The Development of Crash Modification Factors: Highway Safety Statistical Paper Synthesis

PUBLICATION NO. FHWA-HRT-20-069

NOVEMBER 2020

U.S. Department of Transportation
Federal Highway Administration

Research, Development, and Technology
Turner-Fairbank Highway Research Center
6300 Georgetown Pike
McLean, VA 22101-2296

FOREWORD

The research documented in this report was conducted as part of the Federal Highway Administration's (FHWA's) Evaluation of Low-Cost Safety Improvements Pooled Fund Study (ELCSI-PFS). FHWA established this PFS in 2005 to research the effectiveness of the safety improvements identified by the National Cooperative Highway Research Program's Report 500 Series as part of the implementation of the American Association of State Highway and Transportation Officials' *Strategic Highway Safety Plan*.[1] The ELCSI-PFS research studies provide a crash modification factor and benefit–cost economic analysis for each targeted safety strategy identified as a priority by the PFS-member States.

This report identifies opportunities to better understand the relationships between road safety and factors that affect traffic-crash occurrence and severity. In this report, current statistical-analysis methods and data sources used in road-safety research are compared with alternative methods and data sources. Causal-inference methods are compared to observational before–after methods to develop safety-effect estimates for centerline and edgeline rumble strips. Regression trees and Random Forests™ are compared to count regression methods to predict crash frequencies on freeways. Road-safety performance estimates using the Crash Outcomes Data Evaluation System are also discussed with a focus on opportunities to link hospital and crash data to understand the relationship between crashes' contributing factors. The transportation-engineering community is transforming by integrating quantitative methods into the task-development process, and this report will benefit this community by providing insight into more effective analytical tools.

Brian P. Cronin, P.E.
Director, Office of Safety and Operations
Research and Development

TECHNICAL DOCUMENTATION PAGE

1. Report No. FHWA-HRT-20-069	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle The Development of Crash Modification Factors: Highway Safety Statistical Paper Synthesis		5. Report Date November 2020
		6. Performing Organization Code
7. Author(s) Eric Donnell (ORCID: 0000-0002-5315-0614), Ephraim Hanks, Richard J. Porter (ORCID: 0000-0001-8535-3451), Lawrence Cook (ORCID: 0000-0001-9085-0428), Raghavan Srinivasan (ORCID: 0000-0002-3097-5154), Fan Li (ORCID: 0000-0002-0390-3673), Maggie Nguyen, Kimberly Eccles (ORCID: 0000-0001-7522-9609)		8. Performing Organization Report No.
9. Performing Organization Name and Address VHB 8300 Boone Boulevard, Suite 700 Vienna, VA 22182 Penn State University 212 Sackett Building University Park, PA 16802		10. Work Unit No. 11. Contract or Grant No. DTFH61-13-D-00001
12. Sponsoring Agency Name and Address Office of Safety Research and Development Federal Highway Administration 6300 Georgetown Pike McLean, VA 22101		13. Type of Report and Period Covered Final Report; September 2015–November 2017 14. Sponsoring Agency Code HRDS-20

15. Supplementary Notes
The Federal Highway Administration (FHWA) Office of Safety Research and Development managed this study. The FHWA Office of Safety Research and Development Task Order Manager was Roya Amjadi (HRDS-20; ORCID 0000-0001-7672-8485).

16. Abstract
The transportation-engineering community is transforming by integrating quantitative methods into the task-development process. This report identifies opportunities to better understand the relationships between road safety and factors that affect traffic-crash occurrence and severity. In this report, current statistical-analysis methods and data sources used in road-safety research are compared with alternative methods and data sources. Causal-inference methods are compared to observational before–after methods to develop safety-effect estimates of centerline and edgeline rumble strips. Regression trees and Random Forests™ are compared to count regression methods to predict crash frequencies on freeways. Road-safety performance estimates using the Crash Outcomes Data Evaluation System are also discussed, with a focus on opportunities to link hospital and crash data to understand the relationship between crashes and site-specific contributing factors. Methods to account for underreporting in crash-frequency models are also described.

17. Key Words	18. Distribution Statement
Statistical analysis, propensity score, probabilistic link, regression trees, CODES, causal inference	No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161. http://www.ntis.gov

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	126	N/A

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 1,000 L shall be shown in m^3			
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2,000 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 (2,000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	2.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for 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 ACRONYMS AND ABBREVIATIONS

AADT	average annual daily traffic
AASHTO	American Association of State Highway and Transportation Officials
AIS	Abbreviated Injury Scale
BRT	boosted regression tree
CART	classification and regression trees
CLRS	centerline rumble strips
CMC	cross-median crash
CMF	crash modification factor
CMFunction	crash modification function
CODES	Crash Outcomes Data Evaluation Systems
CP	complexity parameter
DCMF	Development of Crash Modification Factors
DID	difference-in-difference
EB	empirical Bayes
ED	emergency department
EMS	emergency medical services
F	free-flow phase
FARS	Fatality Analysis Reporting System
FB	full-Bayesian
FHWA	Federal Highway Administration
GIS	geographic information system
GUIDE	Generalized, Unbiased, Interaction Detection and Estimation
HSM	*Highway Safety Manual*
ICD-9-CM	International Classification of Diseases 9th Revision Clinical Modification
IQR	interquartile range
IRB	Institutional Review Board
J	wide-moving-jams phase
KABCO	injury severity scale, where K is fatal injury, A is incapacitating injury, B is nonincapacitating injury, C is possible injury, and O is no injury
MAIS	Maximum Abbreviated Injury Scale
MASS	Modern Applied Statistics with S
MCMC	Markov Chain Monte Carlo
MLE	maximum-likelihood estimator
MSE	mean-squared error
MSPE	mean-squared prediction error
MV-PDO	multiple-vehicle, property damage only
MVC	motor-vehicle crash
MVPLN	multivariate Poisson lognormal
NASS-CDS	National Automotive Sampling System Crashworthiness Data System
NB	negative binomial
NCHRP	National Cooperative Highway Research Program
NDS	Naturalistic Driving Study
NHTSA	National Highway Traffic Safety Administration
NYSDOT	New York State Department of Transportation

PDO	property damage only
PennDOT	Pennsylvania Department of Transportation
PS	propensity score
RCM	Rubin Causal Model
RID	Roadway Information Database
RMSE	root-mean-square error
S	synchronized-flow phase
SHRP2	second Strategic Highway Research Program
SPF	safety-performance function
SRS	shoulder rumble strips
TRB	Transportation Research Board
UDOT	Utah Department of Transportation
VMT	vehicle miles traveled
WSDOT	Washington State Department of Transportation

CHAPTER 1. INTRODUCTION

BACKGROUND CONTEXT FOR THIS TASK

The Federal Highway Administration (FHWA) established the Development of Crash Modification Factors (DCMF) program in 2012 to address highway safety–research needs for evaluating new and innovative safety strategies (improvements) by developing reliable, quantitative estimates of their effectiveness in reducing crashes. A goal of the DCMF program is to advance highway-safety and related research by establishing a sound foundation for developing highway transportation–specific statistical methodologies in cooperation with the American Statistical Association and other statistician communities. Several efforts have been conducted or are underway in pursuit of that goal. Notably, a 2-d technical experts' meeting brought together researchers from road-safety, statistics, and other statistics-related fields. The product of the meeting was a white paper that identified and discussed opportunities for advancing methodologies to estimate crash modification factors (CMFs) and safety-performance functions (SPFs) based on the expert opinions of those involved in the meeting.(1)

In a more dispersed effort, researchers throughout North America are also working—although largely independently—to test new analytical and methodological ideas and theories within the context of road-safety research, supporting the advancement of CMFs and SPFs or developing alternative methodologies to perform safety evaluations. The Transportation Research Board (TRB) Annual Meeting is the primary forum in which researchers share ideas and findings with the larger transportation-research community. Held each January in Washington, DC, the meeting brings together researchers from around the world to discuss research and advancements in transportation, including highway safety. In recent years, many of the papers submitted to several key committees (i.e., ABJ80 Statistical Methods; ANB10 Transportation Safety Management; ANB25 Highway Safety Performance; and ANB20 Safety Data, Analysis and Evaluation) explored new analytical methods used in developing CMFs and SPFs or developed alternative approaches to assess the safety performance of highways and streets. In addition to TRB, several other organizations have also published papers that address this topic. Identifying and reviewing these papers and bringing their ideas together in a single critical synthesis or similar product will help identify future collaborative opportunities similar to the technical experts' meeting. Identifying such opportunities was the principal goal of this task.

TASK OBJECTIVES

The objectives of this task were the following:

- Review and critically synthesize recent papers that explored refinements to current research methods (including study design and statistical analysis), and propose new methods to assess the safety performance of highways and streets. This review included methods to predict crash frequencies and severities, assess underreporting in crash-frequency models, and consider the use of nontraditional datasets to estimate safety as well as alternative methods for estimating CMFs. The critical synthesis was intended to serve as a resource to researchers and others looking to advance the science of highway safety.

- Disseminate the findings of the task to highway-safety stakeholders. This dissemination included a workshop at the TRB Annual Meeting as well as a technical session at the Joint Statistical Meetings.

ORGANIZATION OF THIS REPORT

This report is organized into five chapters. Chapter 1 provides background information concerning the current state of the practice in road-safety research. Chapter 2 describes methods most commonly used to estimate SPFs and CMFs. Additionally, chapter 2 describes methods used to estimate crash-severity and -type probabilities. Chapter 3 describes key limitations of the existing safety analysis methods. Chapter 4 describes the results of a critical synthesis of the existing literature on road-safety research. Several state-of-the-art statistical methods were applied to existing data sources, and a summary of findings from these analyses is provided in chapter 5. Finally, appendix A through appendix D are standalone summary reports of the statistical-analysis methods undertaken in the current study.

CHAPTER 2. BACKGROUND

This chapter is organized into three parts. The first part provides background information regarding SPF estimation in road-safety research. The second part is a brief overview of methods used to estimate CMFs. These first two parts are not comprehensive, but they provide a brief overview of the most common methods used by road-safety researchers to estimate the safety performance of highways and streets in the United States. Finally, the third part is an overview of statistical methods that are commonly used to estimate crash-severity and -type distributions. It should be noted, however, that crash-severity and -type models have not yet been incorporated into the *Highway Safety Manual* (HSM) or other national guidance documents for assessing safety performance of the road network.[3] As such, this review is brief because a preferred method(s) to evaluate severity-level or crash-type distributions has not yet been established.

The general framework in which road safety–prediction and –countermeasure assessments are determined follows the methodology described in the first edition of the American Association of State Highway and Transportation Officials' (AASHTO's) HSM.[3] Figure 1 shows the functional form of the crash-prediction algorithm.

$$N_{pred} = N_{spf}(CMF_1 \times CMF_2 \times \ldots CMF_h)$$

Figure 1. Equation. Crash-prediction algorithm in the HSM.[3]

Where:

N_{pred} = predicted crash frequency for a specific year.
N_{spf} = predicted crash frequency per yr determined for a set of base conditions (i.e., SPF).
CMF_1, CMF_2, …, CMF_h = CMFs for a set of a number of nonbase conditions (h).

This same form is applied to various site types, including roadway segments, at-grade intersections, and interchanges. The SPF is estimated for a set of base conditions (e.g., 12-ft travel lanes, 6-ft shoulders, no turn lanes at intersections, and no-passing zones). The CMFs are used to modify the base conditions (e.g., 11-ft instead of 12-ft travel lanes) to match the characteristics of the sites being analyzed. Methods to predict crash severity and type are not currently included in the AASHTO HSM framework but are currently being developed in National Cooperative Highway Research Program (NCHRP) Project 17-62, *Improved Prediction Models for Crash Types and Crash Severities*.

SPFs

SPFs are statistical models that relate the expected number of crashes (dependent variable) to site-specific characteristics of a roadway segment, at-grade intersection, interchange, or other roadway type. In road-safety research, these models nearly always include traffic volume (average annual daily traffic (AADT)) but may also include roadway or roadside features, such as lane width, shoulder width, radius or degree of horizontal curves, presence of turn lanes (at intersections), or the presence and location of roadside hardware. The models may also include information about traffic-control type (e.g., signal or stop control) or other operational characteristics (e.g., posted speed limit or volume–capacity ratio).

Figure 2 is an example of an SPF for a roadway segment.

$$N_{spf} = e^{\beta_0} \times L \times AADT^{\beta_1} \times e^{(\beta_2 X_2 + \cdots + \beta_c X_c)}$$

Figure 2. Equation. SPF for a roadway segment.

Where:

L = segment length (miles).
$AADT$ = AADT volume (vehicles per d).
$\beta_0, \beta_1, \beta_2, \ldots, \beta_c$ = regression parameters to be estimated.
$X_2, \ldots, X_c$ = geometric features, traffic-control type, or other site-specific features included in the model.
c = number of covariates.

Researchers most commonly apply generalized linear modeling with a negative binomial (NB) error distribution and log link function to estimate SPFs. The field adopted NB distribution because it is appropriate for nonnegative count data (i.e., crash frequencies) and accounts for the overdispersion commonly found in reported-crash data. Mannering and Bhat reviewed many alternative count-data modeling strategies recently used to overcome limitations of the NB-regression model.[4] Examples of such alternatives include the following:

- Multivariate methods to estimate multiple dependent variables that are related to each other (e.g., crash types and frequency of different severity levels).
- Zero-inflated models that consider the large proportion of zero counts in crash data by splitting the data into two parts—one to estimate the probability of a zero state and a second to estimate the counts.
- Random-parameters models that allow the regression parameters to vary across observations (roadway segments or intersections are types of observations).
- Random-effects models, which are used to address temporal correlation (i.e., multiple observations at the same location) or spatial correlation associated with observations from adjacent sites.

CMFs

A CMF is a multiplicative factor used to estimate the number of crashes that would be expected at a location after a road-safety countermeasure is implemented. A location that receives a countermeasure is often referred to as a treatment location or site because it is treated with the countermeasure. A CMF may also be used to estimate the expected change in crash frequency when considering alternative designs for an existing or a planned roadway. Crash modification functions (CMFunctions) are formulas used to determine a CMF. A CMF greater than 1.0 indicates an expected increase in crashes, while a value less than 1.0 indicates an expected reduction in crashes.

Several methods are used in road-safety research to estimate CMFs. The most common are the empirical Bayes (EB) before–after method and cross-sectional regression models. The following

sections (EB Before–After Studies and Cross-Sectional Regression Models) describe each method in more detail.

EB Before–After Studies

The EB before–after method is commonly used in observational studies to estimate CMFs. The EB methodology accounts for regression to the mean. In this methodology, SPFs are used to accomplish the following:

- Overcome difficulties of using crash rates in normalizing for volume differences between the before and after periods. (Note: Crash rates assume a linear relationship between crashes and traffic volume, but SPFs have shown this relationship is nonlinear.)
- Account for time trends through the reference-group SPFs.
- Reduce the level of uncertainty in safety-effect estimates through the estimation of SPFs.

In the EB approach, the change in safety (*ΔSafety*) for a given crash type at a site is shown in figure 3.

$$\Delta\, Safety = \lambda - \pi$$

Figure 3. Equation. *ΔSafety* in the EB approach.

Where:

λ = expected number of crashes that would have occurred in the after period if a road-safety countermeasure had not been implemented.
π = expected number of reported crashes in the after period.

In estimating λ, the effects of regression to the mean and changes in traffic volume are explicitly accounted for using SPFs. As noted in the EB Before–After Studies section, SPFs are statistical models that relate expected crash frequencies to site-specific features, such as traffic volume, traffic-control type, and geometric elements. A group of reference sites, which are not treated with the road-safety countermeasure, are used to estimate SPFs.

In the EB procedure, the SPF is first used to estimate the expected number of annual crashes in the before period. The sum of these annual SPF estimates for the before period (B) is then combined with the reported count of crashes (Y) in the before period to obtain an estimate of the expected number of crashes before implementing a road-safety countermeasure (m). Figure 4 shows how the estimate of m is computed.

$$m = w(B) + (1 - w)(Y)$$

Figure 4. Equation. Calculation for *m*.

Where w is weight.

As shown in figure 5, w is estimated from the mean and variance of the SPF estimate.

$$w = \frac{1}{1 + kB}$$

Figure 5. Equation. *w* in the EB approach.

Where *k* is an overdispersion parameter from an NB-regression model.

A factor is then applied to *m* to account for the length of the after period and differences in traffic volumes between the before and after periods. This factor is the sum of the annual SPF predictions for the after period divided by *B*. The result, after applying this factor, is an estimate of λ. The procedure also produces an estimate of the variance of λ.

The estimate of λ is then summed over all sites that received the road-safety countermeasure (to obtain the total expected number of crashes that would have occurred in the after period if a road-safety countermeasure had not been implemented across all sites being evaluated that received that countermeasure (λ_{sum})) and compared with the total count of crashes reported during the after period in that group (π_{sum}). The variance (*Var*) of λ is also summed over all sites in the strategy group.

Figure 6 illustrates how CMFs (*CMF*) are estimated.

$$CMF = \frac{\pi_{sum} / \lambda_{sum}}{1 + \left(Var(\lambda_{sum}) / \lambda_{sum}^2 \right)}$$

Figure 6. Equation. Estimated CMF from the EB approach.

The percent change in crashes is calculated as 100(1 − *CMF*).

Cross-Sectional Regression Models

In some instances, identifying a sufficient sample of sites with the same road-safety countermeasure to perform an observational before–after study is difficult. Cross-sectional models can overcome the limitations associated with observational before–after studies. In a cross-sectional study, the relationship between the outcome (i.e., crashes) and roadway and roadside features (e.g., traffic volume, horizontal and vertical alignment, cross-section elements), including the countermeasure, is generally determined using count regression models (i.e., SPFs). These models determine the statistical association between the outcome and the countermeasure of interest. Cross-sectional models offer the benefit of including a large number of sites in the study sample and do not require a time sequence (i.e., before–after periods) to evaluate a road-safety countermeasure.

A CMF is derived from the model parameter(s) associated with the countermeasure, which is often included as a binary variable (with–without) in a cross-sectional model. Figure 7 shows how the CMF is derived from the cross-sectional SPF model.

$$CMF = e^{\beta_{countermeasure}}$$

Figure 7. Equation. Estimated CMF from the cross-sectional model.

Where $\beta_{countermeasure}$ is the estimated coefficient for the countermeasure.

Common statistical methods to estimate CMFs from a cross-sectional study are described in the section SPFs.

CRASH SEVERITY AND TYPE

The crash-prediction algorithm shown in figure 1 does not consider crash-severity or -type distributions (probabilistic models conditioned on crash occurrence); rather, severity and type are currently considered using an SPF (i.e., frequency of crash severity or type). By estimating the probability of occurrences of crashes at different severity levels or types, the crash-prediction algorithm currently used in the HSM could be modified to estimate the predicted number of crashes at different severity levels or types, as shown in figure 8 through figure 10.[3]

$$N_{fatal} = N_{spf,total} \times P_{fatal}$$

Figure 8. Equation. Number of fatal crashes estimated using HSM algorithm.[3]

Where:

N_{fatal} = predicted number of fatal crashes.
$N_{spf,total}$ = predicted number of total crashes from an SPF.
P_{fatal} = probability of fatal crash.

$$N_{injury} = N_{spf,total} \times P_{injury}$$

Figure 9. Equation. Number of injury crashes estimated using HSM algorithm.[3]

Where:

N_{injury} = predicted number of injury crashes.
P_{injury} = probability of injury crash.

$$N_{PDO} = N_{spf,total} \times P_{PDO}$$

Figure 10. Equation. Number of property damage only crashes estimated using HSM algorithm.[3]

Where:

N_{PDO} = predicted number of property damage only (PDO) crashes.
P_{PDO} = probability of PDO crash.

Figure 8 through figure 10 can be modified to include crash types, such as rear end, head on, angle, or single vehicle. The probability of occurrence of different crash-severity levels (i.e., P_{fatal}, P_{injury}, and P_{PDO}) or crash types, when estimated for entities without a countermeasure(s), can be employed along with base predictions ($N_{spf,total}$) to predict the expected number of crashes at each severity level. Similarly, crash severity or type can be estimated at locations with a countermeasure of interest using probabilistic models. The effectiveness of the countermeasure (e.g., change in geometric feature) can be determined by comparing the change

in crash severity at locations with and without the countermeasure of interest. Bonneson et al. proposed a similar approach to considering crash severity in *Safety Prediction Methodology and Analysis Tool for Freeways and Interchanges*.[5] The authors referred to this approach as a severity-distribution function. The function is multiplied by the SPF to obtain the expected number of crashes for different severity levels.

A variety of statistical methods have been used to estimate crash-severity distributions. The most basic method is a binary logit model, which has been used to compare fatal (or fatal and injury) severity levels to less severe (e.g., noninjury) outcomes. Consider the linear function to determine the severity (i) for crash (n) (S_{in}) shown in figure 11.

$$S_{in} = \beta_i X_{in} + \varepsilon_{in}$$

Figure 11. Equation. S_{in} (logit) model.

Where:

X_{in} = vector of explanatory variables used to determine i for n.
β_i = vector of estimable coefficients for i.
ε_{in} = random-error term associated with i.

ε_{in} is used to account for unobserved factors associated with i and n.

In the case of a binary logit model, only two severity levels, i equal to 1 (e.g., fatal or injury) or 0 (e.g., no injury), are possible, and the cumulative density of ε_{in} is a logistic function.[6] The probability of n experiencing i equal to 1 is shown in figure 12, and the probability of n experiencing i equal to 0 is shown in figure 13.

$$P_n(i = 1) = P(\beta_1 X_{1n} + \varepsilon_{1n} > 0) = P(\varepsilon_{1n} > -\beta_1 X_{1n}) = \frac{\exp(\beta_1 X_{1n})}{1 + \exp(\beta_1 X_{1n})}$$

Figure 12. Equation. Probability of n experiencing i equal to 1 (e.g., fatality or injury).

Where:

$P_n(i = 1)$ = probability of n experiencing $i = 1$.
P = probability.
X_{1n} = vector of explanatory variables for n.
ε_{1n} = random-error term associated with $i = 1$.

$$P_n(i = 0) = \frac{1}{1 + \exp(\beta_1 X_{1n})}$$

Figure 13. Equation. Probability of n experiencing i equal to 0 (e.g., no injury).

Where $P_n(i = 0)$ is the probability of n experiencing $i = 0$.

In the case of the multinomial logit model, which is another common severity modeling framework, more than two severity levels are possible, and ε_{in} is independently and identically

distributed per the generalized extreme value. Extending the concepts from figure 12 but with more than two severity outcomes, figure 14 shows the *P* that *n* results in *i* ($P_n(i)$).

$$P_n(i) = \frac{\exp(\beta_i X_{in})}{\sum_i \exp(\beta_i X_{in})}$$

Figure 14. Equation. Probability of *n* resulting in *i*.

The multinomial logit model may be used, for example, when the possible values for *i* are 1 (no injury), 2 (injury), and 3 (fatality). It can also be used for more detailed breakdowns of values for *i*, such as 1 (no injury), 2 (possible injury), 3 (minor injury), 4 (serious injury), and 5 (fatality).

To estimate the multinomial logit model, an outcome is selected to be a base outcome and the term $\exp(\beta_i X_{in})$ is equal to $1(\beta_i = 0)$ if the severity category is the base category.

Bonneson et al. applied the multinomial logit model in NCHRP Project 17-45, which was included in chapter 18 and chapter 19 of a 2014 HSM supplement for freeways and interchanges.[5]

The multinomial logit model is a possible method researchers could use to estimate crash-type outcomes. Other methods that have been used, primarily in crash-severity modeling, include the nested logit, ordered probit, and mixed logit models. For brevity, none of these methods are reviewed here; however, a detailed exposition of each method can be found in Washington et al.[7]

CHAPTER 3. KEY LIMITATIONS OF EXISTING METHODS

Researchers should consider several key limitations when applying the current state-of-the-practice crash-prediction algorithm in road-safety research. Several reviews describing these limitations have been published recently. Lord and Mannering and Mannering and Bhat described several key issues related to crash-frequency modeling (SPFs).[8,4] With regard to severity modeling, Savolainen et al. and Mannering and Bhat identified several key modeling issues.[9,4] This chapter discusses the crash-frequency and -severity modeling issues described in those reviews. Additionally, this chapter describes several limitations associated with the methods most commonly used to estimate CMFs. Finally, this chapter describes limitations related to common road-safety data sources.

CRASH-FREQUENCY MODELS

As noted in the Cross-Sectional Regression Models section, the NB-regression model is commonly used to estimate expected crash frequencies because this method accounts for overdispersion (variance greater than mean) found in reported-crash data. Other issues and associated problems found in crash-frequency data include the following:

- Temporal/spatial correlation—Temporal correlation results when using multiple years (or months) of crash data from the same locations to estimate models of crash frequency (repeated observations). Spatial correlation results when using data from contiguous road segments or adjacent intersections or interchanges. The error term in an SPF will then be correlated over observations, limiting the precision of the standard error of the regression parameter.

- Low sample mean or small sample size—A large proportion of zeros in crash data or a small sample of data can produce biased parameter estimates because the large-sample properties associated with maximum-likelihood estimation methods may not hold.

- Underreporting—Crash data are typically codified using a reporting threshold. For example, a threshold might be that persons involved in a crash must be injured or a vehicle must be towed from the crash location. This threshold would lead to many PDO and minor-injury crashes being unreported. Underreporting can lead to biased parameter estimates.

- Omitted-variables bias—Many SPFs reported in road-safety literature include few independent variables (e.g., traffic-volume only). Omitted-variables bias results when independent variables that are likely to be associated with expected crash frequencies are not included in the model. The result of omitted-variables bias is erroneous (biased) models. Parsimonious models that are estimated using strict statistical decision rules (e.g., p-values < 0.05) introduce omitted-variables bias.

- Endogeneity or selectivity bias—Selectivity bias results when road-safety countermeasures are installed at locations with high crash histories. When estimating an SPF with an indicator variable for the road-safety countermeasure, the independent variable will be correlated with the error term (unobserved factors affecting crash frequency), which introduces bias in the regression parameters.

- Unobserved heterogeneity—Several factors are likely associated with expected crash frequencies that cannot be collected and included in an SPF. Some of these factors may be correlated with the independent variables included in an SPF, which leads to unobserved heterogeneity. Regression parameters are biased in the presence of unobserved heterogeneity because the parameters are serving as a proxy for the unobserved factors that were not included in the model.

In addition to these data issues, crash-frequency models may also benefit from further consideration of the functional form used when estimating the models. Most models in published road-safety literature have been specified using a log-linear form; however, more flexible forms (e.g., generalized additive models) may be considered. Additionally, current crash-frequency modeling methods do not adequately account for measurement errors associated with many independent variables (e.g., traffic-volume estimates may be over- or underestimated annually based on the quality and frequency of traffic data-collection programs within transportation agencies) included in the model. Methods that account for measurement error should be further considered in road-safety research.

CRASH-SEVERITY MODELS

While crash-severity models were not included in the HSM crash-prediction algorithm (severity distributions were, instead, considered fixed), they have received considerable attention in road-safety research. As noted in the Crash Severity and Type section, binary logit/probit, multinomial logit, ordered probit, nested logit, and mixed logit models are common modeling methods. Many of the same issues that affect crash-frequency models also affect crash-severity models. These issues include underreporting, omitted-variables bias, small sample size, endogeneity bias, and spatial or temporal correlation. Other issues associated with crash-severity models include the following:

- Ordered nature of severities—Crash-severity data have a natural ordering (e.g., fatal, injury, and PDO). This ordering can introduce shared unobservable effects when estimating models using adjacent categories. Biased-parameter estimates will result when failing to account for these effects.

- Fixed parameters—Fixed-parameters models will produce severity-level probabilities that do not change across individual severity observations included in the sample. If unobserved heterogeneity is present, individual injury observations could vary in the sample. Failure to account for unobserved heterogeneity can result in biased parameter estimates.

CMFs

As noted in chapter 2, safety researchers commonly use one of two methods to estimate CMFs in road-safety research. These methods are observational before–after studies using the EB method and cross-sectional models. Observational before–after studies require an adequate sample of treatment sites to estimate statistically-reliable CMFs as well as several years of after-period data to compare the safety performance of a site with the countermeasure to the expected safety performance of the site had the countermeasure not been implemented. Further, the reference group used to estimate the SPF used in the EB method is subjective and likely to produce inconsistent expected crash frequencies based on the sample.

Cross-sectional models are subject to the same issues associated with crash-frequency models. In particular, issues related to omitted-variables bias, temporal or spatial correlation, site-selection bias, and unobserved heterogeneity are common.

In the crash-prediction algorithm shown in figure 1, the uncertainty associated with SPFs and CMFs is generally ignored in practice (Lord proposed an approach to consider the variance associated with applying SPFs and CMFs).[8] Statistical models of crash frequency include regression parameters (coefficients) with standard errors, and CMFs also include standard errors, which should be considered in road-safety evaluations to produce a range of expected outcomes.

ROAD SAFETY–RESEARCH DATA SOURCES

Models of crash frequency often involve at least three electronic data sources. These include traffic-volume, roadway-inventory (e.g., roadway and roadside features, traffic control), and crash-event data files. It is common to use the roadway-inventory file as the base unit of analysis (e.g., roadway segment, at-grade intersection, interchange) and then merge traffic-volume and crash data into the roadway-inventory file. This file is then used to estimate expected crash frequencies. Traditional data sources do not offer information about weather, driver behavior, or other factors that could affect crash frequencies.

In crash-severity models, the most severe injury outcome in a vehicle or the injury severity of a driver is often used as the analysis unit. Information about the passengers in the vehicle, the roadway and environmental conditions at the time of the crash, and information about the vehicles are often included in the model specification. These data are extracted from traditional electronic crash-data files.

Crash-data files that are used to estimate CMFs are often built using crash-frequency data. For example, when a cross-sectional model is used to estimate the safety effectiveness of a countermeasure, the model is estimated much like an SPF. In observational before–after studies, the SPF is an important step in the analysis methodology and is subject to the same modeling issues as those described for SPFs.

Other issues related to observational before–after studies include the following:

- Before completing an evaluation, researchers must wait for several years to pass so an adequate sample of after-period data at treatment sites can be compiled.
- Road-safety research has no prescribed scientific method to identify the reference group, a group of locations similar to the treatment sites but without the countermeasure. This absence makes it difficult to confirm whether the reference and treatment sites are similar except for having the road-safety countermeasure.
- The location and installation date of the countermeasure must be known to employ an observational before–after evaluation.

CHAPTER 4. CRITICAL SYNTHESIS OF SAFETY-ANALYSIS METHODS

The background information described in chapter 2 and chapter 3 of this report briefly identifies the current state of the practice in road-safety research and key limitations of current methods. Refereed journal articles related to these limitations were identified and critically reviewed. The purpose of the review was to develop a framework to advance the state of the practice in road-safety research. Key topics of in the review included the following:

- Countermeasure evaluation methods, such as causal-inference and full-Bayesian (FB) methods (CMFs).
- Underreporting models (measurement error).
- Multivariate models.
- Methods to account for selection bias or endogeneity in count regression models (SPFs).
- Regression trees or Random Forests™ for prediction.
- Crash-severity models.
- Models to mitigate spatial or temporal correlation.
- Models to address unobserved heterogeneity.
- Road-safety studies that use alternative data sources, such as Crash Outcomes Data Evaluation Systems (CODES), second Strategic Highway Research Program (SHRP2) naturalistic driving data, Fatality Analysis Reporting System (FARS) data, and National Automotive Sampling System Crashworthiness Data System (NASS-CDS) data.[10–12]

Both the Literature Review Results section and References of this report detail the papers included in the literature review. Many of the papers were published in the last 6 yr and included in the top international road-safety journals (e.g., *Accident Analysis and Prevention*, *Analytic Methods in Accident Research*, and *Transportation Research Record*).

To accomplish the paper-review objectives, the research team was divided into three pairs, each with a transportation-engineering researcher and a statistician. Each review-team pair documented their findings from the paper-review process in an assessment form. The assessment form included the following key elements: article citation, research objectives, description of how paper improved safety science, candidate application(s) in road-safety research, and rating for future use in road-safety evaluations (high, medium, low). The Literature-Review Results section summarizes the findings from the literature review.

LITERATURE-REVIEW RESULTS

The literature-review findings are organized by the themes assigned to each review-team pair. Within each section, a basic summary of the methods that were included in the articles is provided, followed by an assessment of how easily the method can be implemented in future road-safety research. A general impression of the articles reviewed within each topical area is also provided.

Spatial or Temporal Correlation

The research team reviewed papers by Barua et al. and Quistberg et al.[13–15] The papers focused on a conditional autoregressive random-parameters model to account for spatial correlation, a multivariate Poisson lognormal (MVPLN) model to account for spatial correlation, and a multilevel mixed-effects Poisson model to account for spatiotemporal correlation. The former two models were estimated using FB methods, while the latter model was estimated using maximum-likelihood methods. The Quistberg et al. paper introduced spatiotemporal data (e.g., pedestrian crosswalk locations, residential and employment density, land use, bus ridership, and sidewalk presence) to the model, which is not common in road-safety research.[14]

Among these papers, the models proposed by Barua et al. offer the best opportunity for near-term implementation in road-safety research.[13,14] In particular, the multivariate model of crash frequency for different severity levels, estimated in a FB context accounting for heterogeneity and spatial correlation, outperformed univariate models for road-segment crashes. Limitations of this model were that the data were nearly 20 yr old and the sample size was small.

The research team's general impressions of the papers included in this topical area are as follows:

- Reproducible research should be promoted in road-safety publications. As methods for road-safety studies become more complex, it becomes important that researchers make their methods clear. To foster the use of good statistical methods in road-safety research, it would be beneficial to encourage researchers to either post their data on a secure website (e.g., journal or sponsoring agency) or publish the modeling code (if not readily available in commercial software) with published articles. If authors are not encouraged or required to make their methods clear enough to be reproduced, it can be difficult for others to judge the value of the novel method proposed and reproduce the method to compare it to other existing (or new) methods.

- The availability of spatial and temporal data (e.g., weather and geographic-information-system (GIS) layers) is growing rapidly. The potential to leverage such data to improve future road-safety research is great. The creation of GIS layers with characteristics of road segments has the potential to illuminate new strategies for increasing road safety, which is an added value of the Quistberg et al. paper when estimating statistical models of vehicle–pedestrian crashes at midblock crossings and at-grade intersections.[15]

- Standard checks for possible spatial autocorrelation (e.g., Moran's I) and temporal autocorrelation (e.g., partial autocorrelation functions) are applied to the residuals of any road-safety analysis using observations collected across space and time. If spatial autocorrelation is not accounted for when it is present, the result is increased type-1 error rates for hypothesis tests on fixed effects. That is, it is more likely to incorrectly find a statistically significant correlation between a predictor variable and the dependent variable when spatial or temporal autocorrelation is not considered.

- Including spatial and temporal random effects can illuminate important factors that are missing from a study. Modeling spatial and temporal autocorrelation is often done by including a random effect in a model, with the random effect being correlated in space and/or time. One way to further consider the spatial or temporal random effect is as a missing covariate—something that has not been included in any of the models reviewed within this topical area. This missing covariate is true of any random effect, but when an estimated spatial or temporal random effect is plotted in space and/or time, it is often possible to identify important covariates that should have been included in the model but were not. That is, after fitting a model with spatial and/or temporal autocorrelation, one should examine the estimated random effect in an exploratory way. Doing so can provide insight into important variables needed for future studies.

Crash-Severity Models

The research team reviewed papers by Chen et al., Yu and Abdel-Aty, Cerwick et al., and Yasmin et al.[16–19] These papers focused on a multinomial logit-Bayesian hybrid approach to estimate driver-injury severities, a hierarchical Bayesian binary probit model to analyze crash severities, and a latent class and mixed logit model comparison and a latent segmentation-based generalized ordered logit model to examine driver-injury severities. The Chen et al. and Yu and Abdel-Aty methods considered Bayesian estimation techniques.[16,17] The two latent class-modeling papers considered maximum likelihood methods to estimate the model parameters.

Among these papers, the models proposed by Chen et al. and Yu and Abdel-Aty appear to offer the greatest potential for near-term implementation in road-safety research addressing crash severity.[16,17] A Bayesian network was used by Chen et al. to identify relationships between severity and various crash characteristics regarding drivers and vehicles.[16] Based on several comparison measures, the Bayesian network outperformed a traditional multinomial logit model. To effectively implement a Bayesian network, expert knowledge is required and the network must be well trained. In future applications, it would be helpful to use the method to develop predicted probabilities as a function of the independent variables included in the model.

In Yu and Abdel-Aty, a hierarchical Bayesian binary probit model outperformed a maximum-likelihood binary probit model.[17] Another added-value feature of this research relates to the integration of real-time traffic (i.e., average segment speed and standard deviation of speed) and weather (i.e., visibility) data from automatic-vehicle identifiers and weather stations located along a freeway. Not only did their analysis method offer an improvement over more traditional crash-severity modeling approaches, it accounted for unobserved heterogeneity while integrating unique data into the model specification. This modeling approach could be extended to further consider spatial and temporal correlation using a conditional autoregressive model.

The research team's general impressions of the papers included in this topical area are as follows:

- Model comparison is best accomplished by comparing predictive power. There are many methods for comparing models and approaches in the road-safety literature (and in scientific literature in general). In cases when data are abundant (when there are more than 100 observed crashes), it is recommended that any comparison between models be done using out-of-sample predictive error. This approach would involve the following steps: randomly selecting a subset of the data, which is referred to as a "hold-out" or "testing" set, fitting models using the remaining data, which are referred to as the "training" set, and using the fitted models to predict the dependent variables in the hold-out or testing dataset, using the independent predictor variables from the hold-out or testing set. The predictions from different models can be compared using mean-squared prediction error (MSPE), the area under the receiver operating characteristic curve, or other appropriate metrics. Predictive power is attractive as a model selection criterion because it can be applied to any model, allowing, for example, Bayesian methods to be compared with maximum likelihood–estimator (MLE) approaches. Additionally, predictive power can reveal overfitting and other potential pitfalls as road-safety models become more complex.

- In studies with many independent predictor variables, statistical regularization approaches, such as ridge regression and the least absolute shrinkage and selection operator regression, provide a stable approach to parameter estimation. These approaches can minimize the impact of collinear predictor variables and almost always provide better predictive power than standard approaches to estimation, such as MLE.

Underreporting

The research team reviewed papers by Abay and Yasmin and Eluru.[20,21] The papers focused on a bivariate ordered-response probit model to assess reporting bias in injury severity data and a comparison of several unordered and ordered response models to assess underreporting. The former model was developed using both police-reported and emergency-room data, and model performance was judged based on changes in police-reporting practices in Denmark. The latter paper randomly removed PDO-crash records from a sample of General Estimates System data and compared the elasticities between the "true" model and the underreported data to assess model performance.

Among these papers, the mixed generalized ordered logit model appears to offer the best opportunity for near-term implementation in road-safety research. In future research, however, it would be helpful to have a better estimate of the level of underreporting in crash data by considering secondary data sources, such as insurance data (PDO and minor-injury crashes are most likely to be reported as insurance claims) or municipal-level crash data (which often includes information about crashes that do not result in a reportable record in State transportation agency data files).

The research team's general impressions of the papers included in this topical area are as follows:

- Underreporting is clearly an important consideration in road-safety research because a large proportion of PDO and minor-injury crashes are not included in State transportation agency crash records. Underreporting not only biases the estimates of the relationships between measured (or observed) factors and crash severity, but it makes clearly quantifying this bias difficult, if not impossible, without external data.

- Identifying and obtaining an independent data source on individual crashes or aggregated crash rates has great potential to inform crash-severity models by allowing for a rigorous estimation of the underreporting rates. Two potential sources for such data include insurance records, which will contain information on PDO crashes, and police records of PDO crashes (that are not necessarily included in reported-crash records). Developing such a data source has potential for high impact in road-safety research.

- Once such an independent data source is available, multiple existing methods could be employed to account and correct for the bias resulting from underreporting. These methods include propensity-score (PS) matching, common in clinical-trials literature, models for inhomogeneous detection rates in ecological occupancy and species distribution models, the two-step Heckman correction common in econometrics, and FB approaches to modeling imperfect detection.

Countermeasure Evaluation Methods

The research team reviewed papers by Karwa et al., Graham et al., Wood et al., and Sacchi and Sayed.[22–25] These papers discussed a range of methods for evaluating the causal effects of safety countermeasures from observational data, including the standard EB method, PS methods, and causal diagrams. The study designs considered include cross-sectional, before–after, and panel data approaches. Specifically, Karwa et al. compared two causal-inference frameworks—potential outcomes (particularly PS methods) and causal diagrams—in reducing selection bias associated with the safety-effect estimators using cross-sectional data.[22] In the context of panel data, Graham et al. proposed a mixed-effects model for the generalized PSs to adjust for time-invariant unmeasured confounding, and used the method to evaluate the causal effects of road network capacity expansion on traffic volume and density.[23] Wood et al. compared three methods—cross-sectional PS matching, cross-sectional regression, and before–after EB—to evaluate the SafetyEdge℠ treatment.[24] Sacchi and Sayed compared the EB and FB methods in a nonlinear intervention model for before–after evaluation studies, particularly in null cases, when no countermeasure had been implemented.[25]

Among all the methods discussed, the before–after EB method is the most widely used in road-safety research. The cross-sectional PS (both matching and weighting) methods appear to have the highest potential for widespread use because they improve the balance of covariates (i.e., measured confounders) and thus mimic the randomized design better than traditional cross-sectional regression modeling. PS-matching techniques can also be used to identify appropriate reference or comparison sites that are similar to the treatment sites in before–after studies. Though the nonlinear autoregressive FB approach was shown to outperform the EB

method in no-countermeasure cases via simulations, its potential for general use is hampered by the difficulty in implementation and thus offers less potential for widespread application in road-safety countermeasure evaluations.(25) The causal-diagram (Bayesian network) approach appears to have the lowest potential to be widely adopted because it is not as intuitive and easy-to-use as the potential outcomes–PS approach and it ties the definition of causal effect to a parametric model. A data source that can potentially be used to test these methods is the simulated dataset from the FHWA-funded Artificial Realistic Data project.[1]

The research team's general impressions of the papers included in this topical area are as follows:

- Causation has no universal definition, and causal inference has no universal framework. Well-known frameworks for causal inference include the potential-outcomes framework (also known as the Rubin Causal Model (RCM) in statistics), causal diagrams, and Granger causality. Because these frameworks have entirely different definitions of causal effects and assumptions, comparisons between them are not usually meaningful. Instead, researchers should focus on the particular problem at hand and choose the most applicable framework (based on the study design or the research target). Specifically, to evaluate road-safety countermeasures, the potential-outcomes framework, particularly the PS methods, appear to be the most intuitive, applicable, and adaptable.

- A central challenge in road safety–countermeasure evaluation from observational studies is selection bias, usually appearing in the form of significant covariate–confounder imbalance between treatment countermeasure and control groups. The PS is a scalar summary of the multidimensional covariates; balancing the PS leads to balancing all the covariates. Therefore, PS matching (as well as weighting) can ensure covariate balance in observational studies, mimicking a randomized design. The methods are easy to understand and implement and have been popular across a wide range of disciplines. It is not surprising to see road-safety researchers discovering the power of the PS methods. An important caveat is that the validity of the PS methods hinges on the assumption of no unmeasured confounding, which is not always plausible. Consequently, sensitivity analysis should be routinely conducted to assess this assumption when comparing different PS methods in road-safety countermeasure evaluations.

- Though PSs are traditionally used for binary countermeasures (i.e., present or not present) and cross-sectional data, their extensions to more complex settings have been well established. In the context of road-safety studies, Graham et al. extended PSs to continuous countermeasures (dose-response functions) and panel data, where the authors proposed a mixed effects model for the PS that is capable of adjusting for time-invariant unmeasured confounding.(22) Such extensions offer an important direction for further advancing and popularizing these methods in road-safety research. Comparisons to alternative causal methods with the potential-outcomes framework, such as marginal structural models, would be of great value.

[1]This research project was performed under the HSIS contract from August 2015 to May 2017.

- The before–after EB method is the gold standard in before–after designs, but it, in fact, does not have formal causal interpretation within the potential-outcomes framework. The study design (treatment versus control at two time points) of the EB method is similar to the widely used difference-in-difference (DID) method in econometrics, which has a well-defined causal interpretation. Therefore, a potentially fruitful direction for future research is to explore the connections between EB and DID methods, which could provide a formal causal interpretation of the EB method within the potential-outcomes framework.

- Generally, the EB method or a before–after analysis with control groups is preferred over cross-sectional regression models because it is believed to better capture the temporal trend of the underlying system and therefore reduces selection bias. However, the EB method also implicitly relies on the inherently untestable assumption that the temporal trend is identical between the treated and control groups. This point is often overlooked in road-safety research. As such, similarly to PS methods, this assumption should be tested using sensitivity analyses to assess the consequences associated with assumption violations.

- Sacchi and Sayed found that the FB method with a nonlinear intervention model performs better than the EB method (narrower confidence interval and less bias) in the case of biased treatment-site selection.[24] The FB method also leads to more consistent results across different hotspot ranking methods relative to the EB method. However, these results are obtained from a single dataset with a relatively small sample size, so more empirical comparisons would be informative to examine the performance of the FB in more general settings. Additionally, the technical and computational sophistication of the FB method poses challenges for road-safety researchers who are not familiar with Bayesian analysis. The implementation issue is indeed universal with all FB methods, including the multivariate models reviewed in another section. Developing user-friendly open-source software packages and related manuals would be crucial for popularizing the FB method.

Multivariate Modeling

The research team reviewed papers by Eluru et al., Chiou and Fu, El-Basyouny et al., and Sacchi et al.[26–29] These papers developed new multivariate models for simultaneously analyzing several outcome variables from cross-sectional crash data. Specifically, El-Basyouny et al. proposed FB MVPLN models to investigate time and weather effects on counts of different crash types.[28] Sacchi et al. used the same MVPLN models to estimate crash counts of different severity levels at each site; a user of the model results could then rank the sites and, consequently, identify hotspots.[29] Eluru et al. proposed a novel copula-based multivariate model to simultaneously analyze injury severity of multiple occupants in a vehicle.[26] Chiou and Fu developed a multinomial generalized Poisson regression model to analyze crash frequency and severity jointly.[27] For statistical estimation and inference, El-Basyouny et al. and Sacchi et al. used the Bayesian approach via the Markov Chain Monte Carlo (MCMC) algorithms, whereas Eluru et al. and Chiou and Fu adopted the classical (approximate) maximum-likelihood approach. (See references 28, 29, 26, and 27.)

All the multivariate models discussed in these papers appear to have conceptual and practical advantages compared to the corresponding univariate models that are standard in road-safety literature. However, the potential for widespread use of these methods may be hampered by their technical, computational, and programming demand. Also, as these methods are still relatively new, there is a lack of empirical studies comparing the performance between the methods themselves (e.g., copula versus MVPLN) and with other methods (e.g., univariate methods). Therefore, at the current stage, their use in future road-safety research is limited. Nonetheless, if more user-friendly open-source software packages that implement these methods become available, opportunities for more widespread implementation will exist.

The research team's general impressions of the papers included in this topical area are as follows:

- Compared to the standard approach of modeling each crash outcome independently, multivariate modeling incorporates the correlation between different outcomes and, thus, is more efficient (i.e., smaller standard errors). More importantly, multivariate modeling can provide information about the relationships between different outcomes in crash data, such as injury severity of multiple occupants in a vehicle, or crash counts of different severity levels. Such information, which is not available from conventional univariate approaches, may be useful in future road-safety research.

- All the methods focus on regression models for cross-sectional data, and no causal-inference techniques were considered. Therefore, the conclusions from these methods should not be interpreted as causal. Nonetheless, the proposed models can be combined with some causal-inference methods, such as the RCM, in future research to enable causal inferences.

- El-Basyouny et al. showed statistically significant weather effects on crash types.[28] Arguably, weather conditions play a nontrivial role in traffic crashes, but weather information is not commonly used in road-safety studies. Such information could be easily obtained from weather data sources and should be used more often in road-safety studies.

- Implementation of FB multivariate models may be daunting for road-safety researchers who are not familiar with Bayesian statistics. In particular, fitting these models via the MCMC algorithm requires significant statistical and programming skills, and no standard computer package is available to estimate these models. Even the recommended WinBUGS software is semiautomated and may be difficult for researchers to use if not familiar with Bayesian statistics. Though maximum-likelihood methods may be easier to implement, they only give point estimates, not confidence intervals. An urgent task to popularize these multivariate Bayesian methods is to develop corresponding user-friendly, open-source software packages.

- As displayed in Eluru et al., the copula approach—which models a multivariate distribution from prespecified marginal distributions of each outcome variable—is flexible and powerful.[26] However, the concept and implementation of copulas involves nontrivial statistical knowledge, and indeed the copula approach is still relatively unknown in road-safety research. Additional empirical studies comparing the performance between the copula method and other multivariate models, such as the MVPLN model, are needed to further popularize the method.

Selection Bias and Endogeneity in Count Models

The research team reviewed papers by Bhat et al., Chen et al., and Lord and Kuo.[30–32] Though all these papers had a broad theme of selection bias, they described two different types of biases. The first two papers focused on bias due to endogeneity, which is also referred to as unmeasured confounding in the statistical literature, whereas the last paper focused on site-selection bias due to entry criterion. Specifically, Bhat et al. proposed a two-component model for cross-sectional count data to evaluate road-safety countermeasures with endogenous covariates.[30] The first component is a multinomial probit model for the selection mechanism, the second component is a generalized ordered-response model for the outcome, and the error terms of the two models are assumed to follow a joint covariance matrix. Chen et al. applied simultaneous equation models developed by Kim and Washington to examine endogeneity of speed limits in crash-count models for intersections, where one iteratively models each outcome using the other outcome as a predictor.[31,33] Lord and Kuo provided an analytical form of the site-selection bias for estimating the effectiveness of road-safety countermeasures as a function of entry criteria and other factors and proposed a new method to eliminate this bias when a control group was not available in before–after studies.[32]

Among all the site-selection strategies discussed in the papers reviewed by the research team, Lord and Kuo's strategy appears to have the highest potential for immediate widespread use in road-safety research because it is easy to implement, theoretically sound, and before–after analysis is widely used in road-safety research though obtaining control groups with similar characteristics to the treatment sites is usually difficult.[32] Bhat et al.'s strategy also has potential for widespread use in future research because it targets causal inference using cross-sectional data and has a well-established foundation in econometrics literature.[30] Chen et al.'s strategy appears to have the lowest potential for application due to its lack of theoretical support and empirical evidence for its validity in addressing endogeneity.[31]

The research team's general impressions of the papers included in this topical area are as follows:

- Evaluation of road-safety countermeasures from observational data are routinely skewed by various types of bias regardless of the study design (e.g., cross-sectional and before–after). It is important to identify, differentiate, and account for these biases in the analysis. For example, site-selection bias results from the common site entry criterion rendering selected sites as a nonrandom sample of all sites; this bias is distinct from bias due to regression-to-mean or endogeneity. An analytical model to account for site-selection bias as a function of the entry criterion and other factors is available, but other types of biases are generally impossible to qualify. To account for the latter, strong model assumptions are often invoked.

- The two-component model presented by Bhat et al. is a special case of the Heckman selection models in econometrics.[30] Though widely used in economics and statistics, selection models critically hinge on correctly specifying the selection mechanism that is dependent on unobserved variables and, thus, can be sensitive to model misspecification. As a remedy, sensitivity analyses should be routinely conducted in such methods.

- An important direction for future research is toward developing methods that account for multiple sources of bias simultaneously.

Regression Trees and Random Forests

The research team reviewed three papers that demonstrated applications of regression trees, boosted regression trees (BRTs), and Random Forests. Khan et al. explored the feasibility of using classification-tree methods to analyze the severity of cross-median crashes (CMCs) in Wisconsin and examine whether the use of classification-tree methods can reveal additional information about factors that influence crash severity and can improve, replace, or compliment more traditional methods of crash-severity modeling.[34] The authors tested tree models for both multiple-vehicle and single-vehicle CMCs using a default and a user-defined misclassification cost matrix (i.e., a total of four trees). The alternative cost matrices were tested to explore the effects of the misclassification–cost matrix structure on results. Khan et al.'s use of Generalized, Unbiased, Interaction Detection and Estimation (GUIDE) classification trees revealed new variables affecting CMC severity that ordinal logit and ordinal probit models do not even though they use similar CMC datasets. One of the main strengths of regression trees is the improved accuracy in prediction from typical regression models. Khan et al. did not explicitly address this advantage of regression trees, but they focused on uncovering one or two additional right-hand-side predictors of crash severity compared to the ordinal probit and logit models instead.[34]

Saha et al. used the BRT to evaluate the importance of variables identified in the HSM "predictive methods" for urban and suburban arterials and analyze the marginal effects of the variables on crash predictions.[35,3] The authors described regression trees as decision tree–based models formed by dividing a predictor space into a number of mutually exclusive regions and boosting as fitting a number of trees in a sequential process and combining predictions from a series of "weaker models" to produce "strong predictions."[35] They identified BRTs as one of

two “ensemble approaches” based on the use of regression trees, the other being based on Random Forests. The databases used for analysis consisted of 1,791 urban and suburban undivided, 2-lane, arterial segments (616.7 mi) and 4,969 urban and suburban divided, 4-lane, arterial segments (1,400.7 mi) in Florida. The BRT allowed the authors to find that variables exhibited nonlinear and sometimes complex relationships to predicted crash counts, and only a few variables were found to explain most of the variation in the crash data.

Xu et al. evaluated the safety performance associated with freeway-traffic flow in the framework of three-phase traffic theory.[36] This theory classifies traffic flow into three phases: free-flow (F), synchronized flow (S), and wide moving jams (J). It also includes phase transitions (e.g., F→S, S→F, S→J, J→S). A Bayesian conditional logit model was developed to estimate the relative safety performance associated with various traffic phases and phase transitions. The authors implemented the random-forest technique to explore the extent to which traffic-flow variables contribute to predicting crash occurrences within the various traffic phases and phase transitions. The methodological approach was successful, suggesting relationships between traffic phases or phase transitions and safety performance.

Regression trees and Random Forests seem promising in terms of road safety–analysis applications. One of the main strengths of regressions trees is the improved accuracy in prediction over typical regression models. More challenging interpretations that come along with this method is a possible downside of regression trees. Opportunities for more widespread use of these methods are available as these methods use the same type of datasets common to crash frequency– and severity–modeling approaches and the techniques are covered by various statistical software packages. In further exploring and replicating these and other related methods and applications, additional avenues for research found include the following:

- Test alternative misclassification cost matrices to see if the structure or behavior of trees changes drastically.
- Compare interpretability and prediction accuracy of variations on regression-tree algorithms (e.g., classification and regression trees (CART) and GUIDE) and ensemble approaches (e.g., BRTs and Random Forests).
- Explore the correct handling of missing data when implementing tree algorithms.
- Explore the transferability of variable-importance rankings and marginal effects from regression trees and Random Forests to various parts of the country.

Unobserved Heterogeneity

The research team reviewed four papers that attempted to explicitly address issues related to heterogeneity. Kim et al. used a mixed logit model with random intercept and slopes to explore the effects of various observable characteristics on driver-injury severities conditional on a single-vehicle crash having occurred.[37] This model was selected due to its ability to capture heterogeneity through the use of random parameters. Age is an example of why this approach was needed; the authors noted that, as drivers age, they become more fragile (i.e., higher probability for more severe injury in a crash), but this change happens at varying rates across the

population.[37] Data on 18,183 single-vehicle crashes (omitting nonprivate vehicles, heavy truck-involved crashes, and pedestrian crashes) were drawn from all reported crashes in the State of California during 2003 and 2004. One of the most notable challenges with the approach described in the paper was the 4 mo required to develop a model using this method, with no practical advantage over the multinomial logit with interaction terms.

Malyshkina and Mannering implemented a two-state Markov-switching NB-regression model that considered a zero-accident state and a normal-count state.[38] The model was defined so that no accidents occur in the zero-accident state, and accidents follow a standard NB distribution in the normal-count state. The Markov-switching model allowed direct estimation of the safety state that sites are in at specific points in time. The model also allowed sites to change safety states over time. The authors noted that, similarly to traditional zero-inflated models, the Markov-switching model attempts to statistically account for the preponderance of zeros observed in accident-count data (a preponderance of zeros means more zeros than predicted by a fitted standard model, such as Poisson or NB). However, they also believed the ability for sites to switch states addresses previous criticisms that the traditional zero-inflated models are unreasonable because they expect any road segment or intersection to be in the zero state all the time and have a long-term mean accident frequency equal to zero. The ability to statistically account for the preponderance of zeros observed in accident-count data, to directly estimate the state that sites are in at specific points in time, and to allow sites to change states over time is a positive attribute of the approach. Its main negative attribute is method complexity.

Mitra and Washington estimated random-effects NB-regression models and random-parameters models of expected crash frequencies at intersections for two purposes.[39] First, the authors assessed reasonableness of the safety effects of the spatial factors and their contribution to intersection safety–model estimation and, second, estimated the amount and direction of omitted-variable bias in coefficient estimates of commonly included variables and the consequence of omission in overall prediction. One of the most significant contributions of the paper is the use of "non-traditional variables" (and the corresponding sources for the data), more so than the modeling approaches themselves.[39] The random-effects and random-parameters methods were not necessarily advanced in terms of their application to statistical road-safety modeling.

Sacchi and Sayed applied a Koyck model, which was introduced in 1954 in the context of investment analysis, of expected number of crashes to demonstrate how to account for time trends and heterogeneity among treatment sites when developing CMFunctions from an observational before–after study.[40] The before–after dataset the authors used to demonstrate the approach corresponded to a Signal Head Upgrade Program in Surrey, BC, and was provided by the Insurance Corporation of British Columbia.[40] The ability to estimate CMFunctions from before–after studies as well as consider changes in countermeasure effectiveness with time (e.g., to account for possible driver adaptation effects) are valuable advancements this paper offered over existing methods. However, the method is complex to implement. Full explanations were not provided in this paper, and additional exploration would be necessary to fully replicate and evaluate the approach.

Beyond the specific scope of the papers reviewed in this chapter, the research team made the following two observations:

- It is possible that the statistical road safety–modeling field would benefit from a clearer understanding of associated terminology, its intended purpose, and appropriate applications and interpretations.

- The statistical road safety–modeling field could benefit from a clearer understanding of terminology, its intended purpose, and appropriate applications and interpretations of zero-inflated models in general (such as the zero-inflated NB-regression models presented in this paper for comparison papers).

Alternative Data Sources

Five papers that utilized alternative data sources were reviewed, including two that used data from the CODES in Utah and Maryland, one that used a combination of FARS data and an "expanded version" of the NASS-CDS, and two that analyzed Naturalistic Driving Study (NDS) data.[41–45] Burch et al. compared the consistency of distributions between crash-assigned (fatal injury, incapacitating injury, nonincapacitating injury, possible injury, no injury (KABCO)) and hospital-assigned (Maximum Abbreviated Injury Scale (MAIS)) injury severity scoring systems for two States (i.e., Maryland and Utah).[41] The researchers used CODES data from both States for 2006 and 2008 for the analysis. The distributions of both KABCO and MAIS injuries varied between States, but the MAIS was more consistent. This finding was expected since the MAIS system has the advantage of being based on information provided by trained medical professionals following an assessment of the patient at the hospital, while the KABCO determination is made by the police officer at the scene of the crash.[41]

Daniello and Gabler also used Maryland CODES data to analyze 3 yr (2006–2008) of motorcycle collisions to determine the type, relative frequency, and severity of injuries incurred in motorcycle-to-barrier crashes.[42] The researchers compared the motorcyclist-injury distributions for motorcycle-to-barrier crashes to injury distributions for other motorcycle-crash types to identify how such collisions differ. Results showed, for example, that motorcyclists involved in barrier collisions were more likely to suffer serious injuries to the thorax and were at a higher risk of rib fractures than motorcyclists involved in other types of collisions. Such detailed injury findings are not possible with traditional police accident reports.

Clark et al. used emergency-medical-services (EMS) information from FARS as well as from an expanded version of NASS-CDS to estimate the time-varying effects of EMS and hospital intervention on mortality.[43] NASS-CDS does not normally contain EMS and hospital times, but these variables were included in NASS-CDS for 2002 through 2003; therefore, the authors used those years for the analysis. A survival model with time-varying covariates and interval censoring was used in the study. The authors highlighted several key findings. Results showed EMS intervention had a beneficial effect, as expected, until a certain point after the crash but not after.[43] Hospital intervention was beneficial, and the beneficial effect increased with time. A crash in a rural location was associated with a higher baseline hazard, and a 50-percent reduction in rural prehospital time reduced 4-h mortality by approximately 7 percent according to the

models. The authors noted these findings seemed to support clinical intuition of a "golden hour" in EMS care and the importance of timely transport to a hospital.

The authors' use of probabilistically linked crash and medical records to obtain treatment details—as well as the Abbreviated Injury Scale (AIS) or MAIS classification of injury severity by trained medical professionals following an assessment of the patient at the hospital—holds significant promise as an advancement over existing methods. (The specific use of FARS and NASS-CDS by Clark et al. has limited applicability for future use as only 2 yr of NASS-CDS data had EMS information.)[43] Building these types of datasets requires police crash reports, clinical data (EMS or emergency department (ED) and hospital inpatient records), and the ability to link the people from the crash reports to those in the clinical records. This work was being done for some time through cooperative agreements between participating States and the National Highway Traffic Safety Administration (NHTSA). NHTSA has since ended funding of the program, but some States, such as Utah, are continuing to build and maintain linked datasets.

Hallmark et al. used SHRP2 NDS and Roadway Information Database (RID) data to explore relationships between driver behavior, roadway factors, environmental factors, and the likelihood of roadway departures on rural, two-lane curves.[44] The primary analysis method used was binary logistic regression, and the following four binary outcomes were modeled:

1. Travel-lane departure to right.
2. Travel-lane departure to left.
3. Entering a horizontal curve more than 5 mph over the advisory speed (or posted speed limit if an advisory speed was not present).
4. Entering a horizontal curve more than 10 mph over the advisory speed (or posted speed limit if an advisory speed was not present).

The authors included 583 sites in the analysis, encompassing 110 curves and 202 drivers. The sample included 57 right-side encroachments and 40 left-side encroachments.[44] Several of the empirical findings were counterintuitive, and the statistical methods themselves were fairly common. The main contribution of this study is a demonstration of how to manage, supplement, and reduce SHRP2 NDS and RID data.

Using a Bayesian approach to develop an MVPLN model, Wu et al. estimated correlations between crashes, near-crashes, and crash-relevant conflicts at a driver level while controlling for a number of different driver characteristics using data from the Virginia Tech Transportation Institute 100-Car NDS dataset.[45] The authors focused on single-vehicle, run-off-road road safety–related events. The response variable was the number of events per driver by severity level (i.e., crash, near-crash, crash relevant). The approach was successful in quantifying the associations between road safety–related events and crash risk while controlling for driver characteristics, showing that, among other findings, drivers under age 25 yr are significantly more likely to be involved in road safety–related events and crashes and significantly positive correlations exist between crashes, near crashes, and crash-relevant incidents.[45]

The use of NDS and RID data in these studies was novel, and the ability to analyze driver behavior in a naturalistic environment presents opportunities for future advancements. In addition, the use of the Bayesian MVPLN model by Wu et al. was effective in this context.[45]

This model has other applications in the road-safety field, and it seems promising for studying the frequency of and correlation between various types of crash and other noncrash, road safety-related events at the driver level.

The findings from this literature review are organized into three thematic areas, including statistical methodology, data sources, and model-estimation frameworks. In future projects, the following statistical methodologies should be further explored:

- Multivariate or hierarchical Bayesian methods to evaluate crash frequency or severity have potential as prediction or estimation methods in road-safety research. These methods can be adapted to overcome issues associated with temporal and spatial correlation as well as the correlation between dependent variables. These models should be compared to univariate models, and the coding used to estimate multivariate Bayesian methods should be documented to enable easy reproduction of the results.

- PSs with matching (or weighting) and many extensions (e.g., marginal structural models or DID methods) are promising countermeasure-evaluation methods in road-safety research.

- Methods to assess the impact of underreporting in road-safety models of crash frequency and severity are needed. Such methods will likely require two independent datasets: a traditional roadway-inventory crash data file and both reported and unreported crashes (those not involving an injury or towed vehicle). The SHRP2 naturalistic driving data offer an opportunity to be this alternative data source. Insurance data may also provide an alternative data source as a means of estimating the magnitude of underreporting in electronic crash records.

- Site-selection and endogeneity bias are important considerations in road-safety research. Two-step estimation procedures, similar to the Heckman selection model, should be further evaluated. This process often begins with a binary model (e.g., probit) to estimate the probability that a site received the countermeasure of interest. The predicted probability is then entered into a count regression model to assess how expected crash frequencies change as a function of the binary outcome probabilities.

Current road-safety research often relies on roadway-inventory and crash data to estimate SPFs, CMFs, or crash severities and types. Many additional data sources are emerging that should be considered in future road-safety research. These data sources include the following:

- SHRP2 naturalistic driving data and corresponding roadway-inventory data. The data offer information related to the driver that has not traditionally been available in road-safety research. As a result, these data afford an opportunity to explain more of the variability in crash-frequency and -severity data.

- CODES data link medical and crash data and are a source of information that has not been frequently used in road-safety research. These data offer an opportunity to evaluate legislative programs and in-vehicle safety equipment.

- Weather station and GIS data are becoming more readily available, and efforts should be made to link these data to crash and roadway inventory to better understand the association between weather information and crash frequency and severity.

- Real-time traffic-monitoring systems are becoming commonplace, particularly on freeways and expressways in urban and suburban areas. Efforts should be made to link data (e.g., speed and traffic flow) to crash and roadway-inventory data.

OBJECTIVES OF TASK A6-6

The objectives of Task A6-6 were to compile, when possible, existing data files from past FHWA-sponsored research and to further investigate advanced statistical methods in road-safety research. Based on the findings of the literature review and data availability, the following analyses were undertaken to compare current safety-evaluation methods to advanced evaluation methods:

- Comparison of the EB and PS potential-outcomes methods using shoulder and centerline rumble strips (CLRS) data from Lyon et al.[46] These data were supplemented with data from Pennsylvania to provide additional site-specific covariates in the analysis.

- Assessment of the level of underreporting in crash-frequency models using data from the New York State Department of Transportation (NYSDOT). In this evaluation, NYSDOT codified geocoded crashes in each municipality as reportable or nonreportable based on crash-reporting thresholds. This analysis was undertaken to assess how candidate covariates vary when considering nonreportable crashes in a safety-performance evaluation.

- Application of linked hospital billing data to identify injured motor vehicle–crash (MVC) occupants. These data were developed using probabilistic linkage to combine crash-report and hospital billing data from Utah.

- Application of Random Forests and regression trees as a means of predicting the frequency of crashes on data collected by Shea et al. in a ramp–interchange spacing project.[47] The outcome of this effort was compared to a prediction obtained by traditional count regression models.

The conclusions from the independent analyses are summarized in the chapter 5. Full descriptions of the analyses are available in the corresponding appendices.

CHAPTER 5. OVERALL DISCUSSION AND CONCLUSIONS

This report has identified and investigated four research approaches and explored opportunities to further understand the relationship between road-safety performance and factors that affect traffic-crash occurrence and severity. A research team was assigned to each of the four approaches. These four approaches are described in the following section, PS Methods. Associated appendix A through appendix D discuss the assigned research team's findings for each approach.

This report compares current statistical-analysis methods and data sources used in road-safety research with alternative methods and data sources. The intent is to identify opportunities to further understand the relationship between road-safety performance and the factors that affect traffic-crash occurrence and severity.

PS METHODS

The purpose of the first analysis was two-fold: first, apply the PS method to road-safety research, and second, compare the results to those obtained from an EB method. The research team used Monte Carlo simulations based on real data from the Pennsylvania Department of Transportation (PennDOT) to further bolster the realistic settings of the comparison. The dataset included 218 mi of shoulder rumble strips (SRS) and CRLS—common low-cost safety strategies—in Pennsylvania.

The research team found that the two methods led to similar results when using real data. The research team also compared the two methods using an artificial dataset simulated from the real data with a generated unmeasured confounder that was imbalanced between treatment and control sites. The research team compared the results to the underlying truth and found the PS method outperformed the EB method in terms of bias and standard errors.

EFFECT OF UNDERREPORTING ON UNDERSTANDING IN CRASH FREQUENCY

The purpose of the second analysis was to investigate the impact of underreporting on crash frequency. NYSDOT supplied three datasets: a comprehensive roadway-inventory file with 125,000 georeferenced segments, a crash dataset with 2.3 million statewide reportable and nonreportable crashes (2008–2011), and boundaries of over 1,500 municipalities.

The research team defined nonreportable crashes as those to which police responded but did not complete a crash report because the incident did not exceed the reporting threshold. The research team conducted the analysis at the municipality level by georeferencing crashes with municipal boundaries and using vehicle miles traveled (VMT) or functional class. The results indicated that directly modeling reported-crash frequencies without accounting for underreporting can lead to bias in the number of predicted crashes as well as bias in estimates of the correlation between road-segment characteristics and crash frequency.

PROBABILISTIC LINK OF HOSPITAL AND CRASH DATA FROM UTAH

The purpose of the third analysis was to examine the utility of using hospital billing data to identify MVC participants using probabilistic linkage. The research team used three datasets in this analysis: the Utah MVC (Utah Department of Transportation (UDOT)), ED (Utah Department of Health, Office of Health Care Statistics), and hospital inpatient discharge databases from 2001 through 2013 (Utah Department of Health, Office of Health Care Statistics).[1]

The results suggested data collected via hospital-injury datasets are a stronger means of estimating number of crashes and injury severity than crash reports. The research team posited this circumstance results because hospital data quantify injuries beyond simple counts. Future research can also use hospital data as a means for comparing the burden of MVC between States.

EXAMPLE APPLICATIONS OF CART AND RANDOM FORESTS FOR STATISTICAL ROAD-SAFETY ANALYSIS

The purpose of the fourth analysis was to apply tree-based methods—Random Forests and CART—to understand the impacts of traffic, geometric design, and operational features on crash frequency and compare the models to NB-regression models with fixed effects. The research team used datasets from previously published journal articles to explore crash frequencies at locations with a right-hand-side entrance ramp followed by a right-hand-side exit ramp.(48)

The research team found that tree-based models had better prediction accuracy than NB-regression models. Although NB-regression models provided a more quantifiable effect of an explanatory variable, tree-based models were more advantageous in terms of providing easy-to-read graphical model forms, direct display of variable importance, and captured interactions between explanatory variables.

[1]Lawrence Cook compiled these datasets on an annual basis with support from the data owners and administrators in the State of Utah. He received permission to reanalyze this data for the purposes of this report.

APPENDIX A. SUMMARY OF DATA ANALYSIS WITH PS METHODS

Definitions for variables used in this appendix are provided in table 1.

Table 1. Definitions for variables used in appendix A.

Variable	Definition
a_0	Intercept
a_1,....., a_{12}	Coefficients of each predictor
AADT	Average annual daily traffic volume
accidents	Number of intersections and driveways per mile
avgdcurv	Average degree of curvature
avgshwid	Average shoulder width
CI	95-percent confidence interval
crash/year	Number of crashes per yr
E[$Y_j(1)$]	Average of the outcome had all sites been treated
E[$Y_j(0)$]	Average outcome had all sites not been treated
e(X_j)	Propensity score
$\hat{e}$ (X_j)	Estimate of propensity score
j	Site
L	Segment length
N	Number of sites
ncurve	Number of horizontal curves per mi
ndrw	Number of driveways
ninter	Number of intersections
pctrk	Percentage of trucks in traffic
rhr	Roadside hazard rating
splim	Speed limit
W	Simulated confounder
width	Pavement width
X	Observed covariates
X_j	Set of pretreatment characteristics or covariates
Y	Outcome of the regression model
y	Year
Y_j	Observed outcome
Z	Countermeasure status
Z_j	Binary-treatment variable
β	Coefficient of the predictor
$\hat{\beta}$	Estimate of the coefficient of the predictor from data
β_0	Intercept
$\hat{\beta}_0$	Estimated coefficient of the intercepts
β_1	Coefficients of the covariates
$\hat{\beta}_1$	Estimated coefficient of the coefficients of covariates
δ	Coefficient of the countermeasure variable
$\hat{\delta}$	Estimated coefficient of the countermeasure variable

Variable	Definition
γ	Coefficient for the simulated confounder
μ_y	Expected number of accidents
$\hat{\sigma}_\delta$	Estimated standard error associated with the estimated coefficient of the countermeasure variable
τ	Causal CMF
$\hat{\tau}$	Estimate of causal CMF

PURPOSE

In road-safety research, a central goal is to evaluate the effectiveness of road-safety programs and countermeasures. The safety effectiveness of a road-safety countermeasure is generally measured by estimating its CMF—a multiplicative factor used to compute the expected number of crashes after implementing a given countermeasure at a specific site.[47]

Due to ethical and practical constraints with road-safety experimentation, observational studies are far more common than randomized experiments in road-safety research. The state-of-the-art gold standard for estimating a CMF is the EB approach.[49] This approach relies on a before–after design; it focuses on precisely estimating the number of crashes that would have occurred at an individual treatment site in the after period had a countermeasure not been implemented, and then, the CMF is estimated from the change in crash frequency from before until after the implementation of the countermeasure. The EB approach is capable of accounting for observed (i.e., reported) changes in crash counts before and after the countermeasure implementation that may be due to regression to the mean. It is well-established and easy to implement. However, the EB approach requires a before–after study design, which is not always feasible because the implementation data of a countermeasure may be unknown. Moreover, the EB approach hinges on an underlying Poisson–Gamma assumption on the outcome model as well as the assumption of a constant time trend for the reference and treatment sites.

From a statistical point of view, evaluating effectiveness of road-safety countermeasures is a causal-inference problem, which refers to designs and methods for evaluating an intervention. PS methods are the most popular causal-inference methods for observational studies.[50] The PS approach adopts a cross-sectional design; it is well established in nearly all statistical-analysis software. PS methods are model free in the sense that they do not require a specific modeling assumption for the outcome. The PS approach has been widely used in medicine, public health, policy, social sciences and other areas but has rarely been used in road-safety research despite its great promise.[51] The research team aimed to introduce the PS approach to road-safety research and compare it to the results obtained from the EB method. Because EB and PS methods involve different designs as well as different assumptions about the model and structure, a fair comparison between them requires careful planning. In this task, the research team provided a framework to compare EB and PS-matching methods in realistic settings, using Monte Carlo simulations based on real data. The research team illustrated these methods using a real study on the application of rumble strips in Pennsylvania.

ANALYSIS METHODOLOGY

Because the EB method is the standard approach in road-safety research, this section omits details on EB and, instead, focuses on the PS approach.

The PS approach uses data from a cross-sectional study. Suppose a number (N) of sites are in the study. Some sites receive a treatment, in this case, a road-safety countermeasure. These sites are referred to as the treatment sites or treatment group. Some sites do not receive the road-safety countermeasure. These sites are referred to as the control sites or control group. For each site (j) (where j equals 1, …, N), let X_j denote a set of pretreatment characteristics or covariates, Z_j be the binary-treatment variable, which is equal to 1 if j receives the countermeasure and 0 otherwise, and let Y_j be the observed outcome. The PS is the probability of being assigned to the treatment group given the covariates, as shown in figure 15.

$$e(X_j) = \Pr(Z_j = 1 | X_j)$$

Figure 15. Equation. PS.

Under the potential-outcome framework, each site has two potential outcomes. $Y_j(0)$ is the outcome that would be observed had the site been assigned to the control group. $Y_j(1)$ is the outcome that would be observed had the site been assigned to the treatment group. The causal effect of a countermeasure for one site is defined as the comparison between the two potential outcomes, and the average causal effect is the average of the individual causal effects for all sites. Because the outcomes of road-safety studies are usually count data, the target causal effect estimated is the causal CMF (τ); that is, the ratio of the average of the outcome had all sites been treated, $E[Y_j(1)]$, versus the average outcome had all sites not been treated, $E[Y_j(0)]$. Using the potential-outcomes notation, τ is show in figure 16.

$$\tau = E[Y_j(1)] / E[Y_j(0)]$$

Figure 16. Equation. Calculation of τ.

For each site, only the potential outcome corresponding to the countermeasure condition is shown, and the other is missing. This circumstance is the fundamental problem of causal inference. To estimate the causal effects from the observed data, two assumptions are commonly made:

- Assumption 1. Unconfoundedness: $\{Y_j(1), Y_j(0)\} \perp X_j$.
- Assumption 2. Overlap or positivity: $0 < e(X_j) < 1$ for all j.

Assumption 1, also known as the assumption of no unmeasured confounder, states that, for sites with the same observed characteristics, the assignment to either the treatment or control group is effectively randomized. Assumption 2 states that each site has a nonzero probability of being assigned to either the treatment or control group; this assumption restricts the study population to values of covariates for which there are both reference and treatment sites.

Under assumptions 1 and 2, one can determine τ directly from observed data as shown in figure 17.

$$\tau = \mathrm{E}(Y_j|Z_j = 1 =)/\mathrm{E}(Y_j|Z_j = 0)$$

Figure 17. Equation. τ from observed data.

Rosenbaum and Rubin proved that the PS approach has two important properties.[50] First, it is a balancing score, that is, sites with the same PSs also have the same distribution of observed covariates (*X*). Consequently, instead of balancing a large set of covariates between treated and control sites, one only needs to balance the PSs. Second, if the assignment to the treatment or control group is unconfounded given *X*, then it is also unconfounded given the PS. This property implies that, given a vector of covariates that ensure unconfoundedness, adjusting for the difference in PSs between treatment and control sites removes all biases associated with the difference in *X*. Based on these two properties, the PS can be viewed as a univariate summary score of the multivariate *X*.

Estimating Causal Effects

PSs are not known and, thus, need to be estimated in observational studies. PSs are typically estimated using a binary logistic regression model (figure 18).

$$\log\left(\frac{e(X_j)}{1 - e(X_j)}\right) = \beta \cdot X_j$$

Figure 18. Equation. Typical logit PS model.

Where β is the coefficient of the predictors, X_j, at each site.

The model goodness of the fit is measured by the resulting covariate balance. After the PS for each site is estimated, denoted by, $1/(1 - \hat{e}(X_j))$ (where $\hat{e}(X_j)$ is the estimate of $e(X_j)$ from the data), τ can be estimated via several methods: matching, weighting, subclassification, or a combination of these three with regression. The following list briefly describes the methods (exact mathematical forms of these estimators can be found, for example, in Imbens and Rubin):[52]

- Model-free methods. The following three methods are model-free in the sense that they do not require a statistical model for the outcome.
    - Matching. For each treatment site, find one or multiple sites in the control group with the closest PSs (with a prespecified threshold). The unmatched sites are dropped from the analysis. The target estimand, τ, is estimated using the ratio of the average outcome between treated and control sites within each matched pair.
    - Weighting. For each site, define the inverse probability weight. For treatment sites, the weight is $1 / \hat{e}(X_j)$; for control site, the weight is $1/(1 - \hat{e}(X_j))$. τ is estimated using the ratio in the weighted average outcome between the treated and control sites. Other weighting methods, such as the overlap weighting, have been increasingly adopted in transportation safety research.[53]

- Subclassification. Stratify the sites into a small number—usually 5 or 6—of subclasses based on the quantiles of the PSs. τ is calculated by first obtaining the overall mean outcome from the weighted average stratum-specific mean outcomes for both the treatment and control groups, and then calculating the ratio between the two.

- Combination with regression. These three model-free methods can be combined with an outcome regression (e.g., a linear regression for continuous outcome or a NB regression for counted outcomes) to further adjust for the residual covariate imbalance and improve the standard-error estimation. For example, Abadie and Imbens advocated for mixed matching and regression methods.(54) Bang and Robins developed the double-robust method (essentially mixed weighting and regression methods).(55) Imbens and Rubin advocated mixed subclassification and regression methods; this combination has been shown to outperform its respective model-free counterparts in many applications.(52) In this task, the research team adopted the modified matching with regression approach in Wood et al., where an NB outcome model with an indicator of the countermeasure status is run on the matched sample, and the exponential of the estimated coefficient of the countermeasure variable is the estimate of τ defined in figure 16.(24)

Simulation-Based Comparisons Between Methods

The EB and PS methods are entirely different, based on different designs, statistical modeling, and structural assumptions (table 2). Each method has strengths and limitations and is suitable for certain, but not all, scenarios. Therefore, a general comparison between the two methods is not meaningful or feasible. Any comparison should be case specific, tailored to the application.

Table 2. Comparison of EB and PS methods.

Feature	EB	PS
Study design	Before–after	Cross-sectional
Model assumption	Poisson–Gamma	None or NB
Assumption of constant time trend	Yes	No
Assumption of unconfoundedness	No	Yes
Address regression-to-mean	Yes	No

In this task, a general framework is proposed to compare different methods using Monte Carlo simulations based on real data. The core idea is to generate an outcome using the covariates and countermeasure variable with variation from the key, underlying assumptions. For example, the key, underlying assumption of the PS method is unconfoundedness, which is not assumed in the EB method. To compare the performance of the two methods when this assumption is violated, the project team used the following procedure:

1. Fit the suitable regression model (e.g., an NB-regression model) to the real data ($Y \sim \beta X + \delta Z$) (where Y is the outcome, δ is the coefficient of the countermeasure variable, and Z is the countermeasure status), and record the estimated coefficients, $\hat{\beta}$ and $\hat{\delta}$, which are estimates of β and δ from the data.
2. Simulate a confounder, W (a variable that is correlated with both outcome and countermeasure status), preferably mimicking the effect size of a true confounder.

3. Simulate a new outcome variable using the same model as in step 1 with X and Z and the estimated coefficients, as well as W with a fixed coefficient: $Y \sim \hat{\beta} X + \hat{\delta} Z + \gamma W$, where γ, a sensitivity parameter, is a prespecified coefficient for W. Based on the true outcome, calculate the truth (e.g., the true τ).
4. Apply the methods undergoing comparison (e.g., EB and PS) to the simulated data without W information, and compare the results.
5. Repeat this procedure under different values of γ, and check the sensitivity of the results.

DATA

This illustration is based on a dataset that was part of a study on the safety evaluation of SRS and CLRS applications. CLRS and SRS are commonly used, low-cost safety strategies that intend to reduce crash frequency by alerting drivers when they are about to depart from the travel lane. This particular dataset focused on sites in Pennsylvania that contained both CLRS and SRS. PennDOT provided a total of 218 mi where both CLRS and SRS were installed. The reference group included two-lane, rural highway segments without rumble strips. The following characteristics were used to narrow the reference sites:

- No access control.
- Divisor (e.g., none, painted divided, man-made barrier, earth divided).
- Divided width equals to 0 ft.
- Speed limit of 20–55 mph.
- Two lanes (one in each direction).
- Documented AADT data from among 17,981 mi of two-lane, rural highway in Pennsylvania.

The available treatment- and control-site data included 466 and 39,360 segments, respectively. The research team further removed all sites with missing values for covariates, and the final dataset included 334 treatment sites as well as 13,286 control sites. In the final analysis database, the variables that were available to compare the EB and PS methods included the following:

- Roadway data—surface type, pavement width, speed limit, number of lanes, resurfacing year, shoulder type, shoulder width, area type (urban/rural).
- Traffic data—AADT from 2003 through 2011 and percentage of trucks in the traffic stream.
- Crash data—data from 2003 through 2012 were obtained for the following crash types (and excluding intersection- and animal-related crashes):
  - Total—identified as a midblock crash and not deer or animal.
  - Fatal plus injury—if number of fatal or injured persons is greater than zero.

- Run-off-road—crash occurred outside the trafficway in an area not intended for vehicles.
- Head-on—opposite-direction collision type.
- Sideswipe—opposite-direction collision type.

Table 3 shows the mean and standard deviation of several site characteristic covariates for both the treatment and control sites.

Table 3. Mean and standard deviation for several characteristic covariates for treatment and control sites.

Site Characteristics	Control Mean	Control SD	Treatment Mean	Treatment SD
Segment length (mi)	0.47	0.13	0.47	0.16
Pavement width (ft)	22.62	3.55	22.89	1.71
Average shoulder width (ft)	3.00	2.17	4.45	1.82
AADT 2003 (vehicles/d)	3,484.34	2,798.71	3,687.11	2,586.68
AADT 2004 (vehicles/d)	3,502.73	2,817.72	3,681.99	2,682.67
AADT 2005 (vehicles/d)	3,510.08	2,807.82	3,700.46	2,710.25
AADT 2006 (vehicles/d)	3,484.51	2,801.47	3,595.61	2,706.57
AADT 2007 (vehicles/d)	3,482.50	2,809.78	3,570.41	2,714.50
AADT 2008 (vehicles/d)	3,415.61	2,767.16	3,548.46	2,693.13
AADT 2009 (vehicles/d)	3,371.22	2,749.67	3,512.53	2,644.74
AADT 2010 (vehicles/d)	3,336.93	2,741.88	3,483.73	2,617.51
AADT 2011 (vehicles/d)	3,290.86	2,701.40	3,388.56	2,634.14
Average AADT 2003–2011 (vehicles/d)	3,430.97	2,739.37	3,574.32	2,644.12
Sum of total crashes during 2003–2012	4.57	4.78	4.04	3.73
Roadside hazard rating	4.84	0.80	4.84	0.78
Number of driveways	7.81	6.70	6.84	5.55
Number of intersections	0.33	0.58	0.24	0.45
Number of horizontal curves	0.97	1.08	0.88	1.01
Number of horizontal curves per mi	2.08	2.36	1.88	2.29
Average length of horizontal curve (ft)	298.68	416.34	339.17	465.62
Length of curve per mi (ft)	930.57	1,185.27	953.13	1,174.37
Average degree of curvature	3.95	6.96	2.80	3.59
Number of intersections and driveways per mi	18.01	15.59	14.75	11.17

SD = standard deviation.
Note: The average AADT is over the period 2003–2011. The sum of crashes for each type is over the period 2 003–2012.

AADT Extrapolation and SPF Estimation

AADT values were only available for 2003 through 2011, but crash-count data were available through 2012. The research team chose to use the average of the previous 3 yr as an estimate for AADT2012.

After estimating AADT2012, the research team estimated the SPF by fitting an NB-regression model on each crash type for the reference sites. Segment length and AADT were transformed to a logarithmic scale as is common practice in road-safety research. The research team considered several models, each including different covariate combinations. The time variable was entered as an indicator variable in these models for each year. The research team also considered a bidirectional stepwise regression model. Model comparison was done based on a likelihood ratio chi-square test as well as predictive performance. The final model was chosen since it performs as well as other models in terms of average absolute bias and root mean-squared error (MSE) yet is simpler. Figure 19 shows the form of this model.

$$\log\left(\frac{crash}{year}\right) = a_0 + a_1 \cdot AADT + a_2 \cdot y + a_3 \cdot \log(L) + a_4 \cdot width + a_5 \cdot splim + a_6 \cdot avgshwid + a_7 \cdot pctrk + a_8 \cdot rhr + a_9 \cdot ndrw + a_{10} \cdot ninter + a_{11} \cdot ncurve + a_{12} \cdot avgdcurv$$

Figure 19. Equation. NB-regression model for crash frequency.

Where:

crash/year = number of crashes per year.
a_0 = intercept.
$a_{1,\ldots\ldots,} a_{12}$ = coefficients of each predictor.
y = year (factor).
width = pavement width.
L = segment length.
splim = speed limit (mph) (factor).
avgshwid = average shoulder width (ft).
pctrk = percentage of trucks in traffic.
rhr = roadside hazard rating.
ndrw = number of driveways.
ninter = number of intersections.
ncurve = number of horizontal curves per mi.
avgdcurv = average degree of curvature.

PS MATCHING

PS matching with regression was used to estimate the countermeasure effect using only the after-period data, following Wood et al.[24] The PSs were estimated using a logistic regression model as shown in figure 18; the covariates are the same shown in figure 19. The research team used nearest neighbor matching with the estimated PSs as the distance metric. Because of the large number of control sites compared to treatment sites, the research team conducted 1–1, 5–1, and 10–1 matching with replacement. Note that selecting the number of matches is essentially a bias-variance tradeoff. The more control sites matched with each treatment site, the larger the bias tends to be since the later matches are, by definition, not as close to the treatment site as the first match. However, the variance will decrease because of the larger number of matched units. Covariate balance is checked by the absolute standardized difference of each covariate. If severe imbalance is detected for some covariates, then the PS model is refit with interaction and/or higher order terms of these covariates. The research team used an iterative process as described

in Imbens and Rubin until balance was deemed satisfactory at 0.10 level (i.e., *p*-value of the t-test of each covariate between treatment and control sites at each of the five PS strata is above 0.10).[52]

After matching, the research team used an NB outcome regression model, shown in figure 20, similar to the one used for SPF estimation.

$$\log(\mu_y) = \beta_0 + \beta_1 \cdot X + \delta \cdot Z$$

Figure 20. Equation. NB outcome regression model.

Where:

μ_y = expected number of accidents.
β_0 = intercept.
β_1 = coefficients of the covariates.

The research team ran a regression model on the matched data and included a countermeasure-status indicator variable as a covariate. τ was then estimated using the equation in figure 21 with 95-percent confidence interval as shown in figure 22.

$$\hat{\tau} = \exp(\hat{\delta})$$

Figure 21. Equation. Estimation of τ.

Where $\hat{\tau}$ is the estimate of τ.

$$CI = \exp(\hat{\delta} \pm 1.96 \cdot \hat{\sigma}_{\delta})$$

Figure 22. Equation. Calculation of confidence interval.

Where:

CI = 95-percent confidence interval.
$\hat{\sigma}_{\delta}$ = estimated standard error associated with $\hat{\delta}$.

Comparison based on original and simulated data

The four methods (i.e., EB and PS with 1–1, 5–1, 10–1 matches) were estimated on the original dataset five separate times, one for each crash type, to see how the results (i.e., CMF) compared to one another. However, as the true CMF is unknown, it is not possible to know which method produces estimates closer to the truth.

For a fair comparison between the methods, the research team also performed analysis on a simulated dataset. The research team first created a new *W* whose distribution is different between the treatment and control sites as shown in figure 23.

$$W|Z \sim \mathrm{N}(25 + 10Z, (5 + 2Z)^2)$$

Figure 23. Equation. Distribution of *W*.

The research team then generated an outcome from an NB-regression model based on *X*, *Z*, and *W*, as shown in figure 24.

$$\log(\mu_y) = \hat{\beta}_0 + \hat{\beta}_1 \cdot X + \hat{\delta} \cdot Z + \gamma \cdot W$$

Figure 24. Equation. NB-regression model based on *X* and *W*.

Where:

$\hat{\beta}_0$= estimated coefficient of β_0.

$\hat{\beta}_1$= estimated coefficient of β_1.

$\hat{\delta}$equals −0.287 in these data. Therefore, the true CMF is exp(−0.287) equal to 0.751. Because the true outcome model is known, the research team can use the true CMF value as a benchmark for comparison. Table 4 shows the covariates used in the PS and outcome models for both the original and simulated data.

Table 4. Covariates included in the PS models and the SPF/Outcome models performed on the original dataset as well as the simulated set.

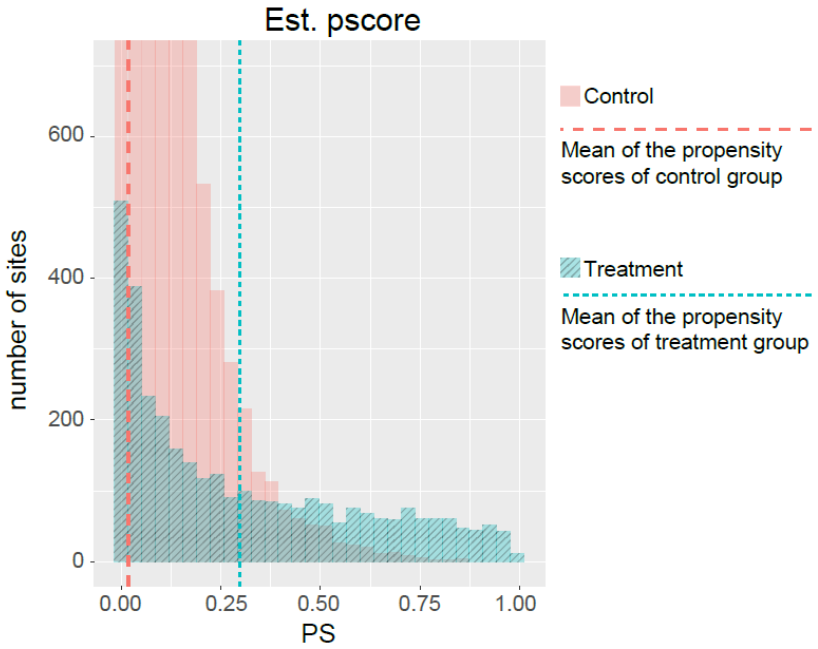
Data Source	PS Model	SPF/Outcome Model
Original data	*L*, *width*, *splim* (factor), *avgshwid*, *pctrk*, *rhr*, *ndrw*, *ninter*, *ncurve*, average length of horizontal curves, length of curves per mile, *avgdcurv*, *accidents*, log(*AADT*)	*y*, log(*L*), *width*, *splim* (factor), *avgshwid*, *pctrk*, *rhr*, *ndrw*, *ninter*, *ncurve*, *avgdcurv*, *accidents*, log(*AADT*)
Data with simulated outcome	*L*, *width*, *avgshwid*, *pctrk*, *rhr*, *ndrw*, *ninter*, *ncurve*, *avgdcurv*, *accidents*, log(*AADT*), log(*W*)	*y*, log(*L*), *width*, *avgshwid*, *pctrk*, *rhr*, *ndrw*, *ninter*, *ncurve*, *avgdcurv*, *accidents*, log(*AADT*), log(*W*)

RESULTS

The research team assumed the true model is known in the sense that the SPF for EB, the PS, and the outcome regression models for the PS-based methods include all of the covariates that were used to generate the data. Different sites have different starting years for the after period (2010, 2011, or 2012). Only data for 2012 were used for the PS method with potential-outcomes analysis. The PSs in the analysis were estimated using this subset of data. Histograms of PSs estimated from the original data, all simulated data, and simulated data for only 2012 are displayed by treatment group in figure 25, figure 26, and figure 27, respectively. Compared to the original dataset, the addition of *W* creates a clear lack of overlap between the control and treatment sites.

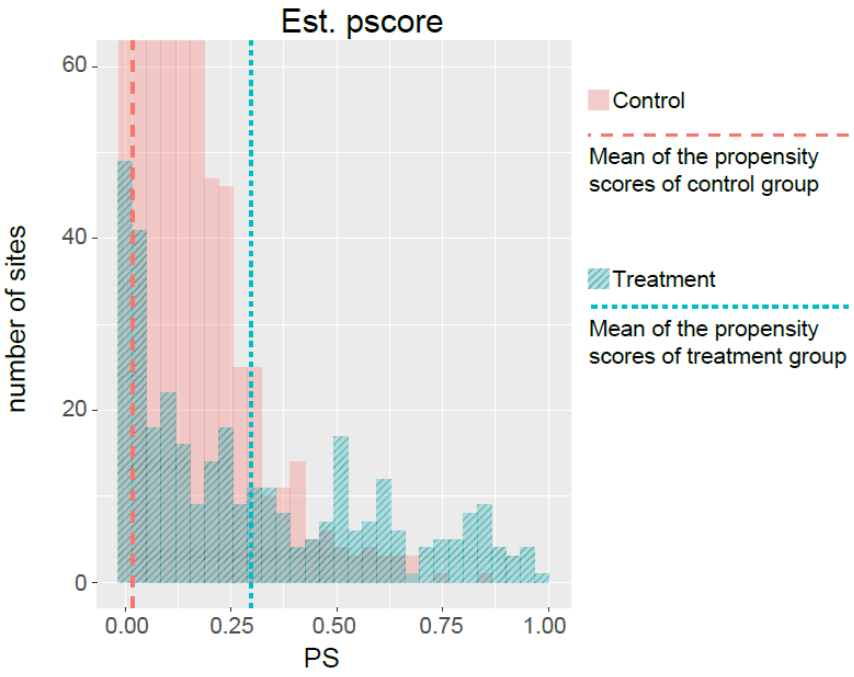
Source: FHWA.
Est. = estimated; pscore = PS.

Figure 25. Histogram. Estimated PSs from the original data.



Source: FHWA.
Est. = estimated; pscore = PS.

Figure 26. Histogram. Estimated PSs from the full data with the additional *W*.



Source: FHWA.
Est. = estimated; pscore = PS.

Figure 27. Histogram. Estimated PSs from the after-only data for year 2012.

Table 5 presents the estimates as well as the 95-percent confidence intervals for each crash type, using the original data. For all crash types, all four methods gave confidence intervals that include the value of 1. Therefore, the research team cannot conclude that the effect of applying the combination of CLRS and SRS is statistically significant at a 95-percent confidence interval. For every observed crash type the estimates were fairly consistent across methods—except for the sideswiped-opposite crash type, for which the EB method provides estimates under 1—while all three PS methods yield estimates larger than 1. The three PS methods gage confidence intervals consistent with expectations: the more reference sites matched to each treatment site, the smaller the width of the confidence intervals became.

Table 5. Original data CMF estimates and 95-percent confidence intervals (lower, upper) by crash type and method.

Method	Total	Injury	ROR	HO	SSOD
EB est	0.91	0.96	0.98	0.92	0.85
EB (95% CI)	(0.79, 1.03)	(0.75, 1.07)	(0.65, 1.17)	(0.35, 1.47)	(0.22, 1.60)
PS 1–1 est	0.93	1.09	0.87	0.84	1.37
PS 1–1 (95% CI)	(0.70, 1.20)	(0.77, 1.55)	(0.46, 1.65)	(0.14, 5.00)	(0.33, 5.60)
PS 5–1 est	0.89	1.03	0.95	1.03	1.42
PS 5–1 (95% CI)	(0.78, 1.01)	(0.85, 1.46)	(0.56, 1.62)	(0.38, 2.80)	(0.47, 4.20)
PS 10–1 est	0.94	1.02	0.98	1.01	1.40
PS 10–1 (95% CI)	(0.72, 1.08)	(0.80, 1.32)	(0.62, 1.55)	(0.42, 2.43)	(0.56, 3.60)

ROR = run-off-road; HO = head-on; SSOD = sideswiped-opposite-direction; CI = confidence interval; est = estimate.

Table 6 presents the CMF estimates and their 95-percent confidence intervals from the simulated data. The EB estimate had the largest distance from the true CMF value of 0.751; its confidence interval neither includes the true estimate nor is it statistically significant. In contrast, all three PS methods gave confidence intervals that include the true CMF value and were statistically significant. These simulation results suggest that, in this simple simulation case, PS-matching methods outperform the EB method in terms of both bias and standard error.

Table 6. Simulated data CMF estimates and 95-percent confidence intervals (lower, upper) by method.

Method	Estimate	95% CI
EB	0.94	(0.78, 1.10)
PS 1–1	0.64	(0.46, 0.89)
PS 5–1	0.83	(0.71, 0.97)
PS 10–1	0.82	(0.74, 0.92)

CI = confidence interval.

DISCUSSION

Evaluating the effectiveness of road-safety countermeasures, from the statistical perspective, is a causal-inference problem. PS methods are the most popular causal-inference methods in observational studies. In this task, the research team compared the PS-matching with the regression method with the gold-standard EB method based on an analysis of a Pennsylvania dataset on rumble strips. In particular, the research team applied the EB method to the whole before–after dataset and the PS method to the after-only data. The two methods led to similar results in the real data. The research team also compared the two methods using an artificial dataset simulated from the real data, with a generated unmeasured confounder that was imbalanced between treatment and control sites. The research team compared the results to the underlying truth and found PS outperformed EB in terms of bias and standard errors.

However, it is important to stress that this specific example only served as a template for conducting simulation-based comparisons between PS and EB methods. As with any

simulation-based analysis, the results may not be generalizable to other situations and should be interpreted with caution. Limitations of the simulation studies include, but are not limited to, the following:

- The simulations are based on the specific Pennsylvania rumble-strip data. Therefore, conclusions are likely to hold for studies with a similar structure and feature. In particular, the Pennsylvania data under study are unique in that the treatment and control sites were carefully chosen to be similar. In such settings, most methods are expected to return similar results because the covariate balance causes the comparisons to be less sensitive to model assumption. However, it is the norm rather than the exception, in road-safety studies, that the treatment and references sites are significantly different in both observed and unobserved characteristics. Therefore, more simulation studies based on real data with severe covariate imbalance would shed more light on the comparative strengths and weaknesses between EB and PS methods.

- The simulation discussed in this appendix focuses on checking the violation of the unconfoundedness assumption. In particular, the research team simulated a scenario with an unmeasured confounder with an arbitrary set of sensitivity parameters (i.e., the correlation between the unmeasured confounder with outcome and countermeasure). Ideally, researchers would conduct simulations under a range of plausible sensitivity parameters and evaluate the results across these settings. The research team did not explore other important aspects of the methods (e.g., model assumptions, such as Poisson–Gamma and constant time trend or misspecification of the PS model).

Nonetheless, simulation-based comparisons and sensitivity analysis are highly relevant and valuable in practice. These methods are standard in statistics literature but are yet widely adopted in road-safety research. It was the primary goal of this research to demonstrate a simple framework for comparing and encouraging more comprehensive simulation studies following the proposed framework described in this study, which provided insights on when either PS or EB methods are preferable.

Going back to the original goal of comparing EB and PS methods, as summarized in table 2, these two methods are entirely different, having different study designs and modeling and structural assumptions. Consequently, each approach has its positives and negatives. The appropriateness of each method is largely case-specific, and one should always choose the method that is most suitable to the data in hand. Regardless, if possible, one should always conduct simulation-based sensitivity analyses, as discussed in the section PS Matching, to check the robustness of the method to the underlying assumptions.

APPENDIX B. EFFECT OF UNDERREPORTING ON UNDERSTANDING VARIATION IN CRASH FREQUENCY

Definitions for variables used in this appendix are provided in table 7.

Table 7. Definitions for variables used in appendix B.

Variable	Definition
a	Overdispersion parameter
Area_SQMI	Area of municipality in squared miles
Binom	Binomial distribution with size equal to total number of crashes in a specific municipality
Density	Population density in municipality in people per squared miles
Divided Miles	Percentage of divided segments within a municipality
e_i	Random effect to allow for over dispersion
*Fun_Class_*1	Percentage of rural principal arterial interstate roadway segments within a municipality
*Fun_Class_*2	Percentage of rural principal arterial freeway/expressway roadway segments within a municipality
*Fun_Class_*4	Percentage of rural principal arterial other roadway segments within a municipality
*Fun_Class_*6	Percentage of rural minor arterial roadway segments within a municipality
*Fun_Class_*7	Percentage of rural major collector roadway segments within a municipality
*Fun_Class_*8	Percentage of rural minor collector roadway segments within a municipality
*Fun_Class_*9	Percentage of rural local roadway segments within a municipality
*Fun_Class_*11	Percentage of urban principal arterial interstate roadway segments within a municipality
*Fun_Class_*12	Percentage of urban principal arterial freeway/expressway roadway segments within a municipality
*Fun_Class_*14	Percentage of urban principal arterial other roadway segments within a municipality
*Fun_Class_*16	Percentage of urban minor arterial roadway segments within a municipality
*Fun_Class_*17	Percentage of urban major collector roadway segments within a municipality
*Fun_Class_*18	Percentage of urban minor collector roadway segments within a municipality
*Fun_Class_*19	Percentage of urban local roadway segments within a municipality
Gamma	Gamma-distributed random variable with scale parameter (overdispersion parameter) and shape parameter of 1
i	Municipality
Intercept	Intercept of the regression model
NB dispersion	Estimate for overdispersion parameter
One-Way Miles	Percentage of one-way segments within a municipality
*Own_Jur_*1	Percentage of roadway segments under NYSDOT jurisdiction
*Own_Jur_*2	Percentage of roadway segments under a county jurisdiction
*Own_Jur_*3	Percentage of roadway segments under a town jurisdiction
*Own_Jur_*4	Percentage of roadway segments under a city or village jurisdiction

Variable	Definition
*Own_Jur_*11	Percentage of roadway segments under a state park jurisdiction
*Own_Jur_*12	Percentage of roadway segments under a local park jurisdiction
*Own_Jur_*21	Percentage of roadway segments under other State agencies' jurisdiction
*Own_Jur_*25	Percentage of roadway segments under other local agencies' jurisdiction
*Own_Jur_*26	Percentage of roadway segments under private (other than railroad) jurisdiction
*Own_Jur_*31	Percentage of roadway segments under NYSDOT thruway jurisdiction
*Own_Jur_*32	Percentage of roadway segments under other toll authority jurisdiction
*Own_Jur_*50	Percentage of roadway segments under Indian Tribal Government jurisdiction
*Own_Jur_*62	Percentage of roadway segments under Bureau of Indian Affairs jurisdiction
*Own_Jur_*80	Percentage of roadway segments under other jurisdiction
p	Reported number for all crashes regardless of municipality
PCT.Miles.Div	Percentage miles divided
PCT.Miles.OW	Percentage miles one way
PCT.Rural	Percentage rural
p_i	Probability a crash is reported
Pois(λ_i)	Poisson distribution with mean of the NB distribution
*POP*2010	Municipality population in 1990, 2000, and 2010, respectively
p_u	Reporting probability in a given municipality in the hold-out set
R_i	Reported number of crashes in a municipality
Road.Density	Road density
R_u	Reported number of crashes in hold-out municipality set
State.Jur	Agency responsible for roadway ownership
T_i	Total number of crashes in a specific municipality
T_u	Total number of crashes in hold-out municipalities
u	Municipalities in the hold-out validation set
Un-Divided Miles	Percentage of undivided segments within a municipality
Urban. MuniType	Urban municipality type
VMT	Vehicle miles of travel within a municipality in a given year
x_i	Road-segment covariates
x_i'	Measured characteristics of a given municipality
x_u'	Measured characteristics of a given municipality in the hold-out set
*year*2008	Indicator variable for the year 2008
*year*2009	Indicator variable for the year 2009
*year*2010	Indicator variable for the year 2010
α	Regression parameters
β	Vector of regression parameters
λ	Expected number of crashes that would have occurred in the after period if a road-safety countermeasure had not been implemented
λ_i	Mean of the NB distribution
μ	Intercept of an intercept-only model

OVERVIEW

The dataset used in the analysis is based on a comprehensive roadway-inventory file maintained by NYSDOT.[1] This inventory contains over 125,000 georeferenced segments (located spatially) and their associated attributes. NYSDOT also maintained the crash dataset that included 2.3 million statewide reportable and nonreportable crashes from 2008 through 2011.[2] A third and final dataset included boundaries for over 1,500 municipalities in New York State.[3] The roadway-inventory file was first aggregated to the municipality level using the New York municipality-boundaries dataset. The municipality-level inventory dataset was then merged with the crash dataset to create an analysis database for all reportable and nonreportable crashes among municipalities in New York, excluding New York City.

The present study defined nonreportable crashes as incidents to which the police responded, but an NYSDOT crash report was not completed because the reporting threshold (perceived injury or property damage exceeding $1,000) was not surpassed. The underreporting analysis performed in this study was based on municipality-level data. In many instances, the nonreportable crashes were georeferenced within municipal boundaries but were not located on a specific roadway at a specific milepost along the roadway. As such, the roadway characteristics included in the analysis were based on aggregate, municipal-level information, such as VMT within a given municipality, or the proportion of roadway mileage classified as certain functional classes. Crashes were also aggregated at the municipal level. The following sections provide a detailed description of the data used in the analysis.

DATA

The NYSDOT geospatial roadway inventory and geospatial crash datasets are described in this appendix. Additionally, the methods used to merge the two data files are described.

NYSDOT Geospatial Roadway-Inventory Dataset

The roadway-inventory dataset contained geospatial attributes of the road segments (i.e., length and location); those attributes helped visualize and map the New York road network. The roadway-inventory dataset contained upward of 50 variables that, in some instances, included missing or inaccurate information. For example, the total pavement and lane-width variables often contained invalid values (0), and therefore, these cell entries were replaced with blank cells, which signified missing data. Additionally, the number of lanes for some segments were contradictory to values found through satellite imagery and was therefore not used in compiling the analysis dataset. The AADT also contained missing data. Because traffic volumes are an important exposure measure in statistical models, missing information was filled using statewide

[1]The researchers downloaded this dataset from a New York GIS website (http://gis.ny.gov/civil-boundaries) on August 10, 2016.

[2]The dataset is unpublished data that were provided to the researchers for this effort via email correspondence on July 19, 2016.

[3]The dataset is unpublished data that were provided to the researchers for this effort via email correspondence on July 19, 2016.

averages for the functional class of roadway (which was always coded). The statewide averages for each functional class included in the roadway-inventory files are shown in table 8.

Table 8. AADT based on functional classification.

Functional Classification	AADT (Vehicles/d)
Rural principal arterial interstate	15,027
Rural principal arterial freeway/expressway	11,585
Rural principal arterial other	5,374
Rural minor arterial	3,694
Rural major collector	2,014
Rural minor collector	831
Rural local	692
Urban principal arterial interstate	35,440
Urban principal arterial freeway/expressway	32,982
Urban principal arterial other	16,252
Urban minor arterial	9,407
Urban major collector	3,770
Urban minor collector	1,911
Urban local	1,670

NYSDOT Geospatial Crash Dataset

Unlike the roadway-inventory dataset, the crash dataset was compiled in a database file format (i.e., without geospatial features), and included more than 2.3 million reportable and nonreportable records over the 4-yr analysis period. The crash dataset included information about the crash event including the date, number of persons injured or killed, and the municipality in which the crash occurred. In the NYSDOT crash dataset, nonreportable crash incidents were indicated by a single-letter prefix (U) in the crash-identification string. Consequently, crash information is summarized in three categories, reportable, nonreportable, and all crashes. Additional subcategories were created for injury and fatal crashes. Summary statistics for the NYSDOT crash dataset are shown in table 9 through table 13.

Table 9. Descriptive statistics for New York crash data for 2008.

Crash Category	Mean	SD	Min	Max
All_Crashes	245.1	871.9	1	18,687
All_Inj+Fat	79.4	348.1	0	7,099
R_Crashes	213.7	679.3	0	14,217
R_Inj+Fat	79.4	348.1	0	7,099
U_Crashes	31.4	215.3	0	4,470

SD = standard deviation; Min = minimum; Max = maximum; Inj+Fat = injury and fatal; R = reportable; U = nonreportable.

Table 10. Descriptive statistics for New York crash data for 2009.

Crash Category	Mean	SD	Min	Max
All_Crashes	256.2	973.3	1	21,304
All_Inj+Fat	78.4	349.2	0	7,272
R_Crashes	215.1	687.6	1	14,242
R_Inj+Fat	78.4	349.2	0	7,272
U_Crashes	41.1	322.2	0	7,062

SD = standard deviation; Min = minimum; Max = maximum; Inj+Fat = injury and fatal; R = reportable; U = nonreportable.

Table 11. Descriptive statistics for New York crash data for 2010.

Crash Category	Mean	SD	Min	Max
All_Crashes	258.2	968.1	1	20,989
All_Inj+Fat	77.7	349.2	0	7,442
R_Crashes	225.7	745.6	1	16,605
R_Inj+Fat	77.7	349.2	0	7,442
U_Crashes	32.6	266.0	0	4,960

SD = standard deviation; Min = minimum; Max = maximum; Inj+Fat = injury and fatal; R = reportable; U = nonreportable.

Table 12. Descriptive statistics for New York crash data for 2011.

Crash Category	Mean	SD	Min	Max
All_Crashes	249.0	894.4	1	20,678
All_Inj+Fat	74.0	326.9	0	6,938
R_Crashes	230.8	799.5	1	18,561
R_Inj+Fat	74.0	326.9	0	6,938
U_Crashes	18.2	142.4	0	4,139

SD = standard deviation; Min = minimum; Max = maximum; Inj+Fat = injury and fatal; R = reportable; U = nonreportable.

Table 13. Descriptive statistics for New York crash data for all years.

Crash Category	Mean	SD	Min	Max
All_Crashes	252.1	927.8	1	21,304
All_Inj+Fat	77.4	343.4	0	7,442
R_Crashes	309.7	6892.9	0	18,561
R_Inj+Fat	77.4	343.4	0	7,442
U_Crashes	30.8	245.7	0	7,062

SD = standard deviation; Min = minimum; Max = maximum; Inj+Fat = injury and fatal; R = reportable; U = nonreportable.

Aggregating Datasets Using Higher Order or Level Variables

As noted in the NYSDOT Geospatial Crash Dataset section, the roadway-inventory dataset contains information related to the location of roadway segments and the length of each individual segment. This information was stored in a geospatial file format (shapefile) and

allowed the roadway-inventory dataset to be spatially merged with other shapefiles. An open-source shapefile that stores the boundaries and location for each of New York's over 1,500 municipalities was then used to append the roadway-inventory dataset to the municipality information. This process was done by spatially merging the two shapefiles using an ArcGIS® spatial-analysis tool and summarizing the roadway-inventory variables by municipality. Because many roadway-inventory variables—such as lane width and number of lanes are segment-based—the percentage of segments with each unique variable value was computed for each of the over 1,500 municipalities. For example, the percentage of segments with each unique number of lanes (e.g., proportion of roadway segments with one lane or two lanes) was computed within each municipality. Because the analysis was performed at the municipal level, some of the roadway-segment variables could not be included in the statistical model. Table 14 provides a descriptive statistics summary for all the municipality-level variables that were considered for the final analysis.

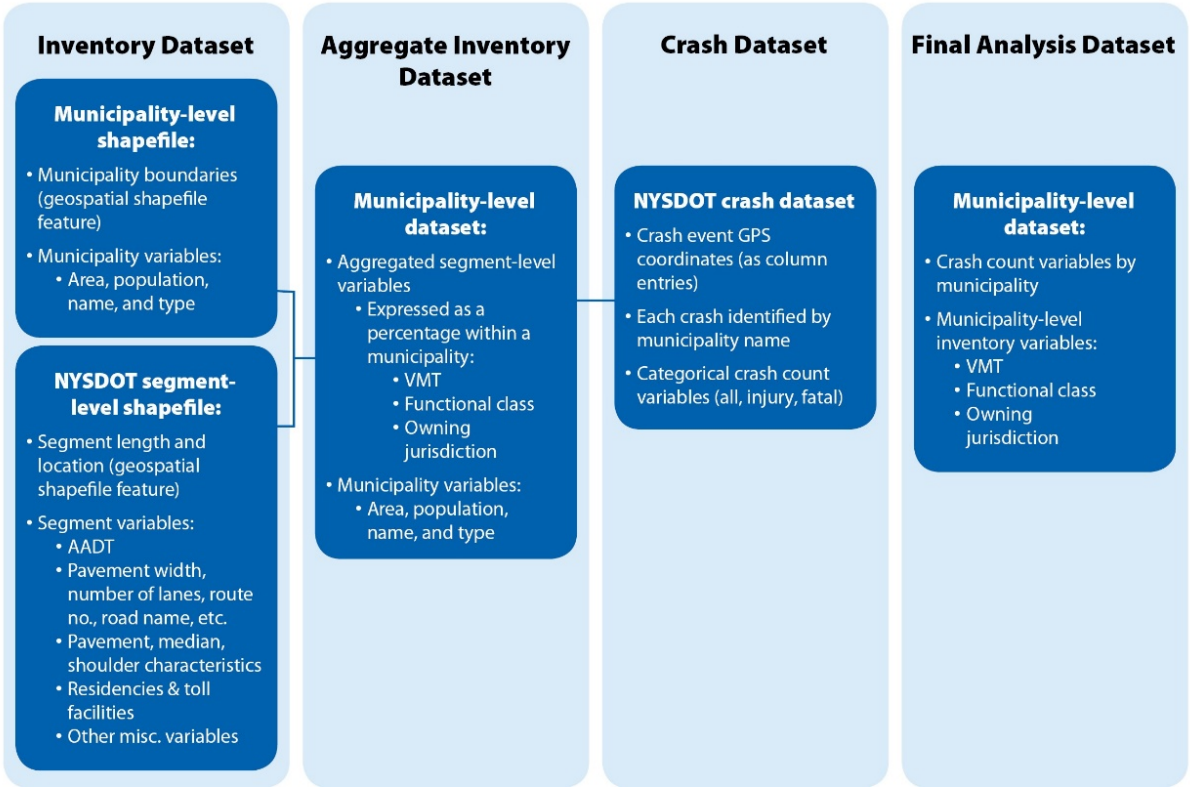
Table 14. Descriptive statistics for compiled municipality-level roadway-inventory dataset.

Variable	Description	Mean	SD	Min	Max
Divided Miles	Percentage of divided segments within a municipality	0.07	0.11	0	1
Un-Divided Miles	Percentage of undivided segments within a municipality	0.93	0.11	0	1
One-Way Miles	Percentage of one-way segments within a municipality (can either be divided or undivided)	0.03	0.07	0	1
*Fun_Class*_1	Percentage of Rural Principal Arterial Interstate roadway segments within a municipality	0.02	0.07	0	1
*Fun_Class*_2	Percentage of Rural Principal Arterial Freeway/Expressway roadway segments within a municipality	0.00	0.02	0	0
*Fun_Class*_4	Percentage of Rural Principal Arterial Other roadway segments within a municipality	0.07	0.16	0	1
*Fun_Class*_6	Percentage of Rural Minor Arterial roadway segments within a municipality	0.10	0.18	0	1
*Fun_Class*_7	Percentage of Rural Major Collector roadway segments within a municipality	0.16	0.21	0	1
*Fun_Class*_8	Percentage of Rural Minor Collector roadway segments within a municipality	0.23	0.24	0	1
*Fun_Class*_9	Percentage of Rural Local roadway segments within a municipality	0.04	0.08	0	1
*Fun_Class*_11	Percentage of Urban Principal Arterial Interstate roadway segments within a municipality	0.03	0.08	0	1
*Fun_Class*_12	Percentage of Urban Principal Arterial Freeway/Expressway roadway segments within a municipality	0.01	0.05	0	1
*Fun_Class*_14	Percentage of Urban Principal Arterial Other roadway segments within a municipality	0.06	0.13	0	1
*Fun_Class*_16	Percentage of Urban Minor Arterial roadway segments within a municipality	0.11	0.19	0	1
*Fun_Class*_17	Percentage of Urban Major Collector roadway segments within a municipality	0.13	0.20	0	1
*Fun_Class*_18	Percentage of Urban Minor Collector roadway segments within a municipality	0.01	0.04	0	1
*Fun_Class*_19	Percentage of Urban Local roadway segments within a municipality	0.03	0.07	0	1
*Own_Jur*_1	Percentage of roadway segments under NYSDOT jurisdiction	0.46	0.26	0	1
*Own_Jur*_2	Percentage of roadway segments under a county jurisdiction	0.34	0.26	0	1

Variable	Description	Mean	SD	Min	Max
*Own_Jur*_3	Percentage of roadway segments under a town jurisdiction	0.05	0.10	0	1
*Own_Jur*_4	Percentage of roadway segments under a city or village jurisdiction	0.14	0.24	0	1
*Own_Jur*_11	Percentage of roadway segments under a state park jurisdiction	0.00	0.02	0	0
*Own_Jur*_12	Percentage of roadway segments under a local park jurisdiction	0.00	0.00	0	0
*Own_Jur*_21	Percentage of roadway segments under other State agencies' jurisdiction	0.00	0.01	0	0
*Own_Jur*_25	Percentage of roadway segments under other local agencies' jurisdiction	0.00	0.01	0	0
*Own_Jur*_26	Percentage of roadway segments under private (other than railroad) jurisdiction	0.00	0.00	0	0
*Own_Jur*_31	Percentage of roadway segments under NYSDOT thruway jurisdiction	0.01	0.05	0	0
*Own_Jur*_32	Percentage of roadway segments under other toll authority jurisdiction	0.00	0.01	0	0
*Own_Jur*_50	Percentage of roadway segments under Indian Tribal Government jurisdiction	0.00	0.01	0	0
*Own_Jur*_62	Percentage of roadway segments under Bureau of Indian Affairs jurisdiction	0.00	0.00	0	0
*Own_Jur*_80	Percentage of roadway segments under other jurisdiction	0.00	0.00	0	0
*POP*2010*	Municipality population in 1990, 2000, and 2010, respectively	8,648	31,370	38	759,757
*Area_SQMI**	Area of municipality in squared miles	34	43	0	510
*Density**	Population density in municipality in people per squared miles	1,057	1,963	0	22,258
VMT	Vehicle miles of travel within a municipality in a given year	188,297	605,364	0	11,900,000

*Variables imported from the original New York municipalities shapefile.
SD = standard deviation; Min = minimum; Max = maximum; Fun_Class = functional class; Own_Jur = owner's jurisdiction; POP = population; SQMI = square miles.

Once a municipality-level inventory dataset was prepared, the research team assembled the final analysis dataset by summarizing crash information (number and severity of crashes) for every municipality (*i*) and year. In aggregating crash counts for each of these categories to the municipality level, municipality name was used as the key aggregating variable since both the NYSDOT crash dataset and the compiled municipality-level inventory dataset identify the municipality name for each of its records. Once aggregation was complete, a final analysis dataset was created that contains municipality-level inventory variables as well as categorical crash count-variables. Figure 28 shows the complete process that was undertaken to develop the analysis data file used in the present study.



Source: FHWA.
GPS = global positioning system.

Figure 28. Graphic. New York State database-development process.

In exploring any potential clustering of nonreportable crashes in certain geographic areas, crash events were mapped using ArcMAP™, which is the central application used in ArcGIS.[(56)]

Fewer nonreportable crashes seem to occur in nonurban areas than in urban areas. For example, considering the crashes mapped in the Buffalo city municipality and its surrounding municipalities, 41.7 percent of total crashes are nonreportable in Buffalo, NY, while 5.04 percent of total crashes were coded as nonreportable in the surrounding municipalities.

METHODOLOGY

In the NYSDOT data, the total number of crashes (T_i) as well as the reported number of crashes (R_i) for each i are known. All other characteristics of i are encoded in the row vector of road-segment covariates (x_i). To compare approaches for estimating the correlation between covariates and total crashes, or underreporting, a standard NB count-regression model was estimated using total crashes, which includes those codified as reported as well as crashes codified as nonreported. The research team estimated a separate NB count-regression model using only reported crashes. Finally, the research team estimated a third NB count-regression model with an underreporting term in the set of linear predictors. This estimation enables exploration of the potential bias in statistical estimation due to underreporting. In the next section, Model 1: NB-Regression Model for Total Crashes, the research team predicted the expected total number of crashes using reported crashes, and the predictive power of the model was based on an out-of-sample validation. The research team randomly selected 1,000 of the 6,017 records in the NYSDOT data and held these data out as a validation set. The remaining 5,017 records were used to estimate model parameters and make predictions on the held-out validation data. This strategy provided an opportunity to compare models based on predictive power. All estimation was done in the R statistical computing environment, with NB-regression models fit using the glm.nb function in the Modern Applied Statistics with S (MASS) package. This function estimates NB regression–model parameters by splitting the parameters into two sets: regression parameters (α) and overdispersion parameter (a). Estimation is done by maximizing the log-likelihood, alternating between treating a as fixed and estimating α using iteratively reweighted least squares (Newton's method for generalized linear models), and then treating those estimates α as fixed and estimating a using score and information iterations. These steps are iterated until numerical convergence results in the MLEs for α and a. More details are found in the documentation for the MASS R-package. After parameter estimates are obtained using the training set, these parameter values are used to predict crashes on the 1,000 held-out records.

Model 1: NB-Regression Model for Total Crashes

To see the effects of underreporting on prediction and estimation, a standard NB-regression model is fit to T_i (figure 29).

$$T_i \sim \mathrm{NB}(\lambda_i, a), \log(\lambda_i) = x_i' \alpha$$

Figure 29. Equation. NB-regression model for T_i.

Where:

λ_i = mean of the NB distribution, or the mean of T_i.
x_i' = measured characteristics of a given municipality.

The NB-regression model is a Poisson–Gamma mixture model for count data with overdispersion, as shown in figure 30.

$$T_i \sim Pois(\lambda_i), \log(\lambda_i) = x_i' \alpha + e_i$$

$$\exp(e_i) \sim Gamma(a, 1)$$

Figure 30. Equation. Poisson–Gamma mixture model for total crashes.

Where:

$Pois(\lambda_i)$ = Poisson distribution with mean λ_i.
e_i = random effect to allow for overdispersion.
Gamma = gamma-distributed random variable with scale parameter, a and shape parameter, 1.

α and a were estimated using maximum-likelihood methods implemented in the R statistical computing environment.

Model 2: NB-Regression Model for Reported Crashes

To explore estimation of α when total crashes are unknown, a standard NB-regression model is fit to R_i, as shown in figure 31.

$$R_i \sim \mathrm{NB}(\lambda_i, a), \log(\lambda_i) = x_i' \alpha$$

Figure 31. Equation. Standard NB-regression model for the reported crashes.

Estimation here is the same as for model 1 (NB-regression model for total crashes), except the response variable is the reported crashes instead of the total crashes. In most scenarios, researchers only have access to reported crashes, and fitting this model and comparing the results to those obtained using the total crashes will provide insights into what might be missed when estimating a model of only the reported crashes.

Model 3: NB-Regression Model With Underreporting

Finally, based on the approach of Cameron and Trivedi, the probability that a crash in *i* is reported (p_i), is shown in figure 32.[57]

$$p_i = \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)}$$

Figure 32. Equation. Probability that each crash at *i* is reported.

Where β is vector of regression parameters, which control the relationship between the reporting rate, p_i, and the road-segment characteristics in x_i.

Figure 33 shows an NB-regression model fit for R_i with an underreporting term in the linear predictor.[56]

$$R_i \sim \mathrm{NB}(\lambda_i, a), \log \lambda_i = \log(p_i) + x_i' \alpha$$

Figure 33. Equation. NB-regression model for reported crashes with underreporting term.

Numerical optimization of the log-likelihood was used to estimate model parameters. In many cases, not all parameters in this model are identifiable. λ_i depends on the predictor variables twice—once in the log-linear model for expected total counts and again in the logistic binary model for reporting probability. For example, consider a simple case in which the only predictor variable is an intercept, and so the reporting probability is p for all crashes (regardless of municipality), and the mean number of total crashes is $\exp(\mu)$ for all i (figure 34), where μ is the intercept of an intercept-only model.

$$R_i \sim \mathrm{NB}(\lambda, a), \log(\lambda) = \log(p) + \mu$$

Figure 34. Equation. NB-regression model for reported crashes including only intercept.

Where λ is expected number of crashes that would have occurred in the after period if a road-safety countermeasure had not been implemented.

In this simple case, no information (R_i) will allow estimation of both p and μ as the model has two intercepts. The inclusion of x_i may improve this identifiability problem but does not completely alleviate it. Covariates are included in the model shown in figure 35.

$$\log(\lambda_i) = x_i'(\beta + \alpha) - \log[1 + \exp(x_i'\beta)]$$

Figure 35. Equation. NB-regression model for reported crashes including covariates.

When $x'_i\beta$ is much less than 0, $\log[1 + \exp(x'_i)]$ is close to 0 and the regression parameters for counts and reporting are not identifiable; only their sum is identifiable. In general, it is likely to be difficult if not impossible to make reliable inferences on both reporting probability and crash frequencies without additional information on reporting probabilities. Thus, obtaining additional information on reporting probabilities whenever possible is recommended.

Predicting Total Crashes From Reported Crashes

Approaches to adjusting R_i to more accurately reflect T_i occurring in a location are now considered. As both T_i and R_i are known, it is natural to consider modeling p_i as a function of covariates through binomial logistic regression. The usefulness of this approach is gauged by comparing this approach to simpler approaches in which the total number of counts is estimated by a simple multiplicative adjustment of the reported number of counts.

Model 4: Binomial Model for Reporting

A binomial logistic regression model for reporting is estimated first (figure 36).

$$R_i \sim Binom(T_i, p_i), p_i = \exp(x_i'\beta) / [1 + \exp(x_i'\beta)]$$

Figure 36. Equation. Binomial logistic regression model for crash reporting.

Where *Binom* is binomial distribution with size T_i.

This model was fit using logistic regression in R. To predict the total number of crashes at *i* values in the hold-out validation set (a different set of municipalities (*u*), which were not used to develop the model), it was assumed that the reported number of crashes in *u* (R_u) was known. The total number of crashes in *u* (T_u) was then predicted by computing figure 37.

$$T_u = \frac{R_u}{p_u}, p_u = \exp\{x_u'\beta\} / [1 + \exp\{x_u'\beta\}]$$

Figure 37. Equation. Prediction function of total number of crashes.

Where:

p_u = reporting probability in a given municipality in the hold-out set.
x'_u = measured characteristics of a given municipality in the hold-out set.

This approach would allow the total number of crashes to be predicted based on known characteristics of a location and the reported number of crashes at that location.

Model 5: Simple Multiplicative Adjustment for Underreporting

A simpler approach to estimating p_u is to assume that all locations have a constant reporting rate for crashes. This strategy is equivalent to the intercept-only model discussed in model 1. In this model formulation, p_u is estimated as the fraction of total crashes that are reported. In the training set, it was found that p_i equals 0.877. If p_u is assumed to equal p_i, T_u can be predicted by using the equation in figure 38.

$$T_u = \frac{R_u}{0.877} = 1.14R_u$$

Figure 38. Equation. Total number of predicted crashes.

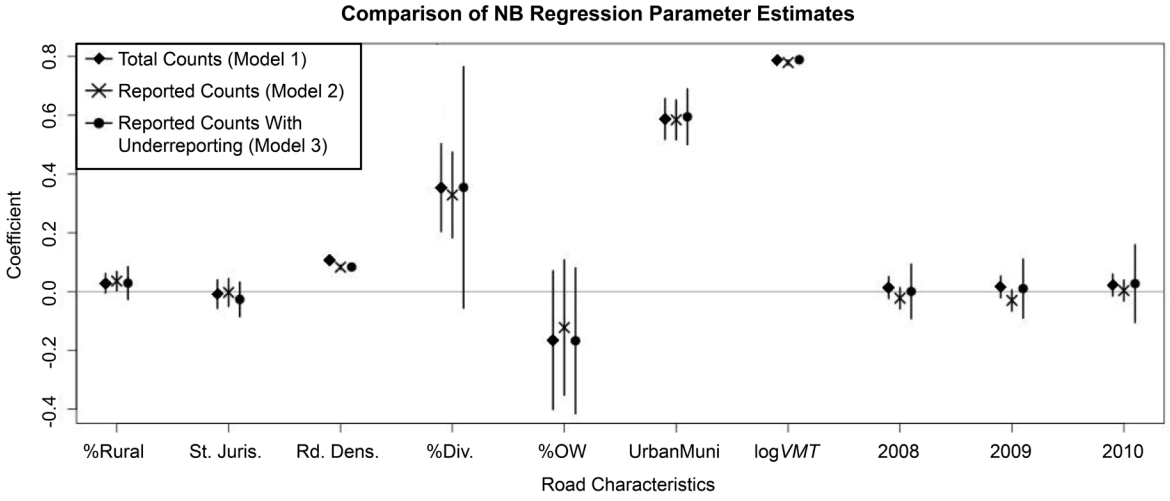
Model 6: No Adjustment to Reported Crashes

As a baseline comparison, the reported crashes were also treated as an estimate of total crashes. This approach allowed for assessing what is gained by accounting for underreporting when attempting to estimate or predict total crash frequency.

RESULTS: CRASH-FREQUENCY MODELS

The estimates for α for models 1–3 are compared in table 15 through table 17. Only model 1 uses the total crashes (reportable and nonreportable) as the dependent variable, while models 2 and 3 show some bias due to underreporting. This same information is presented graphically in figure 39, which plots estimated regression parameters for models 2 and 3 (on the *y*-axis) versus the corresponding regression parameters from model 1 on the *x*-axis. These results show that model 3 (which directly models underreporting in the mean of the NB-regression model for reported crashes) tends to give estimates of regression parameters that are more consistent with estimates from the total crashes (model 1) than model 2 (which does not account for underreporting). Model fit statistics, such as R^2, deviance, and Akaike Information Criterion, are

not comparable between models fit to different sets of data (T_i and R_i) so provide no basis for comparison of models 2 and 3 with model 1.



Source: FHWA.
%Rural = percent rural; St. Juris. = State jurisdiction; Rd. Dens. = roadway density; %Div. = percent divided; %OW = percent one way; UrbanMuni = urban municipality; log*VMT* = logarithm of vehicle miles traveled.

Figure 39. Graph. Comparison of NB regression–parameter estimates (dots) with 95-percent confidence intervals (vertical lines).

The NB α shown in table 15 through table 17 are generally consistent in sign and magnitude across models 1 through 3. The natural logarithm of vehicle miles traveled (*VMT*) (log*VMT*) is similar in magnitude to exposure variables included in traffic-volume (i.e., AADT) variables found in the AASHTO HSM SPFs.[3] None of the year indicator variables are statistically significant and indicate that there is little annual variation in total crashes in the analysis database. The proportion of rural roadway mileage in New York is associated with fewer total and reported crashes. The agency responsible for roadway ownership (*State.Jur*) is not statistically significant nor is the variable that accounts for the proportion of one-way roadways in a New York municipality. The roadway density (roadway mileage per square area of the municipality) and proportion of divided highway mileage are positively correlated with total and reportable crashes. This result is consistent with expectations as the former variable indicates greater exposure to crashes, and the latter variable is representative of travel speeds, which increase the likelihood that crashes are reported.

The results for model 3 show increased variance in some parameter estimates, likely due to identifiability problems as shown in figure 33. The estimates from model 3 do often remove the bias found in estimates from model 2, indicating that including an underreporting term in the NB-regression model has potential to remove bias in parameter estimates due to underreporting at the cost of increased variance.

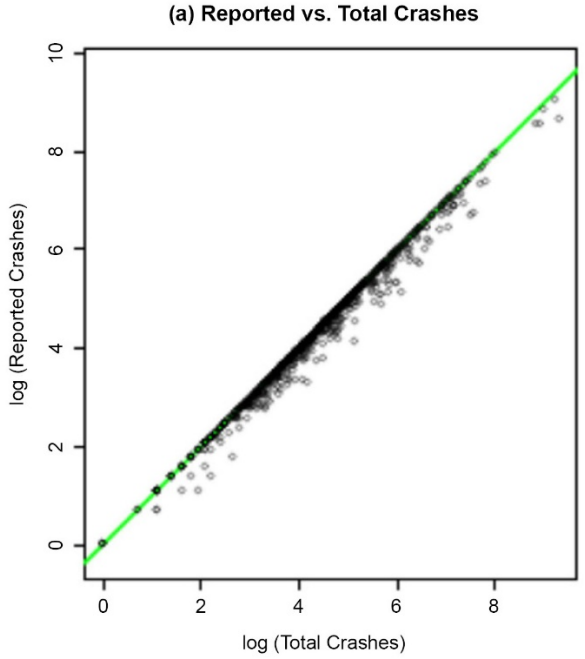
The results for model 4 indicate that all predictor variables are significantly related—either positively or negatively correlated—with crash-reporting probability. Reporting probabilities are significantly higher for roadways under State jurisdiction and significantly lower for cities than

for villages. Reporting probabilities are negatively correlated with road density, log*VMT*, and divided highways. Significant year effects indicate that reporting probabilities may vary between years, suggesting that random-year effects could be included in future models of reporting probability.

The results for logistic regression parameters for α estimated using model 3 show wide divergence from those estimated using the binomial regression approach of model 4. This result is likely due to the potential identifiability problems of model 3. Adjusting for underreporting in an NB-regression model, as is done in model 3, shows some promise for alleviating bias in NB-regression parameters but may not be a reliable way to estimate reporting probabilities.

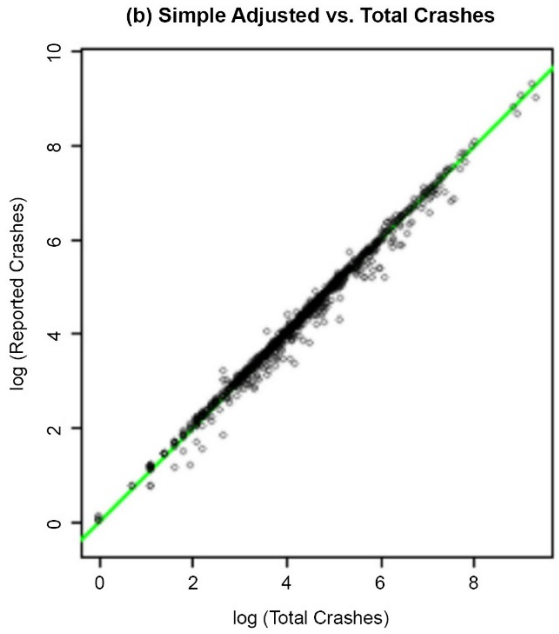
Figure 40 through figure 42 show the predicted total crashes versus the true total crashes, based on the three approaches (models 4–6). The MSPE between the reported number of crashes and total crashes (model 6) is 39,862, the MSPE between the total crashes and adjusted reported number of crashes using the simple multiplicative adjustment in model 5 is 24,752, and the MSPE between the total crashes and adjusted reported number of crashes using the logistic regression adjustment in model 4 is 19,517. These results indicate that a simple adjustment (multiplying reported crashes by a constant underreporting rate) improves predictive accuracy in NYSDOT on held-out data but not as well as applying an adjustment based on covariates and logistic regression (model 4).

Parameters were estimated using NB-regression models fit to the total number of crashes, the reported number of crashes, and the NB-regression model with underreporting from Cameron and Trivedi.[57] Results show only a slight bias due to underreporting and that the approach of Cameron and Trivedi removes much of the bias in this case, at the cost of increased variance.[57]



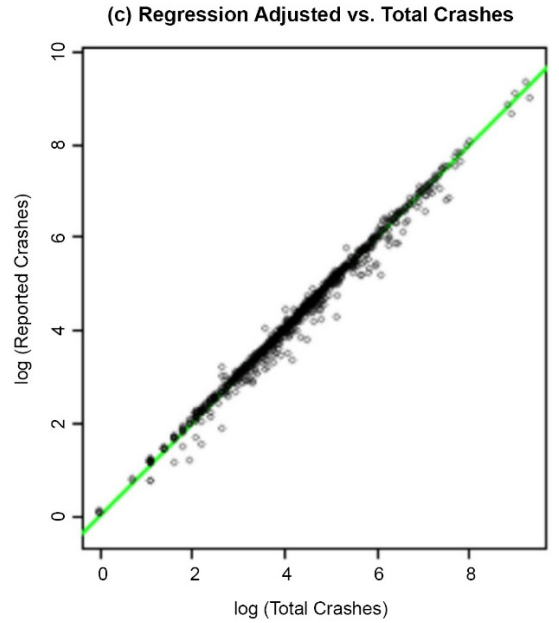
Source: FHWA.

Figure 40. Graph. Comparison of adjustments to reported crashes.



Source: FHWA.

Figure 41. Graph. Comparison of simple adjustments to reported crashes.



Source: FHWA.

Figure 42. Graph. Comparison of regression adjustments to reported crashes.

On average, in the NYSDOT data, 87.7 percent of all crashes were reported. A flat, multiplicative adjustment to all reported crashes (T_i equal to $1.14R_i$) unbiases the reported number of crashes, but a logistic regression–based adjustment of reported crashes results in more accurate predictions of total crashes from reported crashes and location information.

SUMMARY

The results of the NYSDOT analysis provide several noteworthy findings as shown in table 15 through table 19. First, directly modeling reported-crash frequencies without accounting for underreporting can lead to bias in the number of predicted crashes. This bias can be removed by considering a logistic regression-based estimate of the reporting probability dependent on road segment (or other location-specific) characteristics. This process requires a sample (which may be small) of road segments on which both the reported and total crash frequencies are known. This information is often difficult to obtain but critical to fully understanding crash frequencies.

Additionally, directly modeling reported-crash frequencies without accounting for underreporting can lead to bias in estimates of the correlation between road-segment characteristics and crash frequency. This bias was small in the NYSDOT data (figure 39), which may indicate that NB-regression models fit to reported-crash frequencies will provide fairly reliable inference on correlations between road segment characteristics and crash frequency.

Table 15. NB regression on T_i for model 1.

Statistical Variable Notation	Est.	Lower CI	Upper CI
Intercept	−4.322*	−4.435*	−4.209*
PCT.Rural	0.028	−0.005	0.062
State.Jur	−0.009	−0.057	0.040
Road.Density	0.107*	0.099*	0.115*
PCT.Miles.Div	0.354*	0.204*	0.503*
PCT.Miles.OW	−0.166	−0.401	0.070
Urban.MuniType	0.587*	0.518*	0.657*
log*VMT*	0.787*	0.778*	0.796*
*year*2008	0.013	−0.023	0.050
*year*2009	0.016	−0.021	0.052
*year*2010	0.022	−0.015	0.059
NB dispersion	4.095*	3.942*	4.248*
AIC	62,061	—	—
Deviance	6,404.6	—	—

*Parameters significantly different from zero.
—No data.
Est. = estimated; CI = confidence interval; *PCT.Rural* = percentage rural; *Road.Density* = road density; *PCT.Miles.Div* = percentage miles divided; *PCT.Miles.OW* = percentage miles one way; *Urban.MuniType* = urban municipality type; log*VMT* = logarithm of vehicle miles traveled; *year*2008 = indicator variable for the year 2008; *year*2009 = indicator variable for the year 2009; *year*2010 = indicator variable for the year 2010; *NB dispersion* = estimate for *a*; AIC = Akaike Information Criterion.
Note: Upper and lower 95-percent confidence interval bounds are also given.

Table 16. NB regression on R_i for model 2.

Statistical Variable Notation	Est.	Lower CI	Upper CI
Intercept	−4.242*	−4.353*	−4.132*
PCT.Rural	0.036*	0.003*	0.068*
State.Jur	−0.003	−0.050	0.044
Road.Density	0.083*	0.075*	0.091*
PCT.Miles.Div	0.329*	0.183*	0.474*
PCT.Miles.OW	−0.123	−0.353	0.107
Urban.MuniType	0.584*	0.517*	0.652*
log*VMT*	0.779*	0.771*	0.788*
*year*2008	−0.023	−0.059	0.013
*year*2009	−0.030	−0.066	0.006
*year*2010	0.003	−0.033	0.039
NB Dispersion	4.323*	4.160*	4.486*
AIC	61061	—	—
Deviance	6403.6	—	—

*Parameters significantly different from zero.
—No data.
Est. = estimated; CI = confidence interval; *PCT.Rural* = percentage rural; *State.Jur* = State jurisdiction; *Road.Density* = road density*; PCT.Miles.Div* = percentage miles divided; *PCT.Miles.OW* = percentage miles one way; *Urban.MuniType* = urban municipality type; log*VMT* = logarithm of vehicle miles traveled; *year*2008 = indicator variable for the year 2008; *year*2009 = indicator variable for the year 2009; *year*2010 = indicator variable for the year 2010*; NB dispersion* = estimate for *a*; AIC = Akaike Information Criterion.
Note: Upper and lower 95-percent confidence interval bounds are also given.

Table 17. NB regression on R_i with underreporting for model 3.

Statistical Variable Notation	Est.	Lower CI	Upper CI
Intercept	−4.329*	−4.497*	−4.162*
PCT.Rural	0.029	−0.027	0.084
State.Jur	−0.027	−0.085	0.032
Road.Density	0.084*	0.076*	0.091*
PCT.Miles.Div	0.354	−0.056	0.765
PCT.Miles.OW	−0.167	−0.416	0.081
Urban.MuniType	0.595*	0.500*	0.689*
log*VMT*	0.789*	0.777*	0.802*
*year*2008	0.000	−0.093	0.093
*year*2009	0.010	−0.090	0.110
*year*2010	0.027	−0.105	0.159
NB Dispersion	0.231*	0.142*	0.240*
AIC	62082	—	—
Deviance	6410.6	—	—

*Parameters significantly different from zero.
—No data.
Est. = estimated; CI = confidence interval; *PCT.Rural* = percentage rural; *State.Jur* = State jurisdiction; *Road.Density* = road density*; PCT.Miles.Div* = percentage miles divided; *PCT.Miles.OW* = percentage miles one way; *Urban.MuniType* = urban municipality type; log*VMT* = logarithm of vehicle miles traveled; *year*2008 = indicator variable for the year 2008; *year*2009 = indicator variable for the year 2009; *year*2010 = indicator variable for the year 2010*; NB dispersion* = estimate for *a*; AIC = Akaike Information Criterion.
Note: Upper and lower 95-percent confidence interval bounds are also given.

Table 18. Logistic regression on R_i for model 4.

Statistical Variable Notation	Est.	Lower CI	Upper CI
Intercept	6.526*	6.472*	6.579*
PCT.Rural	0.126*	0.113*	0.140*
State.Jur	0.467*	0.451*	0.484*
Road.Density	−0.037*	−0.041*	−0.034*
PCT.Miles.Div	−0.922*	−0.978*	−0.865*
PCT.Miles.OW	2.647*	2.516*	2.779*
Urban.MuniType	−0.243*	−0.262*	−0.225*
log*VMT*	−0.303*	−0.306*	−0.299*
*year*2008	−0.738*	−0.755*	−0.720*
*year*2009	−1.036*	−1.0538*	−1.019*
*year*2010	−0.601*	−0.619*	−0.583*
NB Dispersion	—	—	—
AIC	193224	—	—
Deviance	184884	—	—

*Parameters significantly different from zero.
—No data.
Est. = estimated; CI = confidence interval; *PCT.Rural* = percentage rural; *State.Jur* = State jurisdiction; *Road.Density* = road density*; PCT.Miles.Div* = percentage miles divided; *PCT.Miles.OW* = percentage miles one way; *Urban.MuniType* = urban municipality type; log*VMT* = logarithm of vehicle miles traveled; *year*2008 = indicator variable for the year 2008; *year*2009 = indicator variable for the year 2009; *year*2010 = indicator variable for the year 2010*; NB dispersion* = estimate for *a*; AIC = Akaike Information Criterion.
Note: Upper and lower 95-percent confidence interval bounds are also given.

Table 19. NB regression on R_i with underreporting for model 3.

Statistical Variable Notation	Est.	Lower CI	Upper CI
Intercept	0.029	−0.027	0.084
PCT.Rural	−0.027	−0.085	0.032
State.Jur	0.084*	0.076*	0.091*
Road.Density	0.354	−0.056	0.765
PCT.Miles.Div	−0.167	−0.416	0.081
PCT.Miles.OW	0.595*	0.500*	0.689*
Urban.MuniType	0.789*	0.777*	0.802*
log*VMT*	0.000	−0.093	0.093
*year*2008	0.010	−0.090	0.110
*year*2009	0.027	−0.105	0.159
*year*2010	6.530*	2.240*	10.821*
NB Dispersion	—	—	—
AIC	62082	—	—
Deviance	6410.6	—	—

*Parameters significantly different from zero.
—No data.
Est. = estimated; CI = confidence interval; *PCT.Rural* = percentage rural; *State.Jur* = State jurisdiction; *Road.Density* = road density*; PCT.Miles.Div* = percentage miles divided; *PCT.Miles.OW* = percentage miles one way; *Urban.MuniType* = urban municipality type; log*VMT* = logarithm of vehicle miles traveled; *year*2008 = indicator variable for the year 2008; *year*2009 = indicator variable for the year 2009; *year*2010 = indicator variable for the year 2010*; NB dispersion* = estimate for *a*; AIC = Akaike Information Criterion.
Note: Upper and lower 95-percent confidence interval bounds are also given.

APPENDIX C. PROBABILISTIC LINKAGE OF HOSPITAL AND CRASH DATA FROM UTAH

Definitions for variables used in this appendix are provided in table 20.

Table 20. Definitions for variables used in appendix C.

Variable	Definition
A	Event
current odds	Current odds of selecting a correct match at random
desired odds	Desired odds of selecting a correct match
E	Number of true matches
i	Field
m_i	Reliability power of a specific field
N	A number
P	Probability
p	Probability of selecting a correct match
u_i	Discriminating power of a field
$w_{0.5}$	Weight associated with probability of correct match equal to 0.5
$w_{0.9}$	Weight associated with probability of correct match equal to 0.9
w_i	Agreement weight
w_t	Composite weight

PURPOSE

MVCs are a serious public health problem. In 2015, more than 35,000 persons were killed, and an estimated additional 2.44 million persons were injured in MVCs in the United States.[58] Both of these statistics represent increases after years of decline. State Highway Safety Offices and Departments of Transportation have implemented a number of programs to address the human toll of MVCs. Data are needed to identify at-risk populations and roadway locations or conditions that may pose risky driving environments and lead to high injury and fatality rates. Traditionally, States use two main sources of data to support and evaluate programs and interventions: MVC fatalities and State crash reports.

FARS is a census of all persons and vehicles involved in a reportable MVC in which a person died within 30 d of the event.[12] Among the strengths of FARS data is that death represents the most serious outcome of an MVC and has a consistent definition from State to State. FARS standardizes data between States, making national level analyses and between-State comparisons possible. Some States experience significant delays in the availability of the State crash file, making FARS data an attractive alternative for analysis. Depending on the questions of interest, some limitations can be associated with FARS analyses. First, deaths resulting from MVCs are relatively rare events. Focusing on annual rates among subpopulations or in smaller geographic regions can lead to highly variable results simply due to a small increase or decrease in a year. Further, MVC fatalities are typically not representative of the whole MVC population. For example, those killed in MVCs are more likely to be male, younger, and less likely to use safety restraints than those not killed.[59]

Individual State crash reports are a means of overcoming the limitations associated with FARS. State crash reports capture a person's injury severity using the KABCO scale. As defined by the *Model Minimum Uniform Crash Criteria 3rd Editio*n (MMUCC), KABCO's values are as follows: K is fatal injury, A is incapacitating injury, B is nonincapacitating injury, C is possible injury, and O is no injury.(60) *Traffic Safety Performance Measures for States and Federal Agencies* provides a set of outcome metrics for States to track decreases in serious injuries.(61) Many States have adopted a definition of serious injuries to include all K- and A-level injuries. States use this definition to plan programs for vulnerable populations and identify high-risk crash locations. One limitation of KABCO is that the injury severity is assigned at the scene of the crash before a full medical assessment can be performed at the hospital. Additionally, while State crash files provide a full picture of the MVC population, making comparisons between States difficult. Not all States have chosen to adopt the MMUCC guideline for their crash reports. Even if two States use the MMUCC guidelines, differences may exist in the definitions or operationalization of the coding of injuries.

An alternate definition of serious injury can be made using nontraditional MVC-related data sources. This study examines the utility of State-collected hospital billing data to identify injured MVC participants. Probabilistic linkage is also utilized as a means of combining information from the crash report and hospital billing data. Several examples of potential analyses and outcome measures are presented.

DATA SOURCES

The research team used three datasets in this analysis: the Utah MVC, ED, and hospital inpatient discharge databases from 2001 through 2013. Use of these databases was approved by the University of Utah Institutional Review Board (IRB)[1].

The research team obtained the Utah MVC database from the UDOT Traffic and Safety Division. Police crash reports are completed at the scene of the MVC and compiled by the Utah Department of Public Safety. Additional processing and roadway features are added to the database by UDOT. This database contains information on all reportable MVCs (at least one fatality or injury or more than $1,500 in property damages) on public roadways in Utah. Data include information regarding the time, location, and type of crash; the vehicles and drivers involved; and the age, sex, seating location, safety-restraint use, and injury severity of all persons.

The research team obtained ED and hospital inpatient records from the Utah Department of Health, Office of Health Care Statistics, to which all licensed hospitals in Utah report ED and inpatient data. The ED database contains information about persons treated at the ED, while the inpatient database contains information on persons who are admitted to the hospital as inpatients. Both databases have the same structure and data elements. Data include information about the patient, such as their age and sex, injuries, and healthcare. Each record can include up to nine International Classification of Diseases 9th Revision Clinical Modification (ICD-9-CM) codes to describe a person's injuries. Additionally, two External Cause of Injury codes (E Codes) can be

[1]The three datasets are unpublished data that were provided to the research team.

included to describe the injurious event. Additional information in the ED and inpatient databases includes the person's length of stay, billed hospital charges, and discharge status.

The presence of ICD-9-CM codes allows for the calculation of hospital-based injury severity measures, such as the AIS.[62] AIS scores are derived from information collected during hospital evaluation and have been shown to correlate with mortality. An AIS score can be calculated from ICD-9-CM codes using specialized software. Each ICD-9-CM code is assigned to a body region (head, face, neck, thorax, abdomen, spine, upper extremity, lower extremity, or unspecified) and severity level. To arrive at a single score for a patient visit, the MAIS value is calculated as the maximum value of AIS across all body regions (table 21).

Table 21. AIS levels of injury severity and risk of mortality.

AIS Code	Injury Severity	Probability of Death (%)
1	Minor	0
2	Moderate	1–2
3	Serious	8–10
4	Severe	5–50
5	Critical	5–50
6	Maximum	100
9	Not further specified	—

—No data.

METHODS

Probabilistic Linkage

For the purpose of this research, the research team relied heavily on previously published descriptions of the linking process.[63–65] Probabilistic record linkage is accomplished by comparing data fields, such as the date of birth or gender of a patient, in two different files. The comparison of numerous fields leads to a judgment that two records refer to the same person and event and should be linked or the records do not refer to the same person and event and should not be linked. This judgment is based on the cumulative agreement and disagreement of field values. Data fields that are compared have differing impacts on a judgment that two records should be linked. For instance, agreement of the gender field alone would not suffice to conclude that two records refer to the same patient, while agreement of social security number alone greatly enhances the probability that two records refer to the same individual. By assigning log-likelihood values to comparisons, calculating match weights and computerizing the judgment process is possible. The calculation of match weights is based on two probabilities: reliability (the probability that field *i* will agree given two records are known to be a true match) and discriminating power (the probability that field *i* will agree given that two records are known not to match). It is customary to represent the reliability and discriminating power of field *i* as m_i and u_i, respectively. For a given pair of records, if field *i* agrees, the likelihood the records match is m_i/u_i, and the agreement weight (w_i) will be calculated as shown in figure 43.

$$w_i = \log_2\left({}^{m_i}/_{u_i}\right)$$

Figure 43. Equation. Calculation of agreement weight when fields match.

If field *i* disagrees, the likelihood that the records match is $(1 - m_i)/(1 - u_i)$, and the agreement weight is calculated as shown in figure 44.

$$w_i = \log_2\left({(1 - m_i)}/{(1 - u_i)}\right)$$

Figure 44. Equation. Calculation of agreement weight when fields do not match.

The composite weight (w_t) for a record pair is the sum of agreement weights across all data fields used for comparison. As w_t increases, the probability that two records refer to the same MVC participant increases. As w_t decreases, the probability that the records refer to the same MVC participant also decreases.

Both m_i and u_i are theoretical quantities and are rarely known prior to conducting a linkage. The m_i probabilities can be estimated through a variety of techniques. The first technique requires a historical knowledge of the databases being linked. If no major modifications have been made to either database since a previous year's linkage, then m_i values from that linkage can be carried forward as an estimate for the new linkage. If faced with a major modification to one of the databases or linking a new database, then the past year's linkage will be of little use. In this case, one can usually obtain estimates of m_i values from the data owners. In the absence of any information regarding reliability, an estimate can be obtained by linking 1 mo of data to calculate initial reliability estimates to inform the full year's linkage. The u_i values are estimated from the data being linked. Typically, the set of expected matched pairs is negligible in size compared to the set of all possible matched pairs between two databases. Thus, u_i for a given field can be estimated by taking a sample of random pairs and determining how often field *i* agrees.

Since match weights are derived from log odds it is possible to determine the weight needed to achieve a specified probability that two linked records are a true match. The following figures were adapted from formulas developed by McGlincy and supply the background necessary to determining the specified probability.[65] The odds of an event (A) are defined in figure 45.

$$\frac{Probability\ that\ A\ occurs}{Probability\ that\ A\ does\ not\ occur} = \frac{P(A)}{1 - P(A)}$$

Figure 45. Equation. Odds of *A* occurring.

This equation can be rearranged as shown in figure 46.

$$P(A) = \frac{Odds\ of\ A}{1 + Odds\ of\ A}$$

Figure 46. Equation. Probability of *A* occurring.

Using the equation in figure 46, it is possible to calculate the odds and probability of picking a matched pair at random. Given two files with number of records $N1$ and $N2$, respectively, the number of possible record pairings is $N1 \times N2$. If the number of true matches (E) of $N1 \times N2$ pairings are true matches (note that E must be less than both $N1$ and $N2$ since the number of true matches cannot exceed the minimum of the two file sizes), the probability of picking a true match at random is defined in figure 47.

$$P(E) = \frac{E}{N1 \times N2}$$

Figure 47. Equation. Probability of picking a true match.

The odds of picking a true match at random are shown in figure 48.

$$\frac{P(E)}{1 - P(E)} = \frac{\frac{E}{N1 \times N2}}{1 - \frac{E}{N1 \times N2}} = \frac{E}{N1 \times N2 - E}$$

Figure 48. Equation. Odds of picking a true match.

This equation produces a small numeric value since the number of possible record pairings greatly exceeds the number of valid matches. For example, an analyst has two files: file 1 and file 2. There are 1,000 records in file 1 and 1,000 records in file 2. Every record in file 1 is known to uniquely match a record in file 2. Using the equation in figure 48, one can calculate the odds of picking a true match at random in this scenario to be 0.001 or 1 in 1,000 tries (figure 49).

$$\left(\frac{1{,}000}{1{,}000 \times 1{,}000 - 1{,}000} = 0.001\right)$$

Figure 49. Equation. Odds of picking a true match when each file has 1,000 records and 1,000 matches are expected.

Using equation in figure 47, the probability of picking a true match at random is also approximately 0.001.

Researchers need to know how much information is needed to improve the probability of selecting true matches to 0.90. To attain this result, figure 50 uses the general equation shown in figure 48.

$$\frac{0.90}{1 - 0.90} = 9.0$$

Figure 50. Equation. Odds associated with a probability of 0.90.

The ratio of the desired odds to the current odds reveals how much the odds must improve to obtain the desired probability, 0.90. The ratio of the desired odds and the current odds is shown in figure 51, which illustrates that the current odds must increase by a factor of 9,000 to improve the probability of picking correct matches from 0.001 to 0.90. The $\log_2(9{,}000)$ expresses the needed improvement in odds as an improvement in match weight. To increase the probability of selecting a true match, the match weight must increase from its current value of $\log_2(0.001)$, or −9.97, to $\log_2(9{,}000)$, or 13.14. Therefore, only accepting matched pairs that have a match weight of 13.14 will yield a probability of selecting correct matches of at least 0.90.

$$\frac{9}{0.001} = 9000$$

Figure 51. Equation. Ratio of desired odds to current odds.

Using the same notation as figure 48 for $N1$, $N2$, and E and denoting the desired probability of selecting a correct match as p, the w_t that corresponds to p of a true match can be expressed as shown in figure 52.

$$w_t = \log_2\left(\frac{desired\ odds}{current\ odds}\right) = \log_2 \frac{\frac{p}{1-p}}{\frac{E}{N1 \times N2 - E}}$$

Figure 52. Equation. Weight factor of selecting a correct match.

Where:

desired odds = desired odds of selecting a correct match.
current odds = current odds of selecting a correct match at random.

The equation in figure 52 can now be used to determine cut points for true matches, false matches, and the clerical review region. One option, for instance, would be to use the equation in figure 52 to determine the weights associated with probabilities of correct matches equal to 0.9 and 0.5, $w_{0.9}$ and $w_{0.5}$ respectively. All pairs with a weight above $w_{0.9}$ would then be considered true matches, all pairs with a weight below $w_{0.5}$ would be considered false matches, and all pairs between $w_{0.9}$ and $w_{0.5}$ would be manually reviewed. To eliminate the human element of clerical review, another option is to choose a single cut point, such as a match probability of 0.9, and consider all pairs of records above the threshold to be true matches and all pairs of records below the threshold to be false matches.

For the purpose of this research, the research team retained all pairs attaining a match probability of 0.9 as true matches and rejected all others. Linkage variables available included first and last name, date of birth (or age), sex, date of crash or hospital treatment, county of the crash, and county of hospital.

SOFTWARE

Probabilistic linkage of the crash and hospital databases was performed using Linksolv 8.3.[66] AIS and MAIS were calculated using ICDMAP90.[67,68]

RESULTS

From 2010 through 2013, 494,995 people were involved in MVCs in Utah. Table 22 through table 25 provide a numerical description of injury severity measures for the crash population. Using the crash-report injury severity scale, KABCO, 17.3 percent of crash participants were killed, injured, or possibly injured. Adopting the KABCO definition of severe injury (incapacity or fatal injury), 6,052 (1.2 percent) crash participants were seriously injured. The research team analyzed other outcomes by taking into account the results of the linkage. First, based on the linkage itself, the highest level of care is defined based on the most severe database that the

person's MVC record matched. For instance, if a person was treated at a local ED and then transported and admitted to a trauma center, their highest level of care would be hospital admission. Fatality is defined as those coded as a fatality on the MVC report or coded as having died at the hospital. Table 23 shows that only about 1 in 10 persons involved in an MVC was treated at the hospital or died as result of their injuries. Just under 10 percent (48,114) of persons were treated at the ED, another 0.9 percent (4,332) were admitted to the hospital, and 0.2 percent (958) died. Table 24 displays the results of the MAIS analysis. There were 449,781 persons without an MAIS. These cases represent the 441,591 people who did not match to a hospital record and an additional 8,190 people whose hospital records did not contain enough information to calculate MAIS. A traditional definition of severe injury based on MAIS is a score of two or higher. This definition results in a percentage, 1.7 percent, of severe injuries similar to KABCO.

For those treated at a hospital, table 25 summarizes their observed discharge statuses. The majority of hospital-treated patients were discharged to home (96.4 percent). However, 585 (1.1 percent) were transferred to another healthcare facility, 476 (0.9 percent) were transferred to long-term care, 368 (0.7 percent) were transferred to rehab, and 319 (0.6 percent) died. MVC participant hospitalizations accounted for 21,960 d in the hospital, with a median of 2 d and more than $365 million in billed charges. The interquartile range for the length of stay was 1 and 3 d, while the interquartile range for hospital charges was $904 and $4,724.

Table 22. Crash and hospital-reported injury outcomes for KABCO.

KABCO	Count (People)	Percent
Not injured	409,184	82.7
Possible injury	51,159	10.3
Nonincapacitating injury	28,600	5.8
Incapacitating injury	5,119	1.0
Fatality	933	0.2

Table 23. Crash and hospital-reported injury outcomes for highest level of care.

Highest Level of Care	Count (People)	Percent
No hospital treatment	441,591	89.2
ED	48,114	9.7
Hospital admission	4,332	0.9
Fatality	958	0.2

Table 24. Crash and hospital-reported injury outcomes for MAIS.

MAIS	Count (People)	Percent
No score	449,781	90.9
1	36,692	7.4
2	5,613	1.1
3	2,663	0.5
4	178	0.04
5	51	0.01
6	17	<0.01

Table 25. Crash and hospital-reported injury outcomes for discharge status.

Discharge Status	Count (People)	Percent
Home	49,972	96.4
Transferred	585	1.1
Long-term care	476	0.9
Rehabilitation	368	0.7
Left against medical advice	121	0.2
Died	319	0.6

Core outcome measures described in *Traffic Safety Performance Measures for States and Federal Agencies* are summarized in table 26.[59] Unrestrained occupants had a higher severe-injury rate compared to the general population using both the KABCO scale (5.7 versus 1.2 percent) and MAIS (6.0 versus 1.7 percent). The research team obtained similar estimates for severe injuries using crash and hospital severe-injury definitions for both alcohol- or drug- and speeding-related crashes. The greatest difference in estimates was seen with motorcycle crashes, where KABCO estimates 19.1 percent of motorcyclists were severely injured, while the MAIS estimate is 28.0 percent. The research team used hospital-based outcomes to assess the impact of these core areas in other ways. For instance, while motorcyclists accounted for 1.2 percent of the population, they account for 14.6 percent of all persons who are treated at the hospital or die. Motorcyclists also accounted for 22.1 percent of all hospital days and 17.4 percent of all hospital charges. Similarly, while unrestrained occupants accounted for 2.6 percent of the population their hospitalizations represent 10.1 percent of the population. Analyzing table 26 (based on hospital outcomes), motorcyclists accounted for the highest number of hospital days and charges, followed by speed-related crashes, alcohol- or drug-related crashes, and unrestrained occupants.

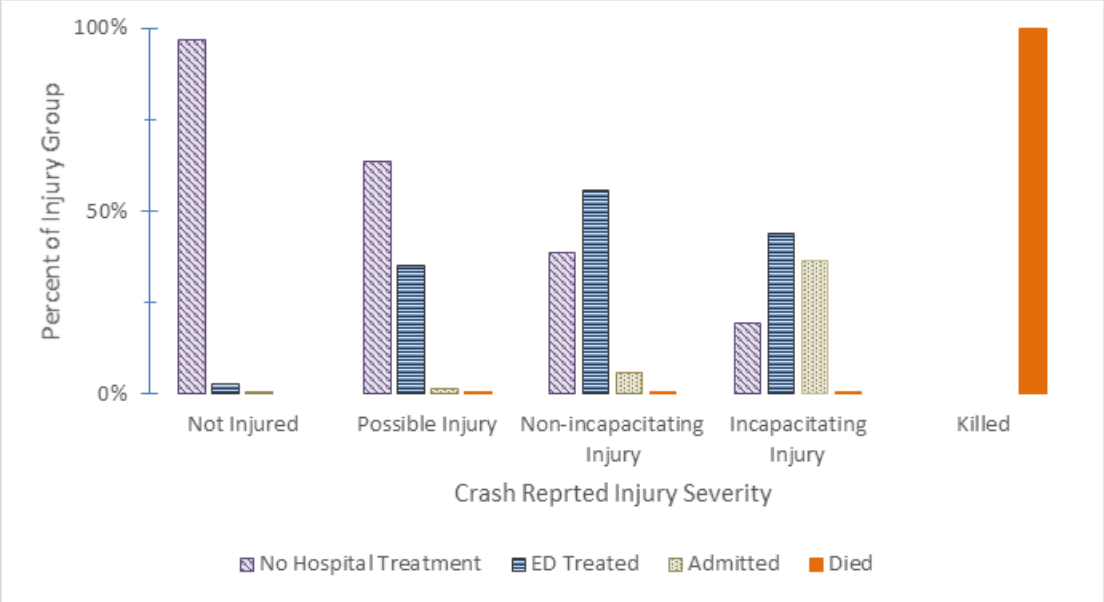
Table 26. Hospital outcomes by performance measures.

Measure	Count (People)	In K or A Crashes	Hospital Treated	AIS (2+)	LOS (d)	Charges ($)
All persons	494,995	6,052 (1.2%)*	53,404 (10.8%)*	8,522 (1.7%)*	21,960	365,323,869
Unrestrained occupants	12,932	742 (5.7%)*	2,815 (22.8%)*	775 (6.0%)*	2,658	36,944,640
Alcohol/drug-related crashes	17,481	982 (5.6%)*	4,221 (24.1%)*	965 (5.5%)*	3,325	49,960,716
Speed-related crashes	64,564	1,157 (1.8%)*	7,802 (12.1%)*	1,399 (2.2%)*	3,772	61,665,167
Motorcycle riders	5,797	1,111 (19.1%)*	3,200 (55.2%)*	1,626 (28.0%)*	4,860	63,373,783
Unhelmeted riders	2,140	498 (23.3%)*	1,292 (60.4%)*	657 (30.7%)*	1,889	25,214,794
Young driver (≤20 yr)–related crashes	29,472	690 (2.3%)*	3,693 (12.5%)*	866 (2.9%)*	2,275	36,358,555
Pedestrians	2,884	525 (18.2%)*	1,583 (54.9%)*	626 (21.7%)*	2,298	29,505,368

*Percentage of count (people) involved.
LOS = length of stay.

Crash and Hospital Oucomes Comparison

While crash-reported injury severity and hospital treatment are related, it was clear that there is variation within individual levels of KABCO (figure 53). While nearly 97 percent of MVC participants coded as not injured did not receive hospital treatment, almost 3 percent were treated and released from the ED, and about 0.5 percent were admited to the hospital. The categories of nonincapcitating and incapacitating injuries were of particular interest. While an injury may appear to be nonincapacitating at the crash scene, over half (55.5 percent) were treated at the ED, and an additional 5.7 percent were admitted to the hospital. For those coded as incapicitated, 43.8 percent were trreated at the ED and 36.4 percent were admitted. One-third (38.7 percent) of nonincapacitating and almost one-fifth (19.6 percent) of incapacitating injuries did not match with a hosptial record.



Source: FHWA.

Figure 53. Graph. Highest level of care by crash-reported injury, KABCO.

As the level of injury severity increased, so did the severity of hospital outcomes (table 27). Of the more than 400,000 persons coded as not injured, 12,000 (3.1 percent) linked to hospital records, and their visits resulted in over $28 million in hospital charges. These hospitalizations appeared to be for minor injuries. The 25th, 50th, and 75th percentiles for MAIS are all one, and the median charge is $1,036. Similarly, a nonsignificant number of persons coded with possible injuries (18,510) were linked to hospital visits. These injuries appear to be similar to those coded as not injured with a 75th percentile for MAIS of one but the median hospital charge has increased by 50 percent to $1,568. Focusing on persons who are considered severely injured using KABCO, those with incapacitating injuries account for the most total charges, more than $140 million, in any of the KABCO levels. The median hospital charge, $12,834, was also more than 10 times that of those coded as not injured. The median MAIS is 2 and the 75th percentile is 3. The majority of persons coded as being fatally injured in the crash appear to never make it to the hospital. However, the visits for those who receive hospital treatment, 363 people (38.9 percent) were among the most expensive, with median charges of $14,114.

Table 27. Hospital outcomes by level of KABCO.

Injury Severity	Count (People)	People Linked to Records (%)	Median MAIS (Q1, Q3)	Severe Injuries (%)	Total Charges ($)	Median Charges (Q1, Q3)
Not injured	409,184	12,319 (3.1%)	1 (1, 1)	544 (0.1%)	28,398,309	$1,036 ($511, $2,033)
Possible injury	51,159	18,510 (10.3%)	1 (1, 1)	1,575 (3.1%)	62,506,906	$1,568 ($848, $3,446)
Nonincapacitating injury	28,600	17,516 (62.3%)	1 (1, 1)	3,654 (12.8%)	115,565,792	$2,647 ($1,221, $6,154)
Incapacitating injury	5,119	4,112 (80.3%)	2 (1, 3)	2,475 (48.3%)	143,821,679	$12,834 ($4,606, $35,982)
Fatality	933	363 (38.9%)	3 (3, 3)	274 (29.4%)*	15,031,182	$14,114 ($6,560, $5,527)

*The calculation of MAIS is only available for the 363 fatally injured persons who linked to a hospital record.
Q1 = first quartile; Q3 = third quartile.

Consistency of MAIS Over Time

Next, the research team investigated how changes to a State crash report can impact the estimation of serious injuries. In 2006, Utah released a MMUCC-compliant crash report. As part of the revision, the labels associated with KABCO were redefined (table 28). The two main changes occurred in levels A and B. Level A changed from "Broken Bones and Bleeding" to "Incapacitating Injury" and level B changed from "Bruises and Abrasions" to "Non-incapacitating Injury."[70,71]

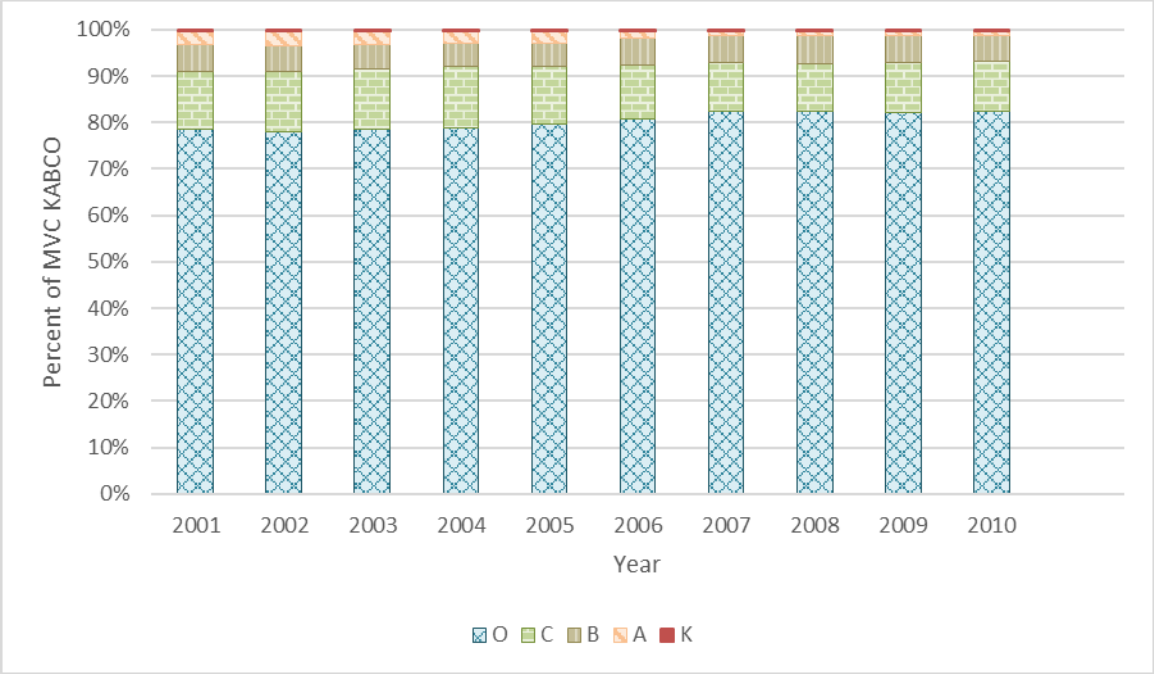
Table 28. Coding of Utah crash-reported injury severity, per KABCO, before and after crash-report revision.

KABCO Level	Pre-2006	Post-2006
K	Fatal injury	Fatal injury
A	Broken bones and bleeding	Incapacitating injury
B	Bruises and abrasions	Nonincapacitating injury
C	Possible injury	Possible injury
O	Not injured	Not injured

The effect of changing the definition of KABCO was immediately apparent (figure 54). Starting in 2006, the percent of A-level injuries dropped by roughly half. Corresponding to this decline was an increase in B-level injuries. The shift in KABCO coding had a dramatic impact on the estimate of serious injuries. Table 29 shows that the KABCO serious-injury rate dropped from 2.9 to 1.7 percent just from the change in forms. By comparison, the MAIS serious-injury rate was consistent over time, regardless of form changes. Because MAIS is not dependent on the specific crash-report in use, and the standardized coding of ICD-9-CM codes, MAIS may prove to be a useful means of comparing outcome measures between States.

Table 29. Percent of crash- and hospital-reported severe injuries by year before and after Utah crash-report revision.

Scale	2001 (%)	2002 (%)	2003 (%)	2004 (%)	2005 (%)	2006 (%)	2007 (%)	2008 (%)	2009 (%)	2010 (%)
KABCO	3.3	3.5	3.2	2.9	2.9	1.7	1.3	1.3	1.2	1.2
MAIS	1.8	1.9	2.1	1.8	1.7	1.7	1.9	1.7	2.1	1.8



Source: FHWA.

Figure 54. Graph. Distribution of crash-reported injury severity, per KABCO, before and after Utah crash-report revision.

Match Probabilities

Any probabilistic linkage study is dependent on the quality of the linked pairs. In this study of crashes from 2010 through 2013, the median probability of a true match between the crash and ED files was 0.9999998 (interquartile range (IQR) 0.9999997, 0.9999999). Linked pairs in the crash to hospital discharge linkage resulted in a median match probability of 0.9999855 (IQR 0.9992621, 0.9999996). By subtracting the match probability from one, the research team obtained an estimate of the probability that each pair of records is a false match. Summing the false match probabilities allowed for the calculation of the false-match rate. For the crash-to-ED linkage, there were an estimated 28 false matches and a false-match rate of <0.1 percent. The crash-to-hospital-discharge linkage resulted in an estimated 21 false matches and a false match rate of 0.1 percent.

CONCLUSIONS

The research team examined the utility of hospital-injury outcome measures as a means of estimating the number and severity of injuries. The results suggested hospital-injury outcomes offer many advantages over crash-reported outcomes.

KABCO is assigned at the scene of the crash before a full medical evaluation has taken place. Many severe internal injuries may not be immediately recognizable without advanced imaging available at hospitals and trauma centers. Additionally, some injuries that initially appear to be quite severe may be classified as minor after the patient is evaluated and bleeding controlled. The research team observed a positive correlation between KABCO-reported severity and

hospital-reported injury severity (MAIS) in the Utah data; however, there were differences within the individual levels of KABCO. For instance, more than 3,000 persons per yr coded as not injured on the crash report still received hospital treatment. Conversely, one in five persons coded as a severe injury by KABCO did not receive hospital treatment.

Hospital outcomes allow for a means of quantifying injuries beyond counts of injured persons. In this study, the research team linked more than 50,000 persons to hospital records. These hospital visits resulted in over 20,000 hospital d and $350 million in billed hospital charges. The linkage of hospital outcomes to the State crash report enables the quantification of the burden of injury due to a number of crash, vehicle, and personal factors. Number of hospitalizations, hospital days, and the amount of billed charges can aid in honing safety messages when trying to quantify the impact of certain crash factors, such as safety belt non-use, alcohol or drugs, and speeding, or subpopulations, such as pedestrians or motorcyclists.

Finally, hospital outcomes can be used as a means for comparing the burden of MVCs between States. While States have control over the design, reporting threshold, and definitions of variables collected on crash reports, hospital billing databases are governed by Uniform Billing Standards, which results in consistently collected data elements with standardized definitions. In this study, the research team demonstrate a nearly 50-percent drop in the crash-reported severe-injury rate simply by a redesign in the data-collection instrument. Over the same time period, the MAIS injury severity rate remained consistent. The stability of MAIS injury severity rates between States has previously been reported.[41]

Using hospital injury outcome measures has limitations. One of the biggest barriers is the linkage of the databases. Probabilistic linkage is a technical process, and its successful completion requires background information to build the linkage model. Acquiring the datasets can sometimes prove challenging. A number of Federal and State privacy regulations regard the use and distribution of healthcare data. Often, the approval process for accessing hospital data can involve gaining approval from multiple IRBs and signing data use agreements. The timeliness of the datasets being linked also can be a barrier to the practicality of a linkage project. For instance, if data related to a crash are available for analysis within 2 weeks of the event but the statewide hospital discharge database is not released for at least 12 mo following the end of the calendar year, then the linked database will by default be at least 1 yr older than the MVCs being studied. A final limitation of analyzing hospital outcomes is that the outcomes only exist for those participants who were linked with a hospital record. In this study, 441,591 persons did not link to a hospital record; therefore, did not have an MAIS score. It is likely that the overwhelming majority of these persons were not injured and, therefore should not have an MAIS score. However, there may be injured persons who chose not to seek hospital treatment for their injuries; therefore, their MAIS severity is unknown. Other persons may not link due to inaccuracies in their data or were transported to an out-of-State hospital.

Despite these limitations associated with probabilistic linkage, the benefits of the added information gained through the use of hospital outcomes make this a method worth pursuing for future studies.

APPENDIX D. EXAMPLE APPLICATIONS OF CARTS AND RANDOM FORESTS FOR STATISTICAL ROAD-SAFETY ANALYSIS

Definitions for variables used in this appendix are provided in table 30.

Table 30. Definitions for variables used in appendix D.

Variable	Definition
ADT	Average daily traffic
ADTen	Average daily traffic volume on entrance ramp
ADTex	Average daily traffic volume on exit ramp
AuxLn	Indicator variable of auxiliary lane connecting entrance and exit ramps
c	Specific leaf
CA_5	Indicator variable for study segments located on I-5 in California
CA_10	Indicator variable for study segments located on I-10 in California
HOVen	Indicator variable for presence of high-occupancy-vehicle lane on entrance ramp
HOVmain	Indicator variable for presence of high-occupancy-vehicle lane on freeway mainline
i	Observation
Intercept	Intercept of the regression model
InvSpa	Inverse of ramp spacing
InvSpaAux	Interaction variable for inverse spacing and presence of auxiliary lane
j	Classes
K	Random, equal divisions of a training observation
L	Segment length
leaves(*T*)	Set of leaves for a tree
MAE	Mean absolute error
Mainline1	Indicator variable for relative vertical position between freeway mainline and cross street associated with entrance ramp
Mainline2	Indicator variable for relative vertical position between freeway mainline and cross street associated with exit ramp
m_c	prediction for a specific leaf
MV-KABC	Number of reported multiple-vehicle crashes resulting in at least one fatality or injury of any level
MV-O	Number of reported multiple-vehicle crashes resulting in PDO
n	Crash
n_c	Number of observations in a specific leaf
p	Probability of selecting a correct match
p_i	Proportion of other classes, not *j*
p_j	Node proportion
p_L	Proportion of data from a parent node that is in the left child node after the split
p_R	Proportion of data from a parent node that is in right child node after the split
RampMet	Indicator variable for presence of ramp meter on entrance ramp
RMSE	Root-mean-square error

Variable	Definition
$R(T)$	Overall misclassification cost for full classification trees
$R_\alpha(T)$	Cost-complexity measure
S	Sum of squared errors
s	Split
Spa_aux	Interaction variable for spacing and presence of auxiliary lane
Spacing	Ramp spacing, from painted gore of entrance ramp to painted gore of exit ramp
T	Tree
t	Parent node
$\lvert\tilde{T}\rvert$	Subtree complexity
T_0	Full, unpruned tree
t_L	Left child node
t_R	Right child node
Upstream_2	Indicator variable of study segments with two travel-way lanes upstream of entrance ramp gore
Upstream_3	Indicator variable of study segments with three travel-way lanes upstream of entrance ramp gore
WA_5	Indicator variable for study segments located on I-5 in Washington
y_i	Value for dependent variable observation
$\hat{y}_i$	Value for observation that is predicted by the model
α	CP

PURPOSE

Tree-based methods are a set of machine-learning and data-mining procedures. They use the form of a binary tree and act as predictive models that map values of a responsive variable as a function of key explanatory variables. Tree-based methods are able to handle nonlinear relationships well and can be applied to both classification (categorical- or discrete-response variable) and regression (continuous-response variable) contexts. Tree-based methods have been implemented in road-safety studies since the 1990s and continue to appear in the literature on a regular basis.[72–76]

This appendix introduces and demonstrates two types of tree-based models: CART and Random Forests. The research team applied the two tree-based methods to assess the impacts of traffic, geometric design, and operational features on expected crash frequency along directional freeway segments that have a right-hand-side entrance ramp followed by a right-hand-side exit ramp. The research team compared the tree-based models to a more traditional analysis approach—NB regression models with fixed effects. The research team used datasets analyzed as part of two previously published journal articles on the safety effects of ramp spacing and auxiliary-lane presence to demonstrate CART and Random Forests.[77,47]

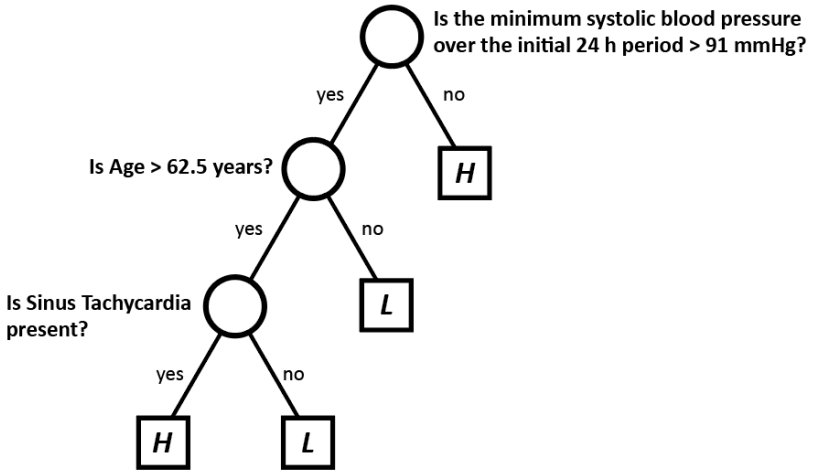
The Methodology, Data, Results, and Discussion sections provide a brief introduction to CART and random-forest methods, a summary of the dataset used for analysis, analysis results, and a discussion of the findings.

METHODOLOGY

Tree-based methods are one of the most popular supervised machine-learning methods. Supervised learning is a machine-learning task of making predictions using labeled training data (i.e., a known dataset).[78] In most road-safety analyses, researchers and practitioners use a historical dataset consisting of observations of a response variable (e.g., crash frequency by crash type and severity) and multiple explanatory variables characterizing traffic, geometric, and operational conditions.

CART

Breiman et al. developed CART, one example of tree-based methods, in the context of a medical study to identify high-risk patients.[79] Specifically, the authors used the following example to illustrate what a tree looked like by grouping heart-attack patients' conditions into two categories: those who will survive 30 d (low risk) or those who will not (high risk). A total of 19 possible explanatory variables were explored, including blood pressure and age. The resulting concept of the classification tree is adapted from Briman et al. and is shown in figure 55.[79]



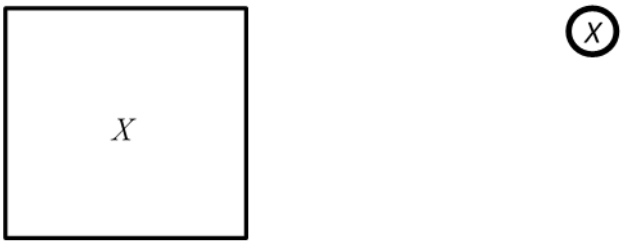
Source: FHWA.
H = high risk; L = low risk.

Figure 55. Illustration. Tree example: heart-attack patients' conditions.

This classification tree ultimately used three explanatory variables to classify a patient's condition. The circle at the very top is called the "root node," and it is split into two "branches" based on the answer to the question about the systolic blood pressure variable next to it. Circles at lower levels are called "internal nodes," with corresponding questions about additional variables (e.g., age). The nodes that consider a split based on a variable are also called "parent nodes," and the nodes following the parent nodes are called "child nodes." The rectangular boxes are called "terminal nodes" or "leaves," showing the terminal classifications that result from following certain paths. This classification tree is a very intuitive tool for estimating whether a patient is at high risk. For example, if a patient's 24-h blood pressure is measured to be less

than 91 following a heart attack, then he/she is in a high-risk condition; if the patient's blood pressure is higher than 91 following a heart attack, and his/her age is less than 62.5, then he/she is in a low-risk condition.

The growing of a tree is a process of recursively splitting the subsets of the study dataset. Starting from the full dataset itself, an explanatory variable is chosen to best split the data into subsets. With the feature space denoted as X (i.e., all possible variables), the splitting of the feature space is shown in figure 56 through figure 58. X can consist of both continuous and discrete explanatory variables. A tree model can handle the mixed variables in a unified manner since it partitions the feature space using binary tests that use thresholds for continuous variables and subset-membership tests for categorical variables.[80]

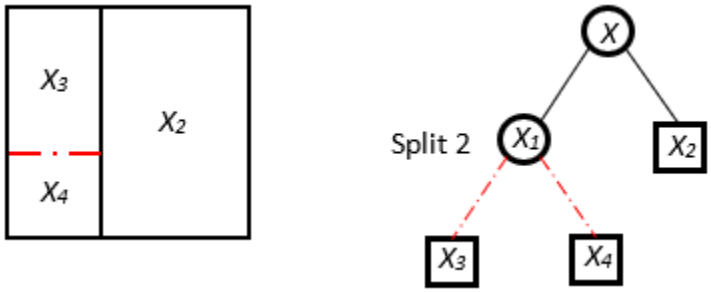


Source: FHWA.

Figure 56. Illustration. Dataset with tree growth.

Source: FHWA.

Figure 57. Illustration. Dataset with tree growth in one split.



Source: FHWA.

Figure 58. Illustration. Splitting of the dataset with tree growth.

The following three steps illustrate how to achieve an appropriate tree using CART:

1. Grow a full tree without restrictions to its size.
2. Use cross-validation to determine the complexity parameters (CPs) associated with different sized subtrees of the full tree.
3. Prune the full tree to the optimal size and corresponding CPs.

Splitting Rule and Stop-Splitting Rule

The splitting rule is a measurement of the goodness of split for every split of the tree. To split the feature space (i.e., to select the proper variable and the appropriate threshold to partition the space), a search over every explanatory variable and every possible threshold will take place, aiming at getting the split that leads to the greatest improvement of a specified score function (or cost function).

For classification trees, commonly used splitting criteria are based on the entropy function and the Gini cost function, which both describe the impurity of a node. The entropy and Gini cost functions are used to characterize the impurity (or purity) of a node when the dependent variable of interest is discrete. If a node has a proportion of p_j of each of the different classes (j), the entropy function of the node is defined in figure 59.

$$i(p) = -\sum_j p_j \log_2 p_j$$

Figure 59. Equation. Entropy function of a regression node.

Where:
i = observation.
p = probability of selecting a correct match.

A node is pure if it has data falling only within a single class (i.e., a single, discrete outcome). In the case of a pure node, p_j is equal to 1 and the entropy function is equal to 0.

The Gini cost function is the most widely used description of impurity and is used by the CART algorithm (figure 60).

$$i(p) = \sum_{i \neq j} p_i p_j = 1 - \sum_j p_j^2$$

Figure 60. Equation. Gini cost function.

Where p_i is the proportion of classes that are not j.

Again, a node is pure if it has data following only within a single class (i.e., a single, discrete outcome). The Gini cost function would also equal 0 in this case.

The goal of splitting the feature space is to minimize the impurity, defined by the entropy function or the Gini cost function. In other words, by adding the tree splits, the goal is to

maximize the goodness of split (i.e., information gained or the reduction of impurity), as shown in figure 61.

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

Figure 61. Equation. Objective function of goodness of split.

Where:

s = split.
t = parent node.
t_L = left child node.
t_R = right child node.
p_L = proportion of data from t that is in the t_L after the split.
p_R = proportion of data from t that is in t_R after the split.

The entropy and Gini cost functions are not meaningful for regression trees of continuous dependent variables (i.e., without discrete outcomes or classes). The impurity is, therefore, defined by the sum of squared errors (S) in figure 62.[81]

$$S = \sum_{c \in leaves\,(T)} \sum_{i \in c} (y_i - m_c)^2$$

Figure 62. Equation. Impurity function of regression trees.

Where:

m_c = prediction for leaf, c.
$leaves(T)$ = a set of leaves for tree (T).
y_i = value for dependent variable i.

m_c is calculated as shown in figure 63.[81]

$$m_c = \frac{1}{n_c} \sum_{i \epsilon c} y_i$$

Figure 63. Equation. Prediction for *c*.

Where n_c is the number of i in c.

Trees heavily depend on the data that are used to build the tree. Thus, if there is no restriction to stop the growth of the trees, they tend to "over fit" the data (i.e., capture not only the patterns in the data, but also the noise specific to the sample dataset). A stop-splitting rule can be applied to tree models to avoid overfitting. Common stop-splitting rules include defining a maximum number of leaves to limit the tree size, defining a minimum count of data points assigned to each leaf, and defining a minimum increase in goodness-of-split measurement.

However, these stopping rules are not necessarily a preferred way of deciding the tree sizes because they can be too shortsighted. In other words, there can be splits that are not informative by themselves but that lead to informative subsequent splits. Therefore, a more effective strategy is to grow a tree, T_0, to its largest extent and then prune it back to obtain a subtree.

Pruning

The pruning method used in CART is minimal cost-complexity pruning (also called weakest-link pruning). For any (T) ≤ T_0 (T_0 is a full, unpruned tree), its complexity (i.e., the size of T or number of terminal nodes in T) is defined as $|\tilde{T}|$. If α is a real number called the CP (i.e., a penalty that defines a cost of each additional c), the cost-complexity measure ($R_\alpha(T)$) is defined in figure 64.

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

Figure 64. Equation. Overall misclassification cost for full classification trees.

Where $R(T)$ is the overall misclassification cost for full classification trees (prior to pruning) or the overall error for full regression trees. If α is set to 0, no penalty is assigned for having a more complex tree, so the fully-grown tree is used. As α increases, the smaller prediction error of a larger tree is penalized more as a result of its higher level of complexity. This circumstance tends to lead to the smallest overall cost, occurring for a subtree that is smaller than the full tree. The best overall subtree for a value of α is the one with the lowest $R_\alpha(T)$. The structure of the best overall subtree will vary with the value selected for α. While α is continuous, Breiman et al. show that the same "best" subtree results from a range of α values.[79] Therefore, only a selected number of α values turn out to be of interest.

To choose an appropriate α, a K-fold cross validation of each subtree can be conducted. A K-fold cross validation consists of randomly dividing the training observations into K equal parts. For each one of the K parts, a full tree is grown using the other K-1 parts of the data, cost-complexity pruning is applied to the full tree to obtain the best subtrees for different levels of α, and the MSPEs of the subtrees corresponding to the different levels of α is assessed.[82] For each sequence of subtrees resulting from a level for α and the K-fold cross-validation process, the prediction error for α is estimated as the average of the prediction errors for that sequence of subtrees. Usually, a 5- to 10-fold cross validation is used. The level of α resulting in the minimum MSPE is then usually selected for final tree pruning, with the corresponding subtree being returned as the pruned tree. One might think that lower values for α (i.e., higher numbers of leaves) would always result in a smaller error. However, this is where the K-fold validation is useful, as it identifies when a low value for α has resulted in an overfitting of the data.

Random Forests

Ho created the first algorithm for Random Forests.[83] Breiman and Cutler extended the algorithm and registered "Random Forests" as a trademark.[84] Random Forests are an ensemble of trees that are created with bootstrapped (randomly selected with replacements) samples. A subset (with the number held constant during the tree growing) of explanatory variables is randomly selected for splitting, allowing for assessments of the importance of variables. Each tree is grown to the largest extent possible with no pruning. The algorithm creates a large number of trees to form a forest. The forest error rate depends on the correlation between any two trees in the forest (as correlation increases, error increases), and the strength of each individual tree in the forest (as strength increases, error decreases). The final classification or prediction is made by averaging or voting across all trees.[85]

Example Application: Expected Crash Frequency on Freeway Segments

To demonstrate CART and Random Forests, the research team applied the models to assess the impacts of traffic, geometric design, and operational features on expected crash frequency along directional freeway segments that have a right-hand-side entrance ramp followed by a right-hand-side exit ramp. The research team used R packages, "rpart" and "randomForest," to implement the analysis and obtain regression tree and random-forest models. The tree-based methods and a more traditional NB-regression model with fixed effects were compared with respect to their predictive power, interpretability, and potential uses in road-safety research and practice.

DATA

The research team analyzed datasets from two previously published journal articles on the safety effects of ramp spacing and auxiliary-lane presence for this analysis.[77,47] The description of the data provided in this section heavily draws on the descriptions in these original papers.

A freeway segment was the basic unit of analysis. The variables in the dataset characterizing each freeway segment consisted of freeway geometric features, traffic characteristics, and the number of multiple-vehicle crashes that had occurred on the segment. The authors of the two previously published papers collected this information in the States of California and Washington. The data-collection period was 2006 through 2008. The paper authors obtained the data using multiple resources, including Google® Earth™ and Google Maps™, Washington State Department of Transportation's (WSDOT's) Interchange Viewer, WSDOT's State Route Web Tool, California DOT's Performance Measurement System, and Federal Highway Administration's Highway Safety Information System databases.[86–91] Segments were located between two successive diamond interchanges (including rural diamonds, compressed diamonds, tight urban diamonds, half diamonds, and single-point diamond interchanges). A study segment was defined as a directional freeway segment beginning at the cross street of the upstream interchange and ending at the cross street of the downstream interchange. Ramp spacing was defined from painted gore to painted gore. A range of traffic and geometric data were collected for each defined freeway segment. Segments were excluded from the dataset if construction activity was identified on or near the segment from 2006 through 2008 (the observation period for each segment). Temporary traffic control devices on the video logs or construction areas present on archived Google Earth photographs were used to identify these segments. Segments with missing volume counts were also excluded as well as segments that included rest-area ramps between entrance and exit ramps associated with two consecutive cross streets. The final dataset consisted of 404 segments, with 154 from Washington State and 250 from California. Both previously published papers contain an illustration of a generic freeway segment, labeling the two cross streets that define the boundaries of each segment, as well as the ramp spacing dimension and the auxiliary lane that is or is not present on each segment.

The variables in the dataset and their definitions are provided in table 30. Descriptive statistics are summarized in table 31.

Table 31. Descriptive statistics geometric, traffic, and crash data for 404 segments used for crash-frequency modeling (adapted from Shea et al.).[47]

Variable	Mean	SD	Min	Max
L (mi)	2.35	1.8	0.5	10.41
ln(*ADT*)	10.672	0.823	8.544	11.9415
ln(*ADTen*)	8.345	1.346	2.833	9.864
ln(*ADTex*)	8.349	1.307	3.219	9.873
Spacing (ft)	9,677.19	9,508.98	316.8	52,219.20
HOVmain	0.07	0.25	0	1
HOVen	0.06	0.23	0	1
RampMet	0.12	0.32	0	1
AuxLn	0.12	0.32	0	1
Upstream_2	0.48	0.5	0	1
Upstream_3	0.27	0.44	0	1
Mainline1	0.43	0.5	0	1
Mainline2	0.43	0.5	0	1
CA_5	0.42	0.49	0	1
CA_10	0.2	0.4	0	1
WA_5	0.2	0.4	0	1
MV-KABC	10.55	13.58	0	96
MV-O	22.04	31.42	0	359

SD = standard deviation; Min = minimum; Max = maximum.

RESULTS

This section presents the modeling results obtained from NB-regression (as used by Shea et al. to analyze this dataset) and tree-based modeling approaches.[47] As the tree-based approaches are quite different from traditional regression methods, their modeling results are presented in very different forms.

NB Regression–Model Estimation Results

The research team replicated the NB-regression models to ensure that the dataset used for this analysis was the exact same dataset used by Shea et al.[47] Shea et al., reported two NB-regression models: one for the expected number of multiple-vehicle fatal and injury (MV-KABC) crashes and one for the expected number of multiple-vehicle, PDO (MV-PDO) crashes.[47] The replicated model results are shown in table 32. The research team checked the estimated coefficients against the model-estimation results in the original paper, and they were exactly the same.

The research team used training and test sets to obtain and evaluate new NB-regression models to ultimately allow a more direct comparison to the tree-based methods. The estimation results are shown in table 33. The training set consisted of 75 percent of the observations from the original dataset, with the remaining 25 percent of the observations making up the test dataset. The research team used a fixed random seed to ensure the same training and test sets were obtained for the NB-regression and tree-based approaches through the random-selection process.

Table 32. Replicated NB regression–model results.

Variable	KABC Coefficient	KABC SE	KABC *p*-value	PDO Coefficient	PDO SE	PDO *p*-value
Intercept	−18.31	1.155	<2.00E−16	−16.37	0.9465	<2.00E−16
ln(*ADT*)	1.722	0.1255	<2.00E−17	1.621	0.1025	<2.00E−17
ln(*ADTen*)	0.2103	0.0407	2.30E−07	0.1176	0.0331	0.0003
ln(*ADTex*)	0.0009	0.0451	0.9840	0.0584	0.0369	0.1133
InvSpa	545.3	231.4	0.0184	577.1	197.4	0.0034
InvSpaAux	−385.6	208.1	0.0639	−333.0	178.6	0.0622
Upstream_2	0.3526	0.1686	0.0365	0.3127	0.1428	0.0285
Upstream_3	0.0305	0.1193	0.7983	0.0532	0.1023	0.6032
*Mainline*1	0.1431	0.0730	0.0499	−0.0187	0.0622	0.7638
*Mainline*2	−0.0415	0.0750	0.5800	0.0224	0.0636	0.7242
RampMet	0.0875	0.1408	0.5344	0.0212	0.1257	0.8660
HOVen	−0.1221	0.1584	0.4407	−0.1323	0.1419	0.3513
HOVmain	−0.0658	0.1533	0.6675	0.1624	0.1367	0.2346
CA_5	−0.4173	0.1177	0.0003	−0.2019	0.0975	0.0384
CA_10	−0.2815	0.1329	0.0340	0.0143	0.1109	0.8974
WA_5	−0.5396	0.1195	6.31E−06	−0.4994	0.1014	8.43E−07

SE = standard error.

Table 33. New NB regression–model results from training dataset.

Variable	KABC Coefficient	KABC SE	KABC *p*-value	PDO Coefficient	PDO SE	PDO *p*-value
Intercept	−18.99	1.302	<2.00E−16	−15.95	1.058	<2.00E−16
ln(*ADT*)	1.785	0.1381	<2.00E−17	1.556	0.1119	<2.00E−17
ln(*ADTen*)	0.2217	0.0458	1.26E-06	0.1451	0.0371	9.09E−05
ln(*ADTex*)	−0.0218	0.0496	0.6605	0.0518	0.0402	0.1969
InvSpa	393.2	249.8	0.1154	506.3	208.9	0.0154
InvSpaAux	−298.8	229.5	0.1927	−378.9	193.3	0.0500
Upstream_2	0.5491	0.1909	0.0040	0.3634	0.1609	0.0239
Upstream_3	0.1514	0.1335	0.2569	0.1463	0.1136	0.1978
*Mainline*1	0.1884	0.0818	0.0212	0.0280	0.0692	0.6856
*Mainline*2	−0.0302	0.0846	0.7210	0.0610	0.0713	0.3921
RampMet	0.2340	0.1608	0.1456	0.0955	0.1431	0.5043
HOVen	−0.1192	0.1919	0.5344	−0.0885	0.1708	0.6043
HOVmain	−0.0779	0.1724	0.6511	0.2297	0.1531	0.1335
CA_5	−0.4950	0.1299	0.0001	−0.1967	0.1071	0.0662
CA_10	−0.3865	0.1482	0.0091	−0.0415	0.1230	0.7358
WA_5	−0.6054	0.1306	3.52E−06	−0.5435	0.1105	8.66E−07

SE = standard error.

Tree-Based Methods

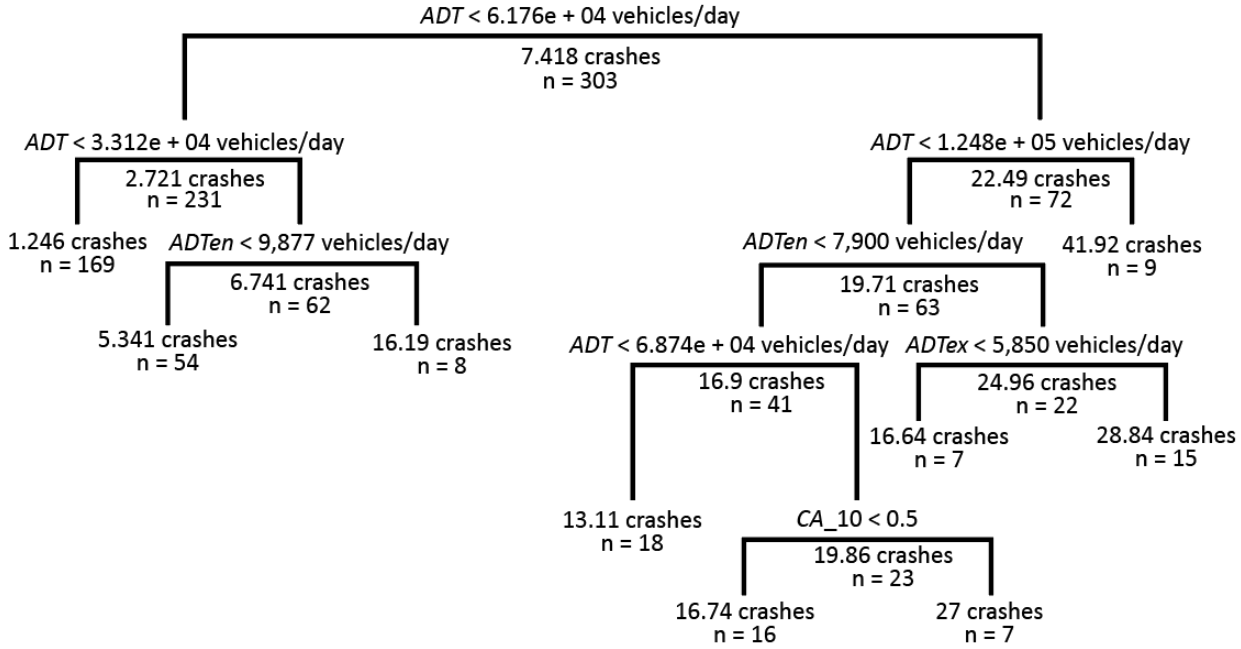
CART and Random Forests were the two approaches used to obtained tree-based models in this report. The research team used the same training dataset used for the NB-regression modeling summarized in the previous section, "NB Regression–Model Estimation Results" to implement the tree-based methods. The variable segment length was used as an offset variable in the NB-regression models. For the tree-based models, segment length was also used. The multiple vehicle–crash counts were divided by segment length, converting the responsive variable to a multiple vehicle–crash frequency per unit length.

The research team used the same explanatory variables from the NB-regression model for tree-based modeling. However, the variables representing traffic volumes and ramp spacing were used in the tree-based models without any type of transformation, as specifying the appropriate functional form of these variables is not critical with tree-based approaches. Tree-based models can capture nonlinear relationships without affecting the partitioning of datasets. The same splitting results should be obtained with or without transforming those variables.

CART

Following the steps of constructing a tree using CART, a full tree was grown first using the training dataset consisting of 75 percent of the observations. Then, a 10-fold cross-validation was carried out to select a proper CP, which was used to prune the tree. The pruned tree was then evaluated using the test dataset, consisting of the remaining 25 percent of the observations.

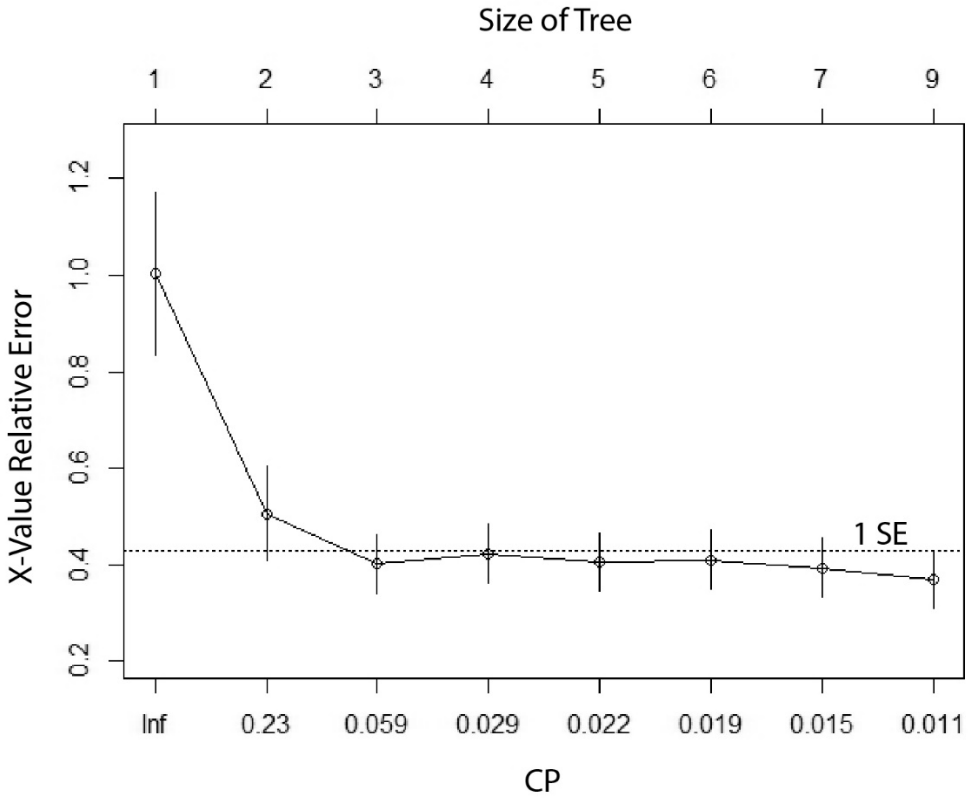
The full regression tree for the MV-KABC crashes is shown in figure 65. There are nine terminal nodes. Associated with each node is a question about a selected explanatory variable and a threshold, with information showing the prediction and the number of data points. The left splits indicate "yes" to the question, and the right splits indicate "no" to the question. Branch length indicates how much information was gained from that split. Longer branch lengths mean more information was gained.



Source: FHWA.

Figure 65. Illustration. Full regression tree for MV-KABC crashes.

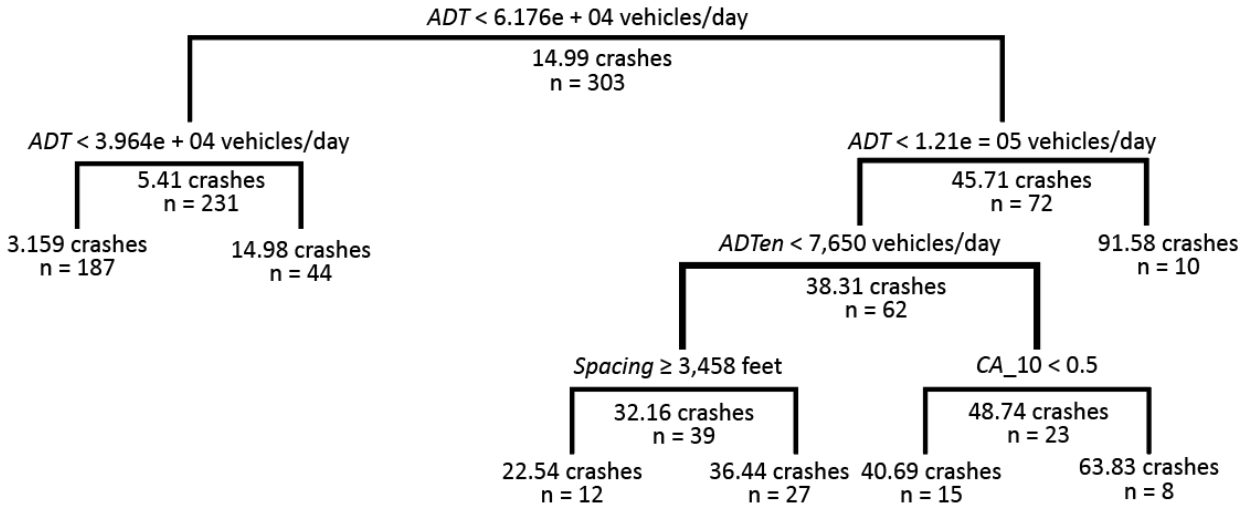
The 10-fold cross-validation results are shown in figure 66. The upper *x*-axis shows the size of the tree, and the lower *x*-axis shows the CP associated with each tree size. The *y*-axis shows the cross-validated error. On the plot, there is a dotted horizontal line marking 1 SE (standard error). The rule for tree pruning is to prune the tree to the smallest tree size where both the mean cross-validated error is lower than both the 1-SE line and all other mean cross-validated errors below the line. As the CP associated with tree size 9 was lower than the 1-SE line and was the smallest, the regression tree for KABC crashes did not need to be pruned.



Source: FHWA.
Inf = infinity.

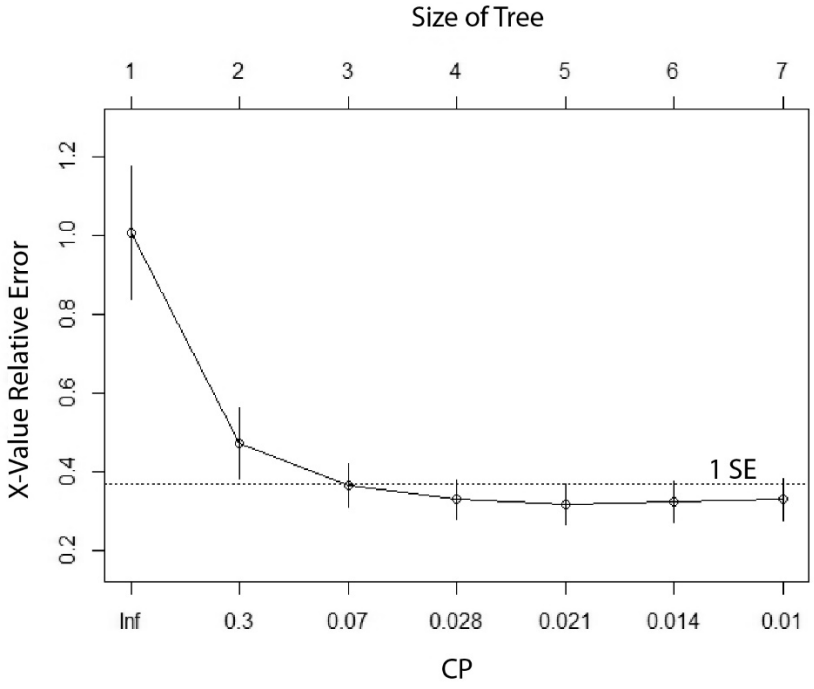
Figure 66. Illustration. Ten-fold cross-validation plot of full regression tree for MV-KABC crashes.

As shown in figure 66, the optimal tree results come from considering only AADT on the mainline, entering volume, exiting volume, and whether the segment is located along I-10 in California. The same process was used to obtain the regression tree for PDO crashes. The full tree, cross-validation plot, and the pruned tree are shown in figure 67, figure 68, and figure 69, respectively. While the ramp spacing does appear in the full tree for MV-PDO crashes (with a threshold of approximately 3,460 ft), the optimal pruned tree considers only AADT on the mainline and entering traffic volumes. The pruned tree consisting of five leaves was selected as it resulted in the lowest mean prediction error resulting from the cross-validation process.



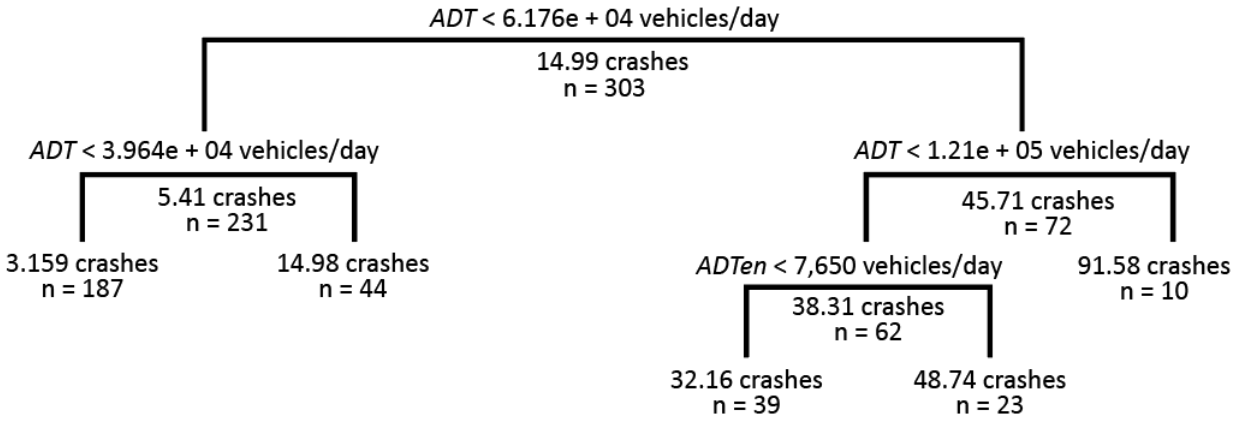
Source: FHWA.

Figure 67. Illustration. Full regression tree for MV-PDO crashes.



Source: FHWA.
Inf = infinity.

Figure 68. Graph. 10-fold cross-validation plot of full regression tree for MV-PDO crashes.

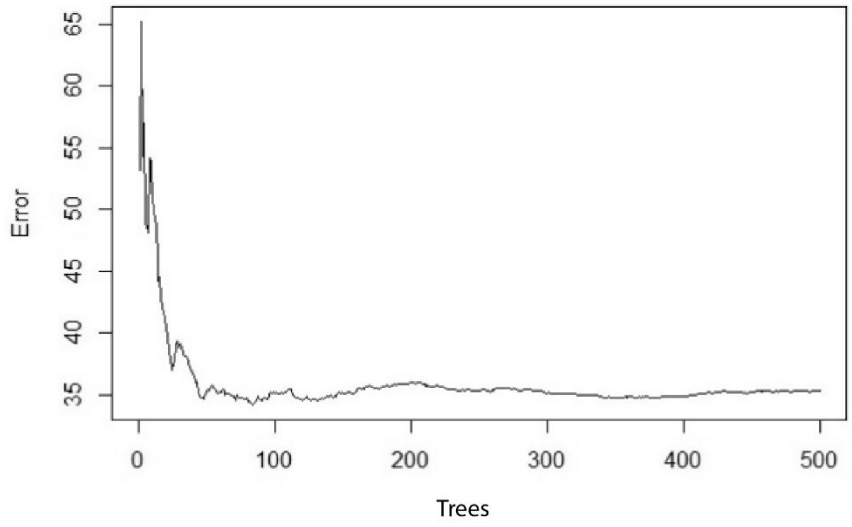


Source: FHWA.

Figure 69. Illustration. Pruned regression tree for PDO crashes.

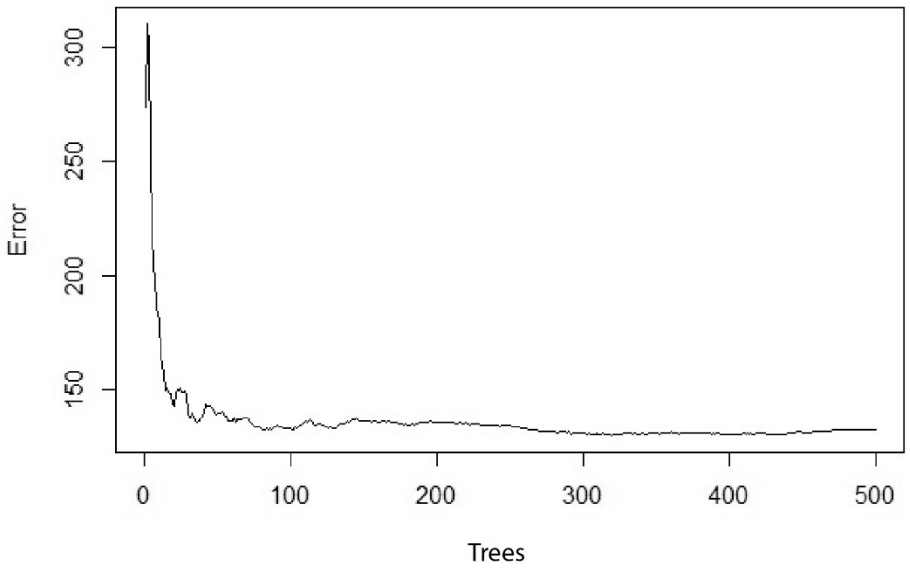
Random Forests

As with the previously discussed NB and CART methods, the research team used the training dataset to grow the random forest, and the test dataset was used to evaluate the results. One thousand trees were created for each random forest. The plots presented in figure 70 and figure 71 show the change of random forest–model error with an increase in the number of trees. The random forest for MV-KABC crashes reached an optimum at 84 trees, while the random forest for MV-PDO crashes reached an optimum at 320 trees.



Source: FHWA.

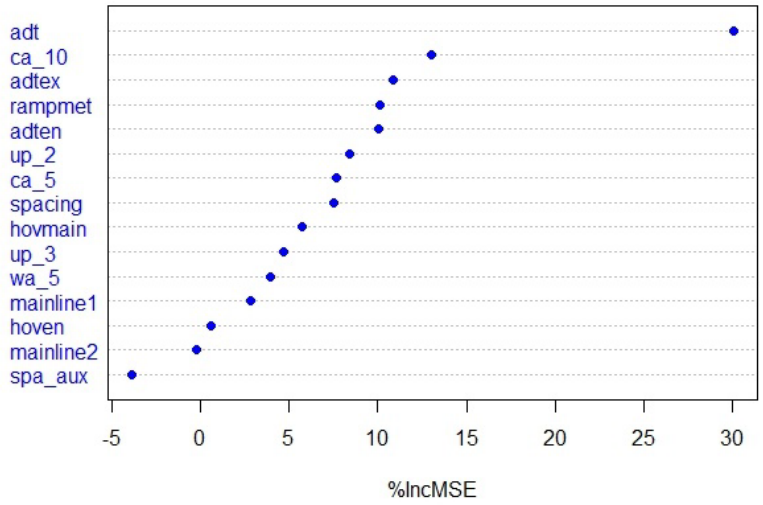
Figure 70. Graph. Random-forest plots showing error against number of trees—KABC crashes.



Source: FHWA.

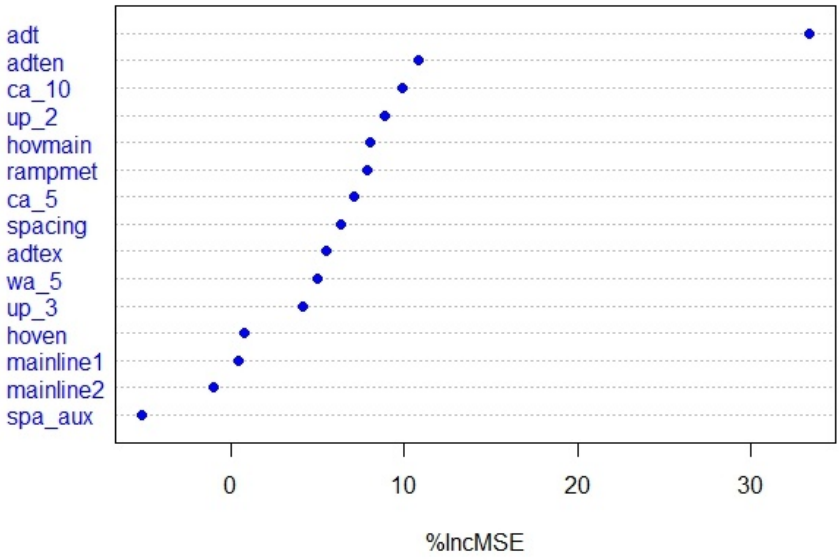
Figure 71. Graph. Random-forest plots showing error against number of trees—PDO crashes.

Random forests are more like a black box and cannot be visualized in the same way as a regression tree. One cannot see the structure of the model or the interaction between the explanatory variables. However, additional analysis can be used to visualize variable importance. Plots in figure 72 and figure 73 show the importance of the explanatory variables used to grow the random forests. The variables are listed according to their level of importance from top to bottom. The corresponding *x*-axis value shows the percentage increase in MSE with the removal of that variable from the random-forest model.



Source: FHWA.
%IncMSE = percent increase of MSE.

Figure 72. Graph. Random-forest variable-importance plot—MV-KABC crashes.



Source: FHWA.
%IncMSE = percent increase of MSE.

Figure 73. Graph. Random-forest variable-importance plot—MV-PDO crashes.

Models by AADT Category

Observing both the NB and tree-based modeling results, the importance of the AADT variables is obvious. However, the strong correlation between AADT and expected crash frequency is already well known by road-safety researchers and practitioners as AADT serves as a key measure of exposure. The research team carried out a variation in the analysis to explore whether other variables influencing expected crash frequency stand out in the tree-based approaches when AADT is accounted for in another way. This variation was achieved by dividing the original dataset into two subsets—one with lower AADT values and the other with higher AADT values. The AADT threshold to make this split was 40,775 vehicles/d, resulting in 248 segments in the lower-AADT subset and 156 observations in the higher-AADT subset. NB-regression models were re-estimated with the same specification reported in table 32 and table 33, but the tree models were created without using the AADT variables (i.e., excluding *ADT*, *ADTen*, and *ADTex* as possible explanatory variables). Since the AADT values were relatively more homogeneous within the subsets than the original dataset, its effects in the NB-regression models should decrease as well, making the NB-regression models comparable to the revised tree-based models.

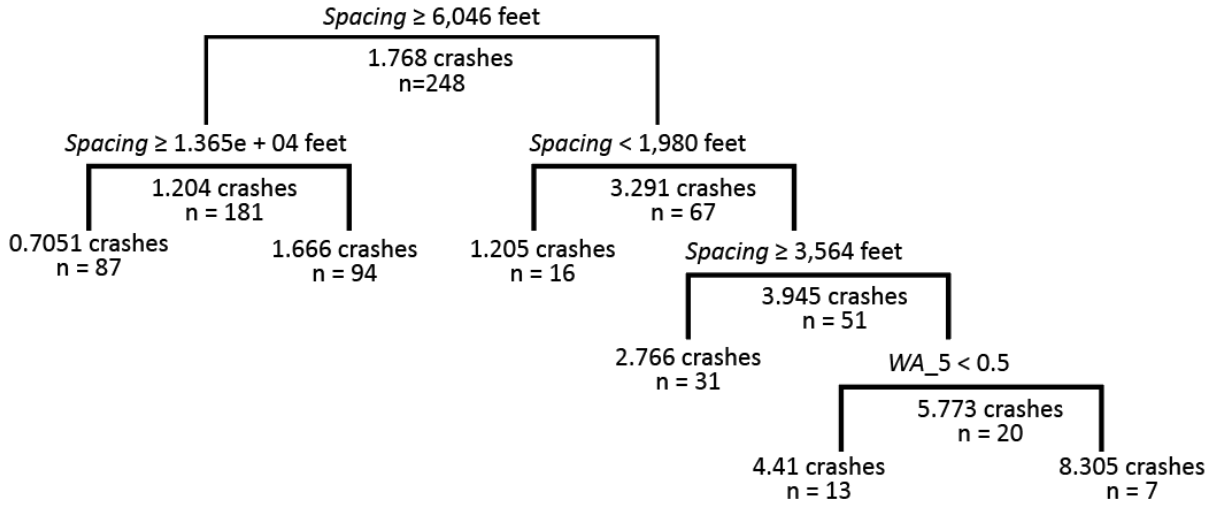
Table 34 shows the NB regression–model estimation results based on the lower AADT subset and the higher AADT subset.

Table 34. NB regression–modeling results with lower and higher AADT subsets.

Variable	Lower-AADT KABC	Lower-AADT PDO	Higher-AADT KABC	Higher-AADT PDO
Intercept	−16.42	−13.31	−17.75	−18.85
ln(*ADT*)	1.557	1.309	1.710	1.850
ln(*ADTen*)	0.2393	0.1612	0.1816	0.1004
ln(*ADTex*)	0.0114	0.0659	−0.0018	0.0848
InvSpa	751.7	688.6	357.3	357.6
InvSpaAux	−1381	−944.1	−157.8	−111.7
Upstream_2	−0.2696	−0.1431	—	—
Upstream_3	−0.6576	−0.4334	0.1037	0.2336
*Mainline*1	0.0769	−0.0963	0.2114	0.0866
*Mainline*2	0.0003	0.1295	−0.0480	−0.0585
RampMet	—	—	0.1246	−0.0102
HOVen	—	—	−0.0813	−0.1312
HOVmain	—	—	−0.2347	−0.0069
CA_5	−0.3188	−0.1101	−0.6556	−0.4160
CA_10	−0.5518	−0.1663	−0.3799	−0.0867
WA_5	−0.4314	−0.3649	−0.5812	−0.5118

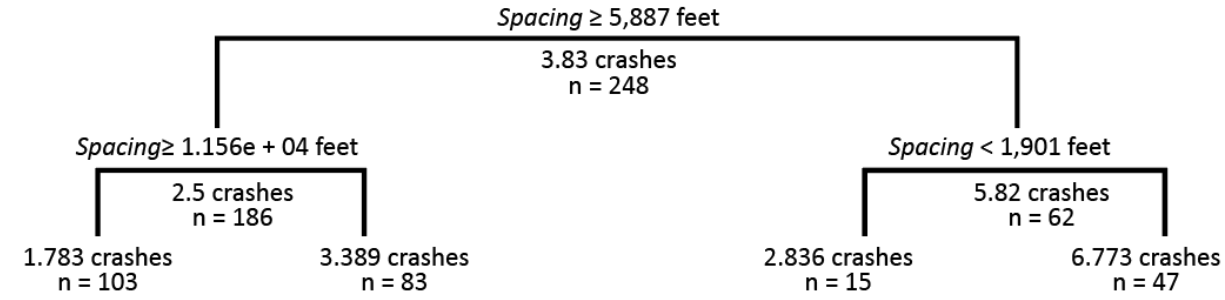
—No data.

The corresponding trees obtained using CART are shown in figure 74 through figure 77. It is noticeable that ramp-spacing and the presence-of-ramp-metering variables play much more important roles in these revised trees.



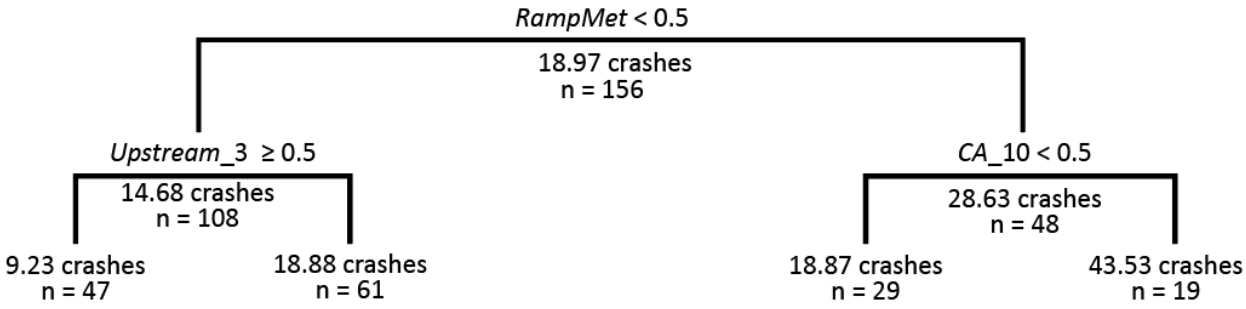
Source: FHWA.

Figure 74. Illustration. Pruned regression tree for MV-KABC crashes on segments with lower AADT.



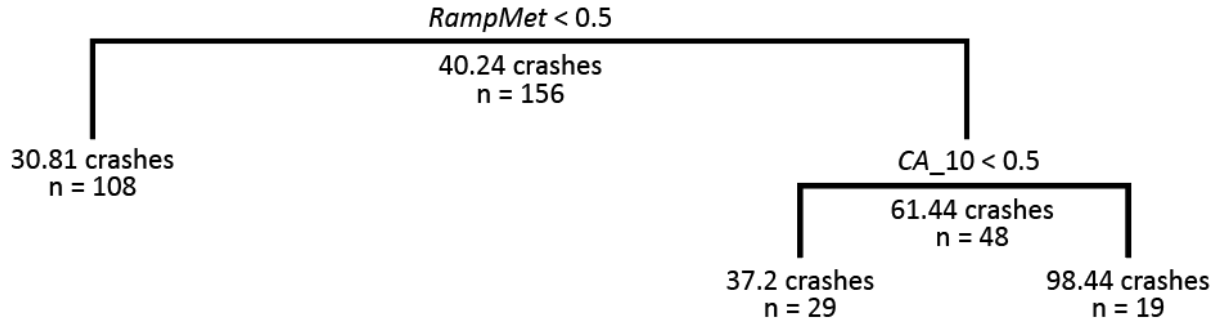
Source: FHWA.

Figure 75. Illustration. Pruned regression tree for MV-PDO crashes on segments with lower AADT.



Source: FHWA.

Figure 76. Illustration. Pruned regression tree for MV-KABC crashes on segments with higher AADT.



Source: FHWA.

Figure 77. Illustration. Pruned regression tree for PDO crashes on segments with higher AADT.

The research team also created random forests using the lower- and higher-AADT subsets. For the lower-AADT subsets, the random-forest model for MV-KABC crashes reached its optimum with 25 trees, and the random-forest model for MV-PDO crashes reached its optimum at 974 trees. For the higher-AADT subsets, the random-forest model for MV-KABC crashes reached its optimum with 560 trees and the random-forest model for MV-PDO crashes reached

its optimum with only one tree. Table 35 shows and table 36 the variable-importance ranking outputs from those random forests. Again, ramp spacing and the presence of ramp metering appear as much more important variables in the revised random forests.

Table 35. Variable-importance rankings based on lower-AADT subsets' random forests.

KABC Variable	KABC %IncMSE	PDO Variable	PDO %IncMSE
Spacing	18.5052	*Spacing*	39.4789
Upstream_2	16.2026	*CA_10*	18.0951
Upstream_3	14.4593	*Upstream_2*	18.0143
CA_5	11.4957	*CA_5*	17.4493
CA_10	10.2168	*Ustream_3*	16.4805
WA_5	6.8507	*WA_5*	14.0277
spa_aux	5.2410	*Mainline1*	13.3080
Mainline2	4.0805	*Mainline2*	11.1860
Mainline1	3.5462	*spa_aux*	4.4423
RampMet	0.0000	*RampMet*	0.0000
HOVen	0.0000	*HOVen*	0.0000
HOVmain	0.0000	*HOVmain*	0.0000
—	—	*HOVmain*	0.0000

—No data.
%IncMSE = percentage increase in MSE if the corresponding variable is removed from the model.

Table 36. Variable-importance rankings based on higher-AADT subsets' random forests.

KABC Variable	KABC %IncMSE	PDO Variable	PDO %IncMSE
RampMet	25.2843	*RampMet*	19.2410
CA_10	20.5629	*CA_10*	14.5960
Upstream_3	17.7928	*HOVmain*	8.6996
CA_5	13.1761	*Upstream_3*	8.2141
HOVmain	11.5366	*CA_5*	7.6381
Mainline1	4.2915	*Mainline1*	6.0809
Spacing	3.4345	*Spacing*	4.7013
Mainline2	3.2119	*Mainline2*	4.4029
WA_5	1.2156	*HOVen*	4.3755
HOVen	0.2222	*spa_aux*	2.2301
Upstream_2	0.0000	*WA_5*	0.1956
spa_aux	−0.6801	*Upstream_2*	0.0000

%IncMSE = percentage increase in MSE if the corresponding variable is removed from the model.

DISCUSSION

The research team evaluated all models obtained using either the tree-based or NB-regression approaches using a test dataset consisting of 25 percent of the observations from the study dataset. In addition to comparing the models' predictive power, interpretability and potential uses were also compared.

Predictive Power

The research team used two criteria to evaluate the predictive power of each model. Root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model and the values observed. *RMSE* is calculated in figure 78.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Figure 78. Equation. Root mean square error.

Where:

n = crash.

$\hat{y}_i$ = value for i that is predicted by the model.

Mean absolute error (*MAE*), calculated in figure 79, was also used to compare model predictions.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

Figure 79. Equation. *MAE*.

The comparison is shown in table 37 and table 38. The tree-based models had better prediction accuracy than the NB-regression models, while the random forests had the best overall model-prediction accuracy. As expected, the differences are quite significant since the tree-based approaches are focused on prediction.

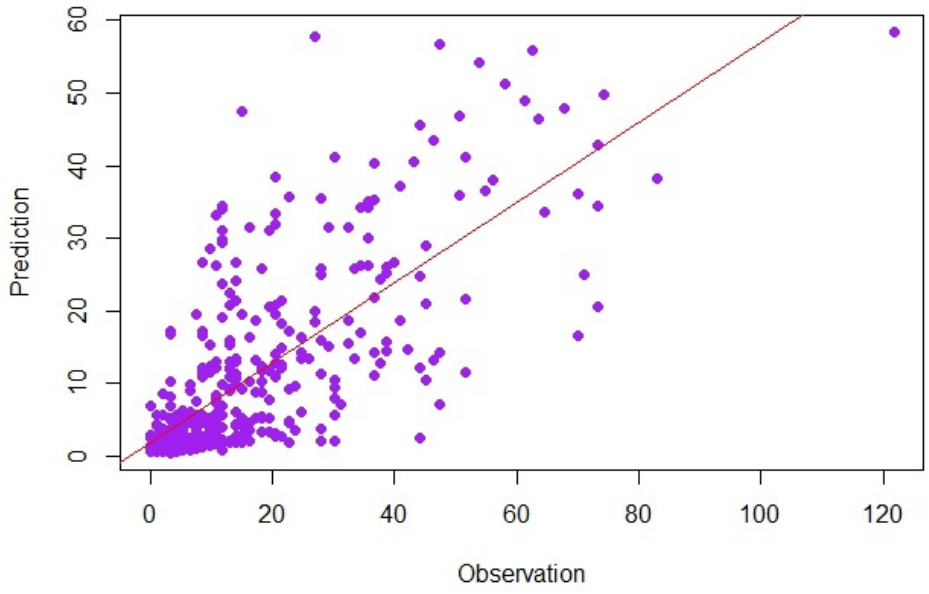
Table 37. Comparison of model predictive power—KABC.

Dataset	Evaluation Measure	NB	CART	Random Forests
Original	RMSE	19.3743	12.9128	13.0541
Original	MAE	11.2901	5.2123	4.7534
Lower-AADT subset	RMSE	4.8703	2.3894	1.8682
Lower-AADT subset	MAE	2.9932	1.2923	0.9679
Higher-AADT subset	RMSE	24.1329	13.3353	9.4378
Higher-AADT subset	MAE	18.2416	9.1073	6.2011

Table 38. Comparison of model predictive power—PDO.

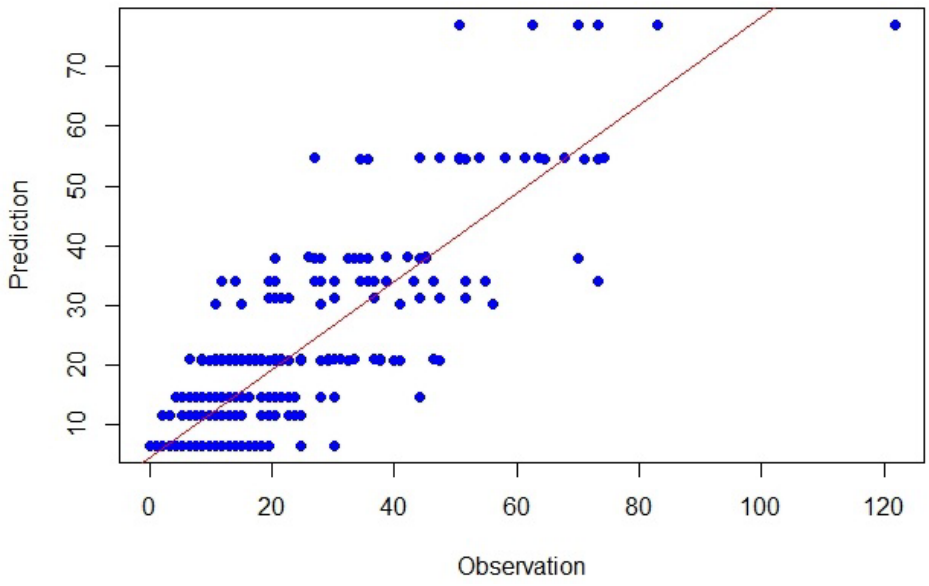
Dataset	Evaluation Measure	NB	CART	Random Forests
Original	RMSE	50.3288	38.2401	39.5668
Original	MAE	25.5780	12.3991	11.4034
Lower-AADT subset	RMSE	8.6747	3.0802	2.2744
Lower-AADT subset	MAE	6.1219	2.0878	1.5266
Higher-AADT subset	RMSE	57.3869	36.3702	24.7727
Higher-AADT subset	MAE	41.0241	21.3122	13.1240

The structure of the predictions is quite different when comparing traditional regression and tree-based models. As shown in figure 80 and figure 81, a regression model gives continuous predictions that spread out in the prediction–observation plot, while a tree gives staged predictions (i.e., groups of segments that have the same predicted average crash frequency represented by the leaves).



Source: FHWA.

Figure 80. Plot. Example of prediction–observation plot from regression models.

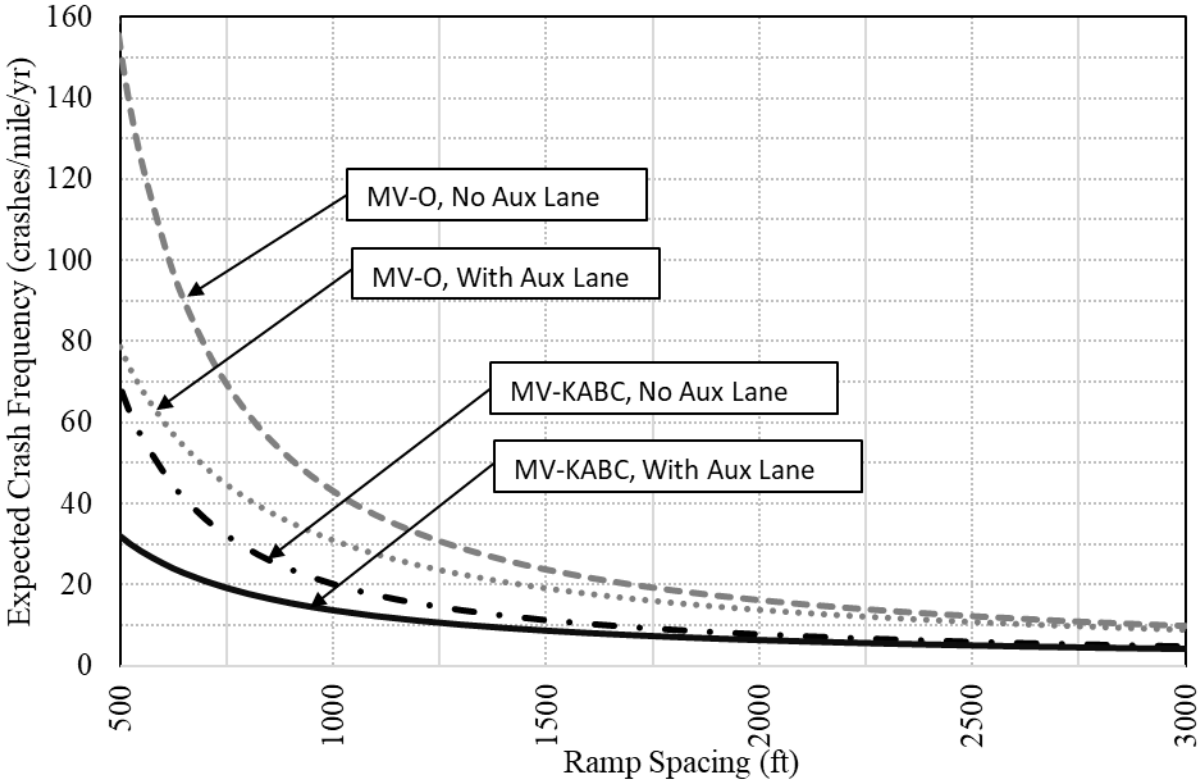


Source: FHWA.

Figure 81. Plot. Example of prediction–observation plot from tree-models.

Interpretability

In terms of interpretability, NB-regression models can provide a quantified, continuous effect of an explanatory variable. The amount of change in the responsive variable with the change of an explanatory variable can be represented using the estimated regression coefficient for the explanatory variable of interest. Specifying the functional form of the model can be particularly important in this case. For example, figure 82 (drawn from Shea et al.) displays the expected annual frequency of MV-KABC and MV-O crashes on a per-mile basis as a function of ramp spacing and auxiliary lane presence.[47] The plot shows that expected crash frequencies increase at a faster and faster rate as spacing gets shorter and shorter. The safety benefit of providing an auxiliary lane (in terms of reductions in expected multiple vehicle–crash frequencies) also gets larger as ramp spacing becomes shorter. Such findings can be particularly useful for planners and designers evaluating new access requests or modifications to existing access on freeways that will result in a change in ramp spacing.



Source: FHWA.
Aux = auxiliary.

Figure 82. Graph. Expected number of MV-KABC and MV-O crashes as a function of ramp spacing and auxiliary lane presence.[48]

Although tree-based models cannot directly provide such intuitive information about safety effects, they have the following advantages related to interpretability:

- Easy-to-read graphical model form (in the case of CART).
- Direct display of variable importance.
- The capture of interactions between explanatory variables.

These features can be useful when conducting safety analysis. Tree-based models are easy and fast to implement since they use relatively simple data-mining algorithms for data partitioning, and no selection or transformation of variables is needed. The structure of a tree-based model can be simple or complex, so it is adjustable according to practical needs. Factors such as tree size and cost functions can all be modified as a function of the study context.

Potential Uses of Tree-Based Approaches

As tree-based models are cost-efficient methods for data analysis, with several advantages over the traditional regression modeling methods, they appear to have strong potential for road-safety analysis. They are particularly effective in making predictions of expected crash frequency, which has applications in multiple contexts (e.g., network screening, alternatives assessment, predicting "what would have been" in before–after studies). Tree-based methods also have potential to inform the specifications that are part of more traditional modeling approaches through identifying the "most important" right-hand-side variables and uncovering informative relationships between left-hand-side and right-hand-side variables. Tree-based methods were demonstrated in this report using expected crash frequencies, which were treated as a continuous variable. Applications for modeling crash severities (i.e., discrete injury outcomes resulting from crashes) may hold even more promise.

Apart from the two tree-based methods applied in this analysis, the CART method and Random Forests, a number of other tree-based models are available in statistical-analysis packages and software, such as ID3, C5.0, CHAID, and GUIDE.[92–94] A number of tree-modeling packages are free, and researchers and practitioners can easily access them.

REFERENCES

1. American Association of State Highway and Transportation Officials. (2005). *AASHTO Strategic Highway Safety Plan.* American Association of State Highway and Transportation Officials, Washington, DC. Available online: https://www.thebreakingnews.com/files/articles/safety-strategichighwaysafetyplan.pdf, last accessed August 12, 2020.

2. Banks, D., Persaud, B., Lyon, C., Eccles, K., and Himes, S. (2014). *Enhancing Statistical Methodologies for Highway Safety Research – Impetus from FHWA*, Report No. FHWA-HRT-14-081, Federal Highway Administration, Washington, DC. Available online: https://www.fhwa.dot.gov/publications/research/safety/14081/14081.pdf, last accessed November 15, 2017.

3. American Association of State Highway and Transportation Officials. (2010). *Highway Safety Manual*, AASHTO, Washington, DC.

4. Mannering, F.L. and Bhat, C.R. (2014). "Analytic Methods in Accident Research: Methodological Frontier and Future Direction." *Analytic Methods in Accident Research*, *1*, pp. 1–22, Elsevier, Amsterdam, Netherlands.

5. Bonneson, J.A., Geedipally, S, Pratt, M.P., and Lord, D. (2012). *Safety Prediction Methodology and Analysis Tool for Freeways and Interchanges*, Transportation Research Board of the National Academies, Washington, DC. Available online: http://onlinepubs.trb.org/onlinepubs/nchrp/docs/nchrp17-45_fr.pdf, last accessed November 15, 2017.

6. Kennedy, P. (2003). *A Guide to Econometrics*, MIT Press, Cambridge, MA.

7. Washington, S.P., Karlaftis, M.G., and Mannering, F.L. (2010). *Statistical and Econometric Methods for Transportation Data Analysis*, Chapman and Hall, CRC Press, Boca Raton, FL.

8. Lord, D. and Mannering, F.L. (2010). "The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives." *Transportation Research Part A*, *44*, pp. 291–305, Elsevier, Amsterdam, Netherlands.

9. Savolainen, P.T., Mannering, F.L., Lord, D., and Quddus, M.A. (2011). "The Statistical Analysis of Crash-Injury Severities: A Review and Assessment of Methodological Alternatives." *Accident Analysis and Prevention*, *43*(5), pp. 1666–1676, Elsevier, Amsterdam, Netherlands.

10. National Highway Traffic Safety Administration. "CODES – NSCA Publication Topic." (website) United States Department of Transportation, Washington, DC. Available online: https://crashstats.nhtsa.dot.gov/#/PublicationList/27, last accessed November 15, 2017.

11. Federal Highway Administration. “SHRP2Solutions: Tools for the Road Ahead.” (website) FHWA, Washington, DC. Available online: https://www.fhwa.dot.gov/goshrp2, last accessed November 15, 2017.

12. National Highway Traffic Safety Administration. (2016). *Fatality Analysis Reporting System (FARS) Analytical User’s Manual 1975–2015*, Report No. DOT HS 812 315, NHTSA, Washington, DC.

13. Barua, S., El-Basyouny, K., and Islam, M.T. (2015). “Effects of Spatial Correlation in Random Parameters Collision Count-Data Models.” *Analytic Methods in Accident Research*, *5*, pp. 28–42, Elsevier, Amsterdam, Netherlands.

14. Barua, S., El-Basyouny, K. and Islam, M.T. (2014). “A Full Bayesian Multivariate Count Data Model of Collision Severity with Spatial Correlation.” *Analytic Methods in Accident Research*, *3*, pp. 28–43, Elsevier, Amsterdam, Netherlands.

15. Quistberg, D.A., Howard, E.J., Ebel, B.E., Moudon, A.V., Saelens, B.E., Hurvitz, P.M. Curtin, J.E., and Rivara, F.P. (2015). “Multilevel Models for Evaluating the Risk of Pedestrian–Motor Vehicle Collisions at Intersections and Mid-Blocks.” *Accident Analysis and Prevention*, *84*, pp. 99–111, Elsevier, Amsterdam, Netherlands.

16. Chen, C., Zhang, G., Tarefder, R., Ma, J., Wei, H., and Guan, H. (2015). “A Multinomial Logit Model-Bayesian Network Hybrid Approach for Driver Injury Severity Analyses in Rear-End Crashes.” *Accident Analysis and Prevention*, *80*, pp. 76–88, Elsevier, Amsterdam, Netherlands.

17. Yu, R. and Abdel-Aty, M. (2014). “Using Hierarchical Bayesian Binary Probit Models to Analyze Crash Injury Severity on High Speed Facilities with Real-Time Traffic Data.” *Accident Analysis and Prevention*, *62*, pp. 161–167, Elsevier, Amsterdam, Netherlands.

18. Cerwick, D.M., Gkritza, K., Shaheed, M.S., and Hans, Z. (2014). “A Comparison of the Mixed Logit and Latent Class Methods for Crash Severity Analysis.” *Analytic Methods in Accident Research*, *3,* pp. 11–27, Elsevier, Amsterdam, Netherlands.

19. Yasmin, S., Eluru, N., Bhat, C., and Tay, R. (2014). “A Latent Segmentation Generalized Ordered Logit Model to Examine Factors Influencing Driver Injury Severity.” *Analytic Methods in Accident Research*, *1*, pp. 23–38, Elsevier, Amsterdam, Netherlands.

20. Abay, K.A. (2015). “Investigating the Nature and Impact of Reporting Bias in Road Crash Data.” *Transportation Research Part A*, *71*, pp. 31–45, Elsevier, Amsterdam, Netherlands.

21. Yasmin, S. and Eluru, N. (2013). “Evaluating Alternate Discrete Outcome Frameworks for Modeling Crash Injury Severity.” *Accident Analysis and Prevention*, *59*, pp. 506–521, Elsevier, Amsterdam, Netherlands.

22. Karwa, V., Slavkovic, A.B., and Donnell, E.T. (2011). “Causal Inference in Transportation Safety Studies: Comparison of Potential Outcomes and Causal Diagrams.” *Annals of Applied Statistics*, *5*(2B), pp. 1428–1455, Institute of Mathematical Statistics, Beachwood, OH.

23. Graham, D.J., McCoy, E.J. and Stephens, D.A. (2014). "Quantifying Causal Effects of Road Network Capacity Expansions on Traffic Volume and Density via a Mixed Model Propensity Score Estimator." *Journal of the American Statistical Association*, *109*(508), pp.1440–1449, American Statistical Association, Alexandria, VA.

24. Wood, J.S., Donnell, E.T., and Porter, R.J. (2015). "Comparison of Safety Effect Propensity Scores-Potential Outcomes Framework, and Regression Model with Cross-Sectional Data." *Accident Analysis and Prevention*, *75*, pp. 144–154, Elsevier, Amsterdam, Netherlands.

25. Sacchi, E. and Sayed, T. (2015). "Investigating the Accuracy of Bayesian Techniques for Before–After Safety Studies: The Case of a 'No Treatment' Evaluation." *Accident Analysis & Prevention*, *78*, pp. 138–145, Elsevier, Amsterdam, Netherlands.

26. Eluru, N., Paleti, R., Pendyala, R., and Bhat, C. (2010). "Modeling Multiple Vehicle Occupant Injury Severity: A Copula-Based Multivariate Approach." *Transportation Research Record: Journal of the Transportation Research Board*, *2165*, pp. 1–11, TRB, Washingtion, DC.

27. Chiou, Y.C. and Fu, C. (2013). "Modeling Crash Frequency and Severity Using Multinomial-Generalized Poisson Model with Error Components." *Accident Analysis and Prevention*, *50*, pp. 73–82, Elsevier, Amsterdam, Netherlands.

28. El-Basyouny, K., Barua, S., and Islam, M.T. (2014). "Investigation of Time and Weather Effects on Crash Types Using Full Bayesian Multivariate Poisson Lognormal Models." *Accident Analysis and Prevention*, *73*, pp. 91–99, Elsevier, Amsterdam, Netherlands.

29. Sacchi, E., Sayed, T., and El-Basyouny, K. (2015). *Multivariate Full-Bayesian Hotspot Identification and Ranking: A New Technique*, Paper No. 15-0159. Proceedings of the 94th Annual Meeting of the Transportation Research Board, Washington, DC.

30. Bhat, C., Born, K., Sidharthan, R., and Bhat, P. (2014). "A Count Data Model with Endogenous Covariates: Formulation and Application to Roadway Crash Frequency at Intersections." *Analytic Methods in Accident Research*, *1*, pp. 53–71, Elsevier, Amsterdam, Netherlands.

31. Chen, W., Wang, J.H., Bryden, G., Ye, X., and Jia, X. (2013). "An Examination of the Endogeneity of Speed Limits and Accident Counts in Crash Models." *Journal of Transportation Safety & Security*, *5*(4), pp. 314–326, Taylor & Francis, London, England.

32. Lord, D., and Kuo, P.F. (2012). "Examining the Effects of Site Selection Criteria for Evaluating the Effectiveness of Traffic Safety Countermeasures." *Accident and Analysis and Prevention*, *47*, pp. 52–63, Elsevier, Amsterdam, Netherlands.

33. Kim, D. and Washington, S.P. (2006). "The Significance of Endogeneity Problems in Crash Models: An Examination of Left-Turn Lanes in Intersection Crash Models." *Accident Analysis and Prevention*, *38*(6), pp 1094–1100, Amsterdam, Netherlands.

34. Khan, G., Bill, A.R., and Noyce, D.A. (2015). "Exploring the Feasibility of Classification Trees versus Ordinal Discrete Choice Models for Analyzing Crash Severity." *Transportation Research Part C*, *50*, pp. 86–96, Elsevier, Amsterdam, Netherlands.

35. Saha, D., Alluri, P., and Gan, A. (2015). "Prioritizing Highway Safety Manual's Crash Prediction Variables using Boosted Regression Trees." *Accident Analysis and Prevention*, *79*, pp. 133–144, Elsevier, Amsterdam, Netherlands.

36. Xu, C., Liu, P., Wang, W., and Li, Z. (2015). "Safety Performance of Traffic Phases and Phase Transitions in Three Phase Traffic Theory." *Accident Analysis and Prevention*, *85*, pp. 45–57, Elsevier, Amsterdam, Netherlands.

37. Kim, J.K., Ulfarsson, G., Kim, S., and Shankar, V. (2013). "Driver Injury-Severity in Single-Vehicle Crashes in California: A Mixed Logit Analysis of Heterogeneity Due to Age and Gender." *Accident Analysis and Prevention*, *50*, pp. 1751–1758, Elsevier, Amsterdam, Netherlands.

38. Malyshkina, N. and Mannering, F. (2010). "Zero-state Markov Switching Count-Data Models: An Empirical Assessment." *Accident Analysis and Prevention*, *42*, pp. 122–130, Elsevier, Amsterdam, Netherlands.

39. Mitra, S. and Washington, S. (2012). "On the Significance of Omitted Variables in Intersection Crash Modeling." *Accident Analysis and Prevention*, *49*, pp. 439–448, Elsevier, Amsterdam, Netherlands.

40. Sacchi, E. and Sayed, T. (2014). "Accounting for Heterogeneity among Treatment Sites and Time Trends in Developing Crash Modification Functions." *Accident Analysis and Prevention*, *72*, pp. 116–126, Elsevier, Amsterdam, Netherlands.

41. Burch, C., Cook, L., and Dischinger, P. (2014). "A Comparison of KABCO and AIS Injury Severity Metrics using CODES Linked Data." *Traffic Injury Prevention*, *15*(6), pp. 627–630, Taylor & Francis, London, England.

42. Daniello, A. and Gabler, H. (2012). "Characteristics of Injuries in Motorcycle-To-Barrier Collisions in Maryland." *Transportation Research Record: Journal of the Transportation Research Board*, *2281*, pp. 92–98, TRB, Washington, DC.

43. Clark, D.E., Winchell, R.J., and Betensky, R.A. (2013). "Estimating the Effect of Emergency Care on Early Survival after Traffic Crashes." *Accident Analysis and Prevention*, *60*, pp. 241–247, Elsevier, Amsterdam, Netherlands.

44. Hallmark, S.L., Tyner, S., Oneyear, N., Carney, C., and McGehee, D. (2015). "Evaluation of Driving Behavior on Rural 2-Lane Curves using the SHRP 2 Naturalistic Driving Study Data." *Journal of Safety Research*, *54*, pp. 17–27, Elsevier, Amsterdam, Netherlands.

45. Wu, K.F., Aguero-Valverde, J., and Jovanis, P.P. (2014). "Using Naturalistic Driving Data to Explore the Association Between Traffic Safety-Related Events and Crash Risk at Driver Level." *Accident Analysis and Prevention*, *72*, pp. 210–218, Elsevier, Amsterdam, Netherlands.

46. Lyon, C., Persaud, B., and Eccles, K. (2015). *Safety Evaluation of Centerline Plus Shoulder Rumble Strips*, Report No. FHWA-HRT-15-048, Federal Highway Administration, Washington, DC.

47. Shea, M.S., Le, T.Q., and Porter, R.J. (2015). "Combined Crash Frequency–Crash Severity Evaluation of Geometric Design Decisions: Entrance–Exit Ramp Spacing and Auxiliary Lane Presence." *Transportation Research Record: Journal of the Transportation Research Board*, *2521*, pp. 54–63, TRB, Washington, DC.

48. Gross F., Persaud, B., and Lyon, C. (2010). *A guide to developing quality crash modification factors*, Report No. FHWA-SA-10-032, Federal Highway Administration, Washington, DC.

49. Hauer, E. (1997) *Observational before-after studies in road safety*, Pergamon, Oxford, England.

50. Rosenbaum, P., and Rubin, D.B. (1983). "The central role of the propensity score in observational studies for causal effects." *Journal of the Royal Statistical Society*: *Series B*, *70*(1), pp. 41–55, Wiley-Blackwell, Hoboken, NJ.

51. Davis, G.A. (2000). "Accident reduction factors and causal inference in traffic safety studies: a review." *Accident Analysis and Prevention*, *32*(1), pp. 95–109, Elsevier, Amsterdam, Netherlands.

52. Imbens, G.W. and Rubin, D.B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, England.

53. Liu, D, Guo, F, and Li, F. (2020). "Evaluating the causal effects of cellphone distraction on crash risk using propensity score methods." *Accident Analysis and Prevention*, *143*, pp. 105–579, Elsevier, Amsterdam, Netherlands.

54. Abadie, A. and Imbens, G.W. (2011). "Bias-corrected matching estimators for average treatment effects." *Journal of Business and Economic Statistics*, *29*, pp. 1–11, Taylor & Francis, London, England.

55. Bang, H. and Robins J.M. (2005). "Doubly robust estimation in missing data and causal inference models." *Biometrics*, *61*, pp. 962–972, International Biometric Society, Washington, DC.

56. ArcGIS. (2020). "ArcMap." (website) Environmental Systems Research Institute, Inc. Available online: https://desktop.arcgis.com/en/arcmap, last accessed April 28, 2020.

57. Cameron, A.C., and Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, England.

58. National Center for Statistics and Analysis. (2016). *2015 motor vehicle crashes: Overview*, Report No. DOT HS 812 318, National Highway Traffic Safety Administration, Washington, DC.

59. National Highway Traffic Safety Administration. (2015). *Traffic Safety Facts 2014: A compilation of motor vehicle crash data from the Fatality Analysis Reporting System and the General Estimates System*, Report No. DOT HS 812 261, NHTSA, Washington, DC.

60. Government Highway Safety Association. (2008). *MMUCC Guideline: Model Minimum Crash Criteria, Third Edition.* United States Department of Transportation, Washington, DC. Available online: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810957, last accessed April 28, 2020.

61. Hedlund, J. (2008). *Traffic Safety Performance Measures for States and Federal Agencies*, Report No. HS-811 025, National Highway Traffic Safety Administration, Washington, DC.

62. Gennarelli, T.A. (1985). *Abbreviated Injury Scale*. American Association for Automotive Medicine, Chicago, IL.

63. Cook, L.J., Olson, L.M., and Dean, J.M. (2001). "Probabilistic record linkage: Relationships between file sizes, identifiers and match weights." *Methods of Information in Medicine*, *40*(3), pp. 196–203, Schattauer, Stuttgart, Germany.

64. Cook, L.J., Thomas, A., Olson, C., Funai, T. and Simmons, T. (2014). *Crash Outcome Data Evaluation System (CODES): An examination of methodologies and multi-state traffic safety applications*, Report No. DOT HS 812 179, United States Department of Transportation, Washington, DC.

65. McGlincy, M.H. (2004). *A Bayesian record linkage methodology for multiple imputation of missing links*. Proceedings of the Joint Statistical Meetings, Toronto, Canada.

66. Strategic Matching, Inc. (2019). "Strategic Matching: Record Linkage Solutions." (website) Strategic Matching Inc., Morrisonville, NY. Available online: http://strategicmatching.com, last accessed April 28, 2020.

67. MacKenzie, E.J., Steinwachs, D.M., and Shankar, B. (1989). "Classifying trauma severity based on hospital discharge diagnoses: validation of an ICD-9CM to AIS-85 conversion table." *Medical Care*, 1989 April 1, *4,* pp. 412-422, Wolters Kluwer, Alphen aan den Rijn, Netherlands.

68. MacKenzie, E.J. and Sacco, W. (1997). *ICDMAP-90: A User's Guide*, The Johns Hopkins University School of Public Health and Tri-Analytics, Inc., Baltimore, MD.

69. SAS Institute Inc. (2020). "SAS 9.4." (website) SAS, Cary, NC. Available online: https://www.sas.com/en_us/software/sas9.html, last accessed April 28, 2020.

70. Utah Department of Public Safety. (2004). *Utah Crash Summary 2004*, Utah Department of Public Safety, Salt Lake City, UT. Available online: https://site.utah.gov/dps-highwaysafe/wp-content/uploads/sites/22/2015/02/2004-Utah-Crash-Summary.pdf, last accessed October 11, 2016.

71. National Highway Traffic Safety Administration. (2006). *Utah Investigators Vehicle Crash Report*, NHTSA, Washington, DC. Available online: https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/ut_di9manual_rev112005_sub012606.pdf, last accessed October 11, 2016.

72. Stewart, J. (1996). "Applications of classification and regression tree methods in roadway safety studies." *Transportation Research Record: Journal of the Transportation Research Board*, *1542*, pp. 1–5, TRB, Washington, DC.

73. Abdel-Aty, M., Keller, J., and Brady, P. (2005). "Analysis of types of crashes at signalized intersections by using complete crash data and tree-based regression." *Transportation Research Record: Journal of the Transportation Research Board*, *1908*, pp. 37–45, TRB, Washington, DC.

74. Yan, X., Richards, S., and Su, X. (2010). "Using hierarchical tree-based regression model to predict train–vehicle crashes at passive highway-rail grade crossings." *Accident Analysis and Prevention*, *42*(1), pp. 64–74, Elsevier, Amsterdam, Netherlands.

75. Khan, G., Bill, A.R., and Noyce, D.A. (2015). "Exploring the feasibility of classification trees versus ordinal discrete choice models for analyzing crash severity." *Transportation Research Part C: Emerging Technologies*, 50, pp. 86–96, Elsevier, Amsterdam, Netherlands.

76. Saha, D., Alluri, P., and Gan, A. (2015). "Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees." *Accident Analysis and Prevention*, *79*, pp. 133–144, Elsevier, Amsterdam, Netherlands.

77. Le, T.Q. and Porter, R.J. (2012). "Safety Evaluation of Geometric Design Criteria for Spacing of Entrance-Exit Ramp Sequence and Use of Auxiliary Lanes." *Transportation Research Record, Journal of the Transportation Research Board*, *2309*, pp. 12–20, TRB, Washington, DC.

78. Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of machine learning*. MIT press, Cambridge, MA.

79. Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A. (1984). *Classification and regression trees*. CRC press, Boca Raton, FL.

80. Hand, D.J., Mannila, H., and Smyth, P. (2001). *Principles of data mining*. MIT press, Cambridge, MA.

81. Shalizi, C. (2006). "Lecture 10: Regression Trees." *Lecture Notes of Course 36-350: Data Mining*, Carnegie Mellon University, Pittsburgh, PA. Available online: http://www.stat.cmu.edu/~cshalizi/350-2006/lecture-10.pdf, last accessed October 11, 2016.

82. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, *6*, Springer, New York, NY.

83. Ho, T.K. (1995). "Random decision forests." *Proceedings of 3rd international conference on document analysis and recognition*, *1*, pp. 278–282, IEEE, New York, NY.

84. Breiman, L. and Cutler, A. (2001). "Random Forests." *Machine Learning*, *45*(1), pp. 5–32, Springer, New York, NY.

85. Brewer, S. (2016). "Classification and Regression Trees." *Lecture Slides of Course GEOG 6000 Advanced Geographical Data Analysis*, University of Utah, Salt Lake City, UT.

86. Google Earth. "Google Earth." (website) Available online: https://www.google.com/earth, last accessed April 28, 2020.

87. Google Maps. (2020). "Google Maps." (website) Available online: https://www.google.com/maps/@42.3636243,-71.1804872,15z, last accessed April 28, 2020.

88. Washington State Department of Transportation. (2016) "Interchange Viewer." (website) WSDOT, Olympia, WA. Available online: https://www.wsdot.wa.gov/mapsdata/tools/InterchangeViewer/default.htm, last accessed April 28, 2020.

89. Washington State Department of Transportation. (2016) "State Route Web Tool (SRweb)." (website) WSDOT, Olympia, WA. Available online: https://www.wsdot.wa.gov/mapsdata/tools/srweb.htm, last accessed April 28, 2020.

90. California Department of Transportation. (2020). "Welcome to PeMS." (website) Caltrans, Sacramento, CA. Available online: http://pems.dot.ca.gov, last accessed April 28, 2020.

91. Federal Highway Administration. "Highway Safety Information System." (website) FHWA, Washington, DC. Available online: https://www.hsisinfo.org, last accessed April 28, 2020.

92. Quinlan, J. (1988). *Programs for Machine Learning*. Morgan Kauffmann Publishers, San Mateo, CA.

93. Statistics Solutions. (2019). "CHAID." (website) Statistics Solutions, Clearwater, FL. Available online: https://www.statisticssolutions.com/non-parametric-analysis-chaid, last accessed April 28, 2020.

94. Loh, W. (2020). *User Manual for Guide ver. 33.1*. Department of Statistics University of Wisconsin-Madison, Madison, WI. Available online: http://pages.stat.wisc.edu/~loh/treeprogs/guide/guideman.pdf, last accessed April 28, 2020.

HRDS-20/11-20(100)E