary crash had ended, but within a window equal to 50 percent of the primary crash duration (referred to as the recovery time in this study).

Table 4. Summary of Candidate Primary-Secondary Crash Pairs and Speed Drop Conditions

State	Candidate Pairs Count	Pairs with Speed Drop (During Clearance Time)	Pairs with Speed Drop (During Recovery Time)
Maryland	28,572	8,093 (28.32%)	4,276 (14.97%)
Virginia	52,310	12,968 (24.79%)	6,889 (13.17%)
Tennessee	98,958	20,605 (20.82%)	8,882 (8.98%)
Florida	637,362	80,848 (12.68%)	23,053 (3.62%)

Figure 2 presents the distribution of speed changes to better illustrate how traffic conditions changed for the candidate pairs that experienced a speed drop. It shows both the average speed reduction over the duration between crashes and the speed reduction during the secondary crash. These distributions are shown separately for pairs that occurred during the clearance time and those that occurred during the recovery time.

According to **Figure 2**, the distribution of speed drop—both in terms of overall average and at the specific time of the secondary crash—is consistent across all four states. In all cases, speed reductions ranged from 0 to nearly 100 percent. One notable pattern is that the speed drop at the time of the secondary crash is generally higher than the average speed drop over the full interval between the primary and secondary events. For example, among candidate pairs occurring during the clearance time of the primary crash, approximately 20 percent experienced an average speed drop of more than 40 percent. In contrast, around 30 percent experienced a speed drop greater than 40 percent, specifically at the time of the secondary crash.

Another important takeaway from this figure is that the choice of a speed drop threshold for filtering candidate pairs directly impacts the number of pairs that remain classified as valid primary-secondary crashes. The higher the threshold, the more confident we can be that the secondary crash was indeed influenced by congestion caused by the primary incident. In particular, for pairs in which the secondary crash occurred during the recovery time, it is reasonable to expect that a more substantial traffic impact would be required for a causal connection to exist. Therefore, applying a stricter speed drop threshold to these cases is justified and may help improve the accuracy of secondary crash identification.

In this study, all candidate pairs in which the secondary crash occurred during the clearance time of the primary crash were considered valid if there was any measurable speed drop along the route, applying a threshold of 0 percent. For pairs where the secondary crash occurred after the primary crash had ended—referred to as recovery time—a more conservative threshold of 10 percent speed reduction was applied to ensure a stronger indication of congestion impact.

For candidate pairs where the secondary crash occurred downstream of the primary crash, no threshold on speed drop was applied. Instead, an alternative filter was used to exclude pairs in which the secondary crash occurred more than 0.5 miles downstream of the primary location. This filter accounts for the fact that downstream secondaries are typically caused by rubbernecking or driver distraction, effects that are unlikely to persist beyond approximately one minute of travel time—roughly equivalent to 0.5 miles at a speed of 30 miles per hour.

As noted previously, the reported geolocation of incidents is prone to error, with many crashes recorded at identical or repeated coordinates. To account for this uncertainty and take a conservative approach, the same speed drop thresholds applied to upstream secondary crashes were also applied to those reported at the same location as the primary crash.

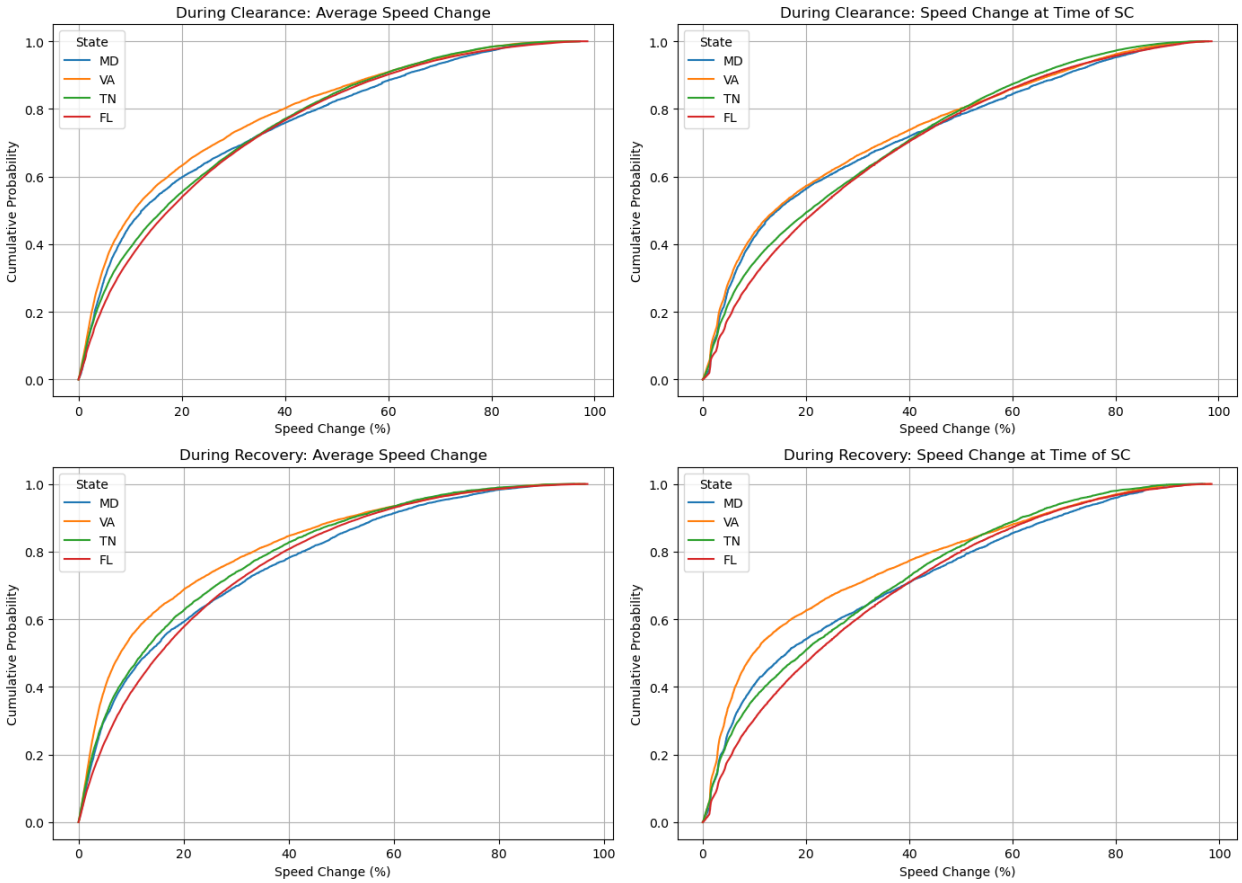


Figure 2. Distribution of Speed Reductions for Candidate Primary-Secondary Crash Pairs by State and Timing Window

Based on these criteria, **Table 5** reports the percentage of unique primary incidents that led to at least one qualifying secondary crash, as well as the percentage of total crashes identified as valid secondaries, for each of the four states.

Table 5. Summary of Candidate Primary-Secondary Crash Pairs and Speed Drop Conditions

State	Total Incidents	Pct Primary Event	Pct Secondary Event
Maryland	214,008	3.86%	3.93%
Virginia	397,964	2.81%	3.07%
Tennessee	268,952	6.5%	7.5%
Florida	1,849,575	4.14%	4.81%

To explore the spatial and temporal relationships within the final set of selected primary-secondary crash pairs, **Figure 3** presents the distributions of the time difference between the start times of the primary and secondary crashes, and the distance between them, across all four states. According to this figure, all four states exhibit similar patterns in the distribution of spatial and temporal gaps between selected primary-secondary crash pairs, with Maryland showing relatively smaller gaps on average. In terms of spatial proximity, the share of secondary crashes occurring within 1 mile of the primary crash ranges from about 50 percent in Florida to around 70 percent in Maryland. For temporal proximity, approximately 40 percent of secondary crashes in

Florida, Virginia, and Tennessee occurred within 30 minutes of the primary crash, compared to nearly 70 percent in Maryland. These results suggest that in Maryland, selected secondary crashes tend to occur more quickly and in closer proximity to the associated primary events.

To further explore the joint relationship between spatial and temporal proximity of secondary crashes, **Figure 4** presents a scatter plot of time difference versus distance between the primary and secondary crash for the final set of selected primary-secondary crash pairs across all states. Overall, there is a weak positive correlation, indicating that secondary crashes occurring later tend to happen farther from the primary crash location. This trend is consistent with the notion that secondary crashes are likely to occur within the congestion queue that builds and propagates upstream over time. The positive correlation is slightly stronger in Maryland, which may be attributed to the fact that the Maryland data is limited to the NHS—comprised of higher-volume roadways—where the impact of a primary crash on queue formation and growth is generally more pronounced.

Two important considerations should be noted. First, although the initial radius search for candidate secondary crashes was limited to five miles based on Euclidean distance, the final matched pairs may reflect graph-based distances greater than five miles due to network routing and segment geometry. Second, while the distances reported in this study are accurately calculated using graph distance and adjusted for segment offsets, they may still be affected by inaccuracies in the reported geolocations of incidents. As discussed earlier, repeated or imprecise location reporting may introduce spatial uncertainty that should be considered when interpreting these values.

To further inform the interpretation of secondary crash timing, **Figure 5** presents a plot of primary incident duration versus the time difference between the start of the primary crash and the start of the associated secondary crash for the final set of matched primary-secondary crash pairs.

This visualization helps investigate how far into the timeline of the primary crash the secondary crash tends to occur. A 45-degree reference line is included to distinguish between secondary crashes that occurred during the clearance time of the primary crash (points to the left of the line) and those that occurred during the recovery window—defined as up to 50 percent of the primary incident's duration after its end time (points to the right of the line).

According to **Figure 5**, in all four states, secondary crashes can occur at any time during the clearance period of the primary crash and even during the recovery period that follows. However, the density of observations—represented by the shading in the heat map—is higher closer to the start of the primary incident. This pattern is particularly evident for longer-duration incidents, where the shading visibly fades as time progresses. This suggests that while secondary crashes may occur throughout the clearance and recovery windows, they are more likely to occur closer to the start of the primary incident than toward its end or beyond.

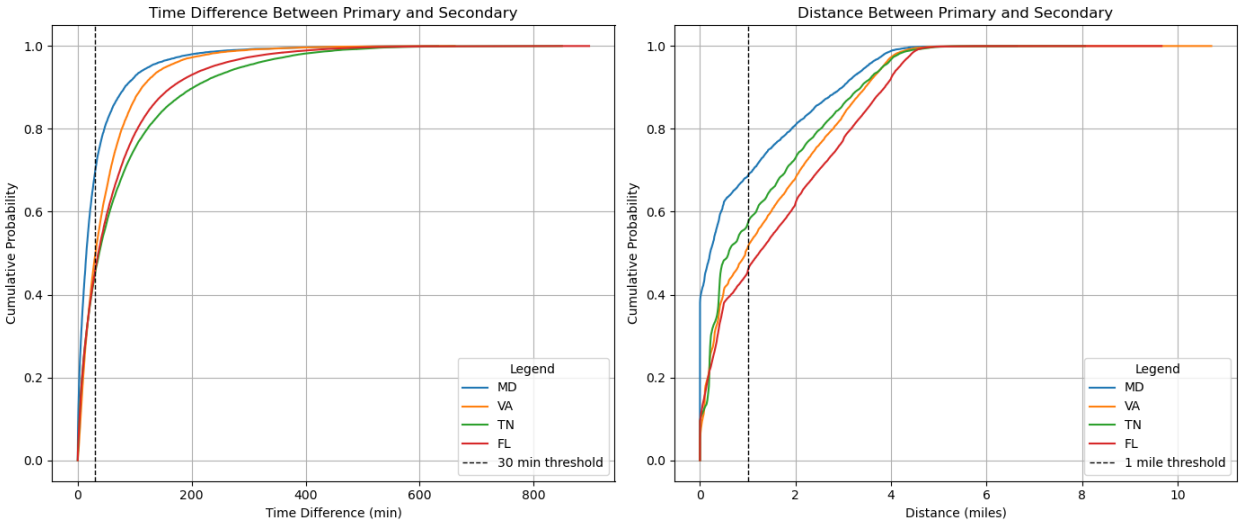


Figure 3. Distribution of Temporal and Spatial Gaps Between Selected Primary-Secondary Crash Pairs by State

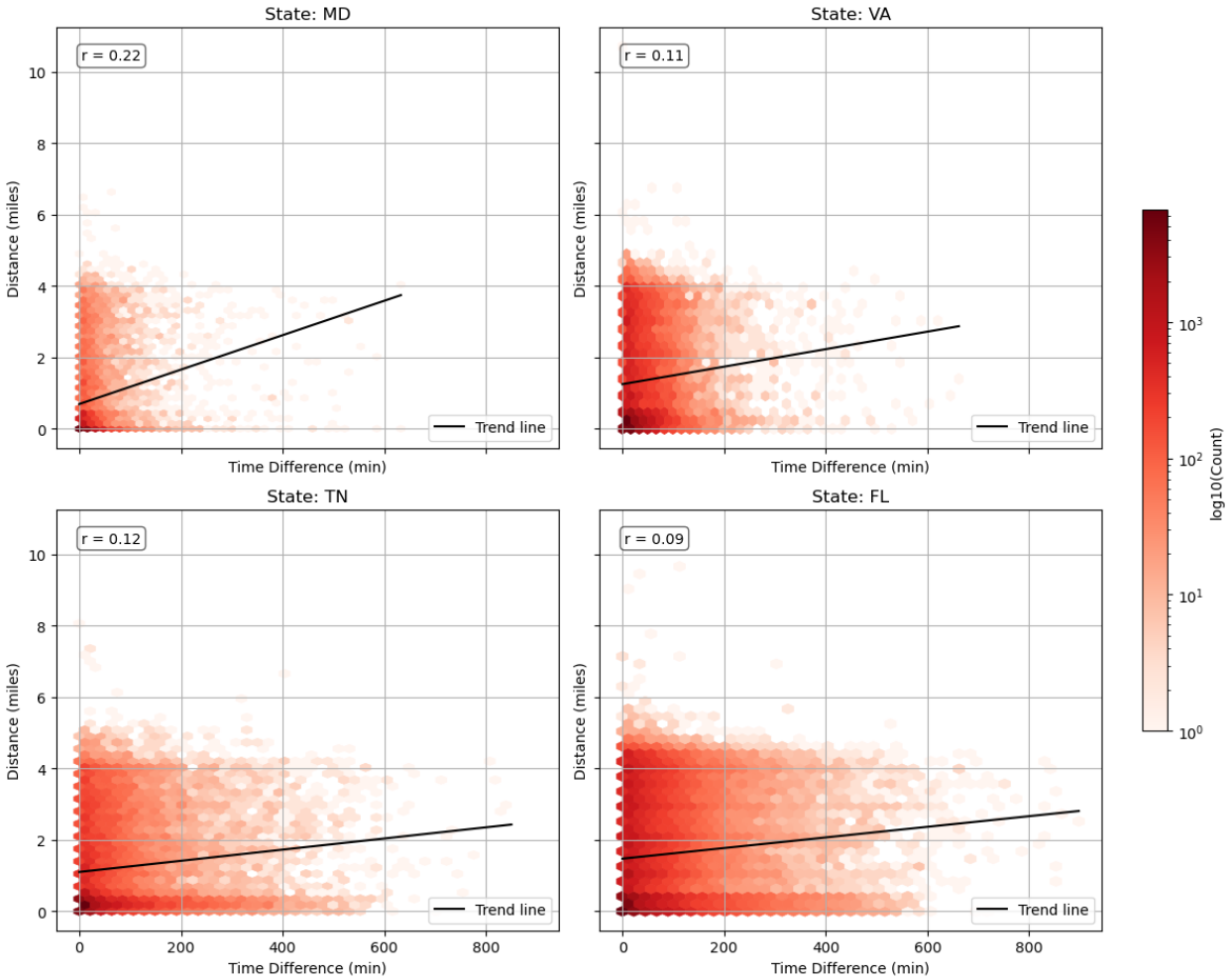


Figure 4. Time Difference vs Distance of Primary-Secondary Pairs (with Trend & Correlation)

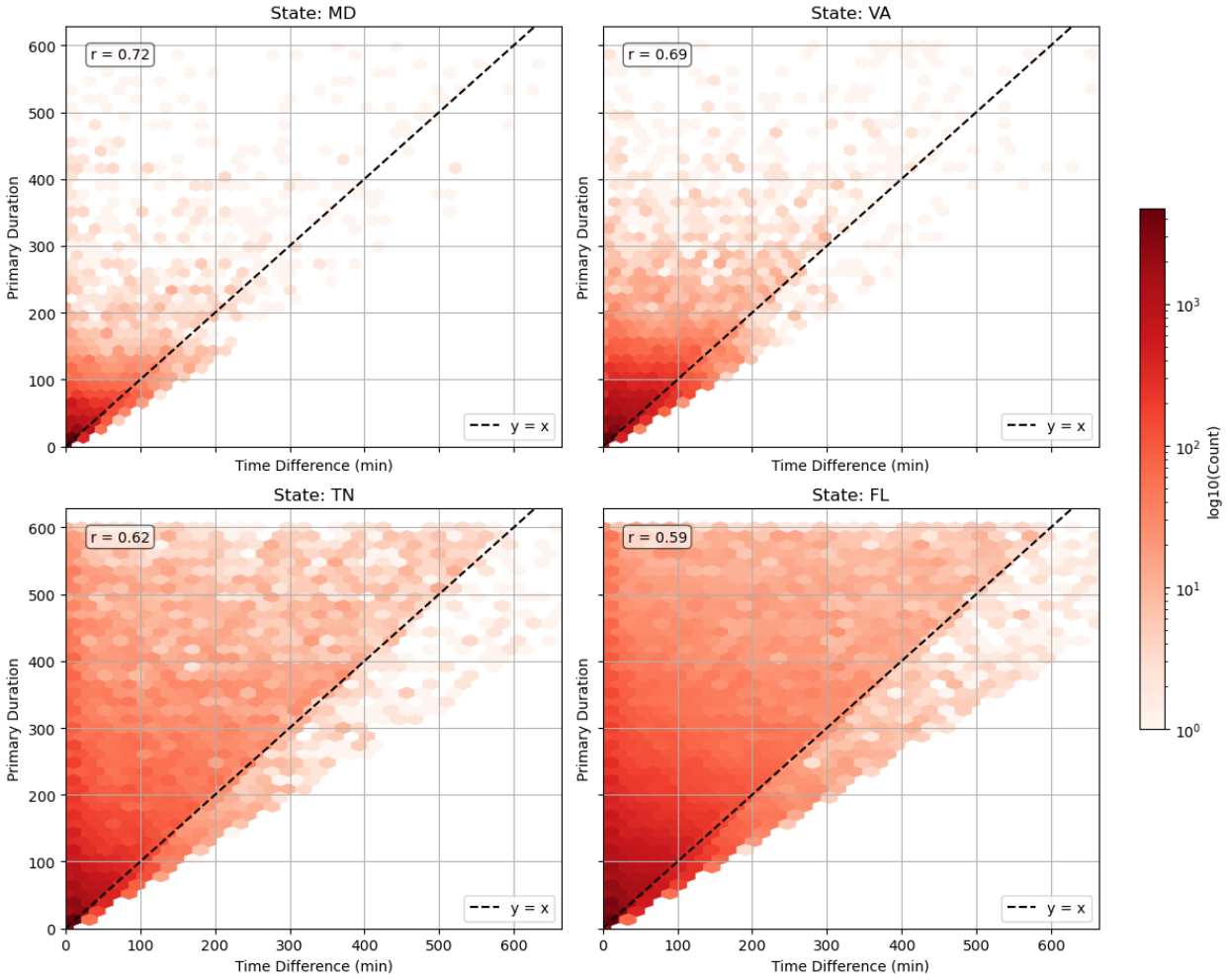


Figure 5. Duration of primary vs. Time Difference between Primary-Secondary Pairs (with 45-degree reference line)

Model Development

This study employed logistic regression, one of the most widely used statistical models in crash prediction and, more specifically, in secondary crash analysis. As a parametric model, logistic regression is particularly well-suited for the inference objectives of this study. The primary goal is not only to predict the occurrence of secondary crashes but also to quantify the impact of key factors—such as incident duration, incident characteristics, roadway geometry, and environmental conditions—on the likelihood that a secondary crash occurs.

Logistic regression is used to model a binary outcome—in this case, whether a secondary crash occurred (1) or did not occur (0). The model estimates the probability of the outcome as a function of a set of independent variables. It does so by modeling the log-odds of the outcome as a linear combination of the predictor variables:

$$log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k \qquad (8)$$

Where:

- $P(Y = 1)$ is the probability of a secondary crash

- $X_1, X_2, X_3, \ldots, X_k$ are the predictor variables (e.g., duration of the primary incident, number of responders, weather condition, etc.)
- β_0 is the intercept and $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ are the model coefficients.

The estimated coefficients can be exponentiated to yield odds ratios, which represent the multiplicative change in the odds of a secondary crash for a one-unit change in the corresponding predictor variable. This interpretability makes logistic regression especially useful for understanding the role of individual features in secondary crash occurrence, beyond their predictive power alone.

Based on the data evaluation and processing steps described earlier, a comprehensive set of independent variables was developed for use in the logistic regression model. **Table 6** presents all variables considered for inclusion in the logistic regression model. The table provides a short description for each variable, the source dataset (incident data, volume data, radar weather data, or segment data), and the percentage of records for which the variable was available in each of the four states. This availability assessment was used to guide variable selection and to ensure consistency in the modeling process across states.

Please note that the availability percentages are based on the filtered incident datasets used for analysis, excluding records without valid geolocation or start/end time, as described in **Table 2**. If a variable is reported as *NA* for a given state, it indicates that the variable was either not available in the incident data provided to CATT Lab, had no variability (i.e., the same value for all observations), or was reported in a format unsuitable for modeling (such as free-text fields or categorical variables with too many unique values).

Logistic Regression Assumptions

An important step in developing a logistic regression model is verifying that the model assumptions are reasonably satisfied. The key assumptions include: (1) independence of observations, (2) a binary (or ordinal) dependent variable, (3) linearity of the independent variables with the log-odds of the outcome, and (4) absence of strong multicollinearity among the independent variables. In this study, the assumption of independence is considered to be met, as traffic incidents are treated as independent events. The dependent variable—whether a given incident results in a secondary crash—is binary by design, satisfying the second assumption. To evaluate the remaining assumptions, standard diagnostic tests were performed to assess multicollinearity and to examine the linearity of continuous variables with respect to the log-odds. These checks are described in detail in the following subsection.

Table 6. Independent Variables Considered for Modeling and Their Availability by State

Variable	Source	Description	Percent Available (%)			
			MD	VA	TN	FL
Clearance Time	Incident Data (Event)	Duration between the start and end time reported for each incident	100.00	100.00	100.00	100.00
Day Type	Incident Data (Event)	Indicates whether the incident occurred on a weekday or during the weekend	100.00	100.00	100.00	100.00
Event Type	Incident Data (Event)	Type of incident, including values such as: disabled vehicle, serious accident, accident, incident, injury accident, medical emergency, vehicle on fire, multi-vehicle accident, disabled semi-trailer, abandoned vehicle, overturned vehicle, accident involving a pedestrian, jackknifed semi-trailer, brush fire, or none	100.00	99.84	100.00	100.00
Severity	Incident Data (Event)	Severity of the incident, categorized as: 'minor', 'intermediate', or 'major'.	NA	NA	100.00	100.00
Lighting	Incident Data (Event)		NA	NA	99.99	83.40
Number of Responders Category	Incident Data (Responders)	Total number of responders, categorized as: 1, 2–3, or more than 3.	90.30	43.34	90.78	87.58
Emergency Vehicle	Incident Data (Responders)	Indicates whether an emergency vehicle was dispatched to the scene or not.	NA	84.95	90.79	NA
Total Vehicles	Incident Data (Vehicle)	Total vehicles involved, categorized as: 1, 2–3, or more than 3	24.57	NA	NA	17.45
Truck Involvement	Incident Data (Vehicle)	Whether an incident involved a truck or not	24.57	NA	NA	17.45
Shoulder Lane	Incident Data (Lane)	Indicating whether a shoulder lane was present at the incident location and whether it was closed.	37.14	99.98	99.93	64.22
Capacity Reduction	Incident Data (Lane)	Calculated as the average lane closure time divided by the total lane time, and categorized as: 0%, 0–10%, 10–20%, 20–30%, 30–50%, and >50%.	37.14	99.98	99.93	64.22
Traffic Flow	Volume and Segment Data	Average expected hourly vehicle count per lane at the incident location during the clearance period, categorized as: '0–500', '500–1000', '1000–2000', and '>2000'.	99.01	98.04	99.79	87.51
Road Curvature	Segment Data	Road curvature, categorized as: 'Straight' or 'Curved'.	99.89	99.90	99.91	99.93
Functional Class	Segment Data	Functional road classification, a system used to group roads based on their intended service, with	99.89	99.90	99.91	99.93

Variable	Source	Description	Percent Available (%)			
			MD	VA	TN	FL
		class 1 being the highest and 5 being the lowest.				
Weather Status	Radar Weather Data	Indicates whether there was rain or snow during the incident clearance period.	100.00	100.00	100.00	100.00

Multicollinearity

Logistic regression requires little or no multicollinearity among the independent variables. This means that independent variables should not be highly correlated with each other. Detecting multicollinearity is important because while multicollinearity does not reduce the model's explanatory power, it does reduce the independent variables' statistical significance. The assumption can be verified with the variance inflation factor (VIF), which determines the correlation strength between the independent variables in a regression model. Cramér's V is another measure for verifying this assumption for categorical variables.

Variance Inflation Factor (VIF)

VIF is a measure of the amount of multicollinearity in regression analysis. A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables. The formula for VIF is:

$$VIF_i = \frac{1}{1-R_i^2} \quad (9)$$

Where R_i^2 is the unadjusted coefficient of determination for regressing the independent variable i on the other variables.

When R_i^2 is equal to 0, and therefore, when VIF or tolerance is equal to 1, the independent variable i is not correlated to the other variables, meaning that multicollinearity does not exist. In general:

- VIF equal to 1 = variables are not correlated.
- VIF between 1 and 5 = variables are moderately correlated.
- VIF greater than 5 = variables are highly correlated.

Cramér's V

Cramér's V correlation is used to measure the association between two categorical variables, and its value varies from 0 (stating no relationship between the variables) to 1 (stating complete association between the variables). It reaches a value of 1 only when an attribute is completely determined by the other attribute. Cramér's V is a normalized measure of association between two categorical variables derived from the Chi-square statistic, but unlike Chi-square, Cramér's V gives a standardized measure of strength. The formula for Cramér's V is:

$$V = \sqrt{\frac{\frac{\varkappa^2}{n}}{min\,(k-1,r-1)}} \quad (10)$$

Where $\varkappa^2$ is derived from Pearson's chi-square test, n is the total number of observations k and r are the number of categories of the two categorical variables.

Perfect Separation

Perfect separation in logistic regression occurs when one or more independent variables can perfectly distinguish between the outcome classes. This means the dependent variable can be completely predicted based on certain values of the independent variables. In such cases, the maximum likelihood estimation used to fit the model fails, as the estimated coefficients tend toward infinity to achieve perfect classification. This results in convergence issues, extremely large or undefined standard errors, and unreliable model outputs. To resolve this issue, the problematic variable was removed.

Once multicollinearity and separation checks were completed for all candidate variables, a final set of independent variables was selected for modeling in each state. The variable selection process ensured that the included predictors met logistic regression assumptions and retained sufficient variability and interpretability. **Table 7** presents the list of variables that qualified for use in the logistic regression models for each state, based on data availability, statistical checks, and relevance to the modeling objectives.

Table 7. Final Set of Independent Variables Used in Logistic Regression Modeling by State

Variable	MD	VA	TN	FL
Clearance Time	✓	✓	✓	✓
Day Type	✓	✓	✓	✓
Event Type	✓	✓	✓	✓
Severity	x	x	✓	✓
Lighting	x	x	✓	✓
Number of Responders Category	x	x	x	x
Emergency Vehicle	x	✓	✓	x
Total Vehicles	x	x	x	x
Truck Involvement	x	x	x	x
Shoulder Lane	x	x	x	x
Capacity Reduction	✓	✓	✓	✓
Traffic Flow	✓	✓	✓	✓
Road Curvature	✓	✓	✓	✓
Functional Class	x	x	x	x
Weather Status	✓	✓	✓	✓

Linearity of Continuous Variables with Log-Odds

A key assumption in logistic regression is that continuous independent variables exhibit a linear relationship with the log-odds of the outcome. In this study, most independent variables were modeled as categorical variables—either by nature (e.g., weather condition, day of week) or by discretization for interpretability and consistency across states. Therefore, this assumption does not apply to those variables.

The only continuous variable retained in its original form was clearance time, which is also a central variable of interest in this study. One of the main research objectives is to assess the marginal effect of unit increases in clearance time on the likelihood of a secondary crash. As such, verifying the linearity of clearance time with the log-odds of the response is necessary.

To assess this, the Box-Tidwell test was applied. This test evaluates whether the logit transformation of the outcome variable is linearly related to the continuous predictor variable. Specifically, the method augments the logistic regression model with an interaction term between the continuous variable and its natural logarithm. For clearance time (x), the model includes an additional term $x \times log\ (x)$ and the significance of this term is tested:

$$log\ \left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x + \beta_2 (x.log\ log\ (x)\ ) + \cdots \quad (11)$$

If the interaction term is statistically significant, this suggests a deviation from linearity. The test was conducted after fitting an initial logistic regression model using the final set of variables selected for each state.

The results of the Box-Tidwell test indicated that clearance time does not exhibit a linear relationship with the log-odds of a secondary crash. Given that clearance time is a key variable in this study, and the objective is to retain it as a continuous predictor, several transformations were explored to address this nonlinearity. Logarithmic and quadratic transformations of clearance time were tested, but did not resolve the nonlinearity issue across all states.

As an alternative, a piecewise modeling approach was adopted. Clearance time was segmented into intervals in which the variable demonstrated an approximately linear relationship with the log-odds of the outcome. Note that the objective of this study was based on inference and utilized historical data. Thus, the duration is known. The following clearance time bins (in minutes) were identified as effective across all four states based on Box-Tidwell diagnostics:

(0–10), (10–30), (30–60), (60–120), (120–300), and (300–600) minutes.

Using this approach, separate logistic regression models were fitted for each clearance time bin within each state. These models were used to estimate the odds ratio of clearance time within intervals where the linearity assumption holds. Additionally, a baseline logistic regression model was fitted for each state using the full set of observations and all independent variables except clearance time. This allowed for estimation of the overall odds ratios associated with other predictors while avoiding the influence of the clearance time's nonlinearity on the model structure.

Modeling Results

This subsection presents the key outputs from the logistic regression modeling phase. As described earlier, separate models were developed for each clearance time bin due to the non-linear relationship between clearance time and the log-odds of a secondary crash. **Table 8** summarizes the odds ratios for clearance time within each defined bin, reported for each state. Alongside the odds ratios, the table includes the number of observations in each bin and the percentage of incidents that resulted in a secondary crash, called primary (i.e., where the dependent variable $Y = 1$).

Across all models, a statistical significance threshold of $p_{value} < 0.05$ was applied. In the results tables, any variable whose coefficient did not meet this significance criterion is labeled as SNS (Statistically Non-Significant). This labeling highlights variables whose impact on secondary crash likelihood was not statistically distinguishable from zero at the 95% confidence level.

Table 8. Odds Ratios for Clearance Time by Duration Bins and State

Duration Bin (minutes)		MD	VA	TN	FL
0-10 minutes	# Observations	96,931	108,794	72,931	370,185
	% Primary	1.61	0.46	1.48	0.82
	Odds Ratio	**1.131**	**1.213**	**1.139**	**1.129**
10-30	# Observations	56,129	112,017	58,908	408,159
	% Primary	4.16	1.81	3.65	1.98
	Odds Ratio	**1.031**	**1.035**	**1.031**	**1.033**
30-60	# Observations	33,020	88,018	41,552.	334,363
	% Primary	6.06	3.69	7.39	3.67
	Odds Ratio	**1.014**	**1.016**	**1.013**	**1.016**
60-120	# Observations	17,782	65,074	40,757	322,009
	% Primary	8.78	6.98	10.99	6.33
	Odds Ratio	**1.008**	**1.013**	**1.007**	**1.009**
120-300	# Observations	7,363	20,644	37,956	292,663
	% Primary	7.96	9.89	15.36	7.94
	Odds Ratio	**1.003**	**1.002**	**1.003**	**1.003**
300-600	# Observations	2,439	2,807	16,458	120,106
	% Primary	7.35	12.54	21.16	7.86
	Odds Ratio	**SNS**	**SNS**	**1.001**	**1.001**

According to **Table 8**, the odds ratio reported for each clearance time bin and state reflects how the odds of an incident leading to a secondary crash change proportionally with each additional minute of clearance time. Across all four states, the odds ratios are consistently higher for incidents with shorter durations, indicating that each minute increase in clearance time for these incidents has a more substantial impact on the likelihood of a secondary crash—an intuitive and expected pattern.

The odds ratios for incidents lasting 0–10 minutes are approximately 1.13 in Maryland, Tennessee, and Florida, and as high as 1.21 in Virginia. This means that each additional minute of clearance time increases the odds of a secondary crash by 13–21% for short-duration incidents. For incidents in the 10–30 minute bin, the odds ratio drops to around 1.03, indicating a 3% increase in odds per minute. This effect diminishes further with longer durations: incidents lasting 30–60 minutes have odds ratios around 1.015 (1.5% increase), those lasting 60–120 minutes show odds ratios of 1.007–1.013, and incidents in the 120–300 minute range have odds ratios close to 1.003.

For very long incidents (over 300 minutes), the odds ratios are either statistically insignificant in Maryland and Virginia or extremely close to 1 (e.g., 1.001 in Tennessee and Florida), suggesting that clearance time has minimal influence on the likelihood of a secondary crash once the incident duration exceeds five hours.

In addition to the bin-specific models for clearance time, a separate logistic regression model was developed for each state using the full dataset, excluding clearance time, to evaluate the effect of all other explanatory variables. The goal of this model was to estimate the overall odds ratios for categorical and other non-continuous variables included in the analysis. As in **Table 8**, a significance threshold of 0.05 was applied, and odds ratios for variables that did not meet this threshold are labeled statistically non-significant (SNS). Variables that were either unavailable in the dataset or excluded due to multicollinearity are marked as *NA*.

For each categorical variable included in the model, a reference category is identified and listed in the table. The odds ratios presented for the remaining categories are interpreted relative to this reference category. That is, an odds ratio greater than 1 indicates a higher likelihood of a secondary crash occurring relative to the reference category. In contrast, an odds ratio less than 1 indicates a lower likelihood.

The results of this analysis are presented in **Table 9**.

According to **Table 9**, several explanatory variables demonstrated consistent patterns across states, while others showed state-specific variations in their association with the likelihood of a secondary crash.

Severity, which was only available in Tennessee and Florida, did not yield a statistically significant odds ratio in Florida. However, in Tennessee, both intermediate and major severity incidents were associated with odds ratios around 1.5, suggesting that more severe incidents increase the odds of a secondary crash by approximately 50% compared to minor severity crashes. This is expected, as severe crashes often take longer to clear and cause more disruption and distraction.

Capacity reduction variables showed mixed results. In Maryland and Tennessee, the 20–30% reduction category had the highest odds ratio (close to 2), implying that such reductions may double the odds of a secondary crash compared to cases with no reduction. In contrast, Virginia and Florida showed the highest odds ratios for the 0–10% reduction range. This may reflect differences in countermeasure deployment (e.g., dynamic message signs) or inaccuracies in lane data entry during incident response.

For the weekday variable, Virginia showed no significant difference between weekday and weekend crashes. Maryland had an odds ratio below 1, suggesting higher secondary crash odds on weekends, while Tennessee and Florida had odds ratios above 1 (1.07–1.17), indicating a slightly greater odds on weekdays.

The presence of emergency vehicles (available in VA and TN) also showed opposite trends: an increased odds of secondary crashes in Virginia and a decreased odds in Tennessee. These differences could reflect variation in response strategies, traffic control practices, or data reporting.

The event subtype categories showed several notable patterns:

- "Abandoned vehicle" had an odds ratio below 1 in Florida (0.8), indicating a lower odds of secondary crashes compared to the reference group.
- "Disabled vehicle" had an odds ratio below 1 across all states, suggesting lower secondary crash risk.
- "Serious accident" in Maryland showed a high odds ratio (2.24), meaning it more than doubled the odds of a secondary crash.
- "Multi-vehicle accident" showed elevated odds (e.g., 1.63 in VA, 1.37 in TN).
- "Medical emergency" and "overturned vehicle" also had higher odds ratios that were significant.

Table 9. Odds Ratios of Non-Continuous Explanatory Variables from Logistic Regression Models for Each State (Excluding Clearance Time).

Variable	MD	VA	TN	FL
const	0.033	0.015	0.029	0.033
Severity, Reference Category: minor				
intermediate	NA	NA	1.562	SNS
major	NA	NA	1.553	SNS
unknown	NA	NA	NA	SNS
Capacity Reduction, Reference Category: 0%				
0 to 10%	1.358	1.927	1.907	1.936
10 to 20%	1.252	2.022	1.211	1.430
20 to 30%	1.815	2.076	SNS	1.535
30 to 50%	1.593	1.662	0.906	1.418
More than 50%	SNS	0.593	1.592	1.395
unknown	0.678	SNS	0.450	0.860
Day Type, Reference Category: Weekend				
weekday	0.853	SNS	1.067	1.169
Emergency Vehicle Involvement, Reference Category: No				
yes	NA	1.163	0.916	NA
Incident Type, Reference Category: Accident				
abandoned vehicle	NA	NA	1.841	0.807
accident involving a pedestrian	NA	NA	0.399	NA
brush fire	NA	NA	NA	SNS
disabled semi-trailer	NA	0.807	NA	NA
disabled vehicle	0.786	0.505	0.846	0.467
incident	SNS	SNS	NA	0.385
injury accident	1.241	NA	NA	NA
jackknifed semi trailer	NA	NA	SNS	NA
medical emergency	0.686	NA	NA	NA
multi-vehicle accident	NA	1.629	1.371	NA
overturned vehicle	NA	NA	1.553	NA
serious accident	2.242	NA	NA	NA
vehicle on fire	SNS	SNS	1.210	SNS
Hourly Flow, Reference Category: 0 to 500 veh/hr/ln				
500 to 1,000 veh/hr/ln	1.429	1.995	1.643	1.585
1,000 to 2,000 veh/hr/ln	2.279	4.195	3.198	3.557

Variable	MD	VA	TN	FL
More than 2,000 veh/hr/ln	2.458	5.988	3.151	5.276
unknown	0.655	0.612	0.070	0.544
Lighting, Reference Category: Daylight				
Dark (No Street Light)	NA	NA	SNS	NA
Dark (Street Light)	NA	NA	SNS	NA
Dawn	NA	NA	SNS	NA
Dusk	NA	NA	SNS	NA
unknown	NA	NA	SNS	NA
Road Curvature, Reference Category: Straight				
curved	SNS	SNS	0.942	0.961
Weather, Reference Category: Clear				
rain	1.934	2.510	2.409	2.364
snow	5.552	4.237	3.710	2.020

Hourly flow, which categorizes expected traffic volume at the incident time, showed strong and consistent effects across states, using the 0–500 veh/hr/ln group as the reference. Higher flow categories had substantially higher odds ratios—up to ~6 in Virginia and ~5.3 in Florida for the >2,000 veh/hr/ln category. This indicates that secondary crashes are far more likely in high-volume traffic environments, likely due to reduced maneuverability and faster congestion buildup following a primary crash. The “unknown” volume category consistently showed odds ratios below 1, possibly reflecting incomplete data or unmeasured low-volume roads.

Lighting and road curvature were mostly not statistically significant, with odds ratios near 1 or labeled SNS (statistically not significant). However, weather conditions were among the most influential variables: compared to clear weather, rain was associated with odds ratios of ~2.0 to 2.5 across all states. Snow showed the strongest effect, with odds ratios ranging from ~2.0 in FL to over 5.5 in MD—the highest of any variable in the model.

These results underscore two critical findings:

- Adverse weather, especially snow and rain, is the most potent predictor of secondary crashes among the evaluated variables.
- High expected hourly traffic flow dramatically increases odds of a secondary crash, making it an essential contextual factor for real-time incident management and risk forecasting.

It is important to note that the exact values of the odds ratios should be interpreted with caution. Several factors can influence the reliability of these estimates:

- Data quality and completeness vary across states and variables. Some features may be underreported, inconsistently defined, or entirely unavailable in certain datasets, which can affect model accuracy.
- Unmeasured factors—such as driver behavior, enforcement presence, or real-time traffic control—were not accounted for but may influence the likelihood of secondary crashes.

- Inaccuracy in incident geolocation is a significant limitation. Since road curvature was derived directly from the reported location of incidents, any spatial inaccuracy can introduce bias, particularly for geometry-related features.

These caveats highlight the need to view the odds ratios as indicative of general trends rather than precise causal estimates and reinforce the importance of improving data quality, especially for geospatial attributes, in future modeling efforts.

Recommendations for Data Collection

A key challenge in this multi-state analysis was handling the inconsistencies between agency databases. For example, the fusion of multiple data sources including traffic speeds, weather, volumes, and roadway geometry—was necessary. The accuracy of time and geolocation attributes can have an impact on the ability to correctly associate an incident with the correct road segments or nearby events, thus affecting the fidelity of both descriptive statistics and modeling outputs.

Similarly, temporal data quality—such as start and end times of incidents, responder arrival and clearance times, or lane closure timestamps—is critical. These time features form the basis of actionable insights into response strategies and are integral for constructing incident timelines. Although this study could not validate these timestamps independently, their accuracy remains essential for drawing reliable statistics and informing incident management practices.

Another important aspect of data collection is the usability of the recorded attributes. One common issue is the treatment of missing values. If a data element is unpopulated due to its perceived irrelevance (for example, no lane closures occurred), this should be explicitly coded (for example, “all lanes open”) rather than left blank. This distinction enables data users to differentiate between genuinely missing values and cases where the feature does not apply, thereby reducing ambiguity in data interpretation.

Additionally, variables with excessive numbers of unique values (especially free-text fields or loosely defined categories) are often unsuitable for direct modeling and require substantial preprocessing. It is highly beneficial for such features to be standardized at the point of collection, using controlled vocabularies or predefined category lists to improve downstream usability.

To improve both efficiency and accuracy, transportation agencies should consider integrating auxiliary data sources directly into the traffic management systems. For instance, data elements such as number of lanes, road curvature, functional classification, or real-time weather conditions can be auto-populated using existing databases (such as datasets already maintained by state DOTS, HERE maps, OpenStreetMap, or NOAA radar feeds). This reduces the burden on field personnel, improves consistency, and allows for real-time validation of reported information.

Conclusion and Future Work

The primary objective of this study was to evaluate the relationship between incident duration and the probability of a secondary crash. However, this research made the following additional tangential contributions:

- Conducted an in-depth review of recent studies in secondary crashes, highlighting methods to identify secondary crashes and methods to model secondary crashes
- Established procedures to fuse disparate data sources into a master database for safety analysis.
- Developed a methodology to identify secondary crashes using real-world speed data.
- Created and evaluated several secondary crash prediction models using rigorous statistical methods to test the assumptions of each model. These models were used to make inferences on the impact of key variables such as incident duration, weather, capacity reduction, and flow rates on the probability of secondary crashes.

- Documented challenges related to best practices in traffic management system crash data collection
- Made recommendations on the critical variables to collect to support secondary crash inference modeling

This project has laid the groundwork for the following proposed future work activities:

- Operationalizing a real-time **secondary crash prediction** capability: This effort will leverage the data processing and fusion techniques developed in this project to build a system for predicting the probability of a secondary crash in real-time at the onset of an incident. Predictions would be updated in real-time as new information is entered into agency traffic management systems such as the arrival of an emergency responder, changing weather conditions, or the reopening of a lane. These prediction algorithms could provide traffic incident management decision makers with valuable insights on an incident's impact soon after detection and throughout the incident management process. This information will enable proactive operational decisions which could improve safety and reduce delays, fuel consumption, emissions, and property destruction.

- Operationalizing a **real-time incident duration prediction** model: Recognizing that incident duration is a critical factor in predicting secondary crashes, a model that could estimate incident duration at the onset of an incident may enhance the accuracy of a real-time crash prediction model. Predictions would be updated in real-time as new information is entered into agency traffic management systems

- Operationalizing a **real-time queue prediction** model: Speeds and associated queues resulting from an incident define the spatial boundaries for searching for secondary crashes. In the proposed real-time application, the max queue length would be predicted at the onset of the incident and updated as new information about the incident is provided by the traffic management system, nearby traffic sensors, weather, and probe-based speed data. Understanding the expected max queue length can inform operational decisions, such as the use of dynamic message signs (DMS) to inform road users of the back of the queue, sudden slowdowns, and the need for possible detour routes or change in operational strategies.

Reference:

[1] M. G. Karlaftis, S. P. Latoski, N. J. Richards, and K. C. Sinha, "ITS Impacts on Safety and Traffic Management: An Investigation of Secondary Crash Causes," *ITS J. - Intell. Transp. Syst. J.*, vol. 5, no. 1, pp. 39–52, Jan. 1999, doi: 10.1080/10248079908903756.

[2] "EXPIRED Reduction of Secondary Crashes." Accessed: Jun. 06, 2025. [Online]. Available: https://www.theiacp.org/resources/resolution/expired-reduction-of-secondary-crashes

[3] N. D. Owens, A. H. Armstrong, C. Mitchell, R. Brewster, and Science Applications International Corporation, "Federal Highway Administration Focus States Initiative : traffic incident management performance measures final report," FHWA-HOP-10-010, Dec. 2009. Accessed: Oct. 09, 2024. [Online]. Available: https://rosap.ntl.bts.gov/view/dot/973

[4] M. Franz, I. Tous, M. Pack, P. Hendren, and E. Strocko, "Identifying and Quantifying the Causes of Congestion - A New Tool for the Nation's States and Counties to Better Understand the Contributing Factors to Traffic," presented at the ITS WC, 2022.

[5] "Congestion Causes." Accessed: May 30, 2025. [Online]. Available: https://congestion-causes.ritis.org/

[6] P. Alluri and Florida International University. Department of Civil and Environmental Engineering, “Strategies to Identify and Mitigate Secondary Crashes Using Real-Time Traffic Data on Florida’s Turnpike System [Summary],” Apr. 2022. Accessed: May 28, 2025. [Online]. Available: https://rosap.ntl.bts.gov/view/dot/61862
[7] R. A. Richard, “Occurrence of Secondary Crashes on Urban Arterial Roadways.” Accessed: May 28, 2025. [Online]. Available: https://journals.sagepub.com/doi/abs/10.3141/1581-07?casa_token=wY_WA9vJrJ4AAAAA:1e6HEhrT3DrgPxQJschYV2r8zQA3vVKwGdO4Pb5vuiIBoaM4S1CIZPLq_MrGYNAjUdRFs16nT8A
[8] M. G. Karlaftis, S. P. Latoski, N. J. Richards, and K. C. Sinha, “ITS Impacts on Safety and Traffic Management: An Investigation of Secondary Crash Causes,” *ITS J. - Intell. Transp. Syst. J.*, vol. 5, no. 1, pp. 39–52, Jan. 1999, doi: 10.1080/10248079908903756.
[9] S. P. Latoski, R. Pal, and K. C. Sinha, “Cost-Effectiveness Evaluation of Hoosier Helper Freeway Service Patrol,” *J. Transp. Eng.*, vol. 125, no. 5, pp. 429–438, Sep. 1999, doi: 10.1061/(ASCE)0733-947X(1999)125:5(429).
[10] J. E. Moore, G. Giuliano, and S. Cho, “Secondary Accident Rates on Los Angeles Freeways,” *J. Transp. Eng.*, vol. 130, no. 3, pp. 280–285, May 2004, doi: 10.1061/(ASCE)0733-947X(2004)130:3(280).
[11] W. Hirunyanitiwattana and S. P. Mattingly, “Identifying Secondary Crash Characteristics for California Highway System,” presented at the Transportation Research Board 85th Annual MeetingTransportation Research Board, 2006. Accessed: May 28, 2025. [Online]. Available: https://trid.trb.org/View/777729
[12] C. Zhan, L. Shen, M. A. Hadi, and A. Gan, “Understanding the Characteristics of Secondary Crashes on Freeways,” presented at the Transportation Research Board 87th Annual MeetingTransportation Research Board, 2008. Accessed: May 28, 2025. [Online]. Available: https://trid.trb.org/View/848231
[13] G.-L. Chang and S. Rochon, “Performance Evaluation and Benefit Analysis for CHART,” 2009. [Online]. Available: http://chartinput.umd.edu/Chart2007final.pdf
[14] L. Kopitch and J.-D. M. Saphores, “Assessing Effectiveness of Changeable Message Signs on Secondary Crashes,” presented at the Transportation Research Board 90th Annual MeetingTransportation Research Board, 2011. Accessed: May 28, 2025. [Online]. Available: https://trid.trb.org/View/1093502
[15] E. R. Green, J. G. Pigman, J. R. Walton, and S. McCormack, “Identification of Secondary Crashes and Recommended Countermeasures to Ensure More Accurate Documentation,” presented at the Transportation Research Board 91st Annual MeetingTransportation Research Board, 2012. Accessed: May 28, 2025. [Online]. Available: https://trid.trb.org/View/1129424
[16] M. Jalayer, F. Baratian-Ghorghi, and H. Zhou, “Identifying and characterizing secondary crashes on the Alabama state highway systems. | EBSCOhost.” Accessed: May 28, 2025. [Online]. Available: https://openurl.ebsco.com/contentitem/doi:10.4399%2F978885488868511?sid=ebsco:plink:crawler&id=ebsco:doi:10.4399%2F978885488868511
[17] Y. Tian, H. Chen, and D. Truong, “A case study to identify secondary crashes on Interstate Highways in Florida by using Geographic Information Systems (GIS). | EBSCOhost.” Accessed: May 28, 2025. [Online]. Available: https://openurl.ebsco.com/contentitem/doi:10.4399%2F978885489182109?sid=ebsco:plink:crawler&id=ebsco:doi:10.4399%2F978885489182109
[18] H. Yang, Z. Wang, K. Xie, K. Ozbay, and M. Imprialou, “Methodological evolution and frontiers of identifying, modeling and preventing secondary crashes on highways,” *Accid. Anal. Prev.*, vol. 117, pp. 40–54, Aug. 2018, doi: 10.1016/j.aap.2018.04.001.
[19] J. Ou, J. Xia, Y. Wang, C. Wang, and Z. Lu, “A data-driven approach to determining freeway incident impact areas with fuzzy and graph theory-based clustering.” Accessed: May 30, 2025. [Online]. Available:

https://onlinelibrary.wiley.com/doi/full/10.1111/mice.12484?casa_token=rrrykQsCnfAAAAAA%3A0iE6WgDsUgq3rDbDAocVSF_rsr88Zh6AXT897wRu-SLmwog6z9esj3vNBkFawhfd_-8Ojtnqod4a6Q
[20] H. Yang, B. Bartin, and K. Ozbay, "Use of Sensor Data to Identify Secondary Crashes on Freeways." Accessed: May 28, 2025. [Online]. Available: https://journals.sagepub.com/doi/abs/10.3141/2396-10?casa_token=Dk-s84PtJMcAAAAA:hPTL_p8-MGdIrfl14dOE6-d2EIXax-bwG9Hzi3ycOZP2k8RX-nk-GPqX6pucRDyd5SHwAc-d8Ys
[21] N. J. Goodall, "Probability of Secondary Crash Occurrence on Freeways with the Use of Private-Sector Speed Data," *Transp. Res. Rec.*, vol. 2635, no. 1, pp. 11–18, Jan. 2017, doi: 10.3141/2635-02.
[22] A. S. Huq, "Identification of Secondary Traffic Crashes and Recommended Countermeasures," Ph.D., Florida International University, United States -- Florida, 2020. Accessed: May 28, 2025. [Online]. Available: https://www.proquest.com/docview/2847792505/abstract/86E19E7E9E054FB6PQ/1
[23] Z. Zhang, Y. Gong, and X. Yang, "Secondary Crashes Identification and Modeling Along Highways in Utah," *Transp. Res. Rec.*, vol. 2678, no. 3, pp. 613–624, Mar. 2024, doi: 10.1177/03611981231182394.
[24] H. Li, Q. Gao, Z. Zhang, Y. Zhang, and G. Ren, "Spatial and temporal prediction of secondary crashes combining stacked sparse auto-encoder and long short-term memory," *Accid. Anal. Prev.*, vol. 191, p. 107205, Oct. 2023, doi: 10.1016/j.aap.2023.107205.
[25] X. Liu, J. Tang, F. Gao, and X. Ding, "Time and Distance Gaps of Primary-Secondary Crashes Prediction and Analysis Using Random Forests and SHAP Model." Accessed: May 28, 2025. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1155/2023/7833555
[26] R. R. Souleyrette, M. Chen, X. Zhang, E. R. Green, and University of Kentucky Transportation Center, "Improving the Quality of Traffic Records for Traffic Incident Management," KTC-18-22/SPR18-567-1F, Dec. 2018. doi: 10.13023/ktc.rr.2018.22.
[27] K. K. Pecheux, G. Carrick, and B. B. Pecheux, "Secondary Crash Research: A Multistate Analysis," 2023.
[28] C. Xu, P. Liu, B. Yang, and W. Wang, "Real-time estimation of secondary crash likelihood on freeways using high-resolution loop detector data," *Transp. Res. Part C Emerg. Technol.*, vol. 71, pp. 406–418, Oct. 2016, doi: 10.1016/j.trc.2016.08.015.
[29] A. E. Kitali, P. Alluri, T. Sando, H. Haule, E. Kidando, and R. Lentz, "Likelihood estimation of secondary crashes using Bayesian complementary log-log model," *Accid. Anal. Prev.*, vol. 119, pp. 58–67, Oct. 2018, doi: 10.1016/j.aap.2018.07.003.
[30] J. H. Salum, L. Reyes, and P. Alluri, "Investigating Factors that Influence the Location and Time Intervals between Primary Incidents and Secondary Crashes," *J. Transp. Eng. Part Syst.*, vol. 150, no. 12, p. 04024083, Dec. 2024, doi: 10.1061/JTEPBS.TEENG-8060.
[31] M. M. Hossain, M. R. Abbaszadeh Lima, and H. Zhou, "Severity Analysis of Secondary Crashes on High-Speed Roadways: Pattern Recognition Using Association Rule Mining," *Transp. Res. Rec.*, vol. 2678, no. 8, pp. 919–931, Aug. 2024, doi: 10.1177/03611981231223194.
[32] J. Chen, Y. Li, C. Ling, Z. Pu, and X. Guo, "Spatiotemporal Prediction of Secondary Crashes by Rebalancing Dynamic and Static Data with Generative Adversarial Networks," Jan. 17, 2025, *arXiv*: arXiv:2501.10041. doi: 10.48550/arXiv.2501.10041.
[33] D. chrank, B. Eisele, T. Lomax, and J. Bak, "Appendix A: Methodology for the 2015 Urban Mobility Scorecard," 2015. [Online]. Available: https://static.tti.tamu.edu/tti.tamu.edu/documents/umr/archive/mobility-scorecard-2015.pdf