RURAL SPEED SAFETY PROJECT

March 2020

Submitted by
Subasish Das, Srinivas Geedipally, Raul Avelar, Lingtao Wu, Kay Fitzpatrick, Mohamadreza Banihashemi, and Dominique Lord

Texas A&M Transportation Institute
400 Harvey Mitchell Parkway South, Suite 300
College Station, TX 77845-4375

Notice

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the use of the information contained in this document.

The U.S. Government does not endorse products or manufacturers. Trademarks or manufacturers' names appear in this report only because they are considered essential to the objective of the document.

Technical Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Rural Speed Safety Project for USDOT Safety Data Initiative		5. Report Date March 2020 6. Performing Organization Code
7. Author(s) Subasish Das, Srinivas Geedipally, Raul Avelar, Lingtao Wu, Kay Fitzpatrick, Mohamadreza Banihashemi, Dominique Lord		8. Performing Organization Report No.
9. Performing Organization Name and Address Texas A&M Transportation Institute The Texas A&M University System College Station, Texas 77843-3135		10. Work Unit No. (TRAIS) 11. Contract or Grant No. DTFH6116D00039L
12. Sponsoring Agency Name and Address Office of the Secretary of Transportation Office of the Assistant Secretary of Transportation for Policy 6300 Georgetown Pike McLean, VA 22101		13. Type of Report and Period Covered Final Report 14. Sponsoring Agency Code
15. Supplementary Notes Projects were performed with the cooperation and participation of the Federal Highway Administration Project Title: Rural Speed Safety Project, for USDOT Safety Data Initiative		
16. Abstract The objective of this project was to examine prevailing operating speeds on a large scale to determine how speed and speed differentials interact with roadway characteristics to influence the likelihood of crashes. The project team conducted three major tasks: (a) developed conflated databases for Ohio and Washington by incorporating the Highway Safety Information System (HSIS) and the National Performance Management Research Data Set; (b) developed static and interactive data visualization tools to show the association between operating speed measures and safety outcomes; and (c) developed best-fit models at annual and daily levels to address the impact of operating speed on safety. The overall finding was that speed-related operational information is an area of opportunity to better understand safety outcomes. This pilot project established the framework of data conflation and an analytical pipeline that will help to address the effect of operation speed measures on safety. The replicability procedure developed in this study can be applied to other HSIS States. The project team developed a weblink that includes descriptive statistics and data visualization tools (both static and interactive). The links provide a more detailed view of the speed measures and descriptive statistics, as well as visualization of the association between speed measures and crashes at a granular level. The team members also developed an interactive decision support tool (https://ruralspeedsafety.shinyapps.io/rss_sdi/) to show annual risk scoring using Washington and Ohio data that contain expected total crashes from the developed models.		
17. Key Words Rural roadways; traffic crashes; operating speed; roadway characteristics; safety performance functions.	18. Distribution Statement No restrictions. This document is available to the public through the National Technical Information Service Alexandria, Virginia 22312	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 122 / 22. Price

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
	NOTE: volumes greater than 1000 L shall be shown in m^3			
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.
(Revised March 2003)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AADT	Annual Average Daily Traffic
AI	After Incident
AR	AI Reference
BI	Before Incident
BR	BI Reference
CMF	Crash Modification Factor
CURE	Cumulative Residual
DI	During Incident
DR	DI Reference
FE	Fixed Effect
FHWA	Federal Highway Administration
FI	Fatal and Injury
GLM	Generalized Linear Model
HPMS	Highway Performance Monitoring System
HSIS	Highway Safety Information System
HSM	Highway Safety Manual
LW	Lane Width
ME	Mixed Effect
MOE	Measure of Effectiveness
MNL	Multinomial Logit
NHS	National Highway System
NPMRDS	National Performance Management Research Data Set
OP	Overdispersion Parameter
PDO	Property Damage Only
RE	Random Effect
RTM	Regression to the Mean
SDF	Severity Distribution Function
SDI	Safety Data Initiative
SMS	Space Mean Speed
SPF	Safety Performance Function
TMC	Traffic Message Channel
TTI	Texas A&M Transportation Institute
USDOT	U.S. Department of Transportation
vpd	vehicles per day

EXECUTIVE SUMMARY

RESEARCH PROBLEM

To save more lives and reduce injuries from roadway crashes, agencies must identify sections of the highways that have an increased risk of crash occurrences. Toward that end, the U.S. Department of Transportation's (USDOT's) vision for the Safety Data Initiative (SDI) includes the integration of big data sources as a focus area to enhance the general understanding of crash risks and mitigate future crash occurrences.

Key Highlights

- Variability in daily average traffic speeds was associated with increased traffic crashes.
- Differences in traffic speeds between weekdays and weekends was correlated with increased traffic crashes
- The beta decision support tool was developed to interactively visualize segment-level risk that includes speed variables.
- The current study is a starting point for more in-depth investigation and continued progress in incorporating speed-related factors into crash predictive models.

Current crash estimation or prediction methods, such as those in the first edition of the Highway Safety Manual (HSM) use annual average daily traffic (AADT) data along with geometric characteristics to predict the annual average crash frequency of roadway segments and intersections. One limitation of the HSM is the omission of speed-related factors from all aspects of safety predictive methods. Recent research has made little substantive progress in incorporating speed-related factors into crash predictive models. To advance the state of practice this study begins the work of investigating the association between crash risk and traffic speeds using traffic speed information from big data.

OVERVIEW OF METHODOLOGY

To address the current research gap and as part of the SDI, the Texas A&M Transportation Institute (TTI) led a pilot project entitled, 'Rural Speed Safety.' This study developed safety performance functions (SPFs) by using geometric and operational characteristics that include speed-related measures. SPF's are the statistical "base" models used to estimate the average crash frequency for a facility type with specific base conditions.

Researchers examined the prevailing operating speeds on a large scale and quantified how traffic speed in rural areas interacts with roadway characteristics to influence the likelihood of crashes. The inclusion of speed information expands upon the existing state of practice by incorporating operational data as risk variables through statistical models. The models developed over the course of the project include speed measures to quantify highway safety risk and better predict crash occurrence.

The research addressed two major research questions: (1) Do different speed measures contribute to crash outcomes? (2) Is there more variability in speeds just prior to a crash? To answer these research questions, the project team developed three types of deliverables:

- Conflated databases for Ohio and Washington state that incorporates crash, roadway, and traffic data from the Highway Safety Information System (HSIS), travel speed data from the National Performance Management Research Data Set (NPMRDS), and roadway information from the Highway Performance Monitoring System (HPMS). The data are from 2015.
- Best-fit models that address the impact of operating speed at the segment and segment-temporal levels.
- Static and interactive data visualization tools to show the association between operating speed measures and safety outcomes.

GENERAL FINDINGS

Certain speed measures incorporated into statistical models were found to be beneficial to quantifying safety risk. This pilot project established a framework of data conflation and an analytical pipeline that will help to include the effects of operation speed measures on crash occurrence frequency. It is important to note that this study is a *starting point* in evaluating the effect of operating speed on crash outcomes. It includes all rural facility types, and the procedures developed can be applied in other states contributing to HSIS. The project team used three different units of analysis in model development:

- **Annual-level crash prediction models:** speed measures and crashes at the roadway segment level (annual).
- **Daily-level crash prediction models:** speed measures and crashes at the roadway segment level (daily).
- **Exploratory examination of time before and after crashes:** speed measures and crashes based on hours around specific roadway locations and points in time.

Annual-level crash prediction models: The project team developed SPFs using aggregated annual data for total (KABCO[1]) crashes, fatal and injury (KABC) crashes, and property damage only (PDO) crashes, for Washington and Ohio separately, as well as for both states together. This also includes different functional classifications (roadway types) within the National Highway System (NHS). Certain speed measures were useful in the development of the annual segment-level statistical models. This study examined aggregated traffic travel speed variation over time. The current study did not examine the speed variability between the vehicles as NPMRDS provides aggregated speed measures.

Table 1 shows a summary of the impacts on the crash frequency of the variables examined for the developed models.

[1] K= Fatal, A= Incapacitating Injury, B=Non-incapacitating Injury, C=Minor Injury, O= No Injury or Property Damage Only (PDO)

Table 1. Impact of Variable Changes on Crash Frequency at Segment Level (Annual Data)

When	Crash Frequency On		
	Rural Interstate	**Rural Two-Lane**	**Rural Multilane**
Traffic volume increases	Increases	Increases	Increases
Segment length increases	Increases	Increases	Increases
Lane width is wider	-	Decreases	–
Percentage of horizontal curves increases	Mostly Increases *(Decreases in OH model)*	Increases	Increases
Intersection is present	–	Increases	Mostly Increases *(Decreases in WA KABC model)*
Road is undivided	NA	NA	Increases
Percentage of days with precipitation increases	–	Decreases	Decreases
Operating speed difference between weekend and weekday increases	Increases	Increases	Increases
Average hourly operating speed variability within a day increases	–	Increases	Mostly Increases *(Decreases: WA KABCO model)*
Operating speed variability by month within a year increases	–	Increases *(OH PDO model only)*	Increases
Average hourly non-peak non-event operating speed increases (free flow) increases	Decreases	–	Increases
Average Hourly Speed increases	–	–	–

Note: at 95% confidence level: Increases (crash frequency goes up), Decreases ((crash frequency goes down), – (not significant), NA= not applicable.

The key findings from the project modeling as shown in the table are the following:

- Increased variability in hourly operating speed within a day and an increase in monthly operating speeds within a year are both associated with increased crashes.
- Multilane, non-freeway roads with higher free-flow speeds are expected to experience a higher crash frequency than those with lower free-flow speeds. However, crash frequency decreases for interstate roadways, which is due to their more robust highway design standards.
- When operational speed difference between weekends and weekdays is greater, all three roadway types experienced a higher number of crashes. Segments experiencing higher speed differentials between weekends and weekdays indicate the nature of roadway use and land use patterns.
- Increased non-peak and non-event speed (average operating speed excluding peak hours and hours with events) is associated with an increase in crash frequencies on rural two-lane roadways. However, the opposite is true for the Interstate model. This finding for decreased crash frequency on the Interstate could be because good design and high standard roads are associated with higher non-peak and non-event speeds.

- As the proportion of horizontal curvature on a segment increases, the crash frequencies are expected to increase.
- In general, segments with intersections are expected to experience more crashes than segments without intersections. This is likely because segments with intersections have a greater number of conflict points. The variable is only significant for multilane roadways.

Daily-level crash prediction models: Prediction based on annual information limits the SPFs' ability to quantify the effects of variables such as operating speeds, operating speed variance, or seasonal differences that fluctuate more often than year-to-year. Agencies require the ability to accurately assess what seasonal or daily changes could affect crash outcomes. To address this, the study developed statistical models for the segment daily level based on crash severity and roadway type. The Tweedie statistical model is applied in the analysis due to the infrequent nature of crashes that requires zero inflation[2]. Table 2 lists the variable changes and affects for the developed models.

Table 2. Impact of Variable Changes on Crash Frequency for the Segment-temporal level Models (Daily)

When	Crash Frequency On		
	Rural Interstate	**Rural Two-Lane**	**Rural Multilane**
Traffic volume increases	Increases	Increases	Increases
Segment length increases	Increases	Increases	Increases
Number of lanes increases	Decreases *(OH KABC model only)*	NA	–
Lane width is wider	Increases *(OH KABC model only)*	Decreases *(WA model only)*	Increases *(WA model only)*
Number of curvatures increases	Decreases *(OH KABCO model only)*	Decreases *(WA KABCO model only)*	Increases *(OH model only)*
Total length of curvatures is higher	–	Increases *(WA KABCO model only)*	–
Percentage of days with precipitation is higher	Increases	Increases	–
Variability of daily average speed increases	Increases	Increases	Increases
Daily average speed increases	Decreases	Increases	–

Note: at 95% confidence level: Increases (crash frequency increases), Decreases (crash frequency increases), – (not significant), NA= not applicable.

The general findings from the modeling that used daily data as shown in the table are:

[2] TMCs with zero annual crashes are also included in the model.

- In all models, a segment with high variation in daily average speeds is expected to experience a higher number of crashes than a segment with a lower variation in daily speeds. ***The strength of this finding is one of the biggest insights gained from this study.***
- Average operating speed increases were associated with increased crashes for rural two-lane roadways. However, average operating speed increases were associated with decreased crashes in the Interstate models. This finding for interstate could be because good design and high standard roads are generally associated with higher average operating speeds.
- As daily average precipitation increases, so do the number of daily crashes.

Exploratory examination of time before and after crashes: This segment–temporal–level analysis examined the speed difference between two scenarios: 1) time around crash events, and 2) time around non-crash condition. This study design examined operating speed measures for 4 hours prior to a crash, and the speeds being traveled in the same location and same hour and day of week, but on a day when no crash occurred. The current analysis is limited to a randomly selected sample dataset (with 150 crashes from Washington interstate roadways). The overall outcome of this analysis is exploratory in nature. The findings are:

- After controlling for other influential factors, as the moment of the crash occurrence approached, the speed trend for the crash-related series decreased and was substantially different in comparison to the trend of the non-crash–related reference series.
- Speed variability increased for the series just prior to a crash, which was also different from the comparison no crash series.

DECISION SUPPORT TOOL

This project team developed an interactive decision support tool[3] that visualizes the Washington state and Ohio data results. The data contain the expected total crashes from the final models to show segment-level high-risk analysis. The tool (see figure 1) will have adaptability options for newer datasets (crash and speed data).

[3] https://ruralspeedsafety.shinyapps.io/rss_sdi/

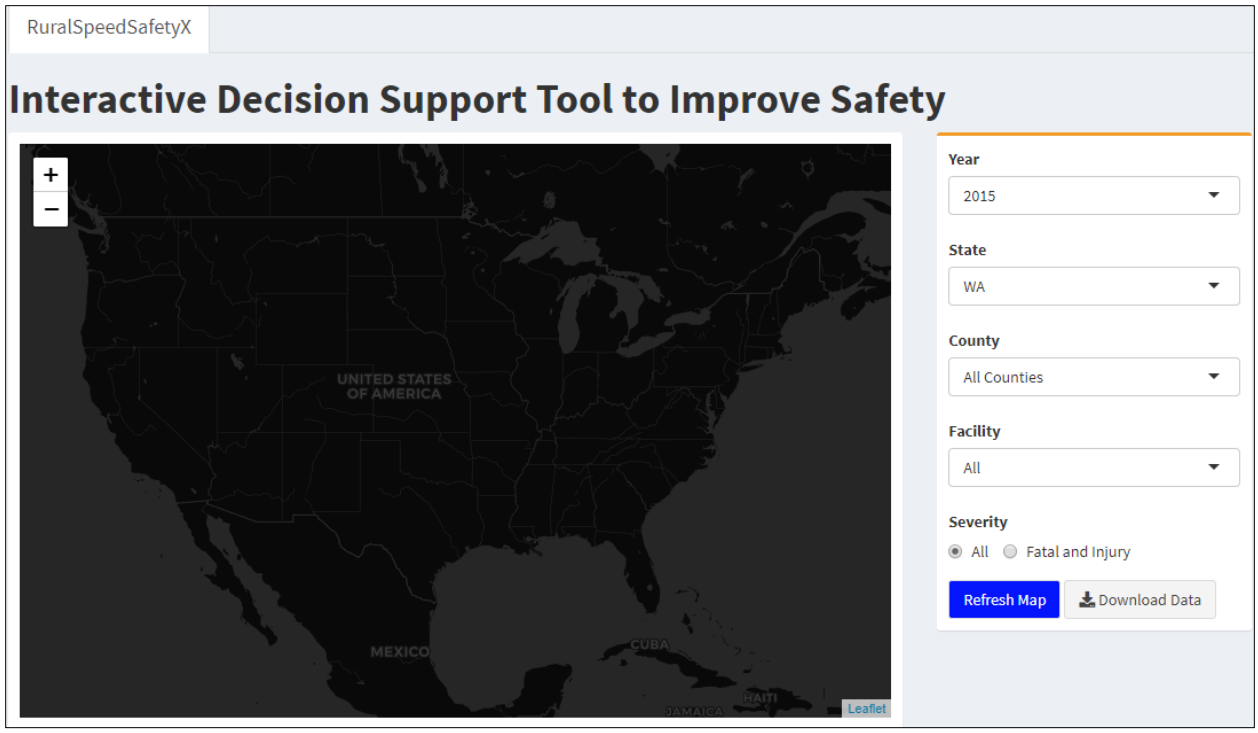


Figure 1. Interface of the Framework of the Decision Support Tool.

The project team also developed a weblink[4] that includes descriptive statistics and data visualization tools (both static and interactive). The clickable links provide a more detailed view of the speed measures and descriptive statistics, as well as visualization of the association between speed measures and crashes at a granular level.

OPPORTUNITIES AND LESSONS LEARNED

The opportunities and lessons learned of the current study and future directions are discussed below:

- Certain speed variables were strongly correlated with crash risk. Locating segments with high variance in hourly or monthly operating speeds, or large differences in operating speeds between weekdays and weekends, could help identify roadways that may warrant additional focus for more enforcement, engineering, and education efforts.
- Current HSM crash prediction models predict crashes for both directions of travel combined; this report used bi-directional prediction models that incorporate the distinct directions of travel when predicting crashes. Directional prediction models will be useful to many State DOTs.
- Horizontal curves are negatively associated with traffic safety, so special attention can be devoted to reducing speeds and improving other traffic conditions on roadways with significant quantities of horizontal curves.

[4] http://subasish.github.io/pages/FHWA_Rural_Speed_T4_1/

- The current study exclusively focused on rural roadways, and further research is needed to consider the association between speed measures, road geometry, and crashes explicitly on urban roadways.
- Newer crowdsourced data fusion can help determine the current research gaps in crash data integration and analysis. Some of the potential crowdsourced data sources are Streetlight traffic flow data, HERE Navigation and Infotainment data, Waycare incident data, Verizon intelligent traffic solution, Strava data, and Miovision Smart City data, amongst many others.

LIMITATIONS

The limitations of the current study and future directions are discussed below:

NPMRDS and HSIS

- The research used roadway segments based on the NPMRDS travel time data Traffic Message Channel, which varied in size with some being quite long. Further examination of the effects of segment length would improve modeling reliability. The NPMRDS provides travel time data in both directions, but HSIS data provide segment-level information.
- Different versions of the NPMRDS (versions 1 and 2) were acquired from different vendors. The current conflation work (using NPMRDS version 1) may need additional conflation for NPMRDS version 2. In many cases, variables of interest are not included uniformly across the road networks examined.
- More robust NPMRDS data with fewer missing values would provide more insightful knowledge on operation speed measures. Recent NPMRDS data may have more complete data as coverage for these big data sources are improving over time.

Zero inflation in crash counts

Model development at a granular level, such as hourly or daily, encounters zero inflation in crash counts. This study used a robust modeling technique (the Tweedie model) to develop daily level models. There is a need for additional experimentation using other advanced modeling techniques that can handle zero inflation at the hourly level.

Model development and refinement

- The geometric variables are limited to several factors (for example, segment length, traffic volume, segment width, median width, median type, and shoulder width). Additional geometric variables (e.g., horizontal and vertical curve, super elevation, and the presence of roadside fixed objects) should be examined in future studies.
- Future expansion of the data conflation to include SHRP2 Roadway Information Database (RID) would open the availability of more detailed mobile data (about 5% of the network) for 6 States.
- The current analysis did not incorporate demographic and driver variables in the statistical model development. However, the decision support tool provides population density and household density at the U.S. Census tract-level on the roadway segments. There is a need for the incorporation of demographic variables in safety evaluation.

There is a need for continued research progress in incorporating speed-related factors into crash predictive models. The current study can be considered as a starting point for the in-depth investigation on the speed-crash association.

CHAPTER 1. INTRODUCTION

Research on the influence of vehicle operating speed, roadway design elements, and traffic volume on crash outcomes will greatly benefit the road safety profession in general. If the relationships between these variables are well understood and characterized, existing techniques and countermeasures for reducing crash frequencies and crash severities can potentially improve, and new methodologies addressing and anticipating crash occurrence will naturally ensue.

PROJECT GOALS AND OBJECTIVES

This study aimed to examine two major research questions: (1) Do different speed measures contribute to crash outcomes? (2) Is there more variability in speeds just prior to a crash? Table 3 displays the questions as well as the dataset structures (described in the following sections) that can be used to explore them.

Table 3. Key Research Questions.

Num	Question	DS1	DS2	DS3
1	Do different operational speed measures contribute to crash outcomes?	✓		
	1a. To what extent can the operational speed be interpreted as directly affecting crash risk or crash severity?	✓	✓	
	1b. To what extent can the decision criteria be developed in assigning the risk measures in the decision support tool?	✓	✓	
	1c. Contingent to the granular level of speed measures, is there a discernible relationship between crash risk/severity, operational speed, geometric characteristics (for example, horizontal and vertical curve information), and traffic volume? i. Is the risk for different crash types dependent on the operational speed, traffic volume, and geometric characteristics? ii. Is there a daily cyclical association of crash risk and operational characteristics? iii. Is there a seasonally cyclical association of crash risk and operational characteristics?		✓	
	1d. How much variability in crash risk/frequency can be explained by cyclical factors?		✓	
2	Is there more variability in speeds just prior to a crash?			✓

DS1 = dataset structure 1 for annual crash prediction. Crash frequency is based on spatial locations (i.e., multiple TMCs are examined for a range of facility types/speeds/etc.).
DS2 = dataset structure 2 for daily crash prediction. Crash frequency is based on spatiotemporal (i.e., a single TMC is examined for a variety of time periods or epochs).
DS3 = dataset structure 3 for exploration of speed difference before crash occurence. Speed behavior is explored prior to and during a crash.

To answer the research questions, the project team had two major goals: (a) develop the conflated dataset with traffic speed, roadway design elements, traffic volume information, and crash frequency; (b) quantify the targeted relationship between crashes and influential variables.

The project team acquired multiple datasets and data conflation strategies applied to previous background work (data conflation for North Carolina) from USDOT to perform the related task. The team conflated the 2015 data (travel speed data from the NPMRDS, as well as crash, roadway, and traffic data from HSIS) and developed several dataset structures as needed to achieve the first major goal and used these datasets to examine statistical and machine learning tools to determine the most suitable approach to address the second goal.

This report describes the methodology developed for characterizing the influence of vehicle operating speed, roadway design elements, and traffic volumes on crash outcomes. Chapter 2 presents the data integration work and key descriptive statistics of the selected variables. Chapter 3 introduces the final iteration results of the statistical model runs. Chapter 4 synthesizes the major findings of this study and provides overall lessons learned about the state of the practice as well as suggestions for future research. This report also contains several appendices.

DATASET STRUCTURES

This study conflated traffic crash data (for the year 2015) for the States of Washington and Ohio at the segment level. The project team identified three dataset structures for the model development to determine the effects of speed on crashes.

Data Structure 1

Table 4 represents a potential dataset structure for the statistical analysis used for annual-level crash prediction models. It does not enlist all potential variables; rather, the intent of the dataset structure is to show the basic structure of the dataset that was used in further analyses. The project team used dataset structure 1 to develop models for total crashes (KABCO), KABC crashes, and PDO crashes, using crash frequency (based on total crashes or crashes for different severity levels) as the response variable. For dataset structure 1, the crash frequency is based on spatial locations, which are represented by TMC segments. Table 4 shows that each TMC segment is conflated with segment information and crash data from the 2015 HSIS data, and each of the TMC segments contains weighted values of the HSIS segment-level information.

Since the speed data are provided at the temporal level, using a suitable speed measure that can capture the impact of operating speed at the segment level is necessary. As part of this research, determining the appropriate speed measure (see the fourth column in table 4 as an example) to use was required. The project team explored and identified representative speed measures that were considered in the segment-level analyses (described in Chapter 3).

Table 4. Dataset Structure 1 (Spatial Locations or Segment Level).

TMC	TMC Length (mi)	TMC Direction	TMC Speed Measure (mph)	HSIS Segment	Traffic Volume (vpd)	Facility Type	# Lanes per Direction	Posted Speed Limit (mph)	Total Crashes (2015)
TMC01	1.03	Northbound	55	HSIS01	24,222	Interstate	2	>50	18
TMC02	1.04	Southbound	56	HSIS01	24,222	Interstate	2	>50	12
TMC03	0.50	Northbound	52	HSIS02	20,563	Multilane Undivided	2	>50	6
TMC04	0.58	Southbound	30	HSIS03	8,000	Two-Lane	1	>25	5
TMC05	0.69	Northbound	22	HSIS04	4,000	Two-Lane	1	>20	0

Note: vpd = vehicles per day.

Data Structure 2

Dataset structure 2 uses a spatiotemporal approach (i.e., a single TMC is examined for various time periods, or epochs) to develop the daily-level crash prediction models. The fourth column (see table 5) is the epoch value that indicates the time of day. It represents a key value from the NPMRDS database. Theoretically, a total of 35,040 time bins (i.e., epochs; 365 days × 96 15-minute bins) or 105,120 epochs (365 days × 288 5-minute bins) level speed measures can be assigned to each TMC. To address the missing value issues, aggregation of temporal bins (hourly or daily) is a potential solution for the model development using this dataset structure. Like dataset structure 1, the response variable for dataset structure 2 is crash frequency (either based on total crashes or crashes for different severity levels) for developing models at different severity levels.

Table 5. Dataset Structure 2 (Spatiotemporal or Segment-Temporal Level).

TMC	TMC Length (mi)	TMC Direction	Epoch	TMC Speed Measure (every 15 min in 2015) (mph)	HSIS Segment	Traffic Volume (vpd)	Crashes in the Associated Time Bins
TMC01	1.03	Northbound	107	55	HSIS01	24,222	0
TMC01	1.03	Northbound	108	57	HSIS01	24,222	0
TMC01	1.03	Northbound	109	NA	HSIS01	24,222	0
TMC01	1.03	Northbound	110	63	HSIS01	24,222	1
TMC01	1.03	Northbound	111	NA	HSIS01	24,222	0

Note: NA = not applicable.

Data Structure 3

Dataset structure 3 was used to answer the second research question (see table 3): Is there more variability in speeds just prior to a crash? This structure considers whether a crash occurred within the given time period and was used to examine speeds before and after crashes (i.e., the

value is either 0 or 1 for that variable). It also considers the consecutive epochs before a recorded crash.

To construct the dataset for analysis, all epochs with crash incidents were identified in the spatiotemporal dataset (dataset structure 2). All epochs within 4 hours of the incidents were also identified and labeled accordingly in a new field named "epoch type" (see table 6). The coding of these incident-related epochs was as follows:

- Before incident (BI): An epoch 4 hours or less before an incident (crash) occurred.
- During incident (DI): An epoch when an incident occurred.
- After incident (AI): An epoch 4 hours or less after an incident occurred.

Table 6. Dataset Structure 3 (Retrospective Time Series or Operational Characteristics in Time Proximity to Crashes at the Segment-Temporal Level).

Epoch Type	Epoch Relative to Incident	TMC	TMC Length (mi)	TMC Direction	Epoch	TMC Speed Measure (every 15 min in 2015) (mph)	HSIS Segment	Traffic Volume (vpd)	Crashes in the Associated Time Bins
BI	−2	TMC01	1.03	Northbound	107	55	HSIS01	24,222	0
BI	−1	TMC01	1.03	Northbound	108	57	HSIS01	24,222	0
DI	0	TMC01	1.03	Northbound	109	NA	HSIS01	24,222	1
AI	1	TMC01	1.03	Northbound	110	63	HSIS01	24,222	0
AI	2	TMC01	1.03	Northbound	111	NA	HSIS01	24,222	0

Note: NA = not applicable.

The project team also developed a companion dataset of reference sets of epochs to be utilized in the analysis. The companion dataset was created as a control to compare normal speed patterns in instances when there was no crash versus speed patterns in instances when there was a crash. The companion dataset used the same TMC and the same month and day of the week to keep the temporal trends similar. The data integration steps of dataset structure 3 are described in appendix E. All the epochs of these reference sets were labeled as follows:

- BI reference (BR): A reference to BI epochs.
- DI reference (DR): A reference to DI epochs.
- AI reference (AR): A reference to AI epochs.

In addition to the epoch type field, a field with relative epochs was added to indicate separation from the incident epoch. For example, a value of −2 indicates a BI or BR epoch that is two 15-minute periods prior to the incident at the corresponding DI or DR epoch. Similarly, a value of +3 indicates an AI or AR epoch that is three 15-minute periods after the incident at the corresponding DI or DR epoch.

CHAPTER 2. DATA INTEGRATION

INTRODUCTION

The project team considered two national databases to explore the research questions for addressing the current research gaps:

- The NPMRDS contains passenger and freight travel time datasets for the National Highway System (NHS) and other roadways.
- The HSIS, a cooperative endeavor funded by FHWA, is a roadway-based system that provides quality data on a large number of crash, roadway, and traffic variables from a group of selected States (California, Illinois, Maine, Michigan, Minnesota, North Carolina, Washington, Utah, and Ohio).

Most of the existing safety models predict crashes for both travel directions combined. With the characteristics of the available databases, it is possible to consider each direction of travel separately, which permits accounting for different directional traffic volumes and operational speeds. Such differences may have distinctive effects on crashes. This study examined the prevailing operating speeds on a large scale and considered directionalities on the rural roadway networks of Ohio and Washington to see how speed and speed differentials interact with roadway characteristics to influence the likelihood of crashes.

DATA DESCRIPTION

Two databases (NPMRDS and HSIS) were used in this study to develop the conflated database for two focus States (Ohio and Washington). Figure 2 shows the data integration flowchart. The data integration work has major three steps:

- Conflate the HSIS roadway network data to NPMRDS directional network
- Determine different speed measures by temporal segregation (for example, annual, month or daily)
- Conflate average precipitation (annual and daily) data (from NOAA) to the NPMRDS network

The 2015 NPMRDS static files were produced more or less on a quarterly basis. The project team used three different static files for 2015: 2014Q3, 2015Q3, and 2015Q4. In an exploration of the three static files, researchers found that over 95 percent of the NPMRDS TMCs are the same in the rural areas of the States across the three NPMRDS static files. A comparison between different static files is shown in a Venn diagram (in figure 3). A comparison between 2015Q4 and 2015Q3 shows that 2015Q4 has 12,356 TMCs in Washington (left side of figure 2). The number of common TMCs (with similar lengths) in both datasets is 11,891. For 247 TMCs (around 5 mi long in total length), the distances are not matched in both files. The comparison also shows that 2015Q4 has an additional 471 TMCs that were not present in 2015Q3 (only six missing in 2015Q4 when compared with 2015Q3). Similar comparisons can be made for 2015Q4 and 2014Q3 (see right side of figure 3).

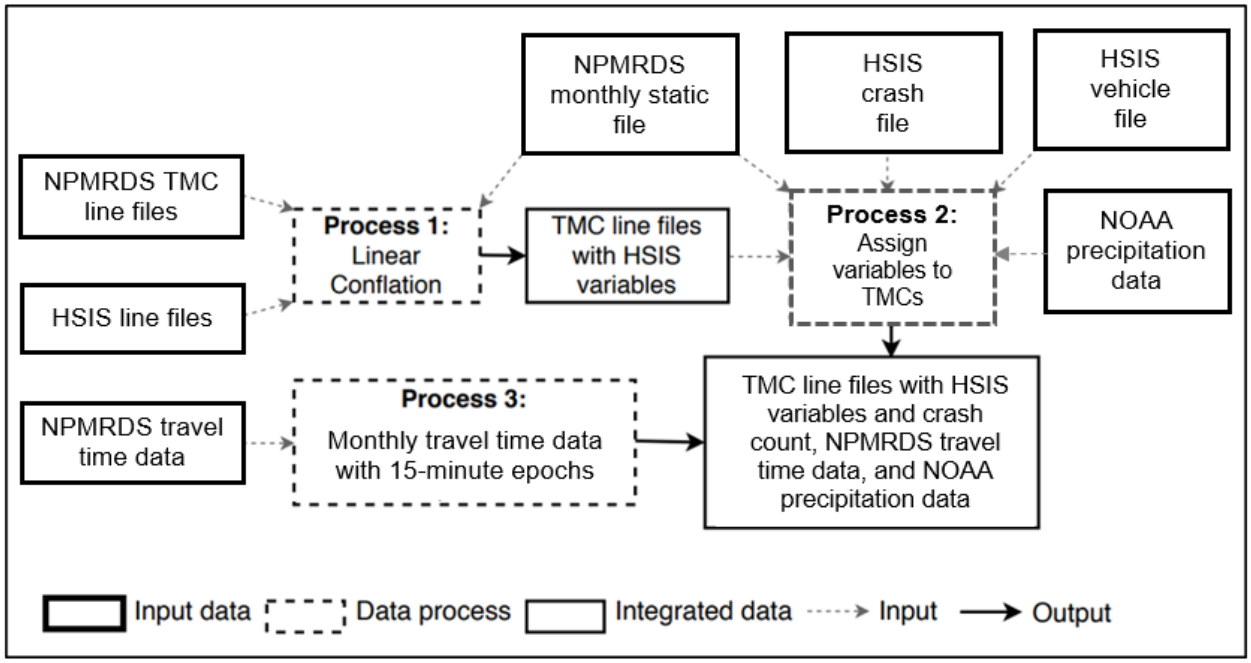


Figure 2. Data Conflation.

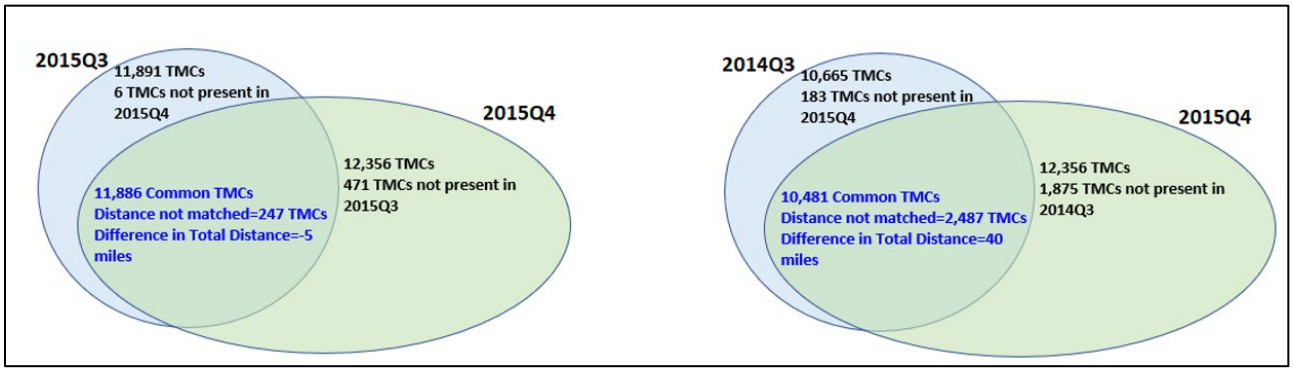


Figure 3. Comparison between Different Static Files Using Washington NPMRDS Data.

Table 7 provides more granular information on distance thresholds in different quarters of the NPMRDS. Based on an examination of the differences, the vast majority of the network was usable for the entire 2015 year.

Table 7. Comparison of Distances between Different Quarters (Static File).

Measures	Washington			Ohio		
	2014Q3	2015Q3	2015Q4	2014Q3	2015Q3	2015Q4
# TMCs	6,624	7,269	8,424	10,665	11,891	12,356
Total length (mi)	10,062	10,803	11,300	16,545	18,028	18,262
# Same distance TMCs (length)	5,531 (8,205 mi)			7,968 (12,025 mi)		
# Same distance TMCs with 2015Q4 (length)	5,535 (8,211 mi)	6,866 (9,940 mi)	8,424 (11,300 mi)	7,994 (12,073 mi)	11,638 (17,518 mi)	12,356 (18,262 mi)
# Different distance TMCs with 2015Q4 (length)	1,089 (1,851 mi)	403 (863 mi)	NA	2,671 (4,472 mi)	253 (510 mi)	NA
Cumulative difference between distances with 2015Q4	147 (1.46% of the total network length)	5 (0.03% of the total network length)	NA	280 (1.69% of the total network length)	8 (0.02% of the total network length)	NA
# TMCs with difference greater than or equal to 0.1 mi (cumulative difference)	154 (145 mi)	12 (3 mi)	NA	268 (267.54 mi)	14 (8 mi)	NA

Note: NA = not applicable.

Data Coverage Using NPMRDS 2015Q4

This project required that the crashes from HSIS be inserted into the base unit of analysis (NPMRDS TMC segment), and further, each of these segments needed a definitive number of crashes associated with it. The project mandated an 85 percent or greater match rate of the base unit of analysis segments. Table 8 shows the data coverage by the final conflated databases. The 2015 crashes were assigned to the associated TMC based on the direction algorithm and the nearest distance between the crash location and surrounding TMCs. To address the missing information in the direction column, some crashes were assigned to the associated TMC based on the shortest distance between the crash and linear segments of the TMCs. If one considers crashes that are correctly linked to the TMCs based on both direction and distance, Ohio and Washington have matching rates of 83 percent and 86 percent, respectively. If one considers both approaches (distance plus direction, and distance-only), Ohio and Washington have match rates of 89 percent and 98 percent, respectively.

Table 8. Data Coverage by the Final Conflated Data.

State	Rural NHS Crashes (2015)	Assigned to TMC Segments			Assigned to Conflated TMC-HSIS Segments		
		Both Distance and Direction	Distance Only	Total	Both Distance and Direction	Distance Only	Total
Ohio	11,547	9,603 (83%)	859[a] (7%)	10,462 (91%)	9,541 (83%)	789 (7%)	10,330 (89%)
Washington	6,017	5,176 (86%)	741[b] (12%)	5,917 (98%)	5,159 (86%)	738 (12%)	5,897 (98%)

[a] For Ohio, "direction of the vehicle involved in crash" was not available. In the absence of this variable, researchers used "direction of reference" to perform the analysis. However, this variable was not well populated. When "direction of reference" was not matched with either of the TMC directions or was not available, researchers used only "distance" to assign the crashes.
[b] For Washington, if the "direction of the vehicle involved in crash" was not matched with either of the TMC directions or was not available, researchers used only "distance" to assign the crashes.

Conflated Data

The final conflated datasets contain TMC-level crash data for the two States. Table 9 shows the basic geometric and traffic characteristics by facility type. Six facility types were considered for preliminary analysis:

- Rural interstate.
- Rural two-lane.
- Rural multilane undivided.
- Rural multilane divided.
- Rural others (for example, roadways with missing information, such as the number of lanes or median width).

Figure 4 illustrates the number of crashes by severity type in these States. It shows that Ohio experienced a significantly higher number of total crashes than Washington in 2015 and that the number of non-injury crashes in Ohio was more than double the amount in Washington.

Table 9. TMCs and Other Key Characteristics.

Functional Class	TMCs	Total TMC Lengths in Both Directions (mi)	Average Segment Length (mi)	AADT (vpd)
Washington				
All Rural	1,122	5,086	4.53	14,824
Rural Interstate	268	948	3.54	37,053
Rural Two-Lane	695	3,552	5.11	5,818
Rural Multilane Undivided	32	83	2.58	14,664
Rural Multilane Divided	107	439	4.10	18,919
Rural Others	20	66	3.29	8,273
Ohio				
All Rural	1,568	5,098	3.25	15,119
Rural Interstate	347	1,532	4.41	36,722
Rural Two-Lane	667	1,907	2.86	5,609
Rural Multilane Undivided	67	105	1.57	12,133
Rural Multilane Divided	472	1,516	3.21	13,370
Rural Others	15	38	2.53	6,594

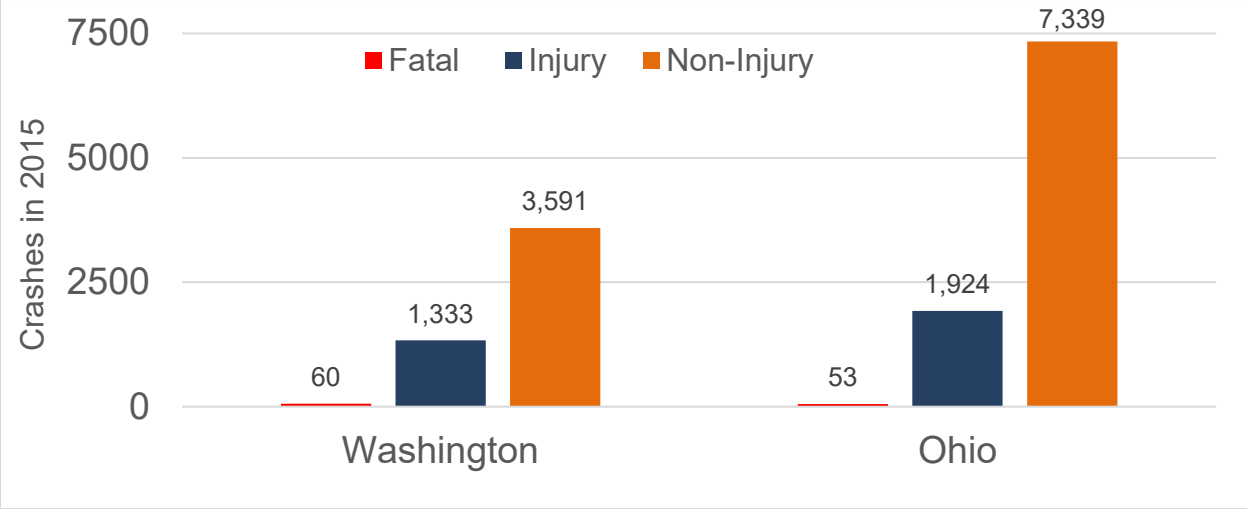


Figure 4. Number of Crashes by Severity Types.

Because the presence of interactions may impact vehicle speed on traffic crashes, understanding the interactions due to the presence of intersections was an important consideration in this study. In the HSIS database, crashes can be identified as segment (non-intersection) or intersection/intersection-related crashes (around 10 percent of total crashes were classified as intersection crashes). The current analysis considered both all crashes and non-intersection crashes to determine the bias associated with the presence of an intersection. Table 10 lists the counts of all crashes and non-intersection crashes based on different facility types. The table shows that Ohio roadways have a lower proportion of crashes involving fatalities than Washington.

Table 10. All Crashes and Non-Intersection Crashes by Facility Type.

Functional Class	All Crashes				Non-Intersection Crashes			
	Total (KABCO)	Fatal (K)	Injury (ABC)	Non-Injury (PDO)	Total (KABCO)	Fatal (K)	Injury (ABC)	Non-Injury (PDO)
Washington								
All Rural	5,897	73	1,678	4,146	2,820	40	776	2,004
Rural Interstate	2,272	20	587	1,665	—*	—	—	—
Rural Two-Lane	2,731	41	835	1,855	2,130	33	582	1,515
Rural Multilane Undivided	120	1	35	84	67	—	25	42
Rural Multilane Divided	643	9	183	451	513	5	139	369
Rural Others	131	2	38	91	110	2	30	78
Ohio								
All Rural	10,251	68	2,264	7,919	3,801	30	815	2,956
Rural Interstate	5,619	23	1,128	4,468	—	—	—	—
Rural Two-Lane	2,233	27	587	1,619	1,761	20	405	1,336
Rural Multilane Undivided	267	1	76	190	191	—	46	145
Rural Multilane Divided	2,091	17	462	1,612	1,820	10	359	1,451
Rural Others	41	—	11	30	29	—	5	24

*Note: * For interstate roadways, crashes are non-intersection-related.*

The project team developed a webpage (see appendix F) that includes descriptive statistics and data visualization graphics in both static and interactive formats. The links provide more detailed insights about the variables of interest.

CHAPTER 3. STATISTICAL MODELS

This chapter provides a brief literature review of studies on the relationship between speed and crashes, followed by an in-depth analysis of the statistical model runs for the three data structures. Each statistical model section provides a brief theoretical introduction, model results, and model inferences.

STUDIES ON SPEED-CRASH RELATIONSHIP

Although speed is considered a major contributing factor of roadway crashes, research findings are inconsistent. While some studies have found that higher speeds are associated with an increased likelihood of collisions, other studies have found the opposite, stating that higher speeds are associated with a lower probability of collisions. A few studies have established statistical models between operating speed and crash occurrence. However, since traffic crashes are random and sporadic events with low occurrence probabilities, spatiotemporal aggregations are needed when formulating the analysis datasets. Findings from the related literature review are summarized in table 11.

Abdel-Aty and Radwan (2000) studied speed in a different form, capturing the magnitude of speeding relative to the posted speed limit. This speeding indicator variable was shown to increase the likelihood of the accident involving male and young drivers. The preliminary analysis of a study conducted by Taylor et al. (2000) based in the United Kingdom revealed that average speed was positively related to crash frequency. The authors attributed this finding to the difference in road quality (urban versus rural) of the road segments sampled; therefore, they created homogenous groups through which the effects of road quality on the relationship between collisions and speed could be captured.

Pei et al. (2012) showed that crash risk is negatively associated with average speed when controlling for distance exposure (the distance traveled on the road), which goes against research that argues that roadway segments designed for higher speeds should deliver better road safety performance. Pei et al. (2012) also revealed that there may be other explanatory factors, such as road design, weather conditions, and temporal distribution, on the relationship between speed and crash risk. Although the information on other possible factors related to crash occurrence, including traffic composition and driver behavior, was not available for this study, these factors are worthy of exploration in future research.

Yu et al. (2013) employed a Bayesian inference method to model crashes using 1 year's worth of crash data on I-70 in Colorado. Their model included real-time weather, traffic, and road geometry variables and indicated that the weather condition variables play a significant role in crash occurrence. This study also suggested that lower speeds at the crash segment and higher vehicle occupancy on the road at the upstream segment 5–10 minutes before the crash time increases the likelihood of crashes and could be an indication of congestion. However, lower speed and higher crash risk can both be the result of severe weather conditions, in which case the relationship between the two would be affected by a confounding variable.

Table 11. Studies Focusing on the Association between Crash and Operating Speeds.

Study	Analysis Level	Roadway/ Location	Speed Measures	Operating Speed Data Source	Key Findings on Speed-Crash Relationship
Abdel-Aty and Radwan (2000)	Segment	Principal arterial, Florida	Speeding relative to posted speed limits	Crash data	Speed measure (speeding relative to posted speed limits) variable was shown to **increase** the crash involvement of male and young drivers.
Taylor et al. (2000)	Segment	Different roadways, UK	Average speed	Road tubes	Excessive speed indicator was strongly and **positively** associated with crashes.
Pei et al. (2012)	Segment	Both urban and rural, Hong Kong	Standard deviation of average speed	Annual traffic census (ATC)	Crash risk was **negatively** associated with average speed when controlling for distance exposure.
Yu et al. (2013)	Segment	Freeways, Colorado	Speed info prior to crash occurrence	Radars	**Negative** relationships were shown between speed and crash occurrence.
Roshandel et al. (2015)	Individual crash	Urban freeways	Average speed	Meta-analysis	Increasing values of average speed were associated with **reduced** crash risk.
Gargoum and El-Basyouny (2016)	Segment	Urban two-lane, Canada	Standard deviation of speed	Speed survey operations	Standard deviation of speed seemed to be **negatively** related to collisions.
Imprialou et al. (2016)	Traffic operation scenarios	Strategic road network, UK	Grouped average speed prior to crash occurrence	Inductive loop detectors	Results of the condition-based approach showed that high speeds trigger crash frequency. The outcome of the segment-based model was the opposite, suggesting that the speed-crash relationship is **negative** regardless of crash severity.
Yu et al. (2018)	Segment	Urban expressway, China	Average speed	Using algorithm	Segment-based crash frequency analysis revealed a **negative** relationship between the crash and speed.
Banihashemi et al. (2019)	Segment	Urban interstate, Washington	Operating and posted speed differential	NPMRDS	Severity of crashes measured by the fatal and injury/total crashes ratio **increased** by increasing the speed differential.

Gargoum and El-Basyouny (2016) conducted a study in which they attempted to model the relationship between average speed and crash counts while considering effects from confounding factors (characteristics of the road, climate, traffic, and vehicle speeds) using structural equation modeling. They collected data from 353 different two-lane urban road segments across the city of Edmonton during 2009–2013 and found that the standard deviation of speed seemed to be negatively related to crash frequencies (i.e., increases in the deviation of speeds from the average were related to decreases in crash frequency, and vice versa); however, this relationship was only statistically significant at the 10 percent significance level (p-value = 0.088). The results of Imprialou et al.'s (2016) segment-based crash-speed relationship study also showed the relationship was negative regardless of crash severity.

In a recent study by Yu et al. (2018), the impacts of aggregation approaches (a segment-based dataset grouped by roadway segment, a scenario-based crash dataset aggregated by traffic operating scenarios, and a disaggregated crash level from individual crashes) on relationship analyses were investigated based on the advanced traffic sensing data of urban expressway systems in Shanghai. Crash frequency analyses with segment-based and scenario-based approaches were first conducted, and then crash risk analyses were developed at the individual crash level. The segment-based crash frequency analysis revealed a negative relationship between speed and crash frequency. Yu et al.'s findings suggested that during congestion periods (i.e., low and moderate speed conditions), an increase in operating speeds is associated with reduced crash likelihoods. Another recent study conducted by Banihashemi et al. (2019) found that the severity of crashes (a ratio of fatal and injury [FI] crashes to total crashes) increased as the speed variability increased.

Based on the differing findings regarding the relationship between different speed measures and crash risks across the literature, an opportunity exists to further advance this debate. This study integrated and analyzed crash and speed data from Ohio and Washington to contribute to this ongoing discussion.

ANNUAL-LEVEL CRASH ANALYSIS

In this study, the project team used dataset structure 1 (segment-level data) to model annual-level crash risk and operating speed for Ohio and Washington State. This section provides a brief overview of exploratory data analysis to provide insights into the data. Table 12 through table 15 list the summary statistics (i.e., mean, standard deviation, minimum, maximum) of the key geometric, traffic, and environmental variables and selected speed measures for Ohio and Washington for different facility types.

The speed data on TMC segments are recorded by epoch (5-minute bins in the raw data). However, the data are not recorded for every epoch; thus, there are a considerable number of missing values. To overcome this issue, the project team averaged the data on a daily or monthly basis. Every observation refers to a monthly average speed at a given epoch, which is calculated as follows:

$$Monthly\ Aveage\ Speed_{epoch\ e,TMC\ i} = \frac{1}{n}\sum_{n=1}^{31} Speed_{day_n,epoch\ e,TMC\ i} \quad (1)$$

where:

$Monthly\ Average\ Speed_{epoch\ e,TMC\ i}$ = the average epoch e speed at segment i over a month.
n = the number of days in a given month.
$Speed_{day_n,epoch\ e,TMC\ i}$ = the NPMRDS speed on day n and epoch e at segment i.

To minimize the missing value issues, the epochs were summed into 15-minute epochs, resulting in 96 speed records per day. However, in preliminary evaluations, both the 5-minute and 15-minute speed data did not provide adequate measures about the relationships among the speed, safety, and operational characteristics of the roadway segment due to a large number of missing values. Therefore, other measures of speed were considered. Since the speed data are autocorrelated, speeds observed at consecutive epochs are not necessarily independent of each other. Because the distributions of the operational speeds vary from facility to facility for different spatial and temporal factors, several speed measures (for example, peak-hour 85th percentile speed) were examined for the model development for different facility types. For the model development documented in this study, the following speed measures were considered:

- Average hourly speed (SpdAvg).
- Average hourly speed during non-peak and non-event (1 hour before and 1 hour after a crash occurrence) periods (SpdNPNE).
- Standard deviation of hourly operating speeds (SDHrSpd).
- Standard deviation of monthly operating speeds (SDMonSpd).
- Differences in the operating speeds during weekdays and weekends (SpdW_W).

Exploratory Data Analysis

Based on the facility types used in the HSM, the project team considered three major facility types for model development: (a) rural interstate highways, (b) rural two-lane highways, and (c) rural multilane highways. Table 12 lists summary statistics for all relevant variables of interest considered for the rural interstate roadways. One interesting finding is that the percentage of curves per segment in Ohio is comparatively lower than in Washington. Both mean values of SpdAvg and SpdNPNE are lower in Washington than in Ohio. The mean values of the speed variability measures (SDHrSpd, SDMonSpd, and SpdW_W) for both States are within the range of 0.9–1.5 mph.

Table 12. Descriptive Statistics of Rural Interstate Roadways (per Segment).

	Code	Mean	SD	Min	Max
Ohio					
Total Crashes	KABCO	16	18	0	147
Fatal and Injury Crashes	KABC	3	4	0	27
PDO Crashes	PDO	13	14	0	120
Segment Length (mi)	Len	4	3	0.1	14
Annual Average Daily Traffic (vehicle per day)	Vol(AADT)	36,722	11,360	4,870	75,040
Lane Width (ft)	LW	55	11	32	73
Presence of Intersection	PIntPre	0	0	0	0
Percentage of Curve	PerHC	2.3	54.2	0	21.2
Percentage of Days with Precipitation	PPrcp	23.5	44	2	67
Average Hourly Speed (mph)	SpdAvg	63.0	5.4	30.5	85.1
Average Hourly Non-Peak Non-Event Speed (mph)	SpdNPNE	65.9	4.7	33.3	85.0
Standard Dev. of Hourly Operating Speeds (mph)	SDHrSpd	1.0	0.7	0.0	7.3
Standard Dev. of Monthly Operating Speeds (mph)	SDMonSpd	1.0	0.8	0.0	7.4
Avg. Spd. Diff. in Weekday/Weekend (mph)	SpdW_W	1.1	0.5	0.7	7.2
Washington					
Total Crashes	KABCO	8	9	0	58
Fatal and Injury Crashes	KABC	2	3	0	18
PDO Crashes	PDO	6	6	0	40
Segment Length (mi)	Len	4	2.7	0.1	11
Annual Average Daily Traffic (vehicle per day)	Vol(AADT)	37,053	24,157	0	128,331
Lane Width (ft)	LW	58	14	31	100
Presence of Intersection	PIntPre	0	0	0	0
Percentage of Curve	PerHC	23.4	27.3	0.0	97
Percentage of Days with Precipitation	PPrcp	36.5	20.4	1	65
Average Hourly Speed (mph)	SpdAvg	59.8	6.3	17.2	67.2
Average Hourly Non-Peak Non-Event Speed (mph)	SpdNPNE	63.8	4.5	23.1	82.2
Standard Dev. of Hourly Operating Speeds (mph)	SDHrSpd	1.5	1.0	0.0	10.0
Standard Dev. of Monthly Operating Speeds (mph)	SDMonSpd	0.9	0.4	0.0	4.1
Avg. Spd. Diff. in Weekday/Weekend (mph)	SpdW_W	1.5	0.8	0.5	10

Table 13 provides summary statistics for all relevant variables of interest considered for the rural two-lane roadways. The statistics show that the percentage of curves and the percentage of days with precipitation for the rural two-lane roadways in Ohio are comparatively lower than for the rural two-lane roadways in Washington. Contrary to the interstate trends, both mean values of SpdAvg and SpdNPNE are higher in Washington than in Ohio. The mean values of the speed variability measures (SDHrSpd, SDMonSpd, and SpdW_W) for rural two-lane roadways in both States are within the range of 0.6–2.1 mph.

Table 13. Descriptive Statistics of Rural Two-Lane Roadways (per Segment).

	Code	Mean	SD	Min	Max
Ohio					
Total Crashes	KABCO	3	4	0	28
Fatal and Injury Crashes	KABC	1	1	0	9
PDO Crashes	PDO	2	3	0	19
Segment Length (mi)	Len	3	2	0.1	15
Annual Average Daily Traffic (vehicle per day)	Vol(AADT)	5,609	2,581	818	15,070
Lane Width (ft)	LW	25	4	18	48
Presence of Intersection	PIntPre	0.3	0.5	0	1
Percentage of Curve	PerHC	6.7	17.2	0	100
Percentage of Days with Precipitation	PPrcp	23	45	5	85
Average Hourly Speed (mph)	SpdAvg	44.9	9.6	13.5	71.3
Average Hourly Non-Peak Non-Event Speed (mph)	SpdNPNE	52.1	8.3	16.9	85.0
Standard Dev. of Hourly Operating Speeds (mph)	SDHrSpd	1.4	1.0	0.0	10.0
Standard Dev. of Monthly Operating Speeds (mph)	SDMonSpd	0.6	0.5	0.0	7.1
Avg. Spd. Diff. in Weekday/Weekend (mph)	SpdW_W	1.9	0.4	1.2	5.6
Washington					
Total Crashes	KABCO	4	5	0	34
Fatal and Injury Crashes	KABC	1	2	0	12
PDO Crashes	PDO	3	3	0	24
Segment Length (mi)	Len	5	4	0.1	25
Annual Average Daily Traffic (vehicle per day)	Vol(AADT)	5,818	4,490	0	26,493
Lane Width (ft)	LW	25	4	20	67
Presence of Intersection	PIntPre	0.4	0.5	0.0	1
Percentage of Curve	PerHC	33.7	27.3	0.0	100
Percentage of Days with Precipitation	PPrcp	37.1	21.9	0.0	70
Average Hourly Speed (mph)	SpdAvg	47.3	11.2	4.5	85.0
Average Hourly Non-Peak Non-Event Speed (mph)	SpdNPNE	55.0	10.0	6.8	85.0
Standard Dev. of Hourly Operating Speeds (mph)	SDHrSpd	1.7	1.6	0.0	10.0
Standard Dev. of Monthly Operating Speeds (mph)	SDMonSpd	0.8	0.6	0.0	4.4
Avg. Spd. Diff. in Weekday/Weekend (mph)	SpdW_W	2.1	0.6	1.9	6.2

Table 14 provides summary statistics for all relevant variables of interest considered for the rural multilane roadways. The statistics show that the percentage of curves and the percentage of days with precipitation for the rural multilane roadways in Ohio are comparatively lower than for the rural multilane roadways in Washington. Contrary to the two-lane trends and similar to the interstate trends, both mean values of SpdAvg and SpdNPNE are lower in Washington than in Ohio. The mean values of the speed variability measures (SDHrSpd, SDMonSpd, and SpdW_W) for rural multilane roadways in both States are within the range of 0.7–1.5 mph.

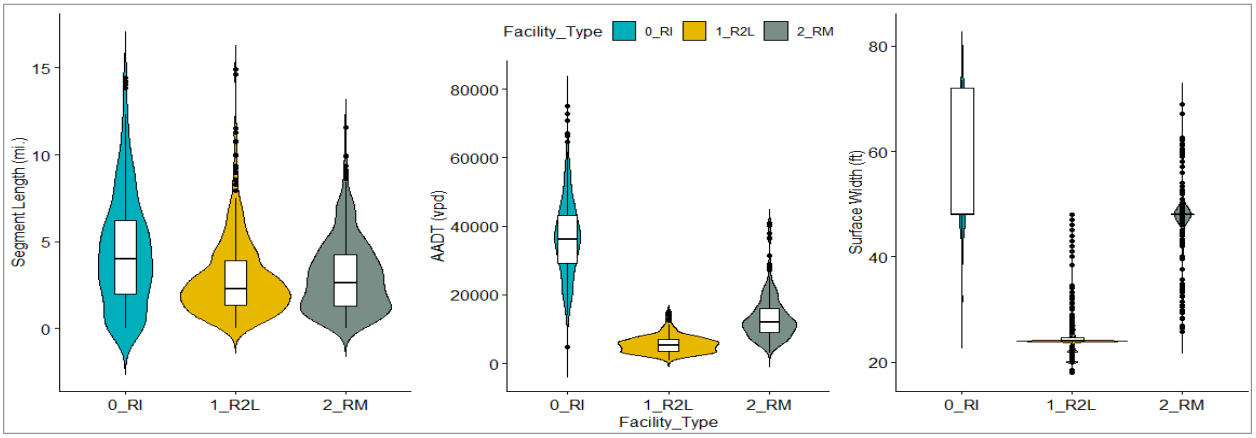
Table 14. Descriptive Statistics of Rural Multilane Roadways (per Segment).

	Code	Mean	SD	Min	Max
Ohio					
Total Crashes	KABCO	4	5	0	36
Fatal and Injury Crashes	KABC	1	2	0	10
PDO Crashes	PDO	3	4	0	27
Segment Length (mi)	Len	3	2	0.1	12
Annual Average Daily Traffic (vehicle per day)	Vol(AADT)	13,216	6,449	3,108	40,840
Lane Width (ft)	LW	48	5	26	69
Presence of Intersection	PIntPre	0.2	0.7	0	1
Percentage of Curve	PerHC	5.1	15.6	0	100
Percentage of Days with Precipitation	PPrcp	22	16.6	5	80
Average Hourly Speed (mph)	SpdAvg	54.3	12.1	18.6	85.0
Average Hourly Non-Peak Non-Event Speed (mph)	SpdNPNE	59.6	9.6	26.4	86.1
Standard Dev. of Hourly Operating Speeds (mph)	SDHrSpd	1.2	1.4	0.0	10.0
Standard Dev. of Monthly Operating Speeds (mph)	SDMonSpd	0.7	0.5	0.0	4.8
Avg. Spd. Diff. in Weekday/Weekend (mph)	SpdW_W	1.2	0.7	0.9	6.1
Washington					
Total Crashes	KABCO	5	6	0	32
Fatal and Injury Crashes	KABC	2	2	0	11
PDO Crashes	PDO	4	5	0	26
Segment Length (mi)	Len	4	3	0.1	12
Annual Average Daily Traffic (vehicle per day)	Vol(AADT)	17,940	12,508	0	77,827
Lane Width (ft)	LW	48	7	29	76
Presence of Intersection	PIntPre	0.5	0.5	0	1
Percentage of Curve	PerHC	30.9	29.3	0	100
Percentage of Days with Precipitation	PPrcp	44.5	23.1	19.2	95.1
Average Hourly Speed (mph)	SpdAvg	52.0	13.3	14.5	85.0
Average Hourly Non-Peak Non-Event Speed (mph)	SpdNPNE	57.8	11.0	20.8	85.0
Standard Dev. Of Hourly Operating Speeds (mph)	SDHrSpd	1.5	1.7	0.3	10.0
Standard Dev. Of Monthly Operating Speeds (mph)	SDMonSpd	0.8	0.8	0.0	4.7
Avg. Spd. Diff. in Weekday/Weekend (mph)	SpdW_W	0.8	1.3	0.0	9.2

Geometric and Traffic Variables

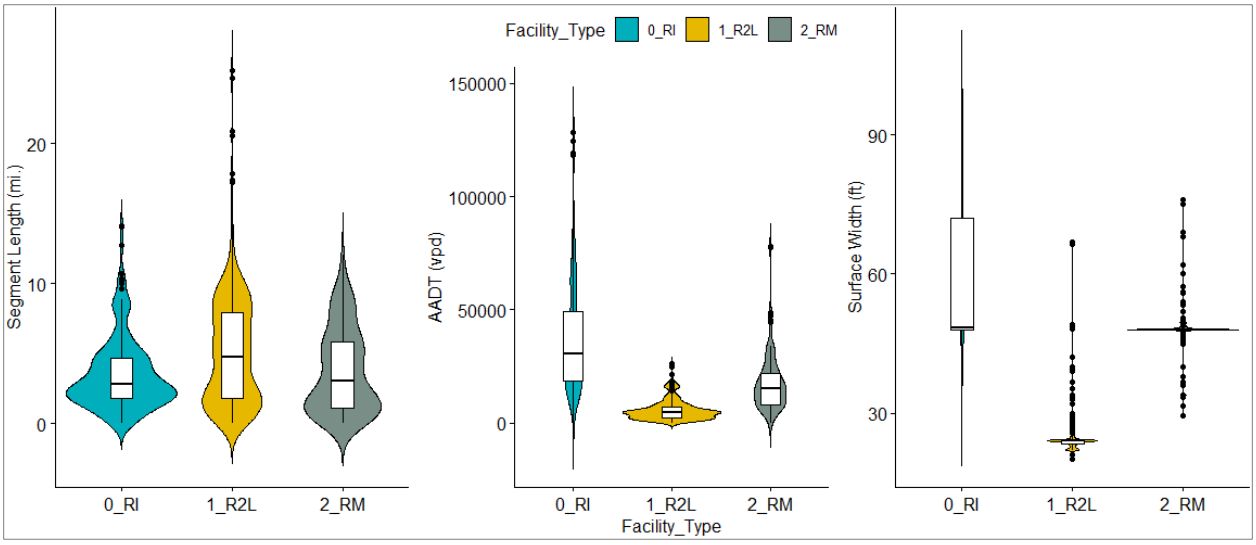
Figure 5 and figure 6 show variabilities between segment length in miles, AADT in vehicles per day, and surface width in feet on the y-axes (from left to right) for different facility types (along the x-axis) in Ohio and Washington. The violin plots compare distributions of quantitative data across categorical levels (like box and whisker plots) and provide a kernel-density estimation by illustrating the distribution of the values in the form of a mirrored histogram or density plot. The shape or width of the violins is the visual display of the frequencies. This plot can be an effective and attractive way to show multiple distributions of data at once; however, the estimation procedure is influenced by the sample size, so violins for relatively small samples might look misleadingly smooth. Additionally, the lower smoothing points of these estimations go beyond zero values, but the actual speed measures are always greater than zero. The line inside the white rectangular box indicates that median and outer edges are interquartile ranges. The whiskers

show a 95 percent confidence interval. Figure 5 and figure 6 provide insights about the key variables and their potential to differentiate between facility types by state.



Note: 0_RI = Rural Interstate; 1_R2L = Rural Two-Lane; 3_RM = Rural Multilane

Figure 5. Distribution of Segment Length, AADT, and Surface Width (Ohio Data).



Note: 0_RI = Rural Interstate; 1_R2L = Rural Two-Lane; 3_RM = Rural Multilane

Figure 6. Distribution of Segment Length, AADT, and Surface Width (Washington Data).

Figure 7 illustrates the density plots of AADT by facility types for Washington and Ohio, respectively. The conflated data have speed measures in each direction. However, AADT and other geometric variables combine both directions. The plots indicate that an increase in AADT is associated with a higher facility type (e.g., rural interstate). The trends in AADT distribution are similar between the States.

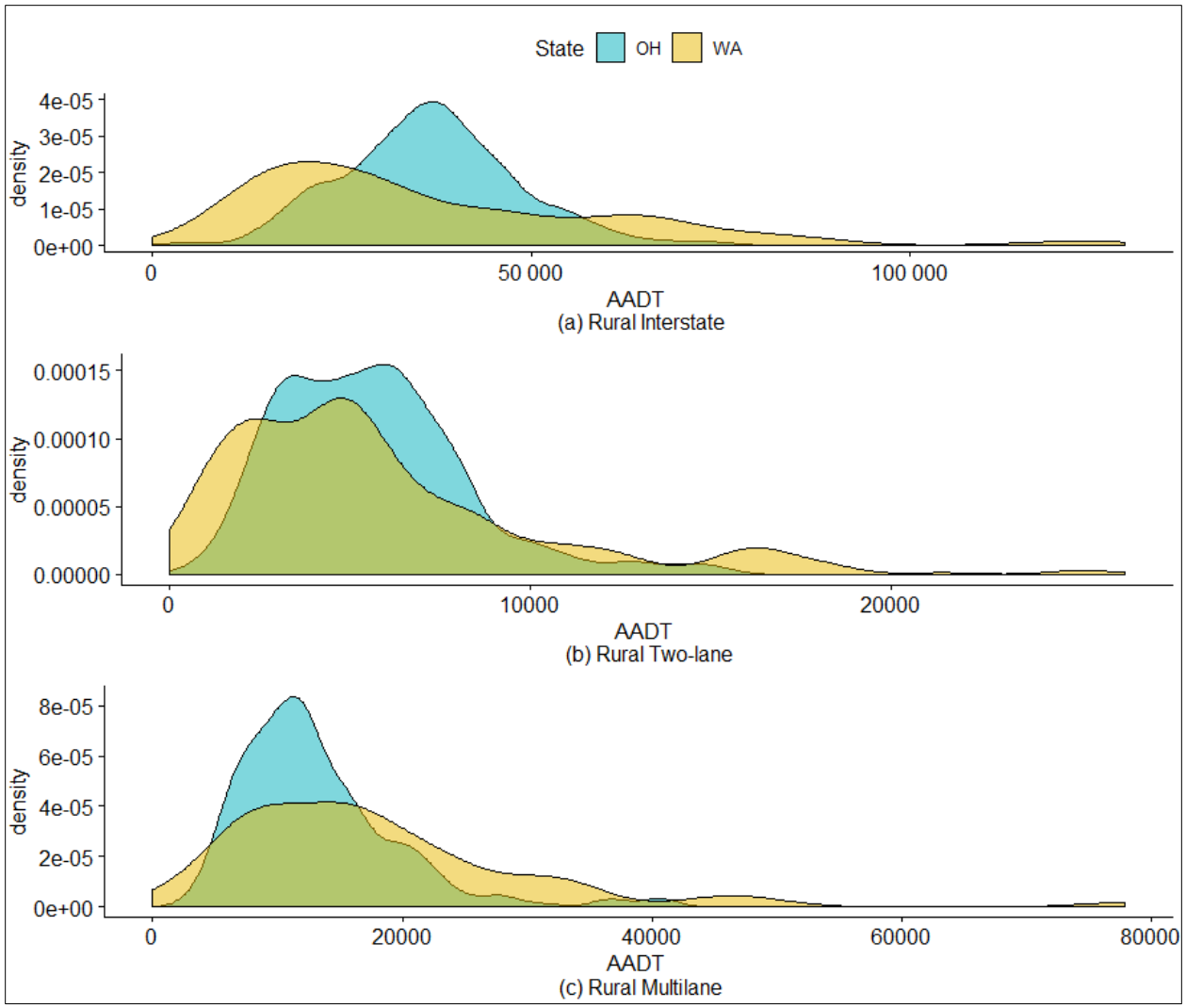
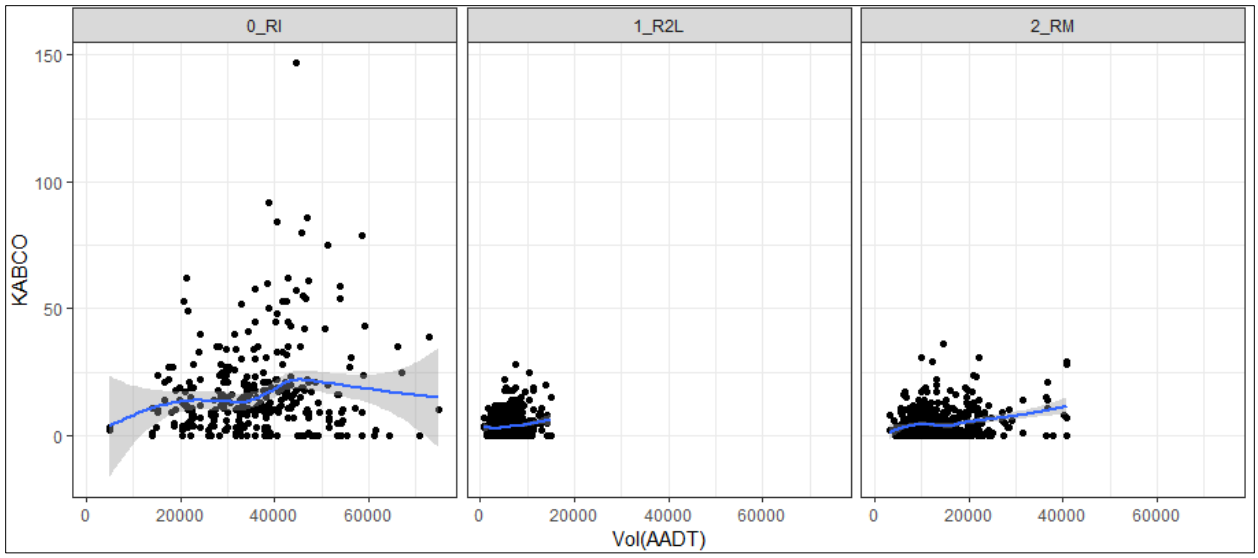


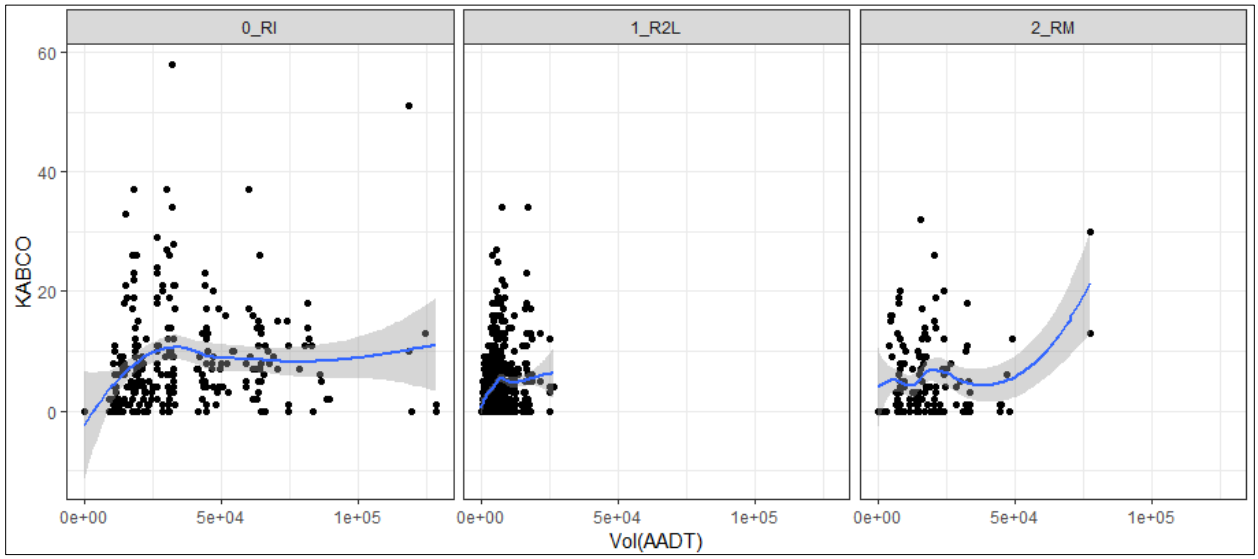
Figure 7. Density of Traffic Volumes by States.

Since the safety literature shows the significance of length and AADT on crashes, the project team plotted AADT against total crashes. Figure 8 and figure 9 show the scatterplots of the annual crash frequency versus the total AADT for each of the TMC segments in Ohio and Washington, respectively. The blue line in these plots indicates a nonparametric scatterplot nonlinear smoother. The gray areas surrounding the blue line indicate the 95 percent confidence interval boundary. Based on these plots, it appears that the relationship holds true up to certain ranges, but it is not as strong as the literature suggests. For example, in Ohio rural interstates, the TMC segments with the highest total crashes are not necessarily the TMC segments with the highest AADT. However, the overall trends are upward.



Note: 0_RI = Rural Interstate; 1_R2L = Rural Two-Lane; 3_RM = Rural Multilane

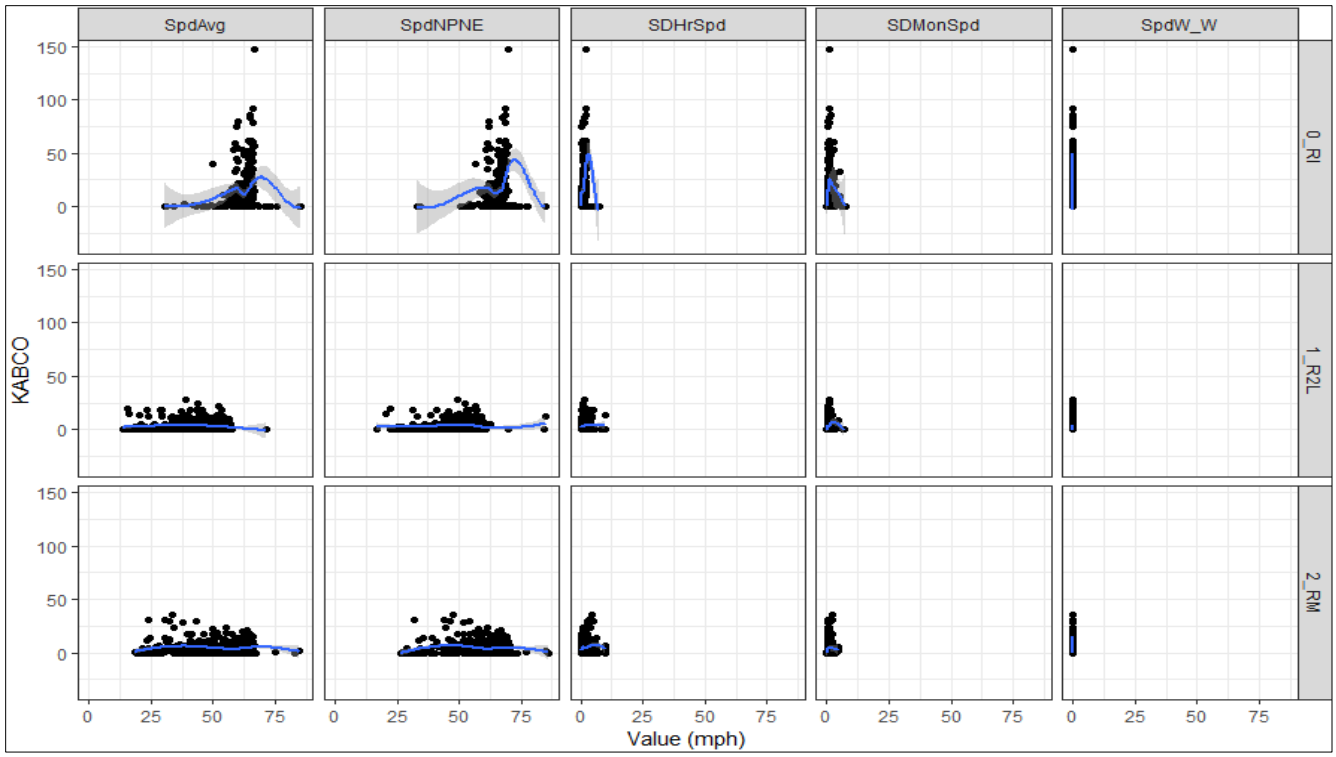
Figure 8. Annual Crash Frequency vs. Total AADT (Ohio).



Note: 0_RI = Rural Interstate; 1_R2L = Rural Two-Lane; 3_RM = Rural Multilane

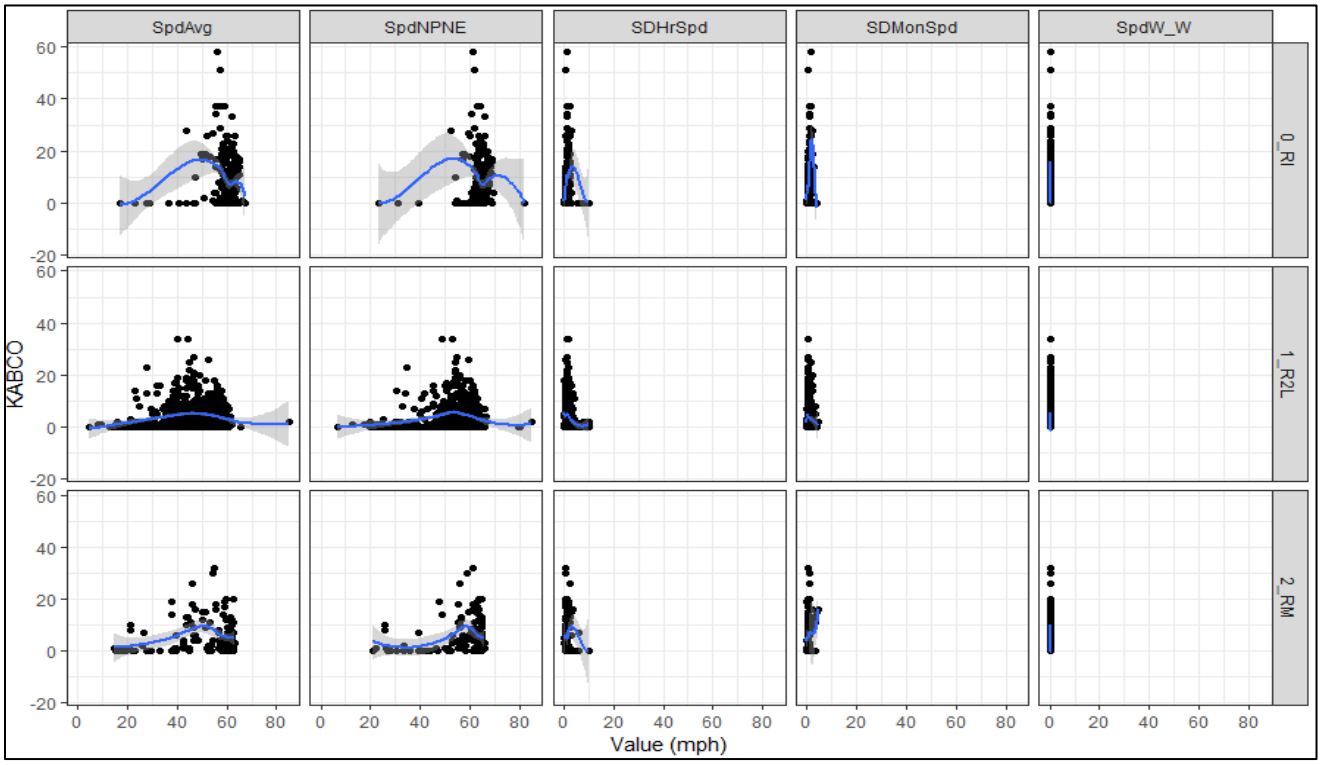
Figure 9. Annual Crash Frequency vs. Total AADT (Washington).

Figure 10 and figure 11 show the scatterplots of the annual crash frequency versus the selected speed measures for each of the TMC segments in Ohio and Washington, respectively. The speed measures range from 0 to 80 mph. The positive assertion between speed and crash frequencies holds true up to certain ranges for the first two speed measures (SpdAvg and SpdNPNE). It is difficult to depict any trend between the speed variability measures (SDHrSpd, SDMonSpd, and SpdW_W) and crashes since the values ranges between 0 and 10 mph. However, slight upward trends are visible for most of the speed variability measures.



Note: 0_RI = Rural Interstate; 1_R2L = Rural Two-Lane; 3_RM = Rural Multilane

Figure 10. Speed Measures vs. Total Crashes (Ohio).

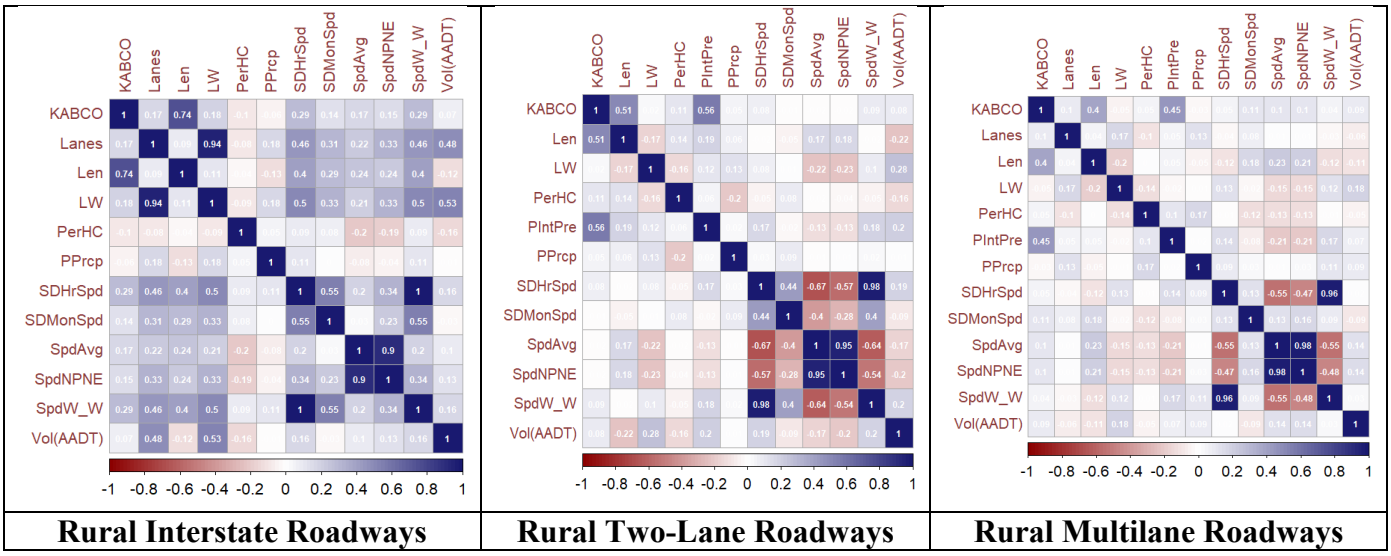


Note: 0_RI = Rural Interstate; 1_R2L = Rural Two-Lane; 3_RM = Rural Multilane

Figure 11. Speed Measures vs. Total Crashes (Washington).

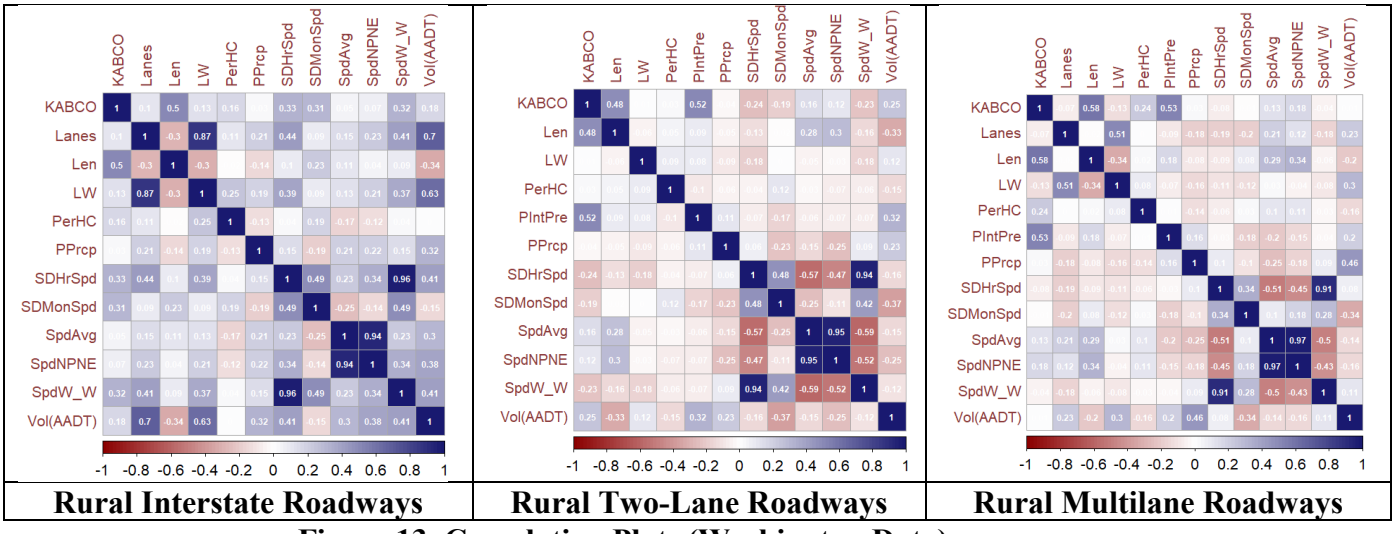
Correlation Analysis

The project team developed and used Pearson correlation plots for each facility type (figure 12 and figure 13) to show the positive or negative correlation between the variables. In these plots, blue means positive and red means negative. The stronger the color, the larger the correlation magnitude. The variables of the rural interstate roadways have higher correlation values than the other two roadways. The correlation plots shown here are based on raw data and total (KABCO) crashes only. The plots show that the segment length, traffic volume, and five speed measures (exception: standard deviation of hourly operating speeds in both rural two-lane and multilane roadways in Washington and standard deviation of monthly operating speeds in rural two-lane roadways in Washington) are positively associated with total crashes. The project team considered all major geometric variables, five speed measures, and weather-related variables after removing some outliers.



Rural Interstate Roadways | **Rural Two-Lane Roadways** | **Rural Multilane Roadways**

Figure 12. Correlation Plots (Ohio Data).



Rural Interstate Roadways | **Rural Two-Lane Roadways** | **Rural Multilane Roadways**

Figure 13. Correlation Plots (Washington Data).

Safety Performance Functions by Facility Types

Separate models were developed for total (KABCO), KABC, and PDO crashes. Experience with the regression-based calibration of SPFs and crash modification factors (CMFs) using total, KABC, and PDO crashes indicates that the calibration coefficients often vary among model types for common variables. Some of this variation is likely due to the fact that geometric elements often have a different effect on KABC crashes than on PDO crashes. Further, it is widely recognized that PDO crash counts vary widely on a regional basis due to significant variations in the reporting threshold. When crash frequency varies systematically from county to county, district to district, and State to State because of formal and informal differences in the reporting threshold, the use of PDO crash data to build PDO crash prediction models may yield inaccurate results about the variable influence. Thus, the project team developed models for three severity levels to understand the difference in variable effects. Except for curve length and radius, the interaction between the variables was not considered. As noted by Srinivasan and Bauer (2013), interactions are not usually considered during SPF development. The authors mentioned that there is no easy way to identify which interactions are important and how they should be included in a model unless there is some theoretical reason for including certain interactions.

Most available crash prediction models currently predict crashes for both directions of travel combined. The project team developed directional prediction models that incorporate the distinct directions of travel when predicting crashes. These directional prediction models (known as SPFs) have the potential to be useful to many State DOTs. The models show that traffic volume and length are positively associated with a high number of crashes. The discussions about these two variables are not provided in the model explanation section due to the nature of their presence and similarity in all models.

Models Developed for Interstate Roadways

The model presented below was informed by findings from several preliminary regression analyses:

$$N_i = L \times e^{b_{0,i}+b_{aadt}\ln(AADT)} \times CMF_{hc} \times CMF_{sdif} \times CMF_{svar1} \times CMF_{svar2} \times CMF_{sff} \times CMF_{prec}; I = 4 \text{ or } 6 \quad (2)$$

with,

$$CMF_{hc} = 1.0 + b_{hc}\left(\frac{5730}{R_{min}}\right)^2\left(\frac{L_c}{L}\right)$$
$$CMF_{sdif} = e^{b_{sd}(SpdDiff)}$$
$$CMF_{svar1} = e^{b_{sv1}(I_{svar1})}$$
$$CMF_{svar2} = e^{b_{sv2}(I_{svar2})}$$
$$CMF_{sff} = e^{b_{sff}(SFF)}$$
$$CMF_{prec} = e^{b_{prec}(p_{prec})}$$

where:

N_i = predicted annual average crash frequency for model *i* (*i* = four or six lanes).
L = segment length, miles.
$AADT$ = average annual daily traffic, vehicles per day.
CMF_{hc} = crash modification factor for horizontal curve.
CMF_{sdif} = crash modification factor for the speed difference between weekends and weekdays.
CMF_{svar1} = crash modification factor for variance in hourly operating speeds.
CMF_{svar2} = crash modification factor for variance in monthly operating speeds.
CMF_{sff} = crash modification factor for free-flow speed.
CMF_{prec} = crash modification factor for precipitation.
R_{min} = radius of the sharpest curve, feet.
L_c = total length of all horizontal curves on the segment.
$SpdDiff$ = percent difference of operating speeds between weekends and weekdays.
I_{svar1} = indicator variable for high variance in hourly operating speeds within a day (= 1 if hourly standard deviation is > 1 mph; = 0 otherwise).
I_{svar2} = indicator variable for high variance in monthly operating speeds within a year (= 1 if monthly standard deviation is > 1 mph; = 0 otherwise).
SFF = free-flow speed, mph.
p_{prec} = percent of days with precipitation.
b_j = calibrated coefficients.

The project team calculated the Pearson correlation coefficient and found that the correlation between independent variables is very small, as shown in table 15. For this reason, the project team decided that the interaction between the variables in the model is not necessary.

Table 15. Correlation Analysis Results.

	AADT	$SpdDiff$	I_{svar1}	I_{svar2}	SFF
AADT	1				
$SpdDiff$	0.23105	1			
I_{svar1}	0.11233	0.28541	1		
I_{svar2}	−0.0826	0.09433	0.23296	1	
SFF	0.1028	−0.0038	0.02121	−0.0222	1

The predictive model calibration process consisted of the simultaneous calibration of four-lane and six-lane models and variable effects using the aggregate model. The simultaneous calibration approach was needed because the AADT and speed-related effects were common to four-lane and six-lane highways. The inverse dispersion parameter, *K* (which is the inverse of the overdispersion parameter), is allowed to vary with the segment length. The inverse dispersion parameter is calculated using:

$$K = L \times e^{k} \quad (3)$$

where:

K = inverse dispersion parameter.
k = calibration coefficient for inverse dispersion parameter.

Table 16 lists the model outputs of rural interstate roadways. Appendix B includes all individual models.

Table 16. Model Estimation Results of Yearly Crash Frequencies at Segments (Rural Interstate).

Variable	Two States			Ohio			Washington		
	KABCO	KABC	PDO	KABCO	KABC	PDO	KABCO	KABC	PDO
Traffic volume (AADT)	0.7613	0.8221	0.7594	0.8028	1.0500	0.7611	0.6358	0.6498	0.7098
Lane width (LW)	—	—	—	—	—	—	—	—	—
Percentage of curve (PerHC)	0.0825	0.0827	0.06865	−0.6258	—	−0.7141	0.0909	0.0780	0.0665
Avg. spd. diff. in weekday/weekend (SpdW_W)	0.1068	—	0.0992	—	—	—	0.1568	0.1063	0.1439
Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—	—	—	—	—	—
Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—	—	—	—	—	—
Average hourly non-peak non-event speed (SpdNPNE)	−0.0378	−0.0551	−0.0406	−0.0547	−0.0549	−0.0578	—	—	—
Presence of intersection (IntPre)	NA	NA	NA	NA	NA	NA	NA	NA	NA
Percentage of days with precipitation (PPrcp)	—	—	—	—	—	—	—	—	—
Added effect of Ohio	0.6284	0.4119	0.6926	NA	NA	NA	NA	NA	NA
Inverse dispersion parameter for 4-lane segments	−0.4359	−0.4875	−0.4972	−0.4634	−0.5888	−0.4935	—	—	—
Inverse dispersion parameter for 6-lane segments	−0.4672	—	−5.0697	−0.7741	−0.6853	−0.7845	—	—	—
Intercept	−4.9668	−5.8519	−5.0697	−3.5814	−7.7923	−3.1117	−6.2446	−4.9994	−7.3084

Note: A dash (—) = not significant at the 95% level; NA = not applicable.

The explanations of the model outcomes are provided below.

Percentage of curve: This variable represents the combination of the presence of horizontal curves and the radius of the sharpest curve on the segment. The coefficient for the model of both States together as well as Washington only shows that as the proportion of horizontal curvature increases, the number of crashes is expected to increase. Moreover, an increase in the sharpness of the curve tends to be associated with high crash frequencies. The coefficient for Ohio data is significant but counterintuitive, which could be due to the correlation with other unknown factors, the curve-related crash reporting issues, or the missing values related to curve

information in the Ohio HSIS data. It is also important to note that Ohio roadways with curves show comparatively lower in proportion than Washington.

Average speed difference in weekday/weekend: This variable represents the percent difference of operating speeds between weekends and weekdays. Generally, this variable is always greater than zero because the operating speeds during weekends are usually higher than weekdays. The variable value is much greater than zero if the road experiences frequent congestion on the weekday or when over-speeding is frequent on weekends due to free-flow conditions. The coefficient is positive in all models, which indicates that as the difference in operating speeds between weekends and weekdays increases for a particular segment, the number of crashes is expected to increase.

Standard deviation in hourly operating speeds: This variable represents the operating speed variation among the hours of a day. The coefficient is not significant in any model. It is possible that the variation among the hours is negligible because rural interstates tend to have similar speeds throughout the day.

Standard deviation in monthly operating speeds: This variable represents the operating speed variation between the months of a year. The coefficient is not significant in any model. It is possible that the variation among the hours is negligible because rural interstates tend to have similar speeds throughout the various months.

Non-peak non-event operating speed: This variable represents the operating speed under the non-peak and non-event conditions. The coefficient is negative in almost all cases and is statistically significant, which means that as the non-peak and non-event speeds increase, crashes are supposed to decrease. This finding could be because well-designed and high-standard roads are generally associated with higher non-peak and non-event speeds.

Percentage of days with precipitation: This variable represents the percent of days with some level of precipitation. It is insignificant in all models, meaning precipitation has no significant effect on crashes occurring on interstates. Generally, precipitation is associated with higher wet-weather crashes, which constitute a minor proportion of all crashes, and this element may be why this model fails to show a significant effect on all crashes.

State effect: When the two States' data are combined, the coefficient for Ohio is positive and significant, which means that—controlling for the other variables—Ohio is associated with more crashes than Washington. This finding could be due to differences in weather, terrain, reporting threshold, and other variables that were not considered in the model.

Models Developed for Two-Lane Highways

Different variable combinations and various model forms were examined to identify the best possible relationship between the number of crashes and independent variables. The model presented below was informed by findings from several preliminary regression analyses. This model form includes variables that are intuitive, are in line with previous findings, and best fit the data.

$$N_{tl} = Len \times e^{b_0 + b_{aadt} \ln(AADT)} \times CMF_{lw} \times CMF_{hc} \times CMF_{sdif} \times CMF_{svar1} \times CMF_{svar2} \times CMF_{sff} \times CMF_{int} \times CMF_{prec} \quad (4)$$

with,

$$CMF_{lw} = e^{b_{lw}(w_l - 12)}$$

$$CMF_{hc} = 1.0 + b_{hc}\left(\frac{L_c}{L}\right)$$

$$CMF_{sdif} = e^{b_{sd}(SpdDiff)}$$

$$CMF_{svar1} = e^{b_{sv1}(I_{svar1})}$$

$$CMF_{svar2} = e^{b_{sv2}(I_{svar2})}$$

$$CMF_{sff} = e^{b_{sff}(SFF)}$$

$$CMF_{int} = e^{b_{int} I_{int}}$$

$$CMF_{prec} = e^{b_{prec}(p_{prec})}$$

where:

N_{tl}	=	predicted annual average crash frequency (rural two-lane roadways).
Len	=	segment length, miles.
$AADT$	=	average annual daily traffic, vehicles per day.
CMF_{lw}	=	crash modification factor for lane width.
CMF_{hc}	=	crash modification factor for horizontal curve.
CMF_{sdif}	=	crash modification factor for speed difference between weekends and weekdays.
CMF_{svar1}	=	crash modification factor for variance in hourly operating speeds.
CMF_{svar2}	=	crash modification factor for variance in monthly operating speeds.
CMF_{sff}	=	crash modification factor for free-flow speed.
CMF_{int}	=	crash modification factor for presence of an intersection on the segment.
CMF_{prec}	=	crash modification factor for precipitation.
w_l	=	average lane width in both directions (ft).
L_c	=	total length of all horizontal curves on the segment.
$SpdDiff$	=	percent difference in operating speeds between weekends and weekdays.
I_{svar1}	=	indicator variable for high variance in hourly operating speeds within a day (= 1 if hourly standard deviation is > 1 mph; = 0 otherwise).
I_{svar2}	=	indicator variable for high variance in monthly operating speeds within a year (= 1 if monthly standard deviation is > 1 mph; = 0 otherwise).
SFF	=	free-flow speed, mph.
I_{int}	=	indicator variable for intersection presence (= 1 if present; = 0 otherwise).
p_{prec}	=	percent of days with precipitation.
b_j	=	calibrated coefficients (j = hc, sd, svar1, svar2, sff, int, prec).

Table 17 lists the model outputs of rural two-lane highways. Appendix B includes all individual models.

Table 17. Model Estimation Results of Yearly Crash Frequencies at Segments (Rural Two-Lane).

Variable	Two States			Ohio			Washington		
	KABCO	KABC	PDO	KABCO	KABC	PDO	KABCO	KABC	PDO
Traffic volume (AADT)	0.6048	0.6435	0.5679	0.3706	0.5967	0.3352	0.6962	0.6744	0.7048
Lane width (LW)	—	—	—	—	—	—	−0.0962	−0.1814	−0.0672
Percentage of curve (PerHC)	0.8681	1.0319	0.8230	—	—	—	1.1538	1.1829	1.1356
Avg. spd. diff. in weekday/weekend (SpdW_W)	—	—	0.02845	—	—	—	0.0444	—	0.0445
Standard dev. of hourly operating speeds (SDHrSpd)	—	0.1654	—	—	—	—	—	0.1878	—
Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—	—	0.3224	—	—	—
Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—	—	—	—	—	—
Presence of intersection (IntPre)	0.3163	0.2769	0.3296	0.4498	0.4074	0.3817	0.2278	0.1801	0.2398
Percentage of days with precipitation (PPrcp)	−0.5372	—	−0.6149	—	—	—	—	—	−1.0547
Added effect of Ohio	0.6332	0.4103	0.6691	NA	NA	NA	NA	NA	NA
Intercept	−5.8138	−7.6705	−5.8097	−3.6772	−6.8237	−3.4111	−6.5573	−7.9683	−6.9051

Note: A dash (—) = not significant at the 95% level; NA = not applicable.

The explanations of the model outcomes are provided below.

Percentage of curve: This variable represents the proportion of the segment with horizontal curves. The two-state and Washington models show positive and significant coefficients, demonstrating that as the proportion of horizontal curvature increases, the number of crashes increases. In preliminary models, the sharpness of the curve was not found to be statistically significant, which does not mean that the curve sharpness has no effect; instead, it is possible that the variability in the data variable may be too low to show a statistical significance for this dataset. Similar reasoning can be attributed to the insignificance of the horizontal curvature variable in Ohio.

Lane width: This variable represents the average of lane widths in both directions. For both States together, the coefficient is negative for KABCO crashes and KABC crashes, but not at the 95 percent significance level (not reported here). For Washington-only data, the variable is

significant and negative in all cases, which means that the increase in lane width on a particular segment is associated with lower crash frequencies.

Average speed difference in weekday/weekend: This variable represents the percent difference of operating speeds between weekends and weekdays. The variable value is much greater than zero if the road experiences frequent congestion during the weekday or if the weekend speeds are much higher due to fewer vehicles on these types of roads. The coefficient is significant and positive for both States (PDO crashes) and for Washington (KABCO and PDO crashes), which means that with higher weekend speeds (compared to weekday speeds), more crashes (especially PDO crashes) are expected to occur, perhaps due to congestion during weekdays or higher speeds on the weekends.

Standard deviation in hourly operating speeds: This variable represents the operating speed variation among the hours of a day, with an indicator variable of 1 for those segments where the standard deviation was greater than 1 mph. The coefficient is positive and statistically significant for KABC crashes in the two-state and Washington-only models. A segment with high variation in hourly operating speeds (i.e., >1.4 mph) is expected to experience a higher number of KABC crashes than a segment with a lower variation in hourly speeds.

Standard deviation in monthly operating speeds: This variable represents the operating speed variation among the months of a year, with an indicator variable of 1 for those segments where the standard deviation was greater than 1 mph. The coefficient is insignificant for both KABCO and KABC crashes in all models but positive and significant for PDO crashes in the Ohio-only model. A segment with high variation in monthly operating speeds (i.e., >1 mph) is expected to experience a higher number of PDO crashes than a segment with a lower variation in monthly speeds.

Non-peak non-event operating speed: This variable represents the operating speed during non-peak and non-event hours. The coefficient is insignificant for all crashes, irrespective of the data used, which could be due to the low variation in the non-peak non-event speeds between the segments considered in the study.

Percentage of days with precipitation: This variable represents the percent of days with some level of precipitation. The coefficient is negative and significant in most of the models. This finding is counterintuitive because it shows that segments with more precipitation tend to have fewer crashes than other segments. However, it is possible that vehicle speeds reduce during wet-weather conditions, which may lead to fewer crashes.

Intersection presence: This variable has a value of 1 if at least one intersection is on the segment. The coefficient is positive and significant in all cases. This finding indicates that rural two-lane segments with intersections are associated with higher crash frequencies than segments without intersections, as expected. With intersections, the conflict points increase; thus, the chance of more crashes increases.

State effect: When the two States' data are combined, the coefficient for Ohio is positive and significant, which means that when controlling for the other variables, Ohio is expected to

experience more crashes than Washington. This finding could be due to differences in weather, terrain, reporting threshold, and other variables that were not considered in the model.

Models Developed for Multilane Roadways

The form considered for the rural multilane roads was:

$$N_{ml} = L \times e^{b_0 + b_u + b_{aadt} \ln(AADT)} \times CMF_{lw} \times CMF_{hc} \times CMF_{sdif} \times CMF_{svar1} \times CMF_{svar2} \times CMF_{sff} \times CMF_{int} \times CMF_{prec} \quad (5)$$

where:

b_u = adjustment for undivided road.
N_{ml} = predicted annual average crash frequency (rural multilane roadways).
Len = segment length, miles.
$AADT$ = average annual daily traffic, vehicles per day.
CMF_{lw} = crash modification factor for lane width.
CMF_{hc} = crash modification factor for horizontal curve.
CMF_{sdif} = crash modification factor for the speed difference between weekends and weekdays.
CMF_{svar1} = crash modification factor for variance in hourly operating speeds.
CMF_{svar2} = crash modification factor for variance in monthly operating speeds.
CMF_{sff} = crash modification factor for free-flow speed.
CMF_{int} = crash modification factor for the presence of an intersection on the segment.
CMF_{prec} = crash modification factor for precipitation.
w_l = average lane width in both directions (ft).
L_c = total length of all horizontal curves on the segment.
$SpdDiff$ = percent difference in operating speeds between weekends and weekdays.
I_{svar1} = indicator variable for high variance in hourly operating speeds within a day (= 1 if hourly standard deviation is > 1 mph; = 0 otherwise).
I_{svar2} = indicator variable for high variance in monthly operating speeds within a year (= 1 if monthly standard deviation is > 1 mph; = 0 otherwise).
SFF = free-flow speed, mph.
I_{int} = indicator variable for intersection presence (= 1 if present; = 0 otherwise).
p_{prec} = percent of days with precipitation.
b_j = calibrated coefficients (j = hc, sd, svar1, svar2, sff, int, prec).

Table 18 lists the model outputs of rural multilane highways. Appendix B includes all individual models.

Table 18. Model Estimation Results of Yearly Crash Frequencies at Segments (Rural Multilane).

Variable	Two States			Ohio			Washington		
	KABCO	KABC	PDO	KABCO	KABC	PDO	KABCO	KABC	PDO
Undivided road	0.2686	0.3903	—	0.4826	0.4598	0.4513	—	0.3376	—
Traffic volume (AADT)	0.4848	0.3573	0.5529	0.4335	0.3381	0.4945	0.6473	0.6593	0.7084
Lane width (LW)	—	—	—	—	—	—	—	—	—
Percentage of curve (PerHC)	2.0307	1.574	2.3082	1.5865	—	—	3.6393	0.9047	4.7683
Avg. spd. diff. in weekday/weekend (SpdW_W)	0.05879	—	0.05048	0.0666	—	0.0599	—	—	—
Standard dev. of hourly operating speeds (SDHrSpd)	—	0.2418	—	—	0.4081	—	−0.292	—	—
Standard dev. of monthly operating speeds (SDMonSpd)	0.3911	—	0.4013	0.2969	—	0.3553	0.8381	1.0588	0.6856
Average hourly non-peak non-event speed (SpdNPNE)	0.0269	0.0239	0.0251	0.0308	0.0238	0.03055	—	0.0598	—
Presence of intersection (IntPre)	0.5714	0.5625	0.5797	0.6052	0.7757	0.5664	0.452	−0.0212	0.5999
Percentage of days with precipitation (PPrcp)	−1.9369	—	−1.8614	−1.7573	—	−2.3932	—	—	—
Added effect of Ohio	0.8282	0.3397	1.0050	NA	NA	NA	NA	NA	NA
Inverse dispersion parameter for undivided roads	−0.5868	−0.2271	−0.6188	−0.601	−0.5042	−0.5522	−0.4212	0.0506	−0.417
Inverse dispersion parameter for divided roads	−0.9955	−0.8616	−0.9469	−1.0595	−0.6138	−1.0097	−0.4212	0.0506	−0.417
Intercept	−6.4938	−6.9752	−7.4546	−5.5104	−6.5777	−6.0308	−6.4405	−11.196	−7.634

Note: A dash (—) = not significant at the 95% level; NA = not applicable.

The explanations of the model outcomes are provided below.

Undivided road: This variable represents whether the segment is undivided or divided. The coefficient is positive and significant in almost all cases (exception: PDO crashes for the two-state model, and KABCO and PDO crashes for the Washington model). This finding indicates that undivided rural multilane roads experience more crashes than divided roads experiencing the same traffic and other conditions. For undivided roads, the likelihood of opposite-direction and turning-related crashes is relatively higher than for divided roads, which may also be a reason this relationship was not significant for PDO crashes.

Percentage of curve: The coefficient is positive and significant in all cases, except for KABC and PDO crashes in the Ohio model. It shows that a higher proportion of horizontal curvature is associated with a higher number of crashes. The sharpness of the curve was not statistically significant in the preliminary models.

Average speed difference in weekday/weekend: The coefficient is significant and positive for KABCO and PDO crashes in the Ohio and two-state models, which means the increase in the difference in speeds between weekends and weekdays is associated with more crashes, perhaps due to occasional congestion during weekdays or higher speeds on weekends.

Standard deviation of hourly operating speeds: The coefficient is positive for KABC in the Ohio and two-state models but insignificant for most of the other conditions. The exception is the coefficient for this variable using the Washington KABCO data, which was negative, thus indicating a counterintuitive result. Additional investigation into this variable is needed. The positive coefficients for KABC in the Ohio and two-state models show that a segment with variation in hourly operating speeds of more than 1 mph is expected to experience a higher number of crashes than a segment with a lower variation in hourly speeds.

Standard deviation of monthly operating speeds: When the coefficient is statistically significant, it is positive. A segment with variation in monthly operating speeds of more than 1 mph is expected to experience a higher number of crashes than a segment with a lower variation in monthly speeds.

Average hourly non-peak non-event speed: The coefficient for non-peak non-event times is significant for most of the cases and positive. This finding means that with the increase in non-peak non-event speeds, crashes increase.

Percentage of days with precipitation: The coefficient is negative and significant in most of the models. This finding is counterintuitive because it shows that segments with more precipitation tend to have fewer crashes than other segments. However, it is possible that the vehicle speeds reduce during the wet-weather conditions, so the result may be fewer crashes.

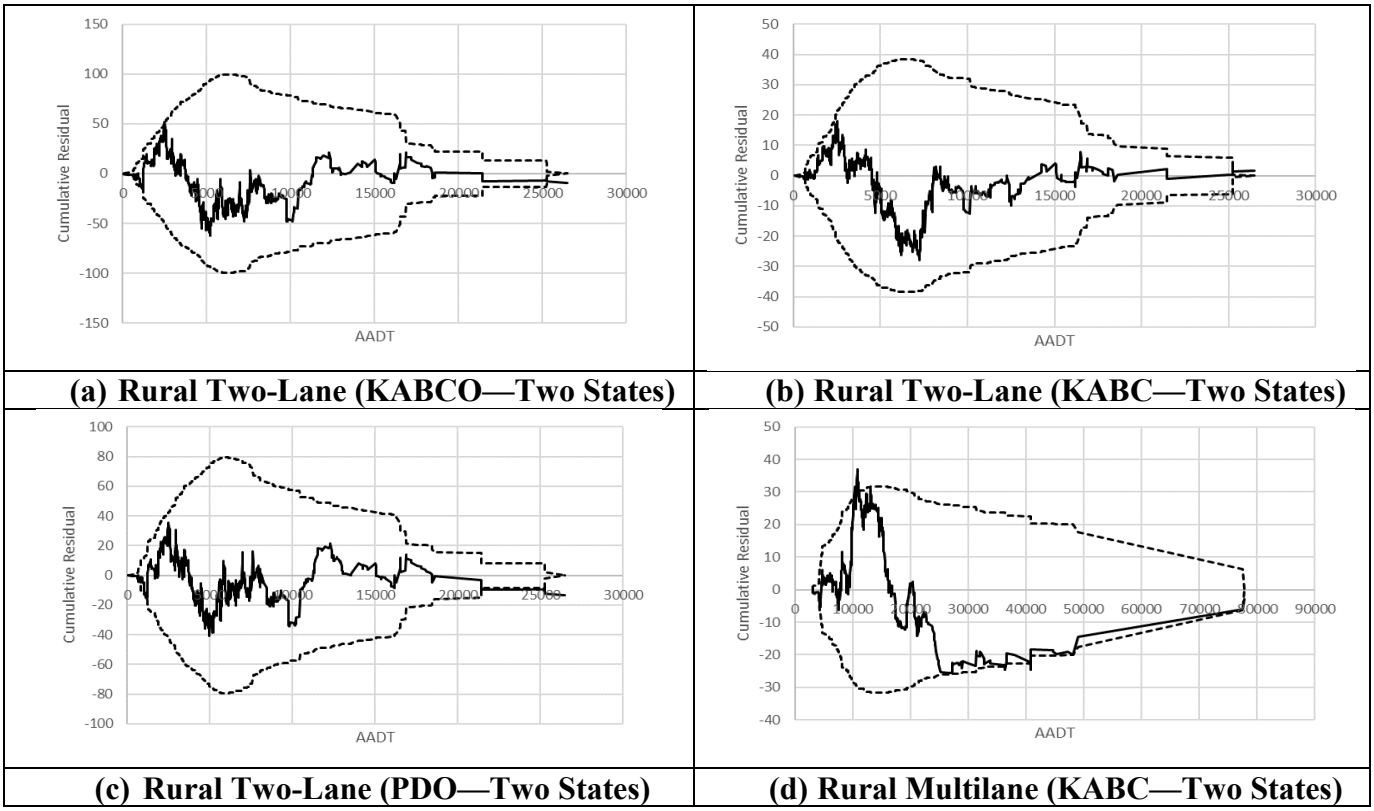
Presence of intersection: The variable is positive and significant in most cases (exception: KABC model for Washington). This finding means that segments with at least one intersection tend to have more crashes than segments without intersections, as expected. With intersections, the conflict points increase, thereby increasing the number of crashes.

State effect: When the two States' data are combined, the coefficient for Ohio is positive and significant for Ohio crashes only, which means that when controlling for the other variables, Ohio is expected to experience the same number of KABC crashes but more PDO crashes than Washington. This finding could be due to the difference in weather, terrain, reporting threshold, and other variables that were not used in the model.

Model Validation

Cumulative residual (CURE) plots were used to conduct the validation for models developed using the two States' combined data. The CURE plots show the performance of the model with respect to a particular variable. Hauer (2015) showed that the model performance is reasonable if the plot of cumulative residuals oscillates around 0, ends close to 0, and does not exceed the ±2*standard deviation bounds. If the plot of residuals shows any systemic drift, then it can be concluded that the model provides biased estimates. Figure 14a–c shows the CURE plots for the rural two-lane highway models. All CURE plots show that the model fits the data along the entire range of AADT values because the cumulative graphs have a random walk oscillating

around zero and end close to zero. Figure 14d shows the best-fit CURE plot for the rural multilane highway models. The CURE plot for KABC crashes shows that the model fits the data along the entire range of AADT values because the cumulative graph has a random walk oscillating around zero and ends close to zero. The project team also developed severity distribution functions (SDFs) for different facility types. Appendix C documents the SDFs for different facility types.



(a) Rural Two-Lane (KABCO—Two States) **(b) Rural Two-Lane (KABC—Two States)**

(c) Rural Two-Lane (PDO—Two States) **(d) Rural Multilane (KABC—Two States)**

Figure 14. CURE Plots.

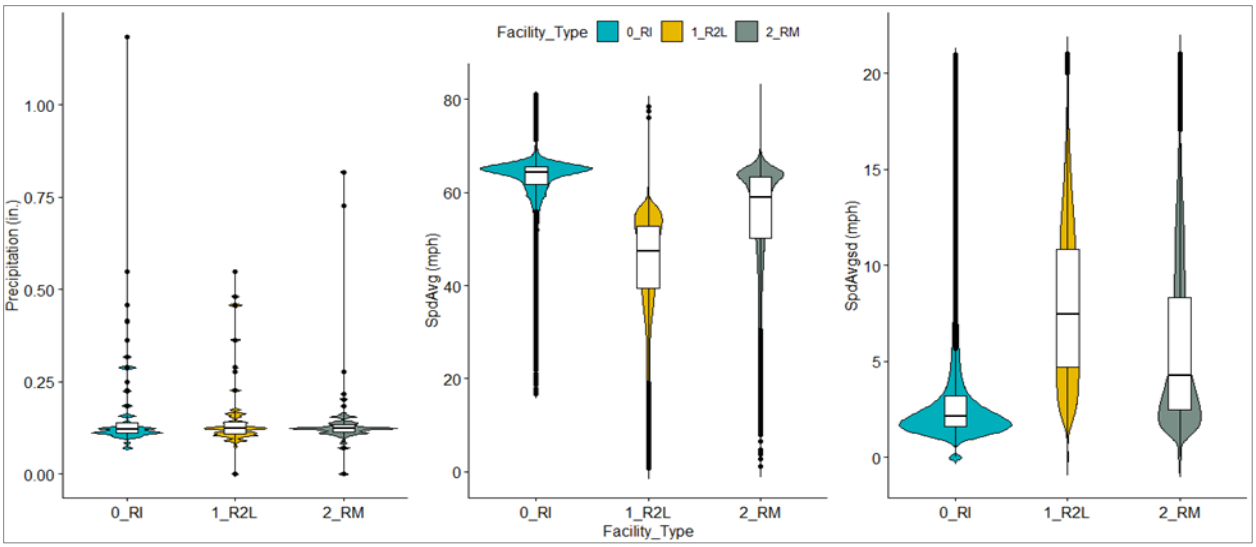
The overall findings from the annual segment-level analyses were:

- Certain speed measures were useful in the development of the annual-level crash prediction models.
    - Increased variability in hourly operating speed within a day and increased monthly operating speeds within a year were both associated with higher crash frequencies.
    - Operational speed differences between weekends and weekdays were positively associated with a higher number of crashes. Segments experiencing higher speed differentials between weekends and weekdays likely indicate the nature of roadway-use and land-use patterns.
    - Non-peak and non-event speed (average operating speed excluding peak hours and hours with events) was positively associated with crash rates on rural two-lane roadways. However, this speed measure was negatively associated with crashes in the interstate model. This finding for interstates could be because well-designed and high-standard roads are generally associated with higher non-peak and non-event speeds.

- As the proportion of horizontal curvature on a segment increased, the number of crashes could be expected to increase. The coefficient for the Ohio data was significant but counterintuitive, which could be due to the correlation with other unknown factors, the curve-related crash reporting issues, or the missing values related to curve information in the Ohio HSIS data.
- In general, segments with intersections could be expected to experience higher crash frequencies than segments without intersections, which is likely because segments with intersections have a greater number of conflict points. The variable was only significant for multilane roadways.
- The precipitation measure was negatively associated with annual crash frequencies. The finding indicates that short-term crash predication models will be suitable in exploring the temporal effect of precipitation measures.

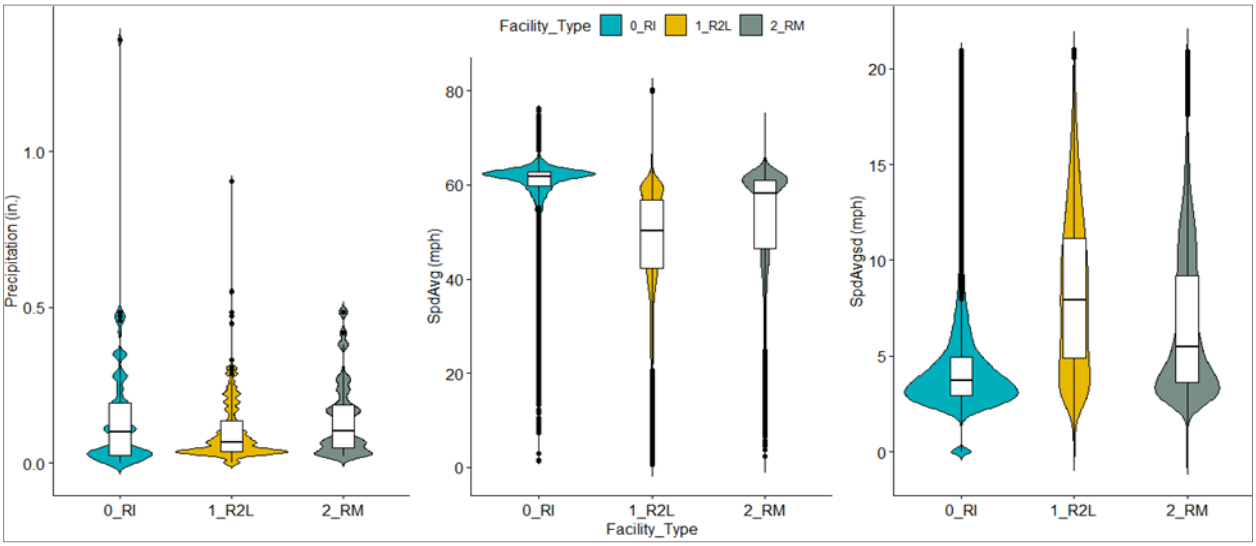
DAILY-LEVEL CRASH PREDICTION ANALYSIS

The annual-level crash prediction models do not allow users to estimate crashes precisely at short durations (e.g., hours, days, weeks, months). The annual estimation methods limit the ability to quantify the effects of variables that fluctuate at a granular temporal scale, such as operating speeds, operating speed variances, or seasonal fluctuations. A need exists to explore the development and functional forms of crash prediction methods using finite exposure measures and representing short-term roadway conditions to better account for these variables and understand short-term fluctuations in highway safety performance. To mitigate the current research gap, the project team developed daily level models separately for Ohio and Washington. Since several of the variables are segment-related, a closer look at the daily level of variables (for example, average daily precipitation, average daily speed [SpdAvgDaily], and standard deviation of daily speed [SDDailySpd]) is provided in figure 15 and figure 16. A key finding from the violin plots shown in figure 15 and figure 16 is that there is a greater distribution for precipitation in Washington compared to Ohio. The average daily speed is lowest for rural two-lane highways compared to the other facility types in both States; however, the medians of standard deviation of average daily speed are the greatest for two-lane highways, with rural interstates having the lowest value.



Note: 0_RI = Rural Interstate; 1_R2L = Rural Two-Lane; 3_RM = Rural Multilane

Figure 15. Box and Violin Plots of Three Daily Level Variables (Ohio Data).



Note: 0_RI = Rural Interstate; 1_R2L = Rural Two-Lane; 3_RM = Rural Multilane

Figure 16. Box and Violin Plots of Three Daily Level Variables (Washington Data).

Functional Form of Tweedie Distribution

Because dataset structure 2 contains a number of zeros where no crash occurs during the hour time period at a given segment, a need exists for an advanced modeling technique. The Poisson-Tweedie family can use the simple estimation of the power parameter to automatically adapt to highly skewed count data with excessive zeros without the need to introduce zero-inflated or hurdle components. The distributions of the spike counts given the predictor are assumed to follow the Tweedie distribution. The Tweedie family includes the Poisson distribution, the Gaussian distribution, and the gamma distribution. However, the case that interests this data structure most in the Tweedie family is the compound Poisson family. The following theorical description is mostly based on Dina and Nelken study (2014).

The project team used the Poisson distribution with rate λ to select n independent variables, and then the identically distributed variables were summed to generate a sample of the compound Poisson distribution. In the Tweedie case, these variables come from the gamma distribution with shape parameter α and scale parameter β.

Therefore, the Tweedie distribution can be written as:

$$\Pr[Y = 0] = e^{-\lambda},$$

$$f_Y(y) = e^{-\beta y} e^{-\lambda} \sum_{n=1}^{\infty} \frac{\beta^{n\alpha}}{\Gamma(n\alpha)} y^{n\alpha-1} \frac{\lambda^n}{n!}, \qquad y > 0 \tag{6}$$

where $y = 0$ if the Poisson variable is null, or the distribution of y is given by a mixture of gamma variables with Poisson weights (Dina and Nelken, 2014).

The Tweedie distribution also fits to the exponential family. It is important for the application of the generalized linear model (GLM) context. The members of the exponential family follow the distribution function:

$$f_Y(y; \theta, \varphi) = exp\left\{\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi)\right\}, \quad y \epsilon R_\psi \tag{7}$$

for some specific functions a(.), b(.), and c(.).

If ϕ is identified, this equation is a 1-parameter exponential family model with canonical parameter θ. The average and the variance of Y can be expressed by $E(Y) = \mu = b'(\theta), var(Y) = b''(\theta)a(\phi)$. Because the parameter θ is connected to the mean (μ), $b''(\theta)$ also depends on the average and is called the variance function. The variance function can be represented by $V(\mu)$. The Tweedie distribution links to the choice of $b''(\theta) = \mu^p$ and $a(\phi) = \phi$ for $1 \leq p \leq 2$. Here, p and ϕ are mutual to all spike counts in each set of measurement, while the parameter μ can vary. It is vital to set $\eta = \log(\mu)$ to be a linear combination of dummy variables that measures the effects of time bins on the responses. Since $b''(\theta)$ is fundamentally the variance, this procedure confirms that the variance (var) and the average μ are related as $var = \phi\mu^p$.

This characteristic of the Tweedie distribution summarizes the overdispersion of spike counts relative to the Poisson case. To achieve these relationships, the connections of ϕ, p, and μ to λ, α, and β are given by:

$$\mu = \frac{\lambda\alpha}{\beta}$$
$$\phi\mu^p = \frac{\lambda\alpha(\alpha+1)}{\beta^2} \text{ with } \mu > 0 \text{ and } \phi > 0 \tag{8}$$

or for other direction:

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \quad \alpha = \frac{2-p}{p-1}, \quad \frac{1}{\beta} = \phi(p-1)\mu^{p-1} \tag{9}$$

The parameterization can be written as follows:

$$\theta = -\beta\varphi = \frac{-1}{(p-1)\mu^{p-1}}, \quad \mu(\theta) = \left(\left(-\theta(p-1)\right)^{-\frac{1}{p-1}}, \quad b(\theta) = \lambda\varphi = \frac{\mu^{2-p}}{2-p}$$

where $c(y, \varphi)$ is defined as the logarithm of the sum in the first equation above. This sum depends on μ and θ only through the product $\lambda\beta^{a}$, which is a function of ϕ only, and therefore depends on y and ϕ, but not on θ (Dina and Nelken, 2014).

After performing the preliminary explorations, the project team selected the variables of interest suitable for the daily level analysis. For example, instead of using the five speed measures used for the segment-level analysis, the current model development applied two speed measures (average daily speed, or SpdAvgDaily, and standard deviation of daily speed, or SDDailySpd) that better convey the daily level analysis. Due to the highly random nature of PDO crashes, the current analysis was limited to KABCO and KABC crashes for Ohio and Washington separately. Since the daily precipitation patterns vary widely between these two States, the two-state model was not developed for the daily level analysis.

Models Developed for Interstate Roadways

Table 19 lists the model outputs of rural interstate roadways. Appendix D includes all individual models.

Table 19. Model Estimation Results of Daily Crash Frequencies at Segments (Rural Interstate).

Variables	Ohio		Washington	
	KABCO	KABC	KABCO	KABC
Traffic volume (AADT)	0.7126	1.0987	0.6470	0.4614
Segment length (Len)	0.2328	0.2256	0.1930	0.2550
Number of lanes (Lanes)	—	−0.0529	—	—
Lane width (LW)	—	0.5191	—	—
Percentage of precipitation (PPrep)	0.2352	0.3090	0.2080	—
Number of curvatures (NCurv)	−0.6125	—	—	—
Total length of curvatures (LCurv)	—	—	—	—
Standard deviation of daily average speed (SDDailySpd)	0.1376	0.3011	0.3520	0.3455
Daily average speed (SpdAvgDaily)	−0.0405	−0.0328	—	—
Intercept	−9.4829	−15.7714	−11.8000	−11.4046

Note: A dash (—) = not significant at the 95% level.

The explanations of the model outcomes are provided below.

Segment length and traffic volume: Both segment length and traffic volume are positively associated with daily level crashes on the roadway segments. These two variables are statistically significant for all models.

Lane width: This variable represents the lane width of a roadway surface. This variable is not statistically significant for most of the models (exception: the KABC crash model for Ohio).

Number of lanes: This variable represents the number of lanes in each direction. This variable is not statistically significant for most of the models (exception: the KABC crash model for Ohio).

Percentage of days with precipitation: This variable represents the percent of days with some level of precipitation. The coefficient is positive and significant in most of the models. This finding is intuitive since it shows that segments with more precipitation tend to have more crashes than segments with less precipitation.

Number and length of horizontal curves: These two variables show the presence of horizontal curves. Both variables are not statistically significant in most of the models. For the KABCO crash model in Ohio, the coefficient of the number of horizontal curvatures is negative and statistically significant. This finding is counterintuitive because it indicates that the segments with more horizontal curvatures tend to have fewer crashes, which could be due to correlation with other unknown factors, curve-related crash reporting issues, or missing values related to curve information in the Ohio HSIS data. It is also important to note that Ohio roadways with curves show comparatively lower in proportion than Washington.

Daily average speed: This variable represents average speed per day on each of the segments. The coefficient is negative in the models developed for Ohio, which means that with the increase in daily average speeds, the crashes decrease. Generally, well-designed and high-standard roads are associated with higher speeds.

Standard deviation of daily average speed: This variable represents the operating speed variation among the hours of a day. The coefficient is positive and significant for all cases. The positive coefficient shows that a segment with high variation in daily average speeds (i.e., >1 mph) is expected to experience a higher number of crashes than a segment with a lower variation in daily speeds.

Models Developed for Two-Lane Roadways

Table 20 lists the model outputs of rural two-lane roadways. Appendix D includes all individual models.

Table 20. Model Estimation Results of Daily Crash Frequencies at Segments (Rural Two-Lane).

Variables	Ohio		Washington	
	KABCO	KABC	KABCO	KABC
Traffic volume (AADT)	0.2610	0.7720	0.7531	0.8316
Segment length (Len)	0.0092	0.2320	0.1609	0.1448
Number of lanes (Lanes)	NA	NA	NA	NA
Lane width (LW)	—	—	−0.0203	−0.0598
Percentage of precipitation (PPrep)	—	—	0.1688	—
Number of curvatures (NCurv)	—	—	−0.0048	—
Total length of curvatures (LCurv)	—	—	0.1748	—
Standard deviation of daily average speed (SDDailySpd)	0.0062	0.0954	0.0490	0.0770
Daily average speed (SpdAvgDaily)	0.0312	—	—	—
Intercept	0.5638	−14.2000	−11.8408	−13.1203

Note: A dash (—) = not significant at the 95% level; NA = not applicable.

The explanations of the model outcomes are provided below.

Segment length and traffic volume: Both segment length and traffic volume are positively associated with daily level crashes on the roadway segments. These two variables are statistically significant for all models.

Lane width: This variable represents the lane width of a roadway surface. This variable is not statistically significant for Ohio models but is significant for Washington models. For Washington models, the coefficient is negative, which is intuitive.

Percentage of days with precipitation: This variable represents the percent of days with some level of precipitation. The coefficient is positive and significant only for the Washington KABC crash model. This finding is intuitive because it shows that segments with more precipitation tend to have more crashes than other segments with less precipitation.

Number and length of horizontal curves: These two variables show the presence of horizontal curves. Both of these variables are not statistically significant in most of the models. For the KABCO crash model in Washington (although the number of curves has a negative coefficient), the coefficient of the length of horizontal curvatures is positive and statistically significant. This finding is intuitive because it indicates that the segments with more horizontal curvatures tend to have more crashes.

Daily average speed: This variable represents average speed per day on each of the segments. The coefficient is positive and statistically significant for the Ohio KABCO crash model. This result is intuitive for two-lane roads because crash frequencies may increase due to the higher average speed due to the presence of higher hazard warnings (e.g., limited sight distance).

Standard deviation of daily average speed: This variable represents the operating speed variation among the hours of a day. The coefficient is positive and significant for all cases. The

positive coefficient shows that a segment with high variation in daily average speeds (i.e., >1 mph) is expected to experience a higher number of crashes than a segment with a lower variation in daily speeds.

Models Developed for Multilane Roadways

Table 21 lists the model outputs of rural multilane roadways. Appendix D includes all individual models.

Table 21. Model Estimation Results of Daily Crash Frequencies at Segments (Rural Multilane).

Variables	Ohio		Washington	
	KABCO	KABC	KABCO	KABC
Traffic volume (AADT)	0.7413	0.7784	0.9219	0.8258
Segment length (Len)	0.2511	0.2259	0.1939	0.2097
Number of lanes (Lanes)	—	—	—	—
Lane width (LW)	—	—	0.0260	0.0268
Percentage of precipitation (PPrep)	—	—	—	—
Number of curvatures (NCurv)	0.0948	0.1038	—	—
Total length of curvatures (LCurv)	—	—	0.1804	—
Standard deviation of daily average speed (SDDailySpd)	0.0806	0.1414	—	0.0694
Daily average speed (SpdAvgDaily)	—	—	—	—
Intercept	−13.1536	−16.1712	−13.8323	−15.6685

Note: A dash (—) = not significant at the 95% level.

The explanations of the model outcomes are provided below.

Segment length and traffic volume: Both segment length and traffic volume are positively associated with daily level crashes on the roadway segments, which aligns with existing highway safety literature. These two variables are statistically significant for all models.

Lane width: This variable represents the lane width of a roadway surface. This variable is only statistically significant (positive coefficient) for Washington models.

Number of lanes: This variable represents the number of lanes in each direction. This variable is not statistically significant for any of the models.

Percentage of days with precipitation: This variable represents the percent of days with some level of precipitation. The coefficient is not statistically significant in any of the models.

Number and length of horizontal curves: These two variables show the presence of horizontal curves. The number of curvatures is positive and statistically significant for both Ohio models.

The total length of curvature is positive and statistically significant for the KABCO crash model in Washington.

Daily average speed: This variable represents average speed per day on each of the segments. The coefficient is not statistically significant in any of the developed models.

Standard deviation of daily average speed: This variable represents the operating speed variation among the hours of a day. The coefficient is positive and significant in most of the cases. The positive coefficient shows that a segment with high variation in daily average speeds (i.e., >1 mph) is expected to experience a higher number of crashes than a segment with a lower variation in daily speeds.

The overall findings from the daily segment-level analyses were as follows:

- This study developed a speed measure that could capture traffic speed variation throughout the day by measuring the standard deviation of daily average speed. The coefficient was positive and significant in all models, which signifies that a segment with high variation in daily average speeds is expected to experience a higher number of crashes than a segment with a lower variation in daily speeds. The strength of this finding is one of the biggest insights gained from this study.
- Average operating speed was positively associated with crashes for rural two-lane roadways. However, average operating speed was negatively associated with crashes in the interstate models. This finding for interstates could be because well-designed and high-standard roads are generally associated with higher average operating speeds.
- Daily average precipitation was positively associated with the number of daily crashes. For the segment-level analysis, the effect of precipitation was mostly insignificant or negative. The finding at the daily level of analysis indicates that precipitation is positively associated with daily crash frequencies.
- Because the geometric variable remained constant for the segment while developing the model at the segment-temporal level, additional insights are needed for the model interpretation. A need exists for further examination of the spatial effects of the geometric variables with the use of advanced modeling techniques.

EXPLORATORY EXAMINATION OF TIME BEFORE AND AFTER CRASHES

Dataset structure 3 was designed to answer the second research question (is there more variability in speeds just prior to a crash?), and to enable the examination of traffic speeds before and after crashes. Dataset structure 3 uses speed as a surrogate form of the response variable rather than the crash frequency because the research question is focused on identifying speed patterns that would signal an imminent crash.

Appendix E provides the data integration steps for this data structure. This dataset structure considers whether a crash occurred within the given time period (i.e., the value is either 0 or 1 for that variable). It considers the consecutive epochs before a recorded crash. To construct the dataset for analysis, all epochs with crash incidents were identified in the spatiotemporal dataset (dataset structure 2). All epochs within 4 hours of the incidents were also identified and labeled

accordingly in a new field named "epoch type." The coding of these incident-related epochs was as follows:

- BI: An epoch 4 hours or less before an incident (crash) occurred.
- DI: An epoch when an incident occurred.
- AI: An epoch 4 hours or less after an incident occurred.

The project team also developed a companion dataset (control group) of reference sets of epochs to be utilized in the analysis. All the epochs of these reference sets were labeled as follows:

- BR: A reference set to BI epochs.
- DR: A reference set to DI epochs.
- AR: A reference set to AI epochs.

In addition to the epoch type field, a field with relative epochs was added to indicate separation from the incident epoch. For example, a value of −2 indicates a BI or BR epoch that is two 15-minute periods prior to the incident at the corresponding DI or DR epoch. Similarly, a value of +3 indicates an AI or AR epoch that is three 15-minute periods after the incident at the corresponding DI or DR epoch.

Functional Form of Mixed-Effects Model

Several models that were suitable for this analysis were considered. The project team found that mixed-effect (ME) modeling would better address the research question. A short overview of the ME modeling is provided below.

By analyzing the impact of variables that vary over time, fixed-effect (FE) models successfully frame the predictor and outcome variables' relationship within an entity. This entity, with its own individual characteristics, might influence the predictor variables. The assumption of an entity's error term and predictors' correlations in an FE model removes the effect of time-invariant features and facilitates measurement of the net effect of the predictors used in the model on the outcome variable. Also, the threat of omitted variable bias is significantly reduced using these models since all time-invariant differences in observables and non-observables are controlled.

The equation for an FE model is:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \quad (10)$$

where:

α_i (i = 1,…..n) = unknown intercept for each entity.
Y_{it} = dependent variable, where i = entity and t = time.
X_{it} = one independent variable.
β_{it} = coefficient for that independent variable.
u_{it} = error term.

Unlike the FE model, a random-effects (RE) model assumes the variation across different entities (epochs) to be random and uncorrelated with the independent variables used in the model. The

RE model permits generalization of inferences outside the sample epochs used in the model. In addition, this model consists of time-invariant variables such as the number of lanes, AADT, median width, shoulder width, and speed limit. This method differs from the FE model, in which all of the time-invariant variables are absorbed by the intercept of the model.

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} + \in_{it} \quad (11)$$

where:

u_{it} = between-entity error.
$\in_{it}$ = within-entity error.

An entity's error term in RE is assumed to not be correlated with the predictor variables used in the model, which allows the time-constant variables mentioned above to act as explanatory variables. An easy interpretation of the above effects can be that fixed effects do not change across individual entities, whereas the random effects do vary, so it is essential to account for both of these effects in order to identify a relationship between the response and predictor variables irrespective of the population sample being used in the model.

When an ME model is applied in the analysis, it accounts for both fixed effects (by including explanatory factors) and random effects (by including exogenous sources of variability). Generally, the modeling structure for MEs accounts for the following types of data characteristics:

$$Y_i = X_i\beta + Z_i b_i + \in_i \quad (12)$$

For i = 1,……n where:

Y_i = i^{th} subject's response vector.
X_i = fixed-effects design matrix, i.e., (n_i x p) matrix of covariates.
β = fixed-effects vector, i.e., (p x 1).
Z_i = random-effects design matrix, i.e., (n_i x q) matrix of covariates.
b_i = random-effects vector, i.e., (q x 1).
$\in_i$ = error vector.
X_i = Zi x A_i, where Z_i contains only within-subject factors and A_i contains only between-subject factors.

This model assumes that:

$$Y_i \sim N(X_i\beta, \sum_i) \quad (13)$$

where:

$\Sigma_i = Z_i \sum Z_i' + \sigma^2 \mathrm{I}\, n$ is the m_i x m_i covariance matrix for the i^{th} subject's data.

Another assumption for an ME model is:

$$Cov[Y_h, Y_i] = 0_{m_h \, x \, m_i} \; if \; h \neq i \quad (14)$$

By using an ME model, individual change across time can be modeled explicitly, and since it is more flexible in terms of repeated measures, the same number of observations per subject are not required. An ME model allows correlation of errors, unlike GLM, and therefore possesses more flexibility when modeling the error covariance structure. Better handling of the missing data and allowance of non-constant variability of error terms is also achieved in the ME model. These characteristics of the model convinced the project team to utilize an ME approach for data analysis.

Model Development

The analysis on data structure 3 was exploratory in nature. Space mean speed (SMS)[5] series for a sample of 150 crashes were selected at random from the Washington interstate/freeway dataset. Along with those series, a set of up to 10 reference series per crash series was selected—also picked at random, but within a month of the crash series, as explained previously. This sample resulted in a set of 16,207 epochs, including 150 SMS crash series and 1,073 reference SMS series.

Initially, a robust set of variables was examined as potential covariates in the model. After stepwise model selection, the project team determined the most significant predictors. The general conclusions were:

- Speed limit, AADT, number of lanes, and median width relate to the general trend of SMS at freeway sites (specific trends are comingled and need a sensitivity analysis for interpretation).
- After controlling for other influential factors, the team found that the trend for SMS before BI tended to be lower than the SMS trend for comparable reference epochs (i.e., BR; lower by as much as 2.89 mph just before the incident).
- After controlling for other influential factors, the team found that the difference in trends for SMS BI and BR was wider immediately before the crash in comparison to earlier epochs. The difference was estimated at –0.49 mph (–2.887 – 0.15 * [–16]) 4 hours prior to the incident but enlarging to 2.89 mph just before the incident.
- The variance of the SMS was found to be larger for the series leading to a crash (i.e., BI) than for the series not leading to a crash (BR). The team estimated that the variance of the BI series was about 20 percent larger than the variance of the reference series BR after accounting for other factors.

The project team anticipated that these precursor differences in operations could be used to construct a procedure to identify an SMS series that could be leading to crashes before the crashes occur. However, a validation effort was necessary using the rest of the conflated data. Before proceeding to the validation phase, the project team tested additional specifications for the differences in variance between the BI and BR series. In the last iteration, the model was fitted so that heteroscedasticities (sub-populations having different variabilities from others) were allowed for BR and BI, with the result being that $[V(SMS)_{BI}] / [V(SMS)_{BR}] = 1.2003$.

[5] Time mean speed (TMS) is the average speed measure of all vehicles passing a point over a period. Space mean speed (SMS) averages the spot speed with spatial weightage instead of temporal.

In a new iteration, different independent heteroscedasticity functions were allowed for each time series, resulting in an improved overall fit with respect to the homoscedastic model. The project team tested different time-sequential epoch thresholds for the heteroscedasticity model (starting 1 hour prior, 2 hours prior, or 4 hours prior to the crash). The best fit was found when the threshold was set at 4 hours prior to the crash incident (p-value < 0.0001 for 138.51 chi-squared statistics on 1 degree of freedom from a log-likelihood ratio test between the two heteroscedasticities model and the two homoscedasticities model).

All other coefficients remained essentially unchanged, but the increased variance for the BI series became more pronounced in this model, as can be seen in figure 17.

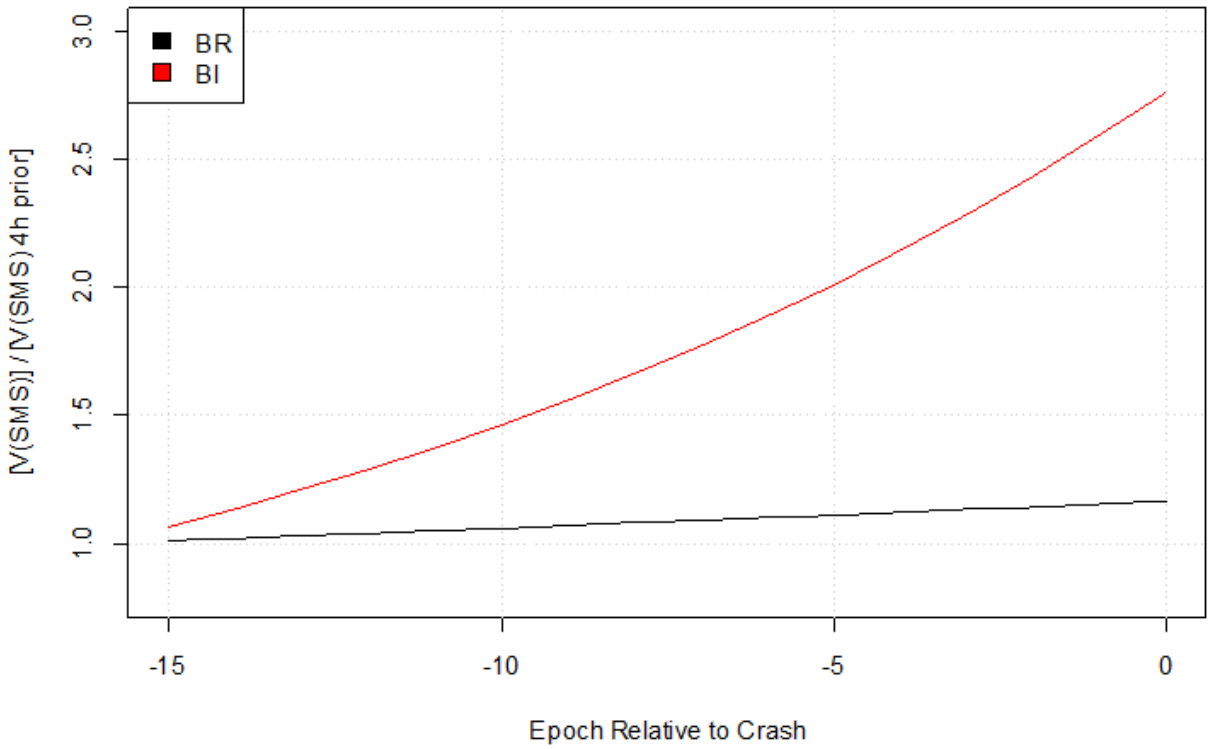


Figure 17. Variance of BI Series.

The reference variance is 4 hours prior to the epoch of the crash (i.e., relative epoch equals –16). Figure 17 shows that the BR and BI variances are essentially the same 4 hours prior to the crash. The variance for BI increases rapidly by an increasing factor up to 2.76 just prior to the crash, while the BR variance remains essentially unchanged. Table 22 lists the model coefficients.

Table 22. Model Coefficients (Dataset Structure 3 for Washington Interstate Roadways).

Design and Traffic Factors	Value	Std. Error	DF	t-value	p-value
MEDWID	−0.00547	0.00171	143	−3.21049	0.0010
SPD_LIMT	−3.4747	2.28453	143	−1.52097	0.1305
log(AADT)	−23.2193	14.59288	143	−1.59114	0.1138
I(NO_LANES - 4)	−25.7943	7.5424	143	−3.41991	0.0008
SPD_LIMT:log(AADT)	0.34973	0.20675	143	1.691536	0.0929
I(NO_LANES - 4):log(AADT)	2.38218	0.70928	143	3.358591	0.0010
Estimates for Difference between Series					
	Value	**Std. Error**	**DF**	**t-value**	**p-value**
BI	−2.74453	0.51717	1072	−5.30681	<0.0001
DR	−0.03469	0.17992	14836	−0.19279	0.8471
Residual Trend by Relative Epoch for BR	−0.00487	0.01799	14836	−0.27073	0.7866
Residual Trend by Relative Epoch for BR	−0.1358	0.0392	14836	−3.46419	0.0005
Scedasticity Model	Estimate				
(Variance for BI_[t])/(Variance for BI_[t-1])	1.06037				
(Variance for BR_[t])/(Variance for BR_[t-1])	1.00989				
Error Covariance Structure: AR(2)	Estimate				
Phi 1	0.34770				
Phi 2	0.13764				

Validation

The project team repeated the model calibration on a similarly selected but non-overlapping random sample of 200 crash events and reference speed series to verify that the coefficient estimates were stable. Similar to naturally expected fluctuation in the coefficients, their magnitudes and directions were confirmed to be essentially unchanged. The next step in the analysis was the development of a procedure to analyze the new speed series and assess if they were likely to result in crashes. It was hypothesized that the following trend signatures from the analysis could help differentiate the crash-associated series (BI) from the non-crash series (BR).

- The trend for SMS BI tended to decrease linearly from the SMS trend for comparable reference epochs (i.e., BR). In other words, the difference in trends for SMS BI and BR was wider immediately before the crash in comparison to earlier epochs. The difference was estimated at –0.54 mph (–2.7445 – 0.1358 * [–16]) 4 hours prior to the incident but enlarging to 2.7445 mph just before the incident.
- The variance of the SMS was found to be larger for the series leading to a crash (i.e., BI) than for the series not leading to a crash (i.e., BR). It was estimated that the variance of the BI series at an epoch T was about 6 percent larger than the variance of the prior epoch in the same series (T-1), after accounting for other factors. In contrast, the variance for the BR series was found constant, on average, for all epochs.

Scoring Methodology

To implement the findings from the modeling into a practical methodology that uses new data (i.e., data not used to fit the models), the project team developed a scoring procedure to quickly assess how a new time series shows the features pointed out above. To obtain a score for the new series, the project team prepared an algorithm to implement the following steps:

1. Extract a family of series from data structure 3 and fit a polynomial trend (fixed at the second order). This order is the same order of the polynomial used in the random effects of the ME model to allow for simple curvatures in the series trend.
2. Fit a differential linear model for each series in the family (i.e., having as the response the difference between the family trend and the individual series observed values). This differential model is such that it imposes the trend signatures listed above as three model features: (a) an intercept shifted negatively by 2.7445 mph; (b) a negative linear slope of –0.1358 mph / [15 minutes]; and (c) a positive heteroscedasticity with a linear envelope with a slope of 1.06037.
3. Apply four hypothesis tests to each differential model and obtain the corresponding p-value. The test is such that the alternative hypotheses state that the three signature adjustments above are incorrect. Thus, a large p-value indicates a lack of evidence against the trend signatures.
4. Obtain a preliminary score for the series simply by multiplying the three p-values.
5. Adjust the preliminary score to penalize the series that have narrower variances relative to the family of series.

The scoring procedure above yields a numeric assessment of the agreement of each series with the findings in the model. However, there are two important but unknown pieces of information:

- In the above methodology, each of the four components of the score is weighted equally. It is likely that there exists a set of weights that would improve the power of the methodology to filter crash-prone series from non-crash series.
- For a given set of weights, it is unclear what threshold to recommend to separate crash-prone series from non-crash series.

The project team developed an experiment to provide answers to the above questions. A sample of 50 randomly selected sets of series such that only series that were not used in the model development were selected (386 series in total, including 336 BR and 50 BI series). Algorithms were developed to execute the scoring procedure described earlier but by applying a set of fixed weights to the four components. Each set of weights was defined as a scheme. For each scheme, scores were obtained for all speed series, and two graphic assessment outputs were developed: (a) a kernel-density plot for the score distribution by either BI or BR; and (b) calculations for four measures of effectiveness (MOEs) at a range of thresholds for the scores—true positives, false negatives, proportion of missed crashes, and total number of positives produced. Since the score density profiles for both types of series look similar, it is expected that this scheme has a limited power in differentiating crashes from non-crashes. Such expectation is confirmed in figure 18, which shows four MOEs.

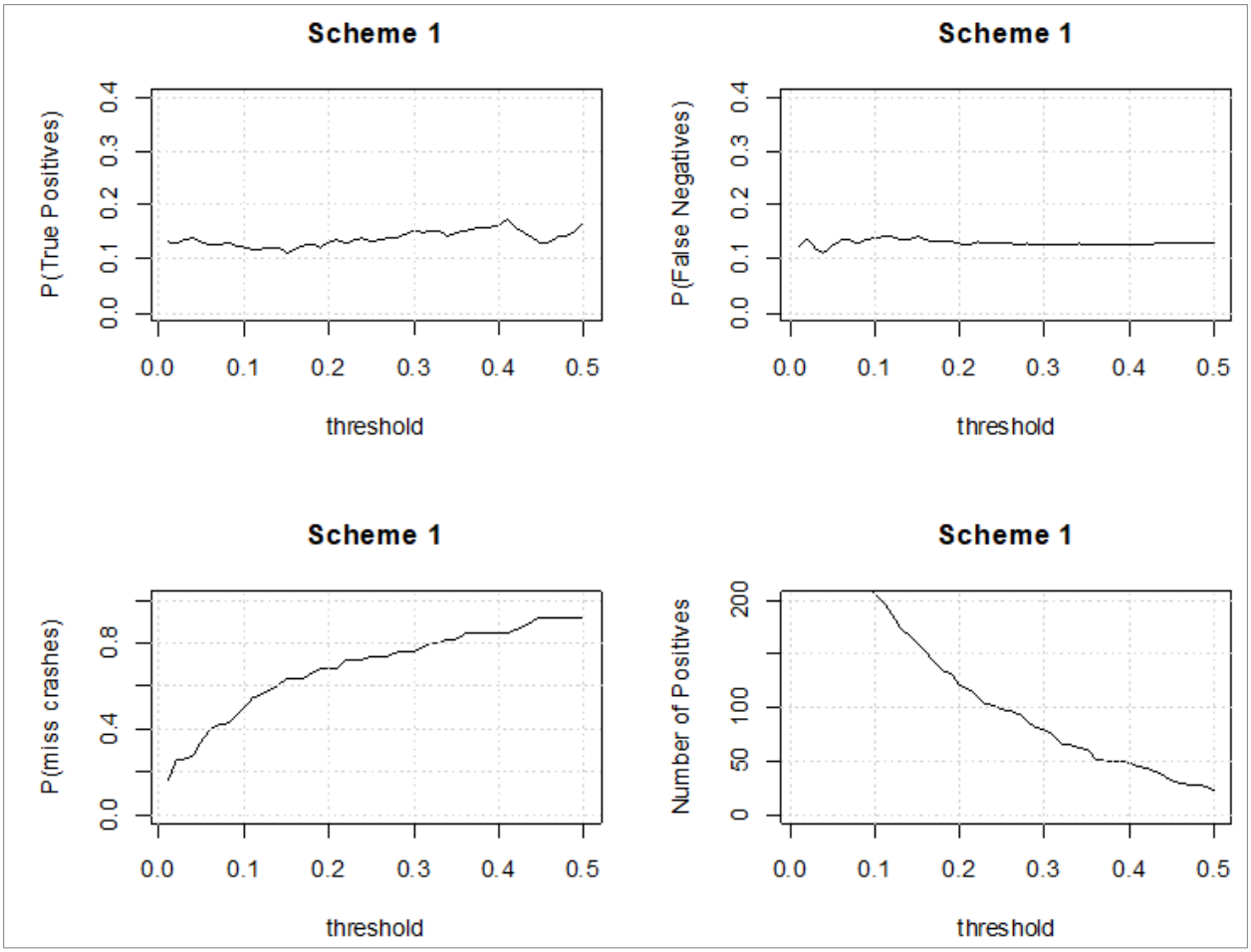


Figure 18. MOEs of Scheme 1.

Figure 18 shows that the proportion of true positives is roughly constant at about 15 percent, and the proportion of false negatives is also essentially constant at a similar level regardless of the threshold picked. The proportion of missed crashes increases with an increasing threshold, from about 20 percent at a threshold of 0.01 to about 93 percent at a threshold of 0.5. However, the number of positives is very large for small thresholds, and it begins to become manageable (i.e., about 50 positives) at thresholds larger than or equal to 0.35. The project team ran the analysis for 20 different schemes, manually adjusting individual weights in the direction that appeared to improve the performance of the classification score.

As mentioned earlier, the attempts to analyze data structure 3 were exploratory in nature. The general findings were as follows:

- After controlling for other influential factors, as the moment of crash occurrence approached, the speed trend for the crash-related series decreased and was substantially different than the trend of the non-crash-related reference series.
- Speed variability increased for the series just prior to a crash, which was also different than the no-crash series.

CHAPTER 4. DECISION SUPPORT TOOL

Based on the results from the statistical runs, the project team developed a prototype decision support tool that uses the statistical modeling results and converts them into a visual and geospatial environment that identifies higher-risk highway segments. This tool is based on the historical data from NPMRDS and HSIS and, if feasible, will expand to include the effects that NPMRDS speed data on a given segment over time have on a given highway segment's risk profile. The project team developed two separate data visualization tools:

- Interactive decision support tool.
- Interactive data visualization tool.

INTERACTIVE DECISION SUPPORT TOOL

The project team developed a beta version of the interactive web-based decision support tool for Washington and Ohio rural NHS roadways (Interactive Decision Support Tool, 2018). The tool was developed through R Shiny, which converts statistical modeling results from segment-level analysis into a visual/geospatial representation of higher-risk highway segments based on roadway design, traffic volumes, and speed data. Figure 19 illustrates the interface of the proposed interactive tool. The features of the interactive tool are the following:

- The tool contains a dashboard with various drop-down lists of steps to evaluate risk scoring at the segment level (direction specific). Users have the flexibility of selecting several options. The beta version has the following drop-down and selection options (see figure 19):
    - Year: 2015.
    - State: WA and OH.
    - County: Counties in Each State.
    - Facility Type:
        - Rural Interstate.
        - Rural Two-Lane.
        - Rural Multilane.
    - Severity: All and FI.
- After selecting the options, the user needs to click the "Refresh Map" button to generate the interactive map. For example, by selecting "Year: 2015; State: WA; County: All Counties; Facility: All; Severity: All," a heatmap based on number of crashes will be generated (see figure 20). The map can also be seen at other smaller spatial units. For example, selection of "Year: 2015; State: WA; County: Whitman County; Facility: All; Severity: All" will allow the user to develop the map and associated data at the county level (see figure 20). Selection of "Year: 2015, State: WA; County: Whitman County; Facility: Multi. Undiv.; Severity: All" will allow the user to develop the map and associated data at the facility level of a county (see figure 21).
- The interactive map has a hovering option. Users can see associated data on a segment by dragging the mouse to that segment (see figure 22).

The current risk rankings are based on the high number of total and FI crashes.

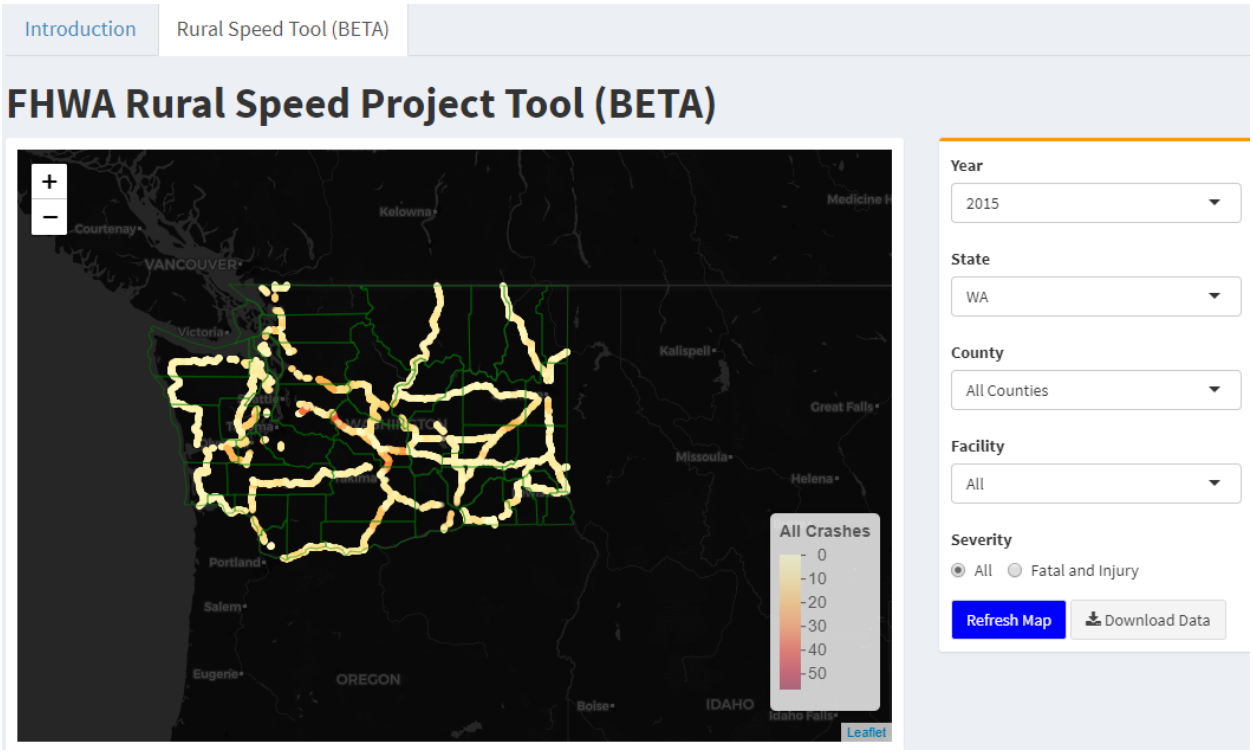


Figure 19. Selection at State Level (Expected Total Crashes at Segments).

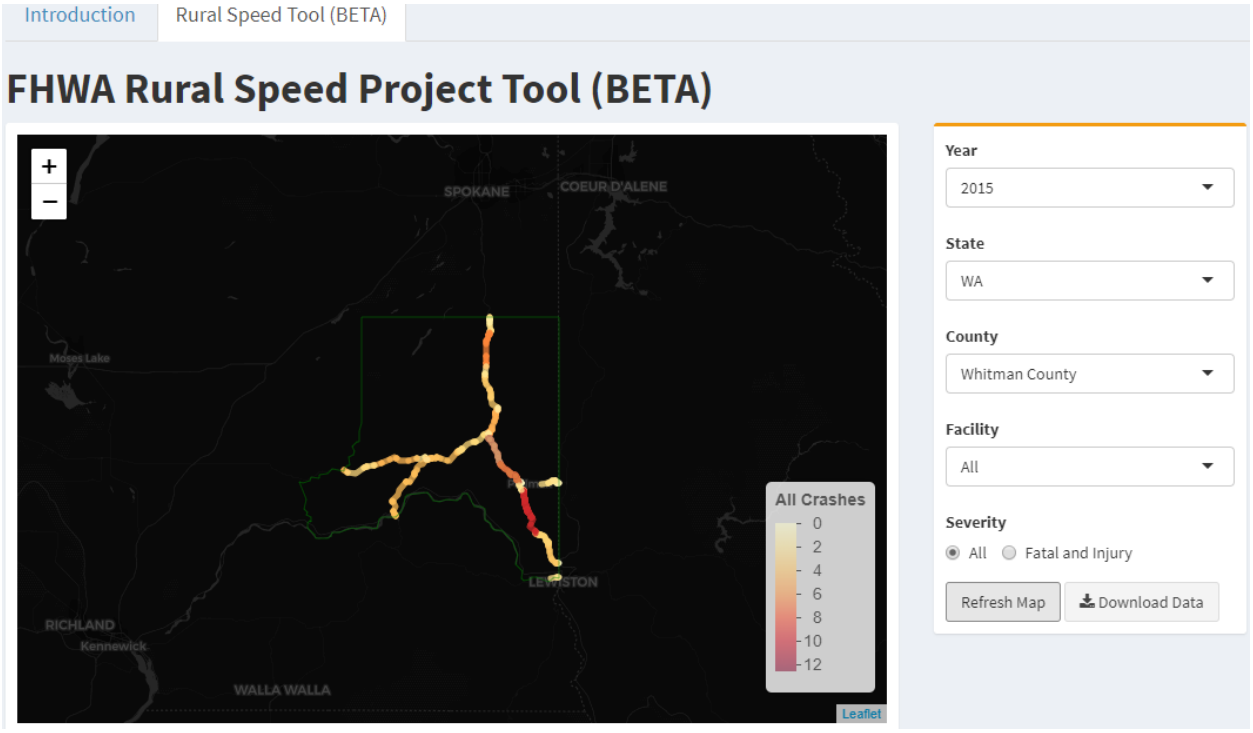


Figure 20. Selection at County Level (Expected Total Crashes at Segment Level).

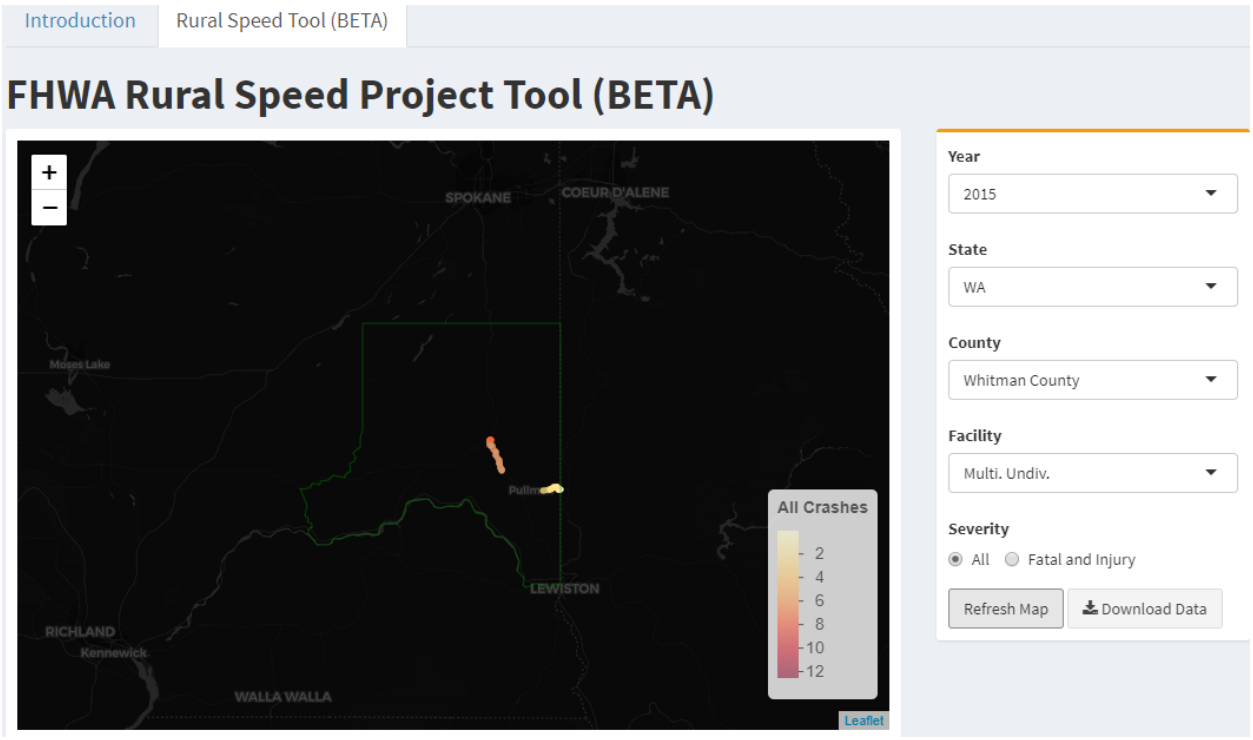


Figure 21. Selection at Facility Level.

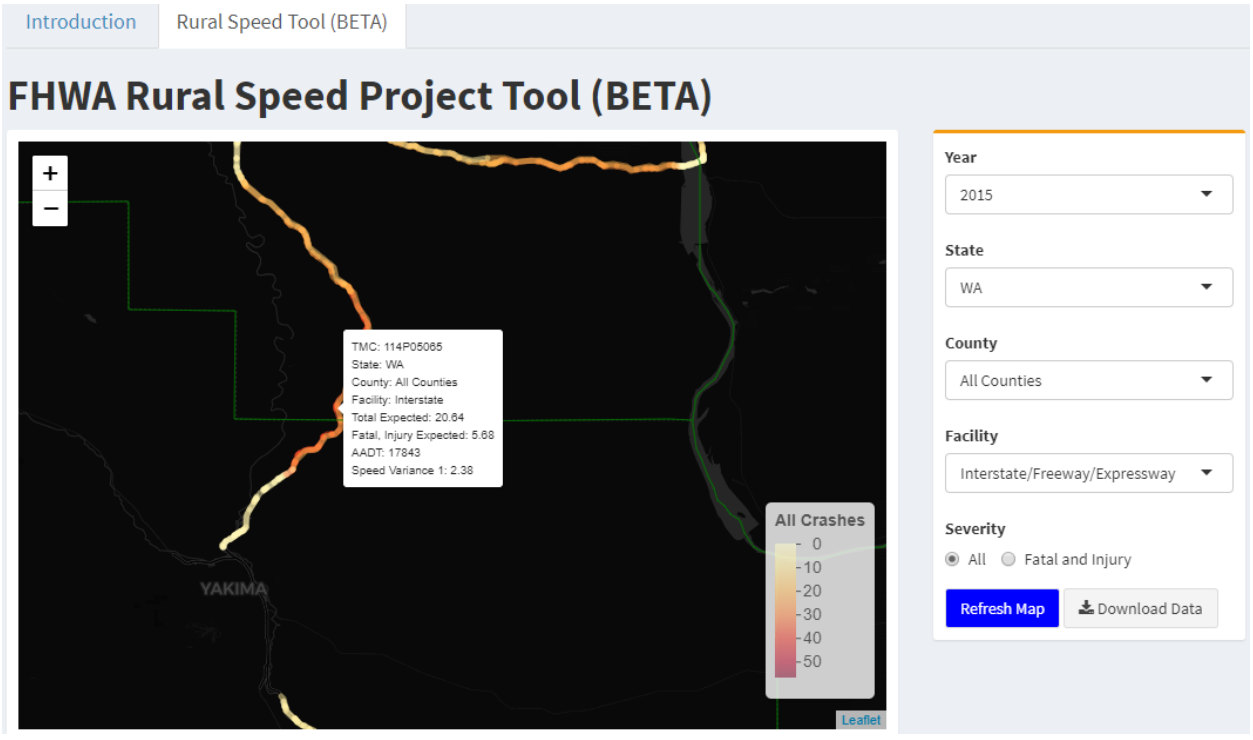


Figure 22. Hovering Option.

Feasibility of Applying Real-Time NPMRDS Data

The current decision support tool is based on 2015 HSIS and NPMRDS data. The tool can be applied to county-level analysis for determining the rural NHS risky segments. It has the functionality of providing expected crashes for different segments (with associated geometric, demographic, and speed measures on the segment in tabular form) based on the user's selection criteria. The current beta version of the tool does not support real-time integration of NPMRDS data. Additionally, it is limited to segment-level risk analysis. However, it can be scalable to segment-temporal-level speed data integration.

INTERACTIVE DATA VISUALIZATION TOOL

Data visualization merits careful consideration for many reasons. One of the prime objectives of this task is to use the most fitting and efficient tools to enhance the understanding of the data presented by creating a clear visual narrative and thinking beyond charts. The project team envisions embracing new visualization techniques that will ensure approachable and accurate interpretations of the association between operational speed, speed differentials, traffic volume, roadway geometry, and crash outcomes.

Appendix F provides the interface of the tool and associated link. The clickable links provide a more detailed view of the speed measures and descriptive statistics, as well as visualization of the association between speed measures and crashes at the granular level. The current version contains analysis at the functional class level. The project team used open-source R data visualization packages (ggplot2, lattice, and htmlwidgets in R) to prepare both static and interactive data visualization plots and tools (Wickham, 2016; Deepayan, 2008; htmlwidgets, 2018). The project team also envisions enhancing the accessibility of data in HTML tables. The project team developed the subset of datasets (based on different dominant clusters) for interactive table viewing. An easily accessible data table helps end-users obtain the necessary information from the table as quickly as possible (Washington data [DV, 2018]). Researchers used DataTables (see table 23), a plug-in for the jQuery JavaScript library (DT, 2018).

Table 23. Example of Data Table View.

Washington 2015 Crashes by Rural Roadway Segments

CSV Excel Search:

	Segment	Length (mi.)	Route	Total_Crash	Fatal	Injury	PSL (mph)	Lanes	AADT (vpd)
1	114N04098	2.8	I-90	17	0	6	70	4	32,975
2	114N04099	2.1	I-90	0	0	0	70	3	31,074
3	114N04101	3.5	I-90	7	0	1	70	3	46,392
4	114N04102	1.19	I-90	16	1	5	70	3	51,532
5	114N04103	3.33	I-90	37	0	14	70	4	60,132
6	114N04104	2.34	I-90	8	0	2	70	3	67,610
7	114N04105	2.24	I-90	15	0	6	70	4	70,433
8	114N04561	6.26	I-90	8	0	1	70	2	19,669
9	114N04677	0.29	I-90	0	0	0	70	4	30,588
10	114N04678	4.21	I-90	28	0	10	70	4	32,308

Showing 1 to 10 of 1,122 entries Previous 1 2 3 4 5 … 113 Next

Note: PSL= Posted Speed Limit, AADT=Annual Average Daily Traffic, vpd= vehicle per day

Example (Interactive GIS Maps)

Figure 23 illustrates the heatmap of the rural NHS crashes (from 2015) in Washington. This interactive plot was created by using mapbox.js. Figure 24 shows the interactive property of figure 23. The user can zoom in and out by clicking on the corresponding circles present in the plot.

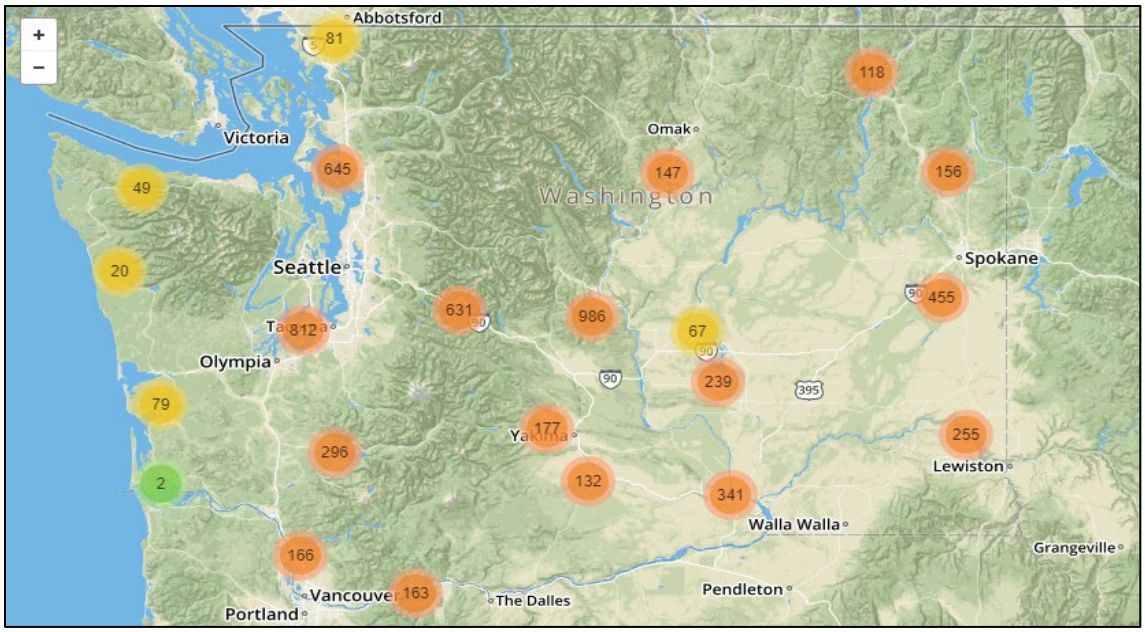


Figure 23. Heatmap of the Rural NHS Crashes in Washington Using Mapbox.js.

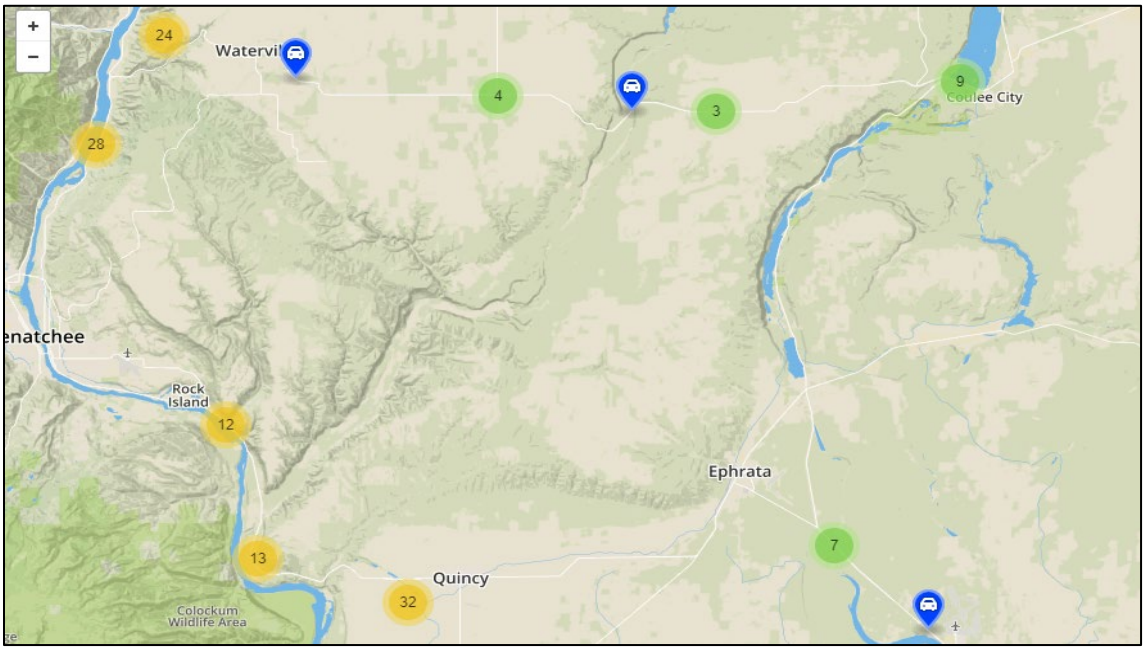


Figure 24. Interactive View of the Plot.

Example (Dygraphs)

Static plots are not suitable for showing the travel time or speed data through the year for all epochs. The project team developed dygraphs, an open-source JavaScript charting library, for the segments with a high number of crashes on different facility types. Figure 25 shows the interactive dygraphs developed for the top two segments with a high number of crashes using Washington conflated data. It shows the pattern of operational speed during the occurrences of crashes. These plots have the following advantages and functionalities:

- Can handle large datasets like NPMRDS travel time data without getting bogged down.
- Have interactivity out of the box, including zoom, pan, range selection (see figure 26), and mouseover, which are on by default.

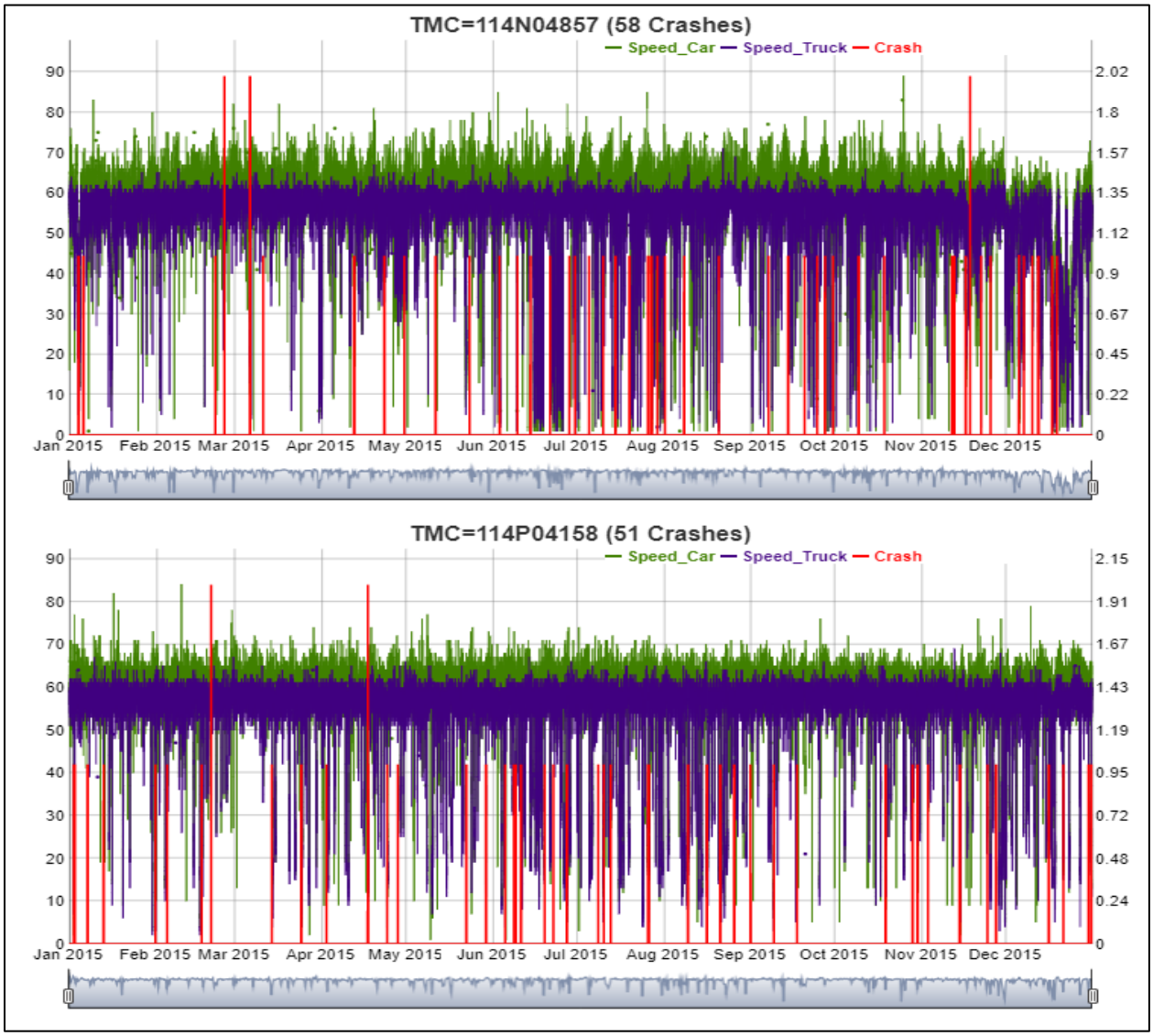


Figure 25. Association between Crash and Operational Speeds on Two Interstate Roadways in Washington.

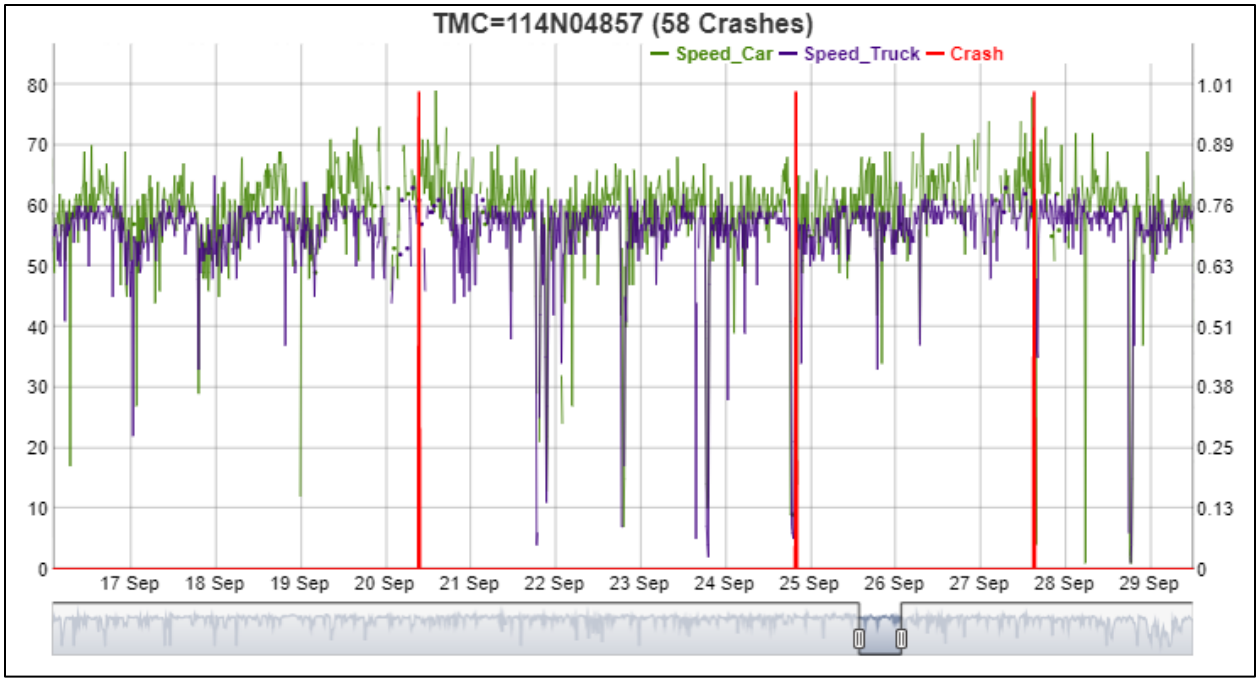


Figure 26. Range Selection Options in Dygraphs.

CHAPTER 5. CONCLUSIONS

This study aimed to determine the association between vehicle operating speed, roadway geometry, traffic volume, and crash occurrences. The project team developed conflated databases for the States of Washington and Ohio by incorporating HSIS and NPMRDS data. The data were then analyzed using three units of analysis (see figure 27):

- Segment level for annual-level crash predictions.
- Segment-temporal level for daily-level crash predictions.
- Segment-temporal level at time before and after crashes.

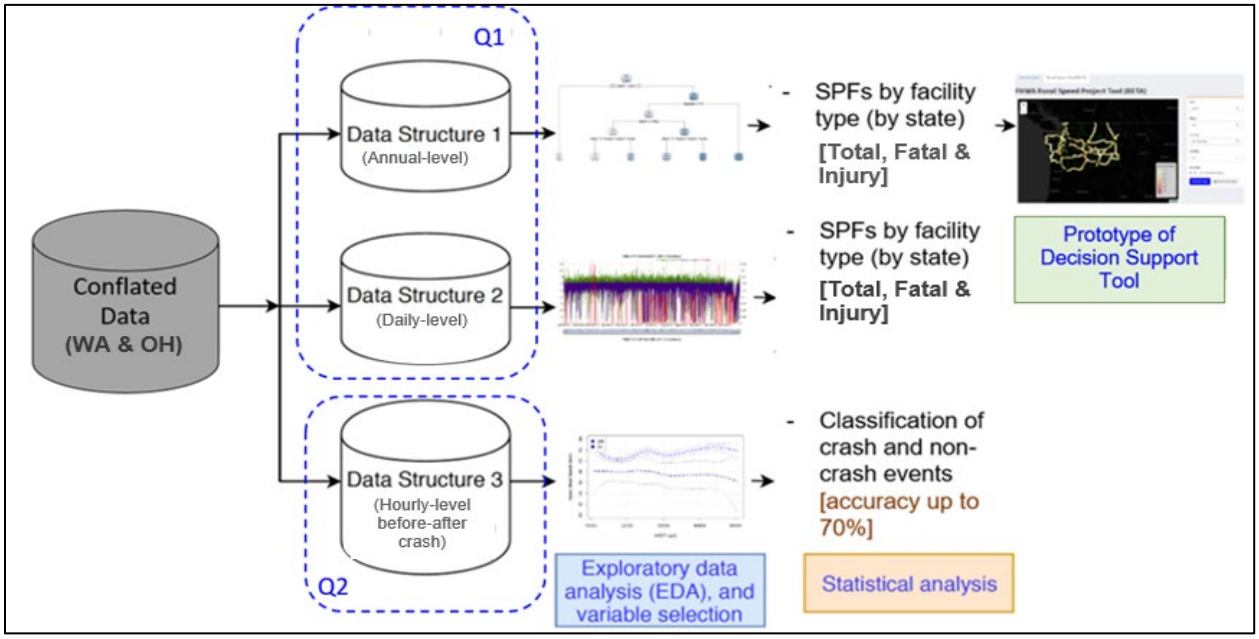


Figure 27. Overall Framework of Analysis.

FINDINGS FROM ANNUAL-LEVEL CRASH PREDICTION MODELING

Segment-level crash models based on the conflated dataset can provide reliable estimates of yearly crash frequencies. The general findings were:

- Certain speed measures were useful in the development of the annual-level statistical models. This study examined aggregated traffic travel speed variation over time.[6]
    - Increased variability in hourly operating speeds within a day and increased monthly operating speeds within a year were both associated with increased crashes.
    - Multilane, non-freeway roads with higher free-flow speeds tended to experience a higher rate of crashes than those with lower free-flow speeds. However, this effect

[6] The current study did not examine the speed variability between the vehicles because NPMRDS provides aggregated speed measures.

was negative for interstate roadways, likely due to their more robust highway design standards.
    - Operational speed differences between weekends and weekdays were positively associated with a higher number of crashes. Segments experiencing higher speed differentials between weekends and weekdays likely indicate the nature of roadway-use and land-use patterns.
    - Non-peak and non-event speed (average operating speed excluding peak hours and hours with events) was positively associated with crash rates on rural two-lane roadways. However, this speed measure was negatively associated with crashes in the interstate model. This finding for interstates could be because well-designed and high-standard roads are generally associated with higher non-peak and non-event speeds.
- As the proportion of horizontal curvature on a segment increased, the number of crashes also increased.
- In general, segments with intersections tended to have more crashes than segments without intersections, which is likely because segments with intersections have a greater number of conflict points. The variable was only significant for multilane roadways.

FINDINGS FROM DAILY-LEVEL CRASH PREDICTION MODELING

Daily-level crash predictions models based on the conflated dataset provided reliable estimates of daily crash frequencies. The general findings were:

- A variable was created to capture traffic speed variation throughout the day by measuring the standard deviation of daily average speed. The coefficient was positive and significant in all models, which signifies that a segment with high variation in daily average speeds is expected to experience a higher number of crashes than a segment with a lower variation in daily speeds. The strength of this finding is one of the biggest insights gained from this study.
- Average operating speed was positively associated with crashes for rural two-lane roadways. However, average operating speed was negatively associated with crashes in the interstate models. This finding for interstates could be because well-designed and high-standard roads are generally associated with higher average operating speeds.
- Daily average precipitation was positively associated with the number of daily crashes.
- Because the geometric variable remained constant at the segment while developing the model at the segment-temporal level, additional insights are needed for the model interpretation.

FINDINGS FROM EXPLORATORY EXAMINATION OF TIME BEFORE AND AFTER CRASHES

This analysis examined the time around crash events to investigate if any significant differences exist between speed series for 4 hours prior to a crash and comparable 4-hour series when no crash occurred (no crash but the same day of the week and hour on a different date). An ME model was fitted to investigate these differences. The analysis was limited to a randomly selected sample dataset (with 150 crashes from Washington interstate roadways). The overall outcome of this analysis was exploratory in nature. The findings were:

- After controlling for other influential factors, as the moment of crash occurrence approached, the speed trend for the crash-related series decreased and was substantially different than the trend of the non-crash-related reference series.
- Speed variability increased for the series just prior to a crash, which was also different than the non-crash series.

The overall finding from this study is that speed-related operational information is an area of opportunity to better understand safety outcomes. Several of the speed measures show positive association with crash outcomes at the segment level (annual or daily). Future replications with different datasets and facility types (for example, urban roadways) are needed to explore the association between operation speed measures, geometric factors, and crash outcomes.

DECISION SUPPORT TOOL

The project team developed an interactive decision support tool to show segment-level high-risk analysis using Washington and Ohio data that contain expected total crashes from the final models. The tool will have adaptability options for newer datasets (crash and speed data). Additionally, the project team developed time series models to forecast speed measures at different temporal units (hour or day) to scale the analysis in the presence of crash data only. The project team also provides recommendations on the integration of the updated NPMRDS data in the tool.

The project team developed a weblink that includes descriptive statistics and data visualization (both static and interactive) tools. The link provides a more detailed view of the speed measures and descriptive statistics, as well as visualization of the association between speed measures and crashes at a granular level. The descriptive statistics and data visualization tools from this weblink can provide new research and safety improvement opportunities.

REFERENCES

Abdel-Aty, M., and Radwan, A. Modeling traffic accident occurrence and involvement. Accident Analysis and Prevention 32, 2000, 633–642.

Anastasopoulos, P., and Mannering, F. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41 (1), 2009, 153–159.

Banihashemi, M., Dimaiuta, M., Zineddin, A., Spear, B., Smadi, O., and Hans, Z. Using Linked SHRP2 RID and NPMRDS Data to Study Speed-Safety Relationships on Urban Interstates and Major Arterials. The 98th TRB Annual Meeting, 2019, Washington DC.

Deepayan, S. Lattice: Multivariate Data Visualization with R. Springer, New York, 2008.

Demo Visualization using Washington Data. http://bit.ly/rss_sdi_demo. Accessed May 9, 2019.

DT. An R interface to the DataTables library. https://rstudio.github.io/DT/. Accessed May 9, 2019.

El-Basyouny, K., Sayed, T. Accident prediction models with random corridor parameters. Accident Analysis and Prevention 41 (5), 2009, 1118–1123.

Gargoum, S., and El-Basyouny, K. Exploring the association between speed and safety: A path analysis approach. Accident Analysis and Prevention 93, 2016, 32–40.

Geedipally, S., Lord, D., and Dhavala S. The negative binomial-Lindley generalized linear model: Characteristics and application using crash data. Accident Analysis and Prevention 45 (2), 2012, 258–265.

Guo, J., and Trivedi, P. Flexible parametric models for long-tailed patent count distributions. Oxford Bulletin of Economics and Statistics 64, 2000, 63–82.

Hauer, E. The Art of Regression Modeling in Road Safety. Springer, 2015.

Highway Performance Monitoring System (HPMS). https://www.fhwa.dot.gov/policyinformation/hpms.cfm. Accessed May 9, 2019.

Highway Safety Information System (HSIS). https://www.hsisinfo.org/. Accessed May 9, 2019.

Htmlwidgets in R. https://www.htmlwidgets.org/. Accessed May 9, 2019.

Imprialou, M., Quddus, M., Pitfield, D., and Lord, D. Re-visiting crash–speed relationships: A new perspective in crash modelling. Accident Analysis and Prevention 86, 2016, 173–185.

Interactive Decision Support Tool (Beta). https://ldwhite.shinyapps.io/RuralSpdSafety_Demo/. Accessed December 31, 2018.

Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34 (1), 1992, 1–14.

Lord, D., and Persaud, B. Accident Prediction Models with and without Trend: Application of the Generalized Estimating Equations (GEE) Procedure. Transportation Research Record, 1717: 6, 2000.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. Winbugs—A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing 10 (4), 2000, 325–337.

Moshitch, D., and I., Nelken. Using Tweedie distributions for fitting spike count data. Journal of Neuroscience Methods, 225, 2014, 13-28.

National Performance Management Research Dataset (NPMRDS). https://ops.fhwa.dot.gov/freight/freight_analysis/perform_meas/vpds/npmrdsfaqs.htm. Accessed August 23, 2018.

Pei, X., Wong, S., and Sze, N. The roles of exposure and speed in road safety analysis. Accident Analysis and Prevention 48, 2012, 464–471.

Roshandel, S., Zheng, Z., and Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. Accident Analysis and Prevention 79, 2015, 198–211.

Srinivasan, R. and K. Bauer. Safety Performance Function Development Guide: Developing Jurisdiction Specific SPFs. Report No. FHWA-SA-14-005, 2013.

Taylor, M., Lynam, D., and Baruya, A. The effects of drivers' speed on the frequency of road accidents. TRL Report 421, 2000.

USDOT. Safety Data Initiative. https://www.transportation.gov/policy/transportation-policy/safety/safetydatainitiative. Accessed August 23, 2018.

Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York, 2016.

Yu, R., Abdel-Aty, M., and Ahmed, M. Bayesian random-effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. Accident Analysis and Prevention 50, 2013, 371–376.

Yu, R., Quddus, M., Wang, X., and Yang, K. Impact of data aggregation approaches on the relationships between operating speed and traffic safety. Accident Analysis and Prevention 120, 2018, 304–310.

APPENDIX A. SAFETY PERFORMANCE FUNCTIONS: BASICS

SAFETY PERFORMANCE FUNCTIONS BY FACILITY TYPES

The project team developed SPFs for different facility types. The major contribution of this task is to incorporate speed measures and effect of precipitation in SPF development.

An SPF is an equation used to predict the average number of crashes per year at a location as a function of exposure and speed measures. For highway segments, exposure is represented by the segment length and AADT associated with the study section, as shown by the following baseline SPF:

$$C_{Predicted} = \exp[\beta_0 + \beta_1 \times \ln(L) + \beta_2 \times \ln(AADT) + \sum \beta_j \times X_j \,] \quad (15)$$

where:

$C_{Predicted}$ = the predicted crash frequency.
B_0 = intercept.
B_1, β_2 = coefficients for segment length and traffic volume, respectively.
L = segment length.
AADT = annual average daily traffic.
B_j = coefficients for other roadway features and speed measurements.
X_j = other roadway features (e.g., shoulder width, median) and speed measurements.

The empirical Bayes method is based on a weighted average principle. It uses a weight factor, w, to combine observed ($C_{Observed}$) and predicted crash frequencies ($C_{Predicted}$) to estimate the expected crash frequency, $C_{Expected}$:

$$C_{Expected} = w \times C_{Predicted} + (1 - w) \times C_{Observed} \quad (16)$$

where:

w = a weight factor that depends on the overdispersion parameter (OP) $= \frac{1}{1 + C_{Predicted} \times OP}$.
$C_{Expected}$ = expected crash frequency.
$C_{Predicted}$ = predicted crash frequency.

With the segment length (L) and the coefficients k and β_1 from the SPF, OP is calculated by:

$$OP = 1/\theta$$

where:

θ = the dispersion parameter of the negative binomial model (i.e., the shape parameter of the Gamma distribution).

Example

Assume a rural interstate segment has a length equal to 1.5 mi and AADT equal to 13,000. The yearly average operating speed is 74.8 mph. Two crashes occurred on this segment in the past year; one was a KABC crash and the other was a PDO crash.

The following steps summarize the process for performing the calculations to evaluate the expected crashes. (Note: for example purposes only; coefficients are not taken from the developed models).

Step 1: Calculate the predicted average crash frequency, $C_{Predicted}$.

Total crashes:

$$\begin{aligned} C_{Predicted\_Total} &= \exp\left[\beta_0 + \beta_1 \times \ln(L) + \beta_2 \times \ln(AADT) + \beta_{SpdAve} \times SpdAve\right] \\ &= \exp[-6.0 + 0.5 \times \ln(1.5) + 0.6 \times \ln(13{,}000) + 0.007 \times 74.8] \\ &= 1.51\ (crashes\ per\ year) \end{aligned}$$

FI crashes:

$$\begin{aligned} C_{Predicted\_FI} &= \exp\left[\beta_0 + \beta_1 \times \ln(L) + \beta_2 \times \ln(AADT) + \beta_{SpdAve} \times SpdAve\right] \\ &= \exp[-7.0 + 0.5 \times \ln(1.5) + 0.6 \times \ln(13{,}000) + 0.007 \times 74.8] \\ &= 0.55\ (FI\ crashes\ per\ year) \end{aligned}$$

Step 2: Calculate the OP and weight factor for the segment.

Total crashes:

$$OP_{Total} = \frac{1}{\theta} = \frac{1}{5.41} = 0.185$$

$$w_{Total} = \frac{1}{1 + C_{Predicted\_Total} \times OP_{Total}} = \frac{1}{1 + 0.185 \times 1.51} = 0.782$$

FI crashes:

$$OP_{FI} = \frac{1}{\theta} = \frac{1}{2.41} = 0.414$$

$$w_{FI} = \frac{1}{1 + C_{Predicted\_FI} \times OP_{FI}} = \frac{1}{1 + 0.414 \times 0.55} = 0.814$$

Step 3: Calculate the expected crashes, $C_{Expected}$.

Total crashes:

$$\begin{aligned} C_{Expected\_Total} &= w_{Total} \times C_{Predicted\_Total} + (1 - w_{Total}) \times C_{Observed\_Total} \\ &= 0.782 \times 1.51 + (1 - 0.782) \times 2 \\ &= 1.62 \text{ crashes/year} \end{aligned}$$

FI crashes:

$$C_{Expected\_FI} = w_{FIl} \times C_{Predicted\_FI} + (1 - w_{Total}) \times C_{Observed\_FI}$$
$$= 0.814 \times 0.55 + (1 - 0.814) \times 1$$
$$= 0.63 \text{ crashes/year}$$

APPENDIX B. DEVELOPED MODELS (ANNUAL-LEVEL DATA)

Table 24. Calibrated Coefficients for KABCO Crashes on Interstates—Two States.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−4.9668	1.0297	−4.82	<.0001
b_{aadt}	Traffic volume (AADT)	0.7613	0.0743	10.24	<.0001
b_{hc}	Percentage of curve (PerHC)	0.0825	0.0327	2.53	0.0118
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.1068	0.0370	2.88	0.0041
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	−0.0378	0.01332	−2.84	0.0047
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
b_{OH}	Added effect of Ohio	0.6284	0.07856	8.0	<.0001
k	Inverse dispersion parameter for 4-lane segments	−0.4359	0.09577	−4.55	<.0001
	Inverse dispersion parameter for 6-lane segments	−0.4672	0.1284	−3.64	0.0003

Note: A dash (—) = very highly insignificant.

Table 25. Calibrated Coefficients for KABC Crashes on Interstates—Two States.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−5.8519	1.3414	−4.36	<.0001
b_{aadt}	Traffic volume (AADT)	0.8221	0.09156	8.98	<.0001
b_{hc}	Percentage of curve (PerHC)	0.08274	0.03688	2.24	0.0253
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.07597	0.0475	1.6	0.1103
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	0.1084	0.09142	1.19	0.2361
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	−0.0551	0.01833	−3.01	0.0028
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
b_{OH}	Added effect of Ohio	0.4119	0.1021	4.04	<.0001
k	Inverse dispersion parameter for 4-lane segments	−0.4875	0.1629	−2.99	0.0029
	Inverse dispersion parameter for 6-lane segments	−0.2615	0.2211	−1.18	0.2374

Table 26. Calibrated Coefficients for PDO Crashes on Interstates—Two States.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−5.0697	1.0658	−4.76	<.0001
b_{aadt}	Traffic volume (AADT)	0.7594	0.07803	9.73	<.0001
b_{hc}	Percentage of curve (PerHC)	0.06865	0.03367	2.04	0.0419
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.09919	0.03975	2.5	0.0129
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	0.08506	0.06708	1.27	0.2053
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	−0.0406	0.01387	−2.93	0.0035
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
b_{OH}	Added effect of Ohio	0.6926	0.08304	8.34	<.0001
k	Inverse dispersion parameter for 4-lane segments	−0.4493	0.1013	−4.43	<.0001
	Inverse dispersion parameter for 6-lane segments	−0.4972	0.1338	−3.71	0.0002

Table 27. Calibrated Coefficients for KABCO Crashes on Interstates—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−3.5814	1.5557	−2.3	0.0219
b_{aadt}	Traffic volume (AADT)	0.8028	0.1157	6.94	<.0001
b_{hc}	Percentage of curve (PerHC)	−0.6258	0.2104	−2.97	0.0031
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.0612	0.0585	1.05	0.2963
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	−0.0547	0.01667	−3.28	0.0011
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
k	Inverse dispersion parameter for 4-lane segments	−0.4634	0.1198	−3.87	0.0001
	Inverse dispersion parameter for 6-lane segments	−0.7741	0.1602	−4.83	<.0001

Table 28. Calibrated Coefficients for KABC Crashes on Interstates—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−7.7923	2.0571	−3.79	0.0002
b_{aadt}	Traffic volume (AADT)	1.05	0.1526	6.88	<.0001
b_{hc}	Percentage of curve (PerHC)	−0.1572	0.5366	−0.29	0.7697
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.0893	0.074	1.21	0.2285
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	−0.0549	0.02177	−2.52	0.0121
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
k	Inverse dispersion parameter for 4-lane segments	−0.5888	0.1975	−2.98	0.0031
	Inverse dispersion parameter for 6-lane segments	−0.6853	0.2453	−2.79	0.0055

Table 29. Calibrated Coefficients for PDO Crashes on Interstates—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−3.1117	1.6184	−1.92	0.0553
b_{aadt}	Traffic volume (AADT)	0.7611	0.1156	6.59	<.0001
b_{hc}	Percentage of curve (PerHC)	−0.7141	0.1793	−3.98	<.0001
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	—	—	—	—
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	0.0967	0.0864	1.12	0.2637
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	−0.0578	0.0177	−3.27	0.0012
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
k	Inverse dispersion parameter for 4-lane segments	−0.4935	0.1254	−3.94	0.0001
	Inverse dispersion parameter for 6-lane segments	−0.7845	0.1655	−4.74	<.0001

Table 30. Calibrated Coefficients for KABCO Crashes on Interstates—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−6.2446	0.8187	−7.63	<.0001
b_{aadt}	Traffic volume (AADT)	0.6358	0.08307	7.65	<.0001
b_{hc}	Percentage of curve (PerHC)	0.0909	0.03139	2.9	0.0041
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.1568	0.04537	3.45	0.0007
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
k	Inverse dispersion parameter for 4-lane segments	−0.2749	0.1688	−1.63	0.1048
	Inverse dispersion parameter for 6-lane segments	0.166	0.238	0.7	0.4862

Table 31. Calibrated Coefficients for KABC Crashes on Interstates—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−4.9994	1.8192	−2.75	0.0065
b_{aadt}	Traffic volume (AADT)	0.6498	0.1083	6.0	<.0001
b_{hc}	Percentage of curve (PerHC)	0.0780	0.0342	2.28	0.0233
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.1063	0.05696	1.87	0.0633
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	−0.0405	0.0287	−1.41	0.1597
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
k	Inverse dispersion parameter for 4-lane segments	−0.2156	0.304	−0.71	0.4788
	Inverse dispersion parameter for 6-lane segments	0.7905	0.5844	1.35	0.1774

Table 32. Calibrated Coefficients for PDO Crashes on Interstates—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−7.3084	0.9487	−7.7	<.0001
b_{aadt}	Traffic volume (AADT)	0.7098	0.09614	7.38	<.0001
b_{hc}	Percentage of curve (PerHC)	0.0665	0.03109	2.14	0.0334
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.1439	0.04983	2.89	0.0042
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	0.1882	0.1206	1.56	0.1202
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
k	Inverse dispersion parameter for 4-lane segments	−0.1788	0.192	−0.93	0.3527
	Inverse dispersion parameter for 6-lane segments	0.0904	0.2575	0.35	0.726

Table 33. Calibrated Coefficients for KABCO Crashes on Two-Lane Highways—Two States.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−5.8138	0.3697	−15.73	<.0001
b_{aadt}	Traffic volume (AADT)	0.6048	0.04399	13.75	<.0001
b_{lw}	Lane width (LW)	−0.0245	0.02119	−1.16	0.2479
b_{hc}	Percentage of curve (PerHC)	0.8681	0.2418	3.59	0.0003
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	—	—	—	—
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{int}	Intersection presence (PIntPre)	0.3163	0.05495	5.76	<.0001
b_{prec}	Percentage of days with precipitation (PPrcp)	−0.5372	0.225	−2.39	0.0171
b_{OH}	Added effect of Ohio	0.6332	0.07294	8.68	<.0001
k	Inverse dispersion parameter	−0.5898	0.08101	−7.28	<.0001

Table 34. Calibrated Coefficients for KABC Crashes on Two-Lane Highways—Two States.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−7.6705	0.5459	−14.05	<.0001
b_{aadt}	Traffic volume (AADT)	0.6435	0.06245	10.3	<.0001
b_{lw}	Lane width (LW)	−0.044	0.031	−1.41	0.1578
b_{hc}	Percentage of curve (PerHC)	1.0319	0.3672	2.81	0.005
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	—	—	—	—
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	0.1654	0.0759	2.18	0.0295
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{int}	Intersection presence (PIntPre)	0.2769	0.07754	3.57	0.0004
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
b_{OH}	Added effect of Ohio	0.4103	0.1013	4.05	<.0001
k	Inverse dispersion parameter	−0.5419	0.1751	−3.1	0.002

Table 35. Calibrated Coefficients for PDO Crashes on Two-Lane Highways—Two States.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−5.8097	0.4087	−14.22	<.0001
b_{aadt}	Traffic volume (AADT)	0.5679	0.04852	11.7	<.0001
b_{lw}	Lane width (LW)	—	—	—	—
b_{hc}	Percentage of curve (PerHC)	0.823	0.264	3.12	0.0019
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.02845	0.01509	1.89	0.0596
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	−0.0689	0.05959	−1.16	0.2476
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{int}	Intersection presence (PIntPre)	0.3296	0.06053	5.45	<.0001
b_{prec}	Percentage of days with precipitation (PPrcp)	−0.6149	0.2535	−2.43	0.0154
b_{OH}	Added effect of Ohio	0.6691	0.08161	8.2	<.0001
k	Inverse dispersion parameter	−0.6856	0.0909	−7.54	<.0001

Table 36. Calibrated Coefficients for KABCO Crashes on Two-Lane Highways—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−3.6772	0.7811	−4.71	<.0001
b_{aadt}	Traffic volume (AADT)	0.3706	0.08698	4.26	<.0001
b_{lw}	Lane width (LW)	0.04195	0.03112	1.35	0.1782
b_{hc}	Percentage of curve (PerHC)	0.3297	0.4105	0.8	0.4222
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	—	—	—	—
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{int}	Intersection presence (PIntPre)	0.4498	0.0889	5.06	<.0001
b_{prec}	Percentage of days with precipitation (PPrcp)	0.6097	0.6492	0.94	0.348
k	Inverse dispersion parameter	−0.5954	0.1132	−5.26	<.0001

Table 37. Calibrated Coefficients for KABC Crashes on Two-Lane Highways—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−6.8237	1.0436	−6.54	<.0001
b_{aadt}	Traffic volume (AADT)	0.5967	0.1223	4.88	<.0001
b_{lw}	Lane width (LW)	0.05122	0.04212	1.22	0.2244
b_{hc}	Percentage of curve (PerHC)	0.6058	0.6394	0.95	0.3437
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	—	—	—	—
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{int}	Intersection presence (PIntPre)	0.4074	0.1149	3.55	0.0004
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
k	Inverse dispersion parameter	−0.3375	0.2781	−1.21	0.2253

Table 38. Calibrated Coefficients for PDO Crashes on Two-Lane Highways—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−3.4111	0.7002	−4.87	<.0001
b_{aadt}	Traffic volume (AADT)	0.3352	0.08329	4.02	<.0001
b_{lw}	Lane width (LW)	—	—	—	—
b_{hc}	Percentage of curve (PerHC)	—	—	—	—
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	—	—	—	—
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	0.3224	0.1433	2.25	0.0248
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{int}	Intersection presence (PIntPre)	0.3817	0.0865	4.41	<.0001
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
k	Inverse dispersion parameter	−0.6361	0.1168	−5.44	<.0001

Table 39. Calibrated Coefficients for KABCO Crashes on Two-Lane Highways—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−6.5573	0.4216	−15.55	<.0001
b_{aadt}	Traffic volume (AADT)	0.6962	0.05081	13.7	<.0001
b_{lw}	Lane width (LW)	−0.0962	0.03186	−3.02	0.0026
b_{hc}	Percentage of curve (PerHC)	1.1538	0.3097	3.73	0.0002
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.04442	0.01539	2.89	0.004
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{int}	Intersection presence (PIntPre)	0.2278	0.0685	3.33	0.0009
b_{prec}	Percentage of days with precipitation (PPrcp)	−0.8338	0.2398	−3.48	0.0005
k	Inverse dispersion parameter	−0.4755	0.1219	−3.9	0.0001

Table 40. Calibrated Coefficients for KABC Crashes on Two-Lane Highways—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−7.9683	0.639	−12.47	<.0001
b_{aadt}	Traffic volume (AADT)	0.6744	0.07368	9.15	<.0001
b_{lw}	Lane width (LW)	−0.1814	0.05974	−3.04	0.0025
b_{hc}	Percentage of curve (PerHC)	1.1829	0.4547	2.6	0.0095
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.0356	0.02236	1.59	0.1118
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	0.1878	0.09764	1.92	0.0549
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{int}	Intersection presence (PIntPre)	0.1801	0.1041	1.73	0.084
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
k	Inverse dispersion parameter	−0.5871	0.2444	−2.4	0.0166

Table 41. Calibrated Coefficients for PDO Crashes on Two-Lane Highways—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−6.9051	0.4772	−14.47	<.0001
b_{aadt}	Traffic volume (AADT)	0.7048	0.05729	12.3	<.0001
b_{lw}	Lane width (LW)	−0.0672	0.0342	−1.96	0.0499
b_{hc}	Percentage of curve (PerHC)	1.1356	0.3471	3.27	0.0011
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.04454	0.0168	2.65	0.0082
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{int}	Intersection presence (PIntPre)	0.2398	0.07629	3.14	0.0017
b_{prec}	Percentage of days with precipitation (PPrcp)	−1.0547	0.2707	−3.9	0.0001
k	Inverse dispersion parameter	−0.5654	0.1411	−4.01	<.0001

Table 42. Calibrated Coefficients for KABCO Crashes on Multilane Highways—Two States.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−6.4938	0.9805	−6.62	<.0001
b_u	Adjustment for undivided road	0.2686	0.1588	1.69	0.0911
b_{aadt}	Traffic volume (AADT)	0.4848	0.09217	5.26	<.0001
b_{cr}	Percentage of curve (PerHC)	2.0307	0.675	3.01	0.0027
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.05879	0.02512	2.34	0.0196
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	0.3911	0.1341	2.92	0.0037
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	0.02695	0.006935	3.89	0.0001
b_{int}	Intersection presence (PIntPre)	0.5714	0.1006	5.68	<.0001
b_{prec}	Percentage of days with precipitation (PPrcp)	−1.9369	0.5511	−3.51	0.0005
b_{OH}	Added effect of Ohio	0.8282	0.1517	5.46	<.0001
k	Inverse dispersion parameter for undivided roads	−0.5868	0.2812	−2.09	0.0373
	Inverse dispersion parameter for divided roads	−0.9955	0.1013	−9.82	<.0001

Table 43. Calibrated Coefficients for KABC Crashes on Multilane Highways—Two States.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−6.9752	1.2166	−5.73	<.0001
b_u	Adjustment for undivided road	0.3903	0.1899	2.06	0.0402
b_{aadt}	Traffic volume (AADT)	0.3573	0.1066	3.35	0.0008
b_{cr}	Percentage of curve (PerHC)	1.574	0.6751	2.33	0.02
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.04402	0.03142	1.4	0.1617
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	0.2418	0.1371	1.76	0.0783
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	0.2159	0.1605	1.34	0.1792
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	0.02393	0.008964	2.67	0.0078
b_{int}	Intersection presence (PIntPre)	0.5625	0.1169	4.81	<.0001
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
b_{OH}	Added effect of Ohio	0.3397	0.1701	2	0.0463
k	Inverse dispersion parameter for undivided roads	−0.2271	0.6194	−0.37	0.7139
	Inverse dispersion parameter for divided roads	−0.8616	0.1959	−4.4	<.0001

Table 44. Calibrated Coefficients for PDO Crashes on Multilane Highways—Two States.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−7.4546	1.0195	−7.31	<.0001
b_u	Adjustment for undivided road	0.2053	0.168	1.22	0.2222
b_{aadt}	Traffic volume (AADT)	0.5529	0.09649	5.73	<.0001
b_{cr}	Percentage of curve (PerHC)	2.3082	0.7815	2.95	0.0033
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.05048	0.02538	1.99	0.0472
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	0.4013	0.1369	2.93	0.0035
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	0.02519	0.007172	3.51	0.0005
b_{int}	Intersection presence (PIntPre)	0.5797	0.1034	5.61	<.0001
b_{prec}	Percentage of days with precipitation (PPrcp)	−1.8614	0.5761	−3.23	0.0013
b_{OH}	Added effect of Ohio	1.005	0.1641	6.12	<.0001
k	Inverse dispersion parameter for undivided roads	−0.6188	0.3123	−1.98	0.048
	Inverse dispersion parameter for divided roads	−0.9469	0.1105	−8.57	<.0001

Table 45. Calibrated Coefficients for KABCO Crashes on Multilane Highways—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−5.5104	1.2142	−4.54	<.0001
b_u	Adjustment for undivided road	0.4826	0.2098	2.3	0.0219
b_{aadt}	Traffic volume (AADT)	0.4335	0.1193	3.63	0.0003
b_{cr}	Percentage of curve (PerHC)	1.5865	1.0349	1.53	0.126
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.06663	0.02749	2.42	0.0158
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	0.2969	0.1651	1.8	0.0728
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	0.03078	0.008153	3.78	0.0002
b_{int}	Intersection presence (PIntPre)	0.6052	0.1256	4.82	<.0001
b_{prec}	Percentage of days with precipitation (PPrcp)	−1.7573	0.9595	−1.83	0.0677
k	Inverse dispersion parameter for undivided roads	−0.601	0.3325	−1.81	0.0714
	Inverse dispersion parameter for divided roads	−1.0595	0.1147	−9.24	<.0001

Table 46. Calibrated Coefficients for KABC Crashes on Multilane Highways—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−6.5777	1.334	−4.93	<.0001
b_u	Adjustment for undivided road	0.4598	0.2491	1.85	0.0655
b_{aadt}	Traffic volume (AADT)	0.3381	0.1328	2.55	0.0112
b_{cr}	Percentage of curve (PerHC)	1.4143	0.8936	1.58	0.1141
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.05036	0.03329	1.51	0.1309
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	0.4081	0.1594	2.56	0.0107
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	—	—	—	—
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	0.02381	0.009788	2.43	0.0153
b_{int}	Intersection presence (PIntPre)	0.7757	0.1346	5.76	<.0001
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
k	Inverse dispersion parameter for undivided roads	−0.5042	0.6187	−0.81	0.4155
	Inverse dispersion parameter for divided roads	−0.6138	0.2682	−2.29	0.0225

Table 47. Calibrated Coefficients for PDO Crashes on Multilane Highways—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−6.0308	1.2302	−4.9	<.0001
b_u	Adjustment for undivided road	0.4513	0.2146	2.1	0.036
b_{aadt}	Traffic volume (AADT)	0.4945	0.1233	4.01	<.0001
b_{cr}	Percentage of curve (PerHC)	0.9358	0.7851	1.19	0.2339
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	0.0599	0.02743	2.18	0.0295
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	0.3553	0.1653	2.15	0.0322
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	0.03055	0.008305	3.68	0.0003
b_{int}	Intersection presence (PIntPre)	0.5664	0.1267	4.47	<.0001
b_{prec}	Percentage of days with precipitation (PPrcp)	−2.3932	0.964	−2.48	0.0134
k	Inverse dispersion parameter for undivided roads	−0.5522	0.3681	−1.5	0.1342
	Inverse dispersion parameter for divided roads	−1.0097	0.1236	−8.17	<.0001

Table 48. Calibrated Coefficients for KABCO Crashes on Multilane Highways—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−6.4405	1.3246	−4.86	<.0001
b_u	Adjustment for undivided road	—	—	—	—
b_{aadt}	Traffic volume (AADT)	0.6473	0.143	4.53	<.0001
b_{cr}	Percentage of curve (PerHC)	3.6393	1.2532	2.9	0.0043
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	−0.1038	0.07085	−1.47	0.145
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	−0.292	0.162	−1.8	0.0737
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	0.8381	0.2444	3.43	0.0008
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{int}	Intersection presence (PIntPre)	0.452	0.1522	2.97	0.0035
b_{prec}	Percentage of days with precipitation (PPrcp)	−1.0708	0.7006	−1.53	0.1288
k	Inverse dispersion parameter for undivided roads	−0.4212	0.2301	−1.83	0.0693
	Inverse dispersion parameter for divided roads	−0.4212	0.2301	−1.83	0.0693

Table 49. Calibrated Coefficients for KABC Crashes on Multilane Highways—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−11.196	2.4944	−4.49	<.0001
b_u	Adjustment for undivided road	0.3376	0.3285	1.03	0.3059
b_{aadt}	Traffic volume (AADT)	0.6593	0.1706	3.86	0.0002
b_{cr}	Percentage of curve (PerHC)	0.9047	0.6937	1.3	0.1944
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	−0.1504	0.08773	−1.71	0.0887
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	1.0588	0.2703	3.92	0.0001
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	0.0598	0.02571	2.33	0.0215
b_{int}	Intersection presence (PIntPre)	−0.0212	0.1919	−0.11	0.9123
b_{prec}	Percentage of days with precipitation (PPrcp)	—	—	—	—
k	Inverse dispersion parameter for undivided roads	0.0506	0.5727	0.09	0.9297
	Inverse dispersion parameter for divided roads	0.0506	0.5727	0.09	0.9297

Table 50. Calibrated Coefficients for PDO Crashes on Multilane Highways—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−7.634	1.4845	−5.14	<.0001
b_u	Adjustment for undivided road	−0.3211	0.243	−1.32	0.1886
b_{aadt}	Traffic volume (AADT)	0.7084	0.1577	4.49	<.0001
b_{cr}	Percentage of curve (PerHC)	4.7683	1.8239	2.61	0.0099
b_{sd}	Avg. spd. diff. in weekday/weekend (SpdW_W)	−0.116	0.07341	−1.58	0.1164
b_{sv1}	Standard dev. of hourly operating speeds (SDHrSpd)	—	—	—	—
b_{sv2}	Standard dev. of monthly operating speeds (SDMonSpd)	0.6856	0.2526	2.71	0.0075
b_{sff}	Average hourly non-peak non-event speed (SpdNPNE)	—	—	—	—
b_{int}	Intersection presence (PIntPre)	0.5999	0.1672	3.59	0.0005
b_{prec}	Percentage of days with precipitation (PPrcp)	−1.1059	0.7611	−1.45	0.1485
k	Inverse dispersion parameter for undivided roads	−0.417	0.2592	−1.61	0.11
	Inverse dispersion parameter for divided roads	−0.417	0.2592	−1.61	0.11

APPENDIX C. SAFETY DISTRIBUTION FUNCTIONS (ANNUAL-LEVEL DATA)

This appendix documents the development of severity distribution functions.

FUNCTIONAL FORM

An SDF is represented by a discrete choice model. It is used to predict the proportion of crashes in each of the following severity categories: fatal = K, injury = I, or property damage only = O. The SDF can be used with the SPFs to estimate the expected crash frequency for each severity category. It may include various geometric, operation, and traffic variables that will allow the estimated proportion to be specific to an individual segment.

The multinomial logit (MNL) model is used to predict the probability of crash severities. Given the characteristics of the data, the MNL is the most suitable model for estimating an SDF. A linear function is used to relate the crash severity with the operational variables. SAS's nonlinear mixed modeling procedure is used for the evaluation of the MNL model.

The probability for each crash severity category is given by the following equations:

$$P_K = \frac{e^{V_K}}{1 + e^{V_K} + e^{V_I} + e^{V_O}} \quad (18)$$

$$P_I = \frac{e^{V_I}}{1 + e^{V_K} + e^{V_I} + e^{V_O}} \quad (19)$$

$$P_O = 1 - (P_K + P_I) \quad (20)$$

where:

P_j = probability of the occurrence of crash severity *j*.
V_j = systematic component of crash severity likelihood for severity *j*.

MODEL DEVELOPMENT

The database assembled for calibration includes crash severity level as a dependent variable and the geometric and operational variables of each site as independent variables. Each row (site characteristics) is repeated to the frequency of each severity level. Thus, a segment with *n* crashes will be repeated *n* number of times. It should be noted that the segments without any crashes are not included in the database. The total sample size of the final dataset for model calibration will be equal to the total number of crashes in the data. During the model calibration, the PDO category is set as the base scenario with coefficients restricted at zero.

Rural Interstate Highways

A model for estimating the systematic component of crash severity V_j for interstate segments is described by the following equations.

$$\begin{aligned} V_K &= ASC_K + b_{sbc,K} \times SpdBefCrSD1 + b_{sv,K} \times SpdVarr1 + b_{pr,K} \times p_{prec} + b_{OH,K} \\ &\quad \times I_{OH} \\ V_I &= ASC_I + b_{sbc,I} \times SpdBefCrSD1 + b_{sv,I} \times SpdVarr1 + b_{pr,I} \times p_{prec} + b_{OH,I} \\ &\quad \times I_{OH} \end{aligned} \quad (21)$$

where:

$SpdBefCrSD1$ = standard deviation of operating speed before crash (1 hour; crash day).
$SpdVarr1$ = standard deviation of operating speed by hour.
p_{prec} = percent of days with precipitation.
I_{OH} = Ohio indicator variable (= 1.0 if Ohio, 0.0 if Washington).
ASC_j = alternative specific constant for crash severity *j*.
$b_{k,j}$ = calibration coefficient for variable *k* and crash severity *j*.

Table 51 summarizes the estimation results of the MNL model for the interstate segments. An examination of the coefficient values and their implication on the corresponding crash severity levels are documented in this section.

Table 51. Parameter Estimation for the Interstate Segments' SDF.

Coefficient	Variable	Fatality (K)		Injury (I)	
		Value	t-statistic	Value	t-statistic
ASC	Alternative specific constant	−4.765	−6.53	−1.024	−8.51
b_{sbc}	Standard Deviation of Operating Speed before Crash (SDBCSpd)	0.017	2.33	0.017	2.33
b_{sv}	Standard dev. of hourly operating speeds (SDHrSpd)	1.024	2.75	−0.134	−2.18
b_{pr}	Percentage of days with precipitation (PPrcp)	−0.046	−2.37	—	—
b_{OH}	Ohio	−0.701	−2.14	−0.309	−4.72
Observations		7,443 crashes (K = 41; I = 1,597; O = 5,805)			

Note: PDO is the base scenario with coefficients restricted at zero.

Standard deviation of operating speed before crash: This variable represents the standard deviation of operating speed before a crash. As this standard deviation of speed before a crash increases, the probability of fatal or injury crash severity increases.

Standard deviation of operating speed by hour: This variable represents the standard deviation of operating speed by the hour. A larger variable value represents that the road segment will experience frequent congestion or the speeds are higher than normal during some hours of a day. As this variable value increases, the probability of a fatal crash increases but the probability of an injury crash decreases.

Precipitation: This variable represents the percent of days with some level of precipitation. The coefficient is negative and significant for fatal crashes only. It shows that, with the increase in precipitation, the chance of fatal crashes decreases, probably due to the decrease in speeds.

State: This variable indicates whether the segment is in Ohio or Washington. The model coefficients indicate that the crashes in Ohio are less severe than in Washington. This finding could be due to differences in weather, terrain, or reporting thresholds.

Rural Two-Lane Highways

A model for estimating the systematic component of crash severity V_j for two-lane segments is described by the following equations.

$$\begin{aligned} V_K &= ASC_K + b_{sfw,K} \times SFW + b_{sv,K} \times SpdVarr1 + b_{sbc,K} \times SpdBefCr1 + b_{OH,K} \times I_{OH} \\ V_I &= ASC_I + b_{sfw,I} \times SFW + b_{sv,I} \times SpdVarr1 + b_{sbc,I} \times SpdBefCr1 + b_{OH,I} \times I_{OH} \end{aligned} \quad (22)$$

where:

SFW = surface width.

Table 52 summarizes the estimation results of the MNL model for the two-lane segments. An examination of the coefficient values and their implications on the corresponding crash severity levels are documented in a subsequent section.

Table 52. Parameter Estimation for the Two-Lane Segments' SDF.

Coefficient	Variable	Fatality (K)		Injury (I)	
		Value	t-statistic	Value	t-statistic
ASC	Alternative specific constant	−3.833	−1.54	0.155	0.29
b_{sfw}	Surface width	−0.072	−0.95	−0.029	−1.82
b_{sv}	Standard dev. of hourly operating speeds (SDHrSpd)	0.178	0.85	0.146	2.74
b_{pr}	Percentage of days with precipitation (PPrcp)	−0.029	1.22	0.011	−2.25
b_{OH}	Ohio	−0.354	−1.17	−0.311	−3.91
Observations		7,443 crashes (K = 53; I = 987; O = 2,851)			

Note: PDO is the base scenario with coefficients restricted at zero.

Standard deviation of operating speed before crash: This variable represents the standard deviation of operating speed before a crash. The coefficient is positive and marginally significant for fatal crashes only. As this standard deviation of speed before a crash increases, the probability of fatal crash severity increases and has no effect on injury crash severity.

Standard deviation of operating speed by hour: This variable represents the standard deviation of operating speed by the hour. The coefficient is positive and marginally significant for fatal crashes only. As this variable value increases, the probability of a fatal crash increases.

Precipitation: This variable represents the percent of days with some level of precipitation. The coefficient is negative for fatal crashes but positive for injury crashes. It shows that, with the increase in precipitation, the chance of fatal crashes decreases but the likelihood of injury crashes increases, probably due to the decrease in speeds.

State: This variable indicates whether the segment is in Ohio or Washington. The model coefficients indicate that the crashes in Ohio are less severe than in Washington. This finding could be due to differences in weather, terrain, or reporting thresholds.

Rural Multilane Highways

A model for estimating the systematic component of crash severity V_j for multilane segments is described by the following equations.

$$\begin{aligned} V_K &= ASC_K + b_{sv,K} \times SpdVarr1 + b_{sbc,K} \times SpdBefCr1 + b_{pr,K} \times p_{prec} + b_{OH,K} \\ &\quad \times I_{OH} \\ V_I &= ASC_I + b_{pr,I} \times p_{prec} + b_{OH,I} \times I_{OH} \end{aligned} \quad (23)$$

Table 53 summarizes the estimation results of the MNL model for the multilane segments.

Table 53. Parameter Estimation for the Multilane Segments' SDF.

Coefficient	Variable	Fatality (K)		Injury (I)	
		Value	t-statistic	Value	t-statistic
ASC	Alternative specific constant	−7.399	−2.12	−1.056	−4.39
b_{sv}	Standard dev. of hourly operating speeds (SDHrSpd)	0.411	1.41	—	—
b_{sbc}	Standard Deviation of Operating Speed before Crash (SDBCSpd)	0.063	1.27	—	—
b_{pr}	Percentage of days with precipitation (PPrcp)	−0.038	−1.10	0.0038	0.63
b_{OH}	Ohio	−0.339	−0.57	−0.4335	−3.89
Observations		2,587 crashes (K = 15; I = 569; O = 2,003)			

Note: PDO is the base scenario with coefficients restricted at zero.

Standard deviation of operating speed before crash: This variable represents the standard deviation of operating speed before a crash. The coefficient is positive and marginally significant for fatal crashes only. As this standard deviation of speed before a crash increases, the probability of fatal crash severity increases and has no effect on injury crash severity.

Standard deviation of operating speed by hour: This variable represents the standard deviation of operating speed by the hour. The coefficient is positive and marginally significant for fatal crashes only. As this variable value increases, the probability of a fatal crash increases.

Precipitation: This variable represents the percent of days with some level of precipitation. The coefficient is negative for fatal crashes but positive for injury crashes. It shows that, with the increase in precipitation, the chance of fatal crashes decreases but the likelihood of injury crashes increases, probably due to the decrease in speeds.

State: This variable indicates whether the segment is in Ohio or Washington. The model coefficients indicate that the crashes in Ohio are less severe than in Washington. This finding could be due to differences in weather, terrain, or reporting thresholds.

APPENDIX D. DEVELOPED MODELS (DAILY-LEVEL DATA)

Table 54. Calibrated Coefficients for KABCO Crashes on Interstate Roadways—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−9.4829	0.8506	−11.1480	<0.0001
b_{aadt}	Traffic volume (AADT)	0.7126	0.0788	9.0450	<0.0001
b_l	Segment length (Len)	0.2328	0.0087	26.6980	<0.0001
b_{nl}	Number of lanes (Lanes)	0.0589	0.1401	0.4210	0.6741
b_{sw}	Lane width (LW)	−0.0093	0.0120	−0.7720	0.4404
b_{prec}	Percentage of precipitation (PPrep)	0.2352	0.0545	4.3150	<0.0001
b_{nc}	Number of curvatures (NCurv)	−0.6125	0.2124	−2.8840	0.0039
b_{lc}	Total length of curvatures (LCurv)	—	—	—	—
b_{ssd}	Standard deviation of daily average speed (SDDailySpd)	0.1376	0.0286	4.8160	<0.0001
b_{avs}	Daily average speed (SpdAvgDaily)	−0.0405	0.0054	−7.5440	<0.0001

Note: Null deviance: 24422; Residual deviance: 23182; Dispersion parameter: 1.4693; Number of Fisher scoring iterations: 6.

Table 55. Calibrated Coefficients for KABC Crashes on Interstate Roadways—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−15.7714	2.0044	−7.8680	<0.0001
b_{aadt}	Traffic volume (AADT)	1.0987	0.1831	6.0020	<0.0001
b_l	Segment length (Len)	0.2256	0.0199	11.3650	<0.0001
b_{sw}	Lane width (LW)	0.5191	0.3053	1.7010	0.0890
b_{nl}	Number of lanes (Lanes)	−0.0529	0.0263	−2.0100	0.0445
b_{prec}	Percentage of precipitation (PPrep)	0.3090	0.1165	2.6530	0.0080
b_{nc}	Number of curvatures (NCurv)	0.0852	0.2768	0.3080	0.7583
b_{lc}	Total length of curvatures (LCurv)	—	—	—	—
b_{ssd}	Standard deviation of daily average speed (SDDailySpd)	0.3011	0.0638	4.7210	<0.0001
b_{avs}	Daily average speed (SpdAvgDaily)	−0.0328	0.0130	−2.5310	0.0114

Note: Null deviance: 7697.2; Residual deviance: 7362.8; Dispersion parameter: 1.7943; Number of Fisher scoring iterations: 8.

Table 56. Calibrated Coefficients for KABCO Crashes on Interstate Roadways—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−11.8000	1.2900	−9.17	<.0001
b_{aadt}	Traffic volume (AADT)	0.6470	0.1240	5.228	<.0001
b_l	Segment length (Len)	0.1930	0.0284	6.815	<.0001
b_{nl}	Number of lanes (Lanes)	−0.1030	0.1850	−0.555	0.579
b_{sw}	Lane width (LW)	−0.0001	0.0152	−0.006	0.9954
b_{prec}	Percentage of precipitation (PPrep)	0.2080	0.1260	1.654	0.0981
b_{nc}	Number of curvatures (NCurv)	0.0137	0.0134	1.023	0.3061
b_{lc}	Total length of curvatures (LCurv)	−0.0882	0.1080	−0.82	0.4123
b_{ssd}	Standard deviation of daily average speed (SDDailySpd)	0.3520	0.0766	4.591	<.0001
b_{avs}	Daily average speed (SpdAvgDaily)	−0.0078	0.0145	−0.535	0.5929

Note: Null deviance: 22,220; Residual deviance: 20,902; Dispersion parameter: 1.842634; Number of Fisher scoring iterations: 7.

Table 57. Calibrated Coefficients for KABC Crashes on Interstate Roadways—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−11.4046	2.3923	−4.7670	<.0001
b_{aadt}	Traffic volume (AADT)	0.4614	0.2311	1.9970	0.0459
b_l	Segment length (Len)	0.2550	0.0506	5.0390	<.0001
b_{sw}	Lane width (LW)	−0.3186	0.3374	−0.9440	0.3451
b_{nl}	Number of lanes (Lanes)	0.0192	0.0273	0.7020	0.4825
b_{prec}	Percentage of precipitation (PPrep)	0.1960	0.2482	0.7900	0.4297
b_{nc}	Number of curvatures (NCurv)	0.0139	0.0256	0.5440	0.5867
b_{lc}	Total length of curvatures (LCurv)	−0.2262	0.2064	−1.0960	0.2732
b_{ssd}	Standard deviation of daily average speed (SDDailySpd)	0.3455	0.1422	2.4300	0.0151
b_{avs}	Daily average speed (SpdAvgDaily)	−0.0080	0.0266	−0.3010	0.7638

Note: Null deviance: 2,583.3; Residual deviance: 2,513.5; Dispersion parameter: 1.889941; Number of Fisher scoring iterations: 9.

Table 58. Calibrated Coefficients for KABCO Crashes on Two Lanes—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	0.5638	0.0562	10.0240	<0.0001
b_{aadt}	Traffic volume (AADT)	0.2610	0.0122	21.3290	<0.0001
b_l	Segment length (Len)	0.0092	0.0087	1.0560	<0.0001
b_{sw}	Lane width (LW)	0.0162	0.0814	0.1990	0.2910
b_{prec}	Percentage of precipitation (PPrep)	0.0156	0.0127	1.2320	0.8425
b_{nc}	Number of curvatures (NCurv)	−0.0417	0.1180	−0.3540	0.2180
b_{lc}	Total length of curvatures (LCurv)	0.0380	0.0073	5.2380	0.7235
b_{ssd}	Standard deviation of daily average speed (SDDailySpd)	0.0062	0.0034	−1.8150	<0.0001
b_{avs}	Daily average speed (SpdAvgDaily)	0.5638	0.0562	10.0240	<0.0695

Note: Null deviance: 22,399; Residual deviance: 21,404; Dispersion parameter: 1.7398; Number of Fisher scoring iterations: 7.

Table 59. Calibrated Coefficients for KABC Crashes on Two Lanes—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−14.2000	1.1780	−12.0570	<0.0001
b_{aadt}	Traffic volume (AADT)	0.7720	0.1105	6.9830	<0.0001
b_l	Segment length (Len)	0.2320	0.0241	9.6300	<0.0001
b_{sw}	Lane width (LW)	−0.0001	0.0175	−0.0040	0.9970
b_{prec}	Percentage of precipitation (PPrep)	0.0279	0.1556	0.1800	0.8580
b_{nc}	Number of curvatures (NCurv)	0.0229	0.0238	0.9600	0.3370
b_{lc}	Total length of curvatures (LCurv)	0.0916	0.2213	0.4140	0.6790
b_{ssd}	Standard deviation of daily average speed (SDDailySpd)	0.0954	0.0136	7.0390	<0.0001
b_{avs}	Daily average speed (SpdAvgDaily)	−0.0027	0.0066	−0.4140	0.6790

Note: Null deviance: 8,330.9; Residual deviance: 7,905.7; Dispersion parameter: 1.9939; Number of Fisher scoring iterations: 9.

Table 60. Calibrated Coefficients for KABCO Crashes on Two Lanes—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−11.8408	0.5244	−22.5780	<.0001
b_{aadt}	Traffic volume (AADT)	0.7531	0.0414	18.1980	<.0001
b_l	Segment length (Len)	0.1609	0.0121	13.2730	<.0001
b_{sw}	Lane width (LW)	−0.0203	0.0082	−2.4830	0.0130
b_{prec}	Percentage of precipitation (PPrep)	0.1688	0.0688	2.4550	0.0141
b_{nc}	Number of curvatures (NCurv)	−0.0048	0.0029	−1.6470	0.0995
b_{lc}	Total length of curvatures (LCurv)	0.1748	0.0354	4.9340	<.0001
b_{ssd}	Standard deviation of daily average speed (SDDailySpd)	0.0490	0.0076	6.4330	<.0001
b_{avs}	Daily average speed (SpdAvgDaily)	−0.0036	0.0037	−0.9600	0.3368

Note: Null deviance: 5,289.4; Residual deviance: 4,828.1; Dispersion parameter: 1.787657; Number of Fisher scoring iterations: 7.

Table 61. Calibrated Coefficients for KABC Crashes on Two Lanes—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−13.1203	0.9968	−13.1620	<.0001
b_{aadt}	Traffic volume (AADT)	0.8316	0.0739	11.2470	<.0001
b_l	Segment length (Len)	0.1448	0.0215	6.7190	<.0001
b_{sw}	Lane width (LW)	−0.0598	0.0194	−3.0830	0.0021
b_{prec}	Percentage of precipitation (PPrep)	−0.0733	0.1567	−0.4680	0.6397
b_{nc}	Number of curvatures (NCurv)	0.0027	0.0053	0.5190	0.6036
b_{lc}	Total length of curvatures (LCurv)	0.0952	0.0667	1.4280	0.1533
b_{ssd}	Standard deviation of daily average speed (SDDailySpd)	0.0770	0.0135	5.7010	<.0001
b_{avs}	Daily average speed (SpdAvgDaily)	0.0020	0.0067	0.2900	0.7718

Note: Null deviance: 9,421.3; Residual deviance: 8,850.2; Dispersion parameter: 2.086383; Number of Fisher scoring iterations: 8.

Table 62. Calibrated Coefficients for KABCO Crashes on Multilane Roadways—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−13.1536	1.0677	−12.3190	<0.0001
b_{aadt}	Traffic volume (AADT)	0.7413	0.0534	13.8740	<0.0001
b_l	Segment length (Len)	0.2511	0.0121	20.8120	<0.0001
b_{nl}	Number of lanes (Lanes)	0.2058	0.2458	0.8370	0.4024
b_{sw}	Lane width (LW)	−0.0125	0.0059	−2.1230	0.0338
b_{prec}	Percentage of precipitation (PPrep)	−0.0956	0.0864	−1.1060	0.2686
b_{nc}	Number of curvatures (NCurv)	0.0948	0.0321	2.9580	0.0031
b_{lc}	Total length of curvatures (LCurv)	−0.2487	0.1699	−1.4640	0.1431
b_{ssd}	Standard deviation of daily average speed (SDDailySpd)	0.0806	0.0083	9.6570	<0.0001
b_{avs}	Daily average speed (SpdAvgDaily)	0.0024	0.0028	0.8640	0.3874

Note: Null deviance: 22,797; Residual deviance: 21,567; Dispersion parameter: 1.7192; Number of Fisher scoring iterations: 7.

Table 63. Calibrated Coefficients for KABC Crashes on Multilane Roadways—Ohio Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−16.1712	2.2335	−7.2400	<0.0001
b_{aadt}	Traffic volume (AADT)	0.7784	0.1073	7.2570	<0.0001
b_l	Segment length (Len)	0.2259	0.0251	8.9840	<0.0001
b_{sw}	Lane width (LW)	0.4187	0.5141	0.8140	0.4154
b_{nl}	Number of lanes (Lanes)	−0.0143	0.0115	−1.2470	0.2125
b_{prec}	Percentage of precipitation (PPrep)	−0.1251	0.1778	−0.7040	0.4817
b_{nc}	Number of curvatures (NCurv)	0.1038	0.0626	1.6590	0.0972
b_{lc}	Total length of curvatures (LCurv)	−0.2054	0.3349	−0.6130	0.5396
b_{ssd}	Standard deviation of daily average speed (SDDailySpd)	0.1414	0.0160	8.8560	<0.0001
b_{avs}	Daily average speed (SpdAvgDaily)	0.0054	0.0054	1.0000	0.3174

Note: Null deviance: 7953.7; Residual deviance: 7474.9; Dispersion parameter: 2.0084; Number of Fisher scoring iterations: 8.

Table 64. Calibrated Coefficients for KABCO Crashes on Multilane Roadways—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−13.8323	1.2798	−10.8080	<.0001
b_{aadt}	Traffic volume (AADT)	0.9219	0.1066	8.6500	<.0001
b_l	Segment length (Len)	0.1939	0.0253	7.6740	<.0001
b_{nl}	Number of lanes (Lanes)	−0.4006	0.2329	−1.7200	0.4854
b_{sw}	Lane width (LW)	0.0260	0.0087	2.9910	0.0028
b_{prec}	Percentage of precipitation (PPrep)	0.1190	0.1133	1.0500	0.2937
b_{nc}	Number of curvatures (NCurv)	0.0125	0.0077	1.6290	0.1034
b_{lc}	Total length of curvatures (LCurv)	0.1804	0.0828	2.1770	0.0295
b_{ssd}	Standard deviation of daily average speed (SDDailySpd)	0.0217	0.0204	1.0650	0.2869
b_{avs}	Daily average speed (SpdAvgDaily)	−0.0103	0.0072	−1.4300	0.1529

Note: Null deviance: 6550.6; Residual deviance: 6378.5; Dispersion parameter: 1.656436; Number of Fisher scoring iterations: 8.

Table 65. Calibrated Coefficients for KABC Crashes on Multilane Roadways—Washington Only.

Coefficient	Variable	Value	Std. Dev	t-statistic	p-value
b_0	Intercept	−15.6685	2.3476	−6.6740	<.0001
b_{aadt}	Traffic volume (AADT)	0.8258	0.1900	4.3470	<.0001
b_l	Segment length (Len)	0.2097	0.0435	4.8220	<.0001
b_{sw}	Lane width (LW)	−0.4010	0.4132	−0.9710	0.3317
b_{nl}	Number of lanes (Lanes)	0.0268	0.0157	1.7050	0.0882
b_{prec}	Percentage of precipitation (PPrep)	0.2143	0.1836	1.1670	0.2432
b_{nc}	Number of curvatures (NCurv)	0.0131	0.0140	0.9370	0.3490
b_{lc}	Total length of curvatures (LCurv)	0.0872	0.1508	0.5780	0.5630
b_{ssd}	Standard deviation of daily average speed (SDDailySpd)	0.0694	0.0371	1.8700	0.0614
b_{avs}	Daily average speed (SpdAvgDaily)	0.0127	0.0143	0.8860	0.3754

Note: Null deviance: 2222.6; Residual deviance: 2066.2; Dispersion parameter: 1.994874; Number of Fisher scoring iterations: 9.

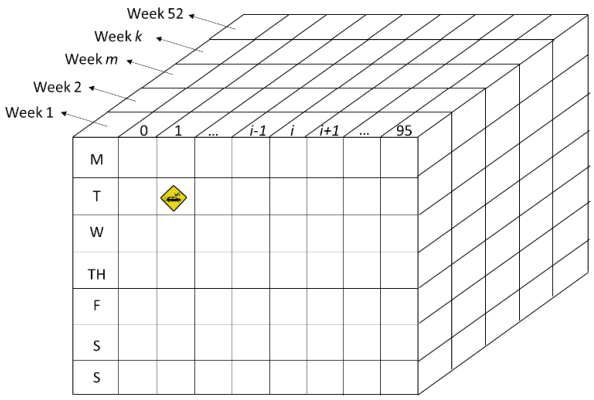
APPENDIX E. DATA PREPARATION FOR DATA STRUCTURE 3

The project team prepared a retrospective time series dataset for analysis from the Washington and Ohio conflated databases. The dataset consisted of sets of consecutive epochs before and after the recorded crashes recorded at the TMCs in the conflated dataset. A maximum of 4 hours before and 4 hours after each crash were included in this set. The epochs in each set were labeled either as BI (before the incident), DI (during the incident of the 15-minute time bin), or AI (after the incident). Using the time stamps of the epochs labeled DI, the conflation team retrieved all available epochs with the same day of the year and same epoch number (for example, if an epoch labeled DI was a Friday at 8:15 a.m., the additional epochs in the database representing every Friday at 8:15 a.m. were retrieved). These reference epochs were labeled DR (during reference, meaning "reference for 'during' epoch"). Starting from the DR epoch, the same maximum 4 hours were retrieved before and after the reference epochs. These additional reference epochs were labeled BR (BI reference) and AR (AI reference).

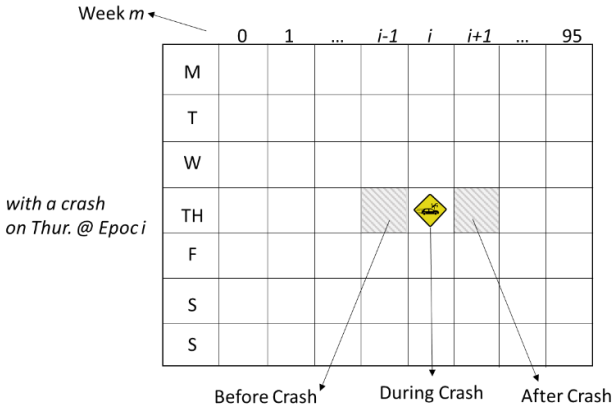
The following steps were taken in preparing the data:

- Step 0: Reformat the epoch data on a weekly basis.
- Step 1: For each crash on a TMC, check if the crash is a single case (i.e., no other crashes occurred 4 hours before or 4 hours after the current epoch). If yes, select this epoch as a valid case, and assign a unique case ID. Flag the current epoch as during crash; flag the epochs within 4 hours before the current epoch as before crash; flag the epochs within 4 hours after the current epoch as after crash.
- Step 2: On the same TMC, screen the same epoch on the same day of the week over all the other 51 weeks. Check if each potential reference epoch is independent (there is no crash within 4 hours before or after the epoch). If yes, select the epochs as references. Particularly, flag the one that has exactly the same epoch time as the crash as during reference; flag those 4 hours prior to it as before reference; and flag those 4 hours after it as after reference. Assign the unique case ID of the corresponding crash case to all these reference epochs.
- Step 3: For each reference epoch, calculate the week difference between it and the corresponding crash case. Negative indicates the reference epoch is prior to the crash, and positive indicates the reference epoch is after the crash.

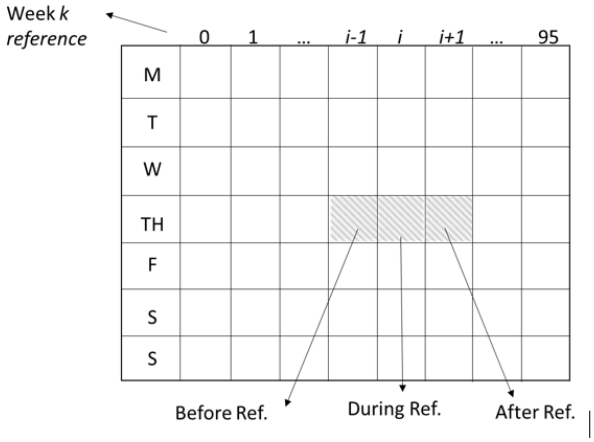
The process is illustrated in figure 28. Figure 28(a) shows the epoch data structure on a weekly basis over the whole year. In figure 28(b), there is a single crash case at epoch *i* on Thursday during week *m*. Figure 28(c) shows that reference epochs are selected on Thursday of week *k*.



(a) Epoch Data by Week over the Whole Year



(b) Epoch Data with a Crash



(c) Reference Epoch without Crashes

Figure 28. Illustration of Reference Epoch Selection.

Using this dataset, the project team intended to develop models comparing the speed trends before and after each recorded crash to the trends of the reference sets of events. The initial hypotheses for these analyses were:

1. By selecting the reference epochs to be the same day and time, a comparison of operations before each crash would allow researchers to draw conclusions about the operational impacts after each crash (after controlling for other factors).
2. If the operations are not influential in crash risk, the comparison of operational trends before the crashes should not yield statistically significant differences (ceteris paribus) since crash occurrence should be statistically independent of operational conditions in this case.
3. Alternatively, if operational differences are found for comparing epochs before crashes to the reference epochs, those differences may indicate operational conditions associated with changes in crash risk.

To explore the research hypotheses above initially, the project team fitted a preliminary set of plots to the freeway data in the Washington conflated database. Nearly 3.2 million records were represented in this dataset, with a total of 1,906 freeway crashes and 94,468 reference events.

The project team prepared a set of plots to visualize the operational differences between the incident and reference sets of epochs. First, a set of 500 incidents was selected at random from TMCs that come from freeways with four lanes (two in each direction). This was done because four-lane freeways are the most common cross-section and to avoid confounding of the trends with operational differences by the number of lanes. Then, a subset of five randomly selected references for each of those incidents was selected such that the maximum difference in time between the incident epoch and the reference epochs did not exceed 4 weeks (to avoid confounding with seasonal changes in the traffic).

Figure 29 shows the trends for before, during, and after epochs across AADT values for all reference epochs. The trends are very close; they track each other and tend to intertwine, which strongly suggests that no significant operational differences exist between the three subsets of epochs. This figure mildly suggests that the DR epochs tend to represent higher speeds than the before and after ones at higher AADTs.

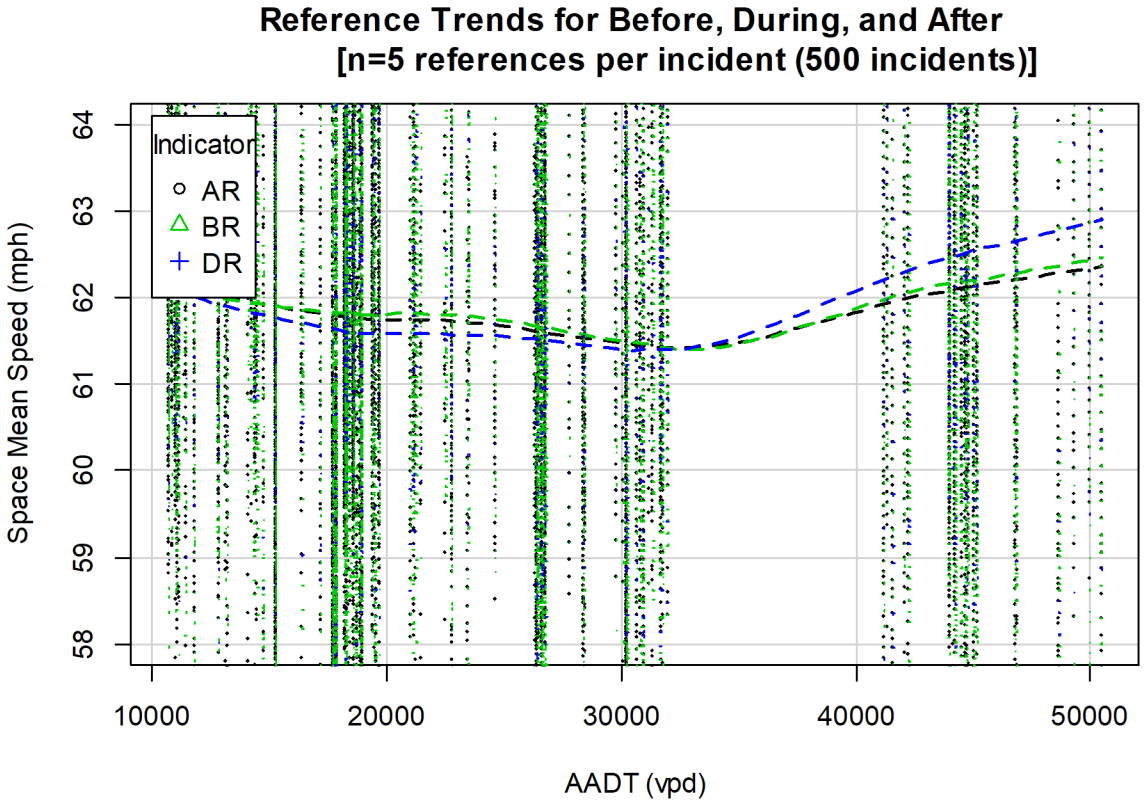


Figure 29. Median Operational Speeds Before, During, and After (Reference Epochs).

However, the trends shown in figure 30 for the epochs around actual crashes are clearly very different from figure 29. First, the three trends are more spread out, strongly suggesting differences in operations attributable to crashes occurring. Second, the operational impact in the 15 minutes around a crash is clearly a reduction of speeds, compared to the before-crash trend. The operational impact of the crashes appears to be higher at freeways with larger AADTs (i.e., a widening gap between the green and blue lines with increasing AADTs). Third, the trend for epochs before crashes tends to stay above the trend for epochs after crashes across the whole range of AADTs, which suggests partial recovery on operations after a crash occurred.

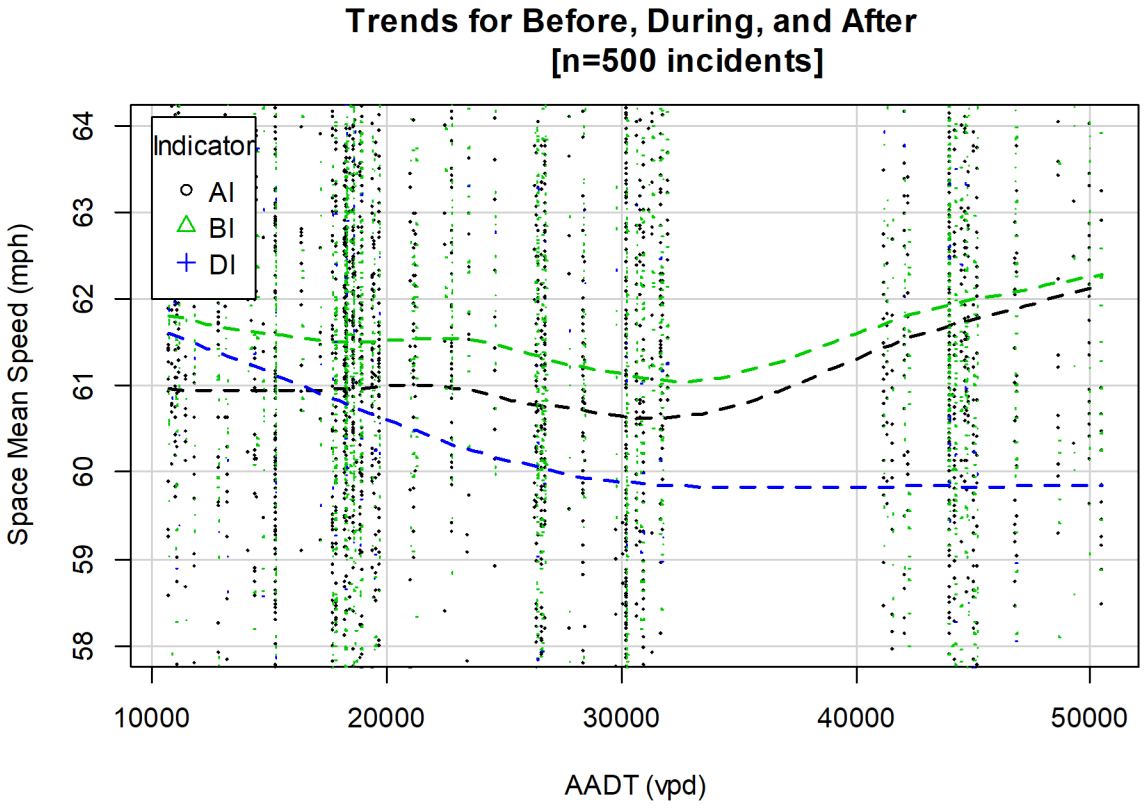


Figure 30. Median Operational Speeds Before, During, and After (Incident Epochs).

Regarding the first hypothesis, it appears that the freeway dataset from Washington may offer preliminary evidence (per figure 30) that:

- The operational impact of crash occurrences may be larger at locations with larger AADTs.
- There may be a hysteresis effect on operations after a crash occurs (i.e., operations may not fully recover to the operational state from before the crash occurred). This may not be the case for all 4 hours after the incident. The black trend being below the green one in figure 30 may be due to a few epochs immediately after the crash but with the potentially full operational recovery still possible if enough time elapses after the crash (the point of recovery can be modeled in this dataset).

To explore the potential of this dataset to test the second and third hypotheses, researchers prepared the following plots comparing pairs of corresponding sets of epochs between the incident and reference subsets. Figures 31–33 have been supplemented with 95 percent confidence envelopes around the trend lines to easily identify significant differences between the trends.

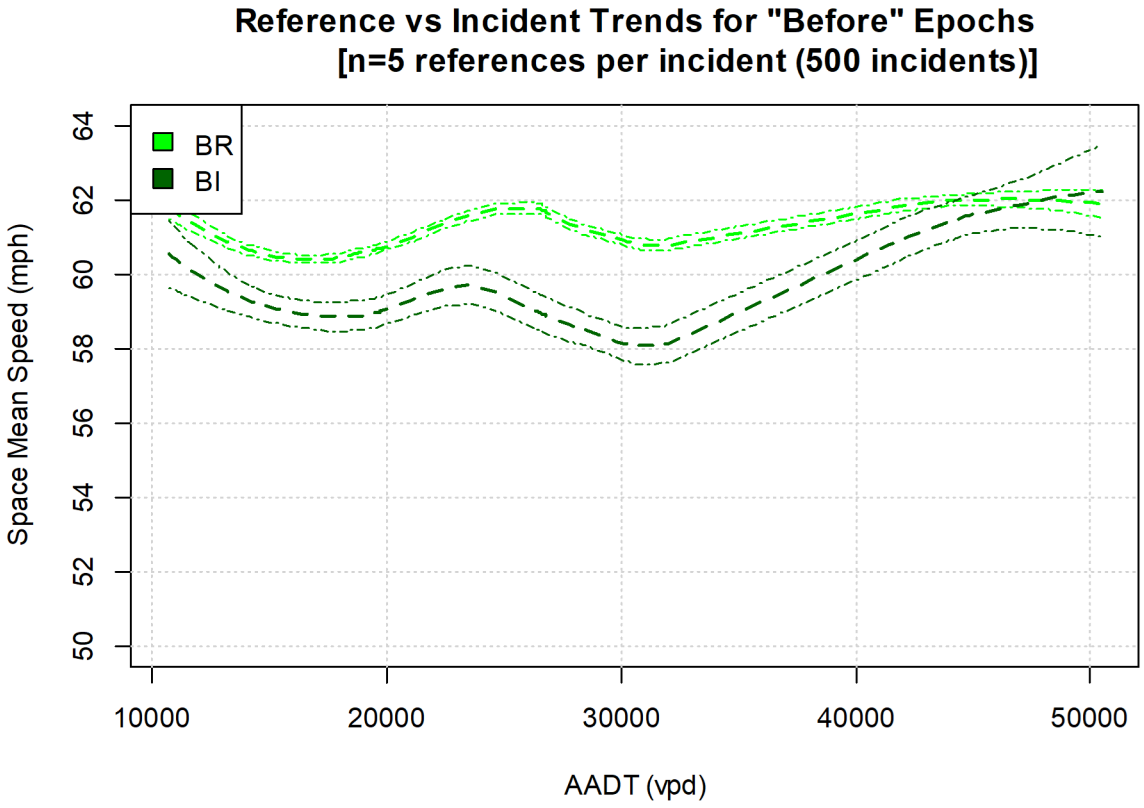


Figure 31. Comparison of Incident and Reference Trends for "Before" Median Operational Speeds.

Figure 31 shows that the before-crash periods tend to have lower speeds compared to their reference counterparts. The gap between the two trends appears to be between 2 and 3 mph maximum (at TMCs with AADTs around 30,000 vpd). However, the trends overlap at TMCs with AADTs larger than 40,000. This "tracking from below" feature of before-crash epochs suggests that congestion may be associated with an increased risk of crashes (congestion probably explains a reduction in operating speed since seasonal and daily fluctuations should be controlled for by having the same times and days of the week within a month of the crashes).

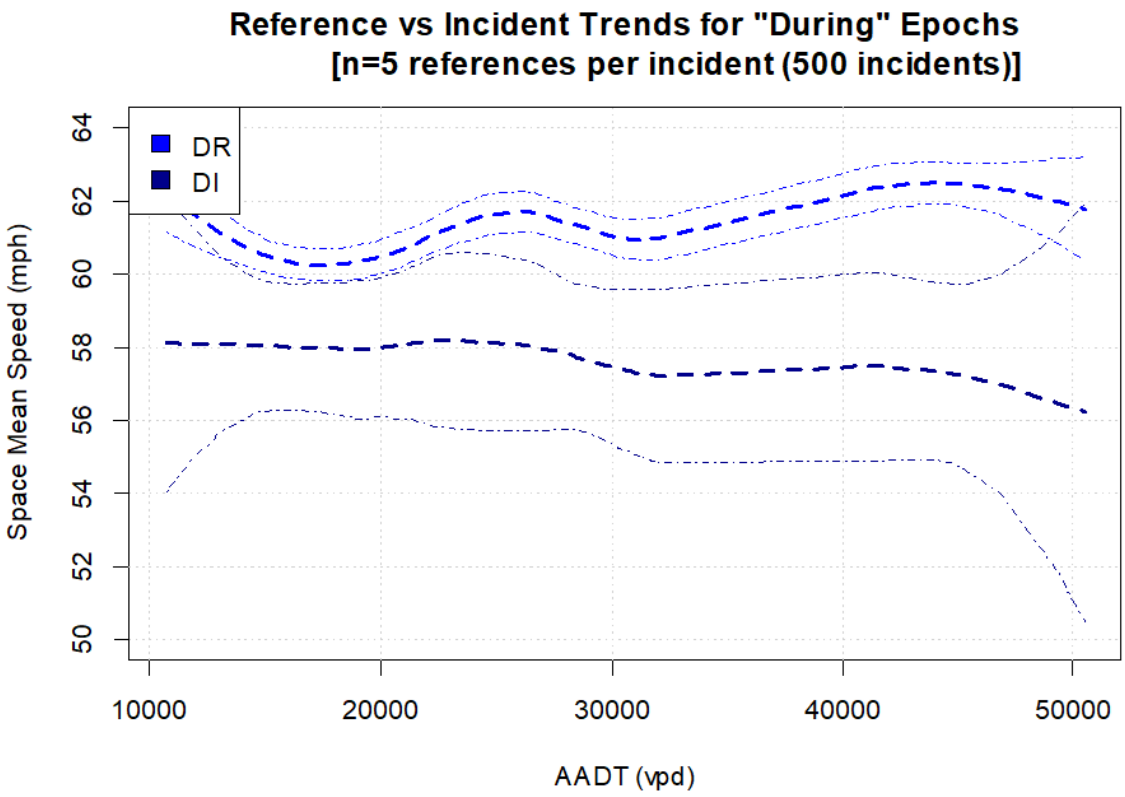


Figure 32. Comparison of Incident and Reference Trends for "During" Median Operational Speeds.

Figure 32 shows that the 95 percent confidence envelopes for the median speed trends for the "during" epochs almost overlap at low AADTs. The difference is clearer at higher AADTs. Like the "before" periods, the epochs when crashes occur tend to have consistently lower operating speeds than their comparable reference epochs. The gap seems to be about 4 mph, slightly wider than that shown in figure 31.

Figure 33 compares "after" trends. Compared to the "before" period plot, this plot suggests that operations may not fully recover to reference levels in the after period.

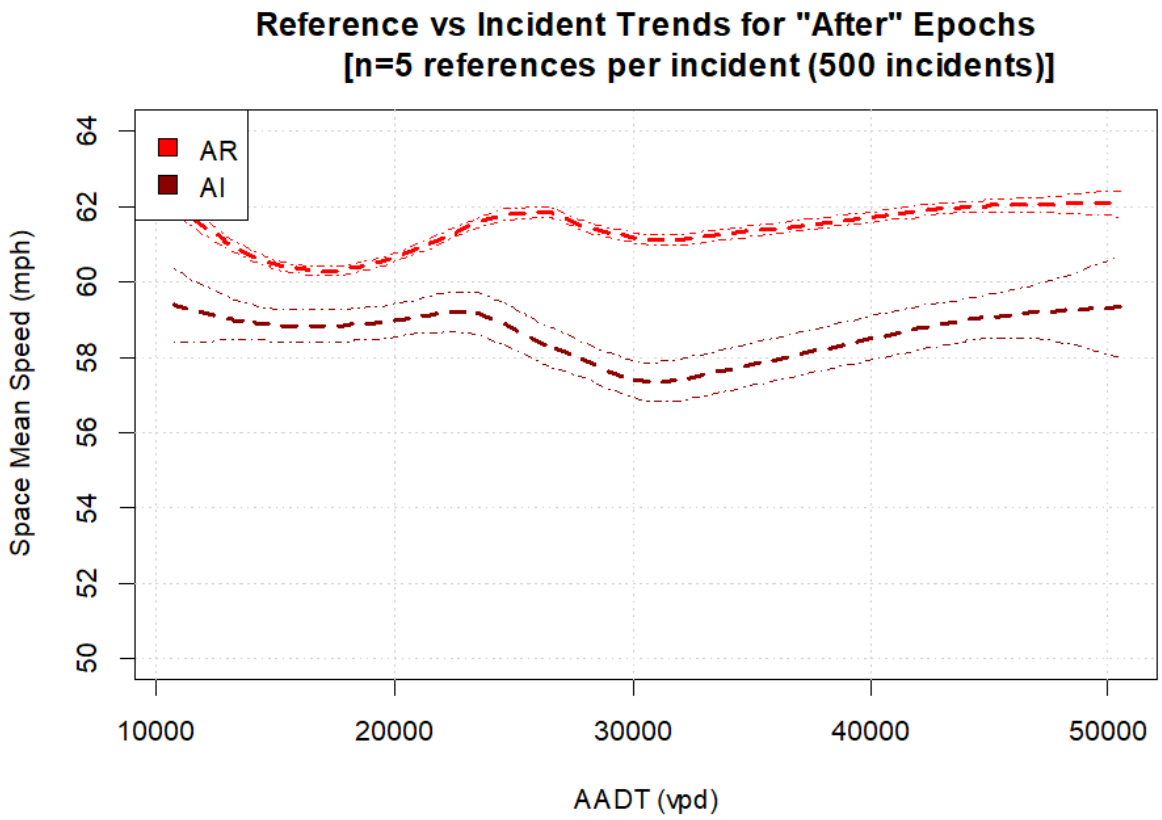
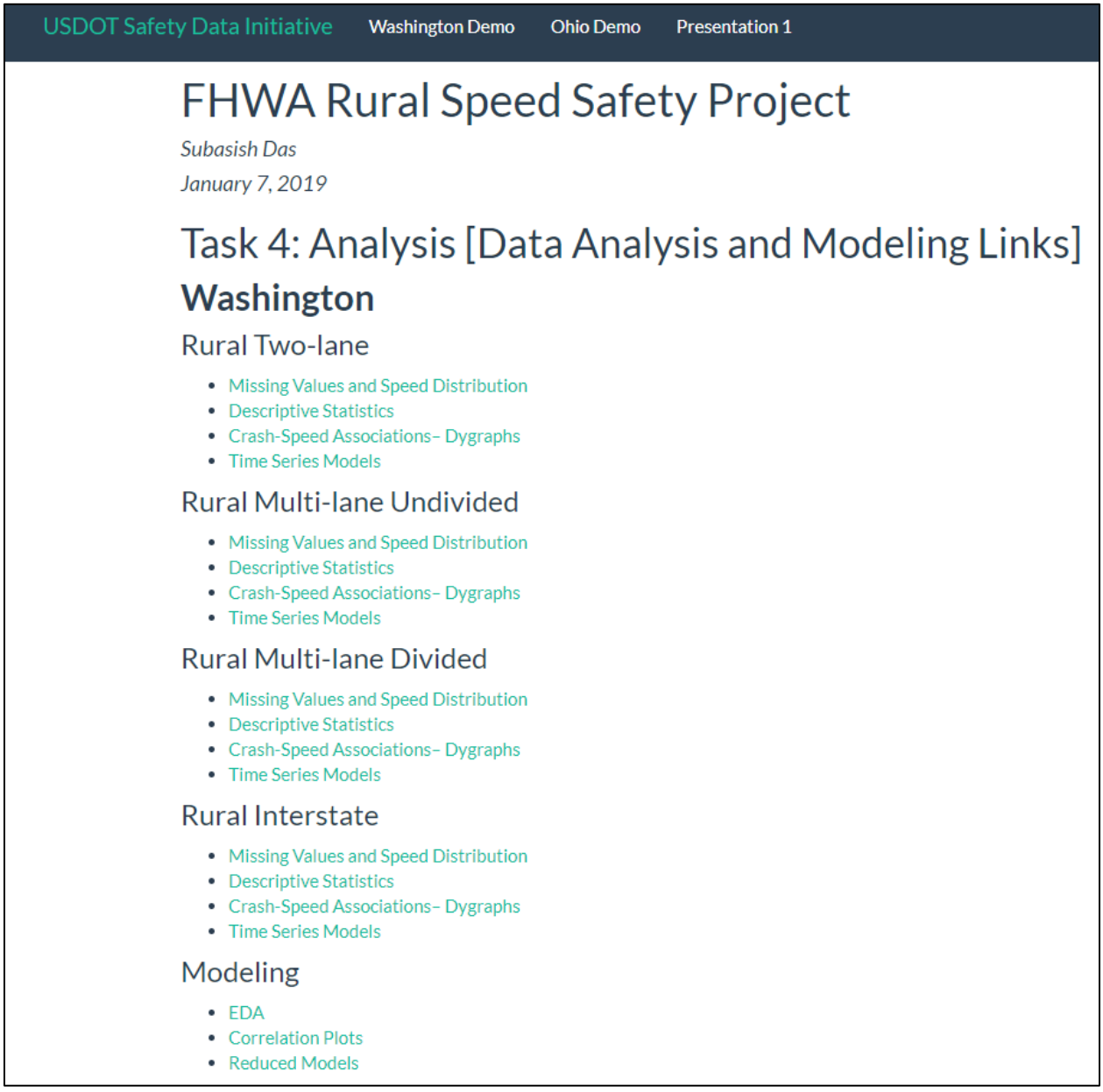


Figure 33. Comparison of Incident and Reference Trends for "After" Median Operational Speeds.

The trends of all three subsets of epochs tend to consistently have lower operational speeds for crashes compared to references. This finding is probably indicative of congestion being more conducive to crash occurrence (especially because this observation is valid for the "before" sets of epochs). The paired comparisons for the "after" sets of epochs suggest that operations may not fully recover to normalcy after crashes, but this observation may also be true for a subset of epochs immediately after the crash occurs, though not for the whole 4-hour period in the analysis dataset.

APPENDIX F. INTERACTIVE DATA VISUALIZATION



Source: http://bit.ly/rss_sdi_dv