Analyzing Highway Safety Datasets: Simplifying Statistical Analyses from Sparse to Big Data

JULY 2019 | Final Report

SAFE-D
SAFETY THROUGH DISRUPTION

Texas A&M Transportation Institute

San Diego State University
Leadership Starts Here

VT | TRANSPORTATION INSTITUTE
VIRGINIA TECH

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. 01-001	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Analyzing Highway Safety Datasets: Simplifying Statistical Analyses from Sparse to Big Data		5. Report Date July 2019 6. Performing Organization Code:
7. Author(s) Dominique Lord Srinivas Reddy Geedipally Feng Guo Arash Jahangiri Mohammadali Shirazi Huiying Mao Xinwei Deng		8. Performing Organization Report No. Report 01-001
9. Performing Organization Name and Address: Safe-D National UTC Texas A&M Transportation Institute Virginia Tech Transportation Institute San Diego State University		10. Work Unit No. 11. Contract or Grant No. 69A3551747115/Project 01-001
12. Sponsoring Agency Name and Address Office of the Secretary of Transportation (OST) U.S. Department of Transportation (US DOT) State of Texas		13. Type of Report and Period Final Research Report 14. Sponsoring Agency Code

15. Supplementary Notes

This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program, and, in part, with general revenue funds from the State of Texas.

16. Abstract

Data used for safety analyses have characteristics that are not found in other disciplines. In this research, we examine three characteristics that can negatively influence the outcome of these safety analyses: (1) crash data with many zero observations; (2) the rare occurrence of crash events (not necessarily related to many zero observations); and (3) big datasets. These characteristics can lead to biased results if inappropriate analysis tools are used. The objectives of this study are to simplify the analysis of highway safety data and develop guidelines and analysis tools for handling these unique characteristics. The research provides guidelines on when to aggregate data over time and space to reduce the number of zero observations; uses heuristics for selecting statistical models; proposes a bias adjustment method for improving the estimation of risk factors; develops a decision-adjusted modeling framework for predicting risk; and shows how cluster analyses can be used to extract relevant information from big data. The guidelines and tools were developed using simulation and observed datasets. Examples are provided to illustrate the guidelines and tools.

17. Key Words Safety, Big Data, Sparse Data, Heuristics Method, Cluster Analysis, Finate Sample Bias Adjustment, Aggregated Data, Disaggregated Data	18. Distribution Statement No restrictions. This document is available to the public through the Safe-D National UTC website, as well as the following repositories: VTechWorks, The National Transportation Library, The Transportation Library, Volpe National Transportation Systems Center, Federal Highway Administration Research Library, and the National Technical Reports Library.

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 30	22. Price $0

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

Abstract

Data used for safety analyses have characteristics that are not found in other disciplines. In this research, we examine three characteristics that can negatively influence the outcome of safety analyses: (1) crash data with many zero observations; (2) the rare occurrence of crash events (not necessarily related to many zero observations); and (3) big datasets. These characteristics can lead to biased results if inappropriate analysis tools are used. The objectives of this study are to simplify the analysis of highway safety data and develop guidelines and analysis tools for handling these unique characteristics. The research provides guidelines on when to aggregate data over time and space to reduce the number of zero observations; uses heuristics for selecting statistical models; proposes a bias adjustment method for improving the estimation of risk factors; develops a decision-adjusted modeling framework in predicting risk; and shows how cluster analyses can be used to extract relevant information from big data. The guidelines and tools were developed using simulation and observed datasets. Examples are provided to illustrate the guidelines and tools.

Acknowledgements

The research team would like to thank everyone at the Safe-D program, especially Dr. Sue Chrysler and Mrs. Martha R. Taylor at the Texas A&M Transportation Institute, as well as Mrs. Melissa Hulse and Mrs. Leslie Harwood at the Virginia Tech Transportation Institute, for the helping with the management of the project. The team members would also like to thank Dr. Soma Dhavala for providing very useful help on the statistical methods related to the heuristic approach developed in this research, Dr. Gerardo Flintsch for sharing the data, and Dr. John Ivan for reviewing this report.

This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program, and, in part, with general revenue funds from the State of Texas. The research in this report was developed as part of a collaborative project between San Diego State University, the Texas A&M Transportation Institute, and the Virginia Tech Transportation Institute.

Table of Contents

List of Figures

List of Tables

Introduction

Data used for safety analyses have characteristics that are not typically found in other disciplines. In this research, we examine three characteristics of transportation safety data that pose challenges for analysis.

The first important characteristic is related to datasets that include a large amount of zero responses. Modeling crash data with many zero observations requires two critical precautions:

Assembling and formatting data. As documented in Lord and Geedipally (*1*), excess zero observations are often attributed to how data are assembled or formatted in spatial or temporal scales. For example, more zero observations are expected in data that are aggregated weekly rather than monthly or yearly. Finding a balance in aggregation is a critical task in data preparation. On the one hand, using disaggregated data may result in having excessive zero observations, in which the traditional negative binomial (NB) model may not be appropriate for the safety analysis (*1*). On the other hand, too much aggregation may result in loss of information (*2*), although it may make the NB model a better alternative. Although several researchers have encountered this issue, there is no proper guidance on whether an aggregated data model is better than a disaggregated data model or vice-versa (*3*, *4*). In this study, we address this issue by conducting a simulation study and measuring the information loss as a function of the precision or accuracy in estimation of the coefficients.

Selection of an appropriate distribution or model. Various distributions and models have been proposed to model crash data (*5*, *6*). Selecting the most appropriate distribution plays a crucial role in safety analyses. Often, the comparison of distributions (or models) is accomplished during the post-modeling phase, using measures such as the goodness-of-fit (GoF) statistics. However, these metrics are neither easy to compute nor practically attainable in some instances when many alternatives exist and/or the analyst deals with big data or excess zero observations. In addition, and most importantly, these metrics do not provide any intuition into why one distribution may be preferred over another or the logic behind the model selection (goodness-of-logic or GoL, as illustrated by *7*). For example, what is a more appropriate distribution to model a skewed dataset or one with excess zero responses? In this research, we address this issue by seeking a model selection method based on the characteristics of the data.

The second important characteristic is the rare occurrence of crash events, which is related to the first characteristic (although this does not necessarily mean that the dataset contains a large percentage of zero responses). The NB regression is a fundamental statistical analysis tool for traffic safety modeling. As crashes are rare events, a limited number of crashes and/or imbalanced data could lead to finite sample bias (i.e., a biased regression parameter estimation). In this study, we propose a bias-correction procedure for more accurate estimation when evaluating the impacts of crash risk factors based on the approximated bias derived by McCullagh and Nelder (*8*) for generalized linear models. We evaluated the finite sample bias issue through a simulation study and infrastructure safety evaluation case study. Furthermore, we developed a decision-adjusted framework to develop risk prediction models tailored for a specific goal, such as rare event prediction.

The third important characteristic is related to big datasets, which are now becoming more prevalent with the use of naturalistic data. There are many methodologies for handling large datasets, such as distributed cloud computing, parallel algorithms, machine learning techniques, and data mining methods (*9*, *10*). The present report focuses on cluster analysis methods, which are often categorized under data mining approaches. With crash data analysis, the challenge is that the data have become so large and complex (e.g., naturalistic data) that data storing, processing, and modeling are cumbersome. When many variables are involved, cluster analysis as a sampling strategy can be applied to select a representative subset of data to deal with this challenge. More variables can be used in cluster analysis compared to traditional sampling methods, such as stratified sampling, without having to increase the sample size. This makes cluster analysis a suitable alternative when the population is too large (*11*). Hence, in this study, we show how cluster analysis can be used for extracting information from big datasets.

Research Objectives

The objectives of this study are to provide guidelines and tools for the analysis of highway safety data characterized by excess zero responses, rare events, and big data. The objectives are divided into three categories: (1) analyzing data with many zero observations, (2) the rare event issue inherent in transportation safety analyses, and (3) handling big datasets. For the first objective, we develop guidelines for aggregating data over time and space, as well as heuristics to determine when the Poisson-lognormal (PLN) is preferred to the NB model and when the negative binomial Lindley (NB-L) is preferred to the NB model. For the second objective, we propose bias adjustment for more accurate estimation of the safety impact of a risk factor and develop a decision-adjusted modeling framework to consider the study objective in predicting crash risk. For the third objective, we utilize cluster analysis methods to classify data into groups with similar characteristics and create predictors using cluster analysis to potentially produce insights or reduce the number of random variables.

Research Outline

The research report is divided into eight sections. The first section provides the background information related to the data and safety modeling issues. The second section describes the analyses related to the spatial and temporal aggregation of the data. The third section covers the results related to the heuristics methods for selecting models. The fourth section presents the results for the finite sample bias adjustment method. The fifth section shows the characteristics of the decision-adjusted modeling framework. The sixth section presents the results of the cluster analysis. The seventh section documents the proposed guidelines. The final section provides the summary and conclusions.

Background

This section briefly summarizes the key literature on (1) models that have been proposed for excess zero responses; (2) spatial and temporal aggregation; (3) issues associated with the bias caused by sparse data; and (4) cluster analysis methods.

Models for Excess Zero Responses

Several mixture models have been proposed to address data with excess zero observations, for example the NB-L (12, 13), NB generalized exponential (14) and NB-Dirichlet process (15). Here, we introduce the NB-L model, which has been used in several recent research studies and is a preferred model for dealing with excess zero responses and/or high dispersion. The probability density function of the Lindley distribution (16) is defined as:

$$\text{Lindley}(v|\theta) = \frac{\theta^2}{\theta+1}(1+v)e^{-\theta v} \qquad \theta > 0, v > 0 \tag{1}$$

The random variable y is distributed by the NB-L (ϕ, θ) distribution (*12*, *17*):

$$y \sim \text{NB}(\phi, p = 1 - e^{-\lambda}) \text{ and } \lambda \sim \text{Lindley}(\theta) \tag{2}$$

The Lindley distribution, in fact, is a mixture of two gamma distributions as follows:

$$\lambda \sim \frac{1}{1+\theta}\text{amma}\ \ (2,\theta) + \frac{\theta}{1+\theta}\text{amma}\ \ (1,\theta) \tag{3}$$

Therefore, the NB-L distribution can be written in the following hierarchical representation:

$$y \sim \text{NB}\left(y\middle|\phi, p = 1 - e^{-\lambda}\right) \tag{4a}$$

$$\lambda \sim \text{amma}\ \ (1+z,\theta) \tag{4b}$$

$$z \sim \text{Bernoulli}\left(\frac{1}{1+\theta}\right) \tag{4c}$$

The mean of the NB-L distribution is equal to (17):

$$\mu = \phi\left(\frac{\theta^3}{(\theta+1)(\theta-1)^2} - 1\right) \tag{5}$$

Lord and Geedipally (12) showed that using the NB-L distribution to fit the data performs better than the NB distribution when the dataset has many zeros or is characterized by a heavy (or long) tail. However, it is not clear at what point the NB-L distribution should be used instead of the NB distribution (18). In this research, we design model selection heuristics to select the distribution closest to the true one for modeling crash data between these two distributions.

Spatial and Temporal Aggregation

Excess zero observations in data can be attributed to four major factors (19): (1) using spatial or time scales that are too small; (2) under-reporting or misreporting of the number or severity of crashes; (3) sites characterized by low exposure and high risk; and (4) bias caused by omitting important variables in the crash data process. The first factor can potentially be overcome by adjusting the time and scale while compiling the datasets (1). On the other hand, the second, third, and fourth factors should be addressed by applying appropriate statistical models (19). Often, researchers use statistical tests or GoF statistics to decide on the level of aggregation (see 20 as an example). However, these comparison criteria are not appropriate since the nature of the data, as well as the sample size, will change as the data are aggregated. In this research, we examine the influence of the level of aggregation using an extensive simulation study by accounting for the accuracy in the coefficient estimation. In the end, guidelines that are based on characteristics of the data are provided.

Bias Correction for Poisson and NB Regression

The Poisson and NB models are generally estimated using the maximum likelihood method. When the sample size is small and/or when the number of events is limited (e.g., small number of

crashes), the maximum likelihood estimators (MLEs) are biased and the bias could be substantial. This finite sample bias could lead to incorrect estimation of the impacts of risk factors and jeopardize traffic safety improvement efforts. Generally, there are two approaches to reduce MLE bias. One approach is based on applying the Jefferys invariant prior to the likelihood function to directly generate an improved estimator (21, 22, 23). The other approach reduces the bias by subtracting the approximated bias from the regular MLE (8, 24, 25). Both approaches can reduce the bias from $O(n^{-1})$ to $O(n^{-2})$. While the method based on the Jefferys invariant prior does not have a closed-form expression, McCullagh and Nelder (8) provide a specific correction formula for the coefficient estimation.

The finite sample bias of Poisson and NB regression models has been sporadically investigated (26, 27). Saha and Paul (27) studied the bias-corrected dispersion parameter estimation of the NB regression, which showed less bias and superior efficiency compared to regular MLE, the method of moments estimator, and the maximum extended quasi-likelihood estimators in most instances. Giles and Feng (26) derived a bias-correction formula for the parameter estimation of a Poisson regression from the general definition of residuals by Cox and Snell (28). However, research has been limited on identifying the situations where the bias correction is necessary and what factors affect the magnitude of bias. This study addresses this gap by studying the finite sample bias for the parameter estimation of Poisson and NB regression models in the context of traffic safety modeling.

Decision-adjusted Crash Risk Prediction

Predicting crash risk and identifying high-risk drivers are critical for developing appropriate safety countermeasures, driver education programs, and user-based insurance. However, predicting driver risk is a challenging task because crashes are rare events and many factors contribute to individual crash risk. As in-vehicle data collection becomes more prevalent and cost-effective, it has become more feasible to improve risk prediction by utilizing kinematics information. Currently, there are several challenges to implementing kinematics-based driver risk prediction models. We focus on two primary issues: (1) the decision rule and (2) the optimal threshold values for kinematics predictors.

One approach is to choose the thresholds that can maximize the area under the curve (AUC) for the receiver operating characteristic (ROC) curve (29, 30). However, maximizing the AUC of an ROC curve is derived with respect to the entire range of risk and is not necessarily optimized with a specific objective, such as predicting a small percentage of high-risk drivers. In this study, we propose a decision-adjusted modeling approach, where the thresholds are chosen to optimize the particular decision. The Second Strategic Highway Research Plan (SHRP 2) naturalistic driving data were used for model development and calibration.

Cluster Analysis Methods

Cluster analysis, also known as data clustering, segmentation analysis, and taxonomy analysis, is a classification problem in which all observations are classified into distinct categories. The aim is to place observations in different clusters in a way that observations in the same cluster are as similar to each other as possible while observations in different clusters are as dissimilar as possible. It is important to note that in cluster analyses when the categories are unknown a priori, an unsupervised classification approach should be applied. On the other hand, if the categories are

known, a supervised classification approach would be appropriate (11). There are several categories of clustering methods, such as partitional, hierarchical, density-based, grid-based, and model-based (31, 32). The present study focuses on the first three categories: partitional algorithms identify all data clusters simultaneously in an iterative process. Each cluster usually has a centroid or an actual observation that is the most representative member of the cluster. Hierarchical algorithms either start with the entire data as one large cluster and recursively partition this big cluster into smaller ones (i.e., divisive), or start with many small clusters each having only one observation and recursively merge these small clusters to create larger ones (i.e., agglomerative). In contrast to the partitional algorithms that produce data clusters in a single level, hierarchical algorithms produce a dendrogram structure with each leaf representing a data cluster (33, 11). Density-based algorithms identify clusters by finding regions with high object density in the data space (34). This makes these algorithms capable of identifying arbitrary shaped clusters and handling outliers.

Methods and Results

Spatial and Temporal Aggregation

Crash data at a site are usually defined as a count number over the space and time scales. Therefore, the number of zero observations in the compiled dataset is directly correlated with the selected spatial and/or temporal scales. By adjusting the time and spatial scales, the number of zero responses observed in the dataset can increase or decrease. For example, by changing the segment length of a site from 0.1 mile to 1 mile, the number of zero observations in the complied dataset will be reduced since the new segment will include all of the crashes on the segments now aggregated. Similarly, changing the time scale from monthly durations to yearly periods will result in a reduction of the number of zero responses in the dataset.

Simulation Protocol

The simulation protocol in this part of the study used three main steps. The detailed steps are described in Appendix A, but they are briefly summarized as follows:

Step 1: The mean number of crashes at each site i and time period m is estimated using the following functional form: $\mu_i^m = e^{\sum_{j=1}^{d} \beta_j x_{ij}^m}$. Here, the index j=1 to d represents the independent variables. βs, the "true" parameters, are taken from a previous study.

Step 2: Crash counts are simulated using an NB distribution. First, a disaggregated dataset is created for each site and time period. Then, datasets are combined into one dataset for all time periods. Second, an aggregated dataset is created for each site i. Step 2 is repeated for n times (500 simulation runs for this study).

Step 3: For each simulation run, an NB regression is estimated for the disaggregated and aggregated datasets, and the standard deviations of the coefficients between the two models are compared.

Simulation Analysis

An existing dataset was considered and two variables, ADT and skid number, from it were used for the simulation analysis. Variables were collected for 5 years, in one-year duration (i.e., each

year is a unique observation). The data for skid number are recorded for at least 3 years out of the 5. Consequently, for some sites the data for skid number are not complete. The inverse dispersion parameter (φ^m) was directly calculated from observed crash data for each year. The average value over the 5 years is around 0.2, which means the data are highly dispersed. Two major scenarios for highly dispersed data were created: (1) data that involve around 90% zero observations and (2) data with 50% zero observations. Each major scenario was divided into seven sub-scenarios, based on year-to-year variation of the skid number. The sub-scenario (1-1) only includes records in which the skid number variation from year to year is always less than 20%. Recursively, sub-scenario (1-2) assumes 30% variation. Sub-scenario (1-3) assumes 40% variation, etc. The last sub-scenario (sub-scenario 1-7) includes the full data.

Table 1 and Table 2 indicate the results of the simulation for different scenarios. Note that even though the ADT variable is used in the analysis, those results are not presented here because the primary focus is on the skid number variable only. As shown in these tables, as the change in variation of the skid number when data are aggregated (CV_{Skid} change) increases, the standard deviation of the estimated parameter in the aggregated data becomes larger than in the disaggregated data. Therefore, aggregation of data becomes less reliable. The decision point can be quantified by the change in coefficient of variation (CV) of the variable in the dataset. For example, in Scenario 1-3, the change in CV of the skid number when data are aggregated is equal to 6.8%. In that regard, it seems that a change in CV by 7% in a variable is a decision point to stop the aggregation, when the data have a high percentage of zero observations. On the other hand, when the percentage of zero observations is small, the aggregation can be stopped when the change in CV of a variable is greater than 4%.

Table 1. Simulation Results for Scenario with About 90% Zeros

Scenario	True Value	Skid Number Year-to-Year Variation	**CV_{Skid}** Change	Disaggregated Data Mean	Disaggregated Data Std.	Aggregated Data Mean	Aggregated Data Std.
1-1	-0.005914	< = 20% (n_1=1570; n_2=6270)	0.1%	-0.005590	0.004308	-0.005649	0.003838
1-2	-0.005914	< = 30% (n_1=2410; n_2=9602)	3.6%	-0.005939	0.003162	-0.005963	0.002878
1-3	-0.005914	< = 40% (n_1=3112; n_2=12368)	6.8%	-0.005854	0.002601	-0.005965	0.002510
1-4	-0.005914	< = 50% (n_1=3664; n_2=14528)	10.5%	-0.006039	0.002219	-0.006011	0.002278
1-5	-0.005914	< = 60% (n_1=4042; n_2=16047)	14.0%	-0.005944	0.002213	-0.005886	0.002257
1-6	-0.005914	< = 80% (n_1=4295; n_2=17083)	17.1%	-0.005827	0.002050	-0.005960	0.002153
1-7	-0.005914	Full Data (n_1=4402; n_2=17504)	18.7%	-0.005945	0.001913	-0.005898	0.002291

Note: n_1 = sample size of the aggregated data, n_2 = sample size of the disaggregated data; Bold numbers represent the preferred values. For the disaggregated data, zeros = 90.07% and crash mean = 0.163. For the aggregated data, zeros = 61.30% and crash mean = 1.933.

Table 2. Simulation Results for Scenario with About 50% Zeros

Scenario	True Value	Skid Number Year-to-Year Variation	CV_{Skid} Change	Disaggregated Data Mean	Disaggregated Data Std.	Aggregated Data Mean	Aggregated Data Std.
2-1	-0.005914	< = 20% (n1=1570; n2=6270)	0.1%	-0.005939	0.002983	-0.005950	0.0027644
2-2	-0.005914	< = 30% (n1=2410; n2=9602)	3.6%	-0.005789	0.002152	-0.006018	0.001996
2-3	-0.005914	< = 40% (n1=3112; n2=12368)	6.8%	-0.005843	0.001639	-0.005996	0.001662
2-4	-0.005914	< = 50% (n1=3664; n2=14528)	10.5%	-0.005882	0.001586	-0.006043	0.001644
2-5	-0.005914	< = 60% (n1=4042; n2=16047)	14.0%	-0.005899	0.001422	-0.006045	0.001484
2-6	-0.005914	< = 80% (n1=4295; n2=17083)	17.1%	-0.005925	0.001401	-0.005987	0.001503
2-7	-0.005914	Full Data (n1=4402; n2=17504)	18.7%	-0.005884	0.001275	-0.005982	0.0014620

Note: n1 = sample size of the aggregated data, n2 = sample size of the disaggregated data; Bold numbers represent the preferred values. For disaggregated data, zeros = 50.04% and crash mean = 9.82. For aggregated data, zeros = 3.81% and crash mean = 49.15.

Case Studies

To identify high-risk segments based on fatal (K) and incapacitating injury (A) crashes, Geedipally et al. (35) conducted spatial aggregation because there were many segments with zero crashes. They aggregated adjacent segments when the change in the ADT was less than a certain threshold. Although the number of zero observations in the data decreased after aggregation, it was not clear to the them when to optimally stop the aggregation. Since the disaggregated data had about 50% zeros, the simulation results suggest stopping the aggregation when the change in CV is above 4%. Table 3 shows the spatial aggregation of adjacent interstate segments in Texas. The adjacent segments were aggregated if they were on the same highway and all other variables remained the same. As per the simulation results, it is recommended to stop the aggregation when the change in ADT is 25% or less between adjacent segments.

Table 3. Spatial Aggregation of Interstate Segments

Aggregation Criteria	Number of segments	% sites with no crashes	$\mathbf{CV_{ADT}}$	Change in $\mathbf{CV_{ADT}}$
Existing	2321	54%	0.58	--
ADT within +10%	519	25%	0.57	2%
ADT within +15%	483	23%	0.56	3%
ADT within +20%	463	23%	0.56	3%
ADT within +25%	451	22%	0.56	3%
ADT within +50%	426	22%	0.55	5%

Pratt et al. (3) developed statistical models with both the disaggregated and temporally aggregated data to evaluate the effect of skid resistance on traffic crashes using data from about 40,000

horizontal curves for a 5-year period in Texas. Two datasets yielded different results for the skid number variable, and it is unknown which one represents the true value. In the disaggregated data, each year is considered as a separate observation. However, in the aggregated data, the dependent variable is the sum of crashes over a 5-year period and the skid number is the average over the time period. For this analysis, two scenarios were considered, as shown in Table 4. First, we considered all sites even if the skid number variable is missing for some years. Second, we considered only those sites where the skid number variable is available in all years. Since this dataset had more than 90% zeros, for the first scenario, using the aggregated data is recommended because the change in CV is 6.2%, which is the less than the 7% threshold calculated above. However, for the second scenario, disaggregated data are recommended because the change in CV is 15.6%.

Table 4. Temporal Aggregation of Crashes on Horizontal Curves

Scenario	Data type	CV_{Skid}	Change in CV_{Skid}
I	Disaggregated	0.318	--
	Aggregated	0.299	6.2%
II	Disaggregated	0.306	--
	Aggregated	0.258	15.6%

Heuristics

This section is divided into two parts. First, the methodology used to design heuristics (18) is briefly described. Second, this methodology is used to design heuristics to select a model between the NB and PLN, as well as between the NB and NB-L.

Methodology

At the heart of the proposed methodology lies a paradigm shift in how model selection is both viewed and treated. We view model selection as a classification problem (see 36). To clarify the strategy, let us assume the analyst is interested in choosing between the Poisson and NB distributions based on the population mean and variance. The mean and variance of the population would create a two-dimensional predictor space (Ω). Now, the analyst's task is to partition the predictor space and assign a label to each partition. We know that if the population variance-to-mean ratio (VMR) is greater than 1.0 (VMR > 1), we may choose the NB distribution, and if it is equal to 1.0 (VMR = 1), the Poisson distribution will be the preferred sampling distribution to use. Hence, the predictor space (Ω) can be classified between the Poisson and NB distributions in a way that is shown in Figure 1.

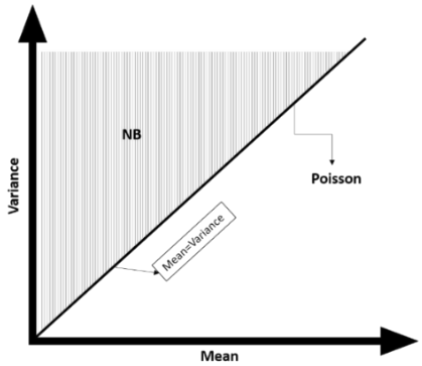

Figure 1. Classifying the NB and Poisson distributions based on the mean and variance of the population.

The decision based on the VMR statistic, in this case, serves as a heuristic to select the most-likely-true sampling distribution between the Poisson and NB distributions. It does not require fitting the models, estimating the model parameters, computing test statistics, etc. It simply uses the descriptive statistics to arrive at a model recommendation. In the case of Poisson versus NB, we know, theoretically, how the two-dimensional predictor space should be partitioned between the Poisson and NB distributions; however, what if such insight was not available to us? In the absence of readily available analytical insights to guide the model selection, we resort to computational approaches (18). It will be assumed that the distributions under consideration can be classified by m summary statistics. These summary statistics would create an m-dimensional predictor space; then, the analyst can benefit from two analytic tools, (1) Monte-Carlo simulations, and (2) machine learning classifiers, to partition the assumed m-dimensional predictor space between the competitive distributions. Using Monte-Carlo simulations, it is possible to simulate numerous datasets (say 100,000 datasets) from each of these distributions (or models) indexed by a label and record the assumed m summary statistics for each. Next, a machine learning classifier can be trained to classify each simulated dataset to predict a model label. Interested readers are referred to the work of Shirazi et al. (18) for the detailed steps of the methodology.

Results

This section is divided into two subsections. First, heuristics for model selection between the NB and PLN are presented. Then, heuristics for model selection between NB and NB-L are documented.

NB versus PLN Heuristics

Simulation is a key step in designing model selection heuristics (18). It is essential to first make sure that the simulated datasets represent the characteristics of the target population, and then ensure that the alternative distributions have fair representations among the simulated data. The first concern can be addressed by simulating data given the most common range observed in the context population, in our case, the crash data population. The second concern can be addressed by ensuring that some summary statistics (referred to as control factors) are distributed similarly among the simulated datasets from alternative distributions. In other words, the analyst seeks to discriminate the distributions based on factors such as the kurtosis and/or skewness, while the control factors such as the mean or VMR are distributed similarly among simulated datasets.

In our problem design, we ensure that the mean and VMR of the data are uniformly distributed among the generated datasets from both of these distributions, simply, by simulating the mean and the VMR from a uniform distribution with a range that is the most commonly observed range in crash data, as shown in Eq. (6):

$$\mathbf{m \sim uniform(0.1, 20)}\ ;\ \mathbf{VMR \sim uniform(1, 25)} \tag{6}$$

Next, the parameters of the NB (μ, ϕ) distribution can be estimated as:

$$\boldsymbol{\mu = m} \quad ; \quad \boldsymbol{\phi = \frac{\mu}{VMR - 1}} \tag{7}$$

Similarly, we have:

$$\boldsymbol{\mu_\lambda = m} \quad ; \quad \boldsymbol{V_\lambda = (VMR - 1) \times \mu_\lambda} \tag{8}$$

Then, the parameters of the PLN (ν, σ^2) distribution can be derived as:

$$\boldsymbol{\nu = \log\left(\frac{\mu_\lambda^2}{\sqrt{v_\lambda + \mu_\lambda^2}}\right)\ ;\ \sigma = \sqrt{\log\left(\frac{v_\lambda}{\mu_\lambda^2} + 1\right)}} \tag{9}$$

It is possible to simulate a dataset with a size of n = 5,000 from both the NB (μ, ϕ), and PLN (ν, σ^2) distributions. The above procedure can be repeated for N = 100,000 iterations. For each one of these distributions, 22 types of summary statistics are recorded. These summary statistics include the mean (μ), variance (σ^2), standard deviation (σ), VMR, CV, skewness (skew), kurtosis (K), percentage of zeros (zeros), quantiles (or percentiles) in 10% increments, the 10th, 20th, 30th, and 40th inter-quantiles (or inter-percentiles), and the range (R). A decision tree classifier (37, 38) was used to partition the 22-dimensional predictor space that is created by the summary statistics and assign a label—either NB or PLN—to each partition. Figure 2 shows the outcome of the decision tree classifier. Note that the tree can be used only for data with the characteristics of 0.1 < mean < 20 and 1 < VMR < 25. Also, it is assumed that the sample size is large.

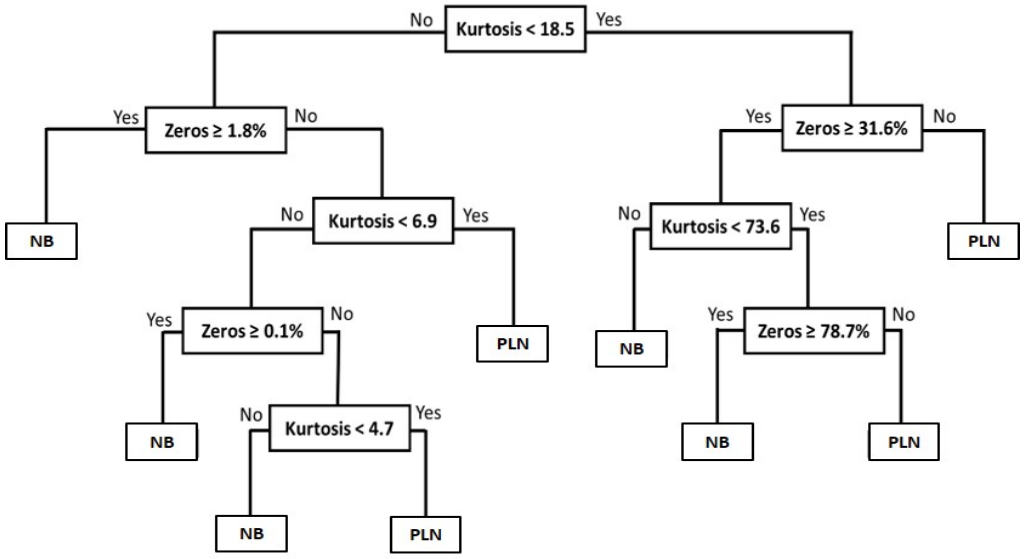

Figure 2. Heuristic to select a model between the NB and PLN distributions.

As shown in Figure 2, the population kurtosis and the percentage of zeros play a substantial role in deciding between the NB and PLN distributions. Overall, the PLN is recommended for situations when data are more skewed but have fewer zero responses, while the NB distribution is a better option otherwise; these results confirm the trends observed and/or reported in previous studies in the literature (5, 39, 40). Unlike previous studies, Figure 2 provides a more perspicuous characteristics-based guidance on selecting a sampling distribution between these two alternatives. The output of a binary classifier can be either true (T) when it correctly classifies the label of the distribution, or false (F) when it misclassifies the label of the correct distribution. Let the PLN and NB distributions, respectively, be labelled as the positive (P) and negative (N) outputs of the binary classification. Such definitions represent a test when the analyst assumes the NB distribution as a base model, while he or she seeks to know when a shift to the PLN distribution is recommended. Table 5 shows the confusion matrix of the binary classification given such assumptions.

Table 5. NB vs. PLN: Confusion Matrix Based on the Results of the **Decision Tree Classifier**

Predicted	PLN True	NB True
PLN	41.50% (TP)	1.18% (FN)
NB	8.50% (FP)	48.82% (TN)

The overall misclassification error is equal to 9.68% and the sensitivity [Note: Sensitivity=TP/(TP+FN)] and specificity [Note: Specificity=TN/(TN+FP)] of the classification are equal to 97.24% and 85.12%, respectively. The sensitivity of the classification is very high, indicating that when the outcome of the binary classifier is the PLN distribution, there is a very high chance that the classifier has correctly detected the label of the distribution. However, the specificity of the classification is not as high as its sensitivity, meaning that when the outcome of the classifier is the NB distribution, there are still some chances that the output label was detected incorrectly. When the output of the classifier is the NB distribution, the analyst may consider other tests as well to decide between these two distributions and/or can decide to choose an alternative tolerance threshold to decide between the NB and PLN.

NB versus NB-L Heuristics

For this comparison, the experiment was designed for datasets with the following range for the mean and VMR of the population that is the most common range observed in crash data:

$$0.1 < \text{mean} < 20 \quad ; \quad 1 < \text{VMR} < 100$$

In total, 100,000 datasets (N = 100,000), each with 5,000 data points (n = 5,000), were simulated from the NB and NB-L distributions. The following uniform distributions were used to simulate the NB and NB-L parameters at each iteration of the simulation:

$$\mu \sim \text{Uniform}(0.1, 20); \text{ for both NB and NB-L}$$

$$\frac{1}{1+\theta} \sim \text{Uniform}(0, 0.5); \text{ for NB-L}$$

$$\phi \sim \text{Uniform}(0.1, 10); \text{ for NB}$$

By simulating the mean of the NB and NB-L distributions from a uniform distribution, we make the distribution of the mean of the simulated datasets generated from both these distributions uniformly distributed. For each simulated dataset, the 22 summary statistics described in the previous section are recorded. A decision-tree classifier is used to partition the predictor space into regions that are most likely to be covered by either the NB or NB-L distributions. As described above, the decision-tree classifier is used for a simple and easy-to-interpret classification (although it is less accurate than other classification methods). Figure 3 shows the results of applying the decision-tree method to partition the 22-dimensional predictor space between the NB and NB-L distributions. Out of the 22 summary statistics used for the analysis, only the skewness of the population was used by the classifier in the decision tree to separate the NB-L distribution from the NB. As shown in Figure 3, the tree involves only one splitting rule. Starting at the top of the tree, it is divided into two sections based on the value of the skewness. The observations that have a skewness of less than 1.92 are assigned to the left branch and the "NB" label is assigned to them. On the other hand, when the value of the skewness is greater than 1.92, the NB-L distribution is

recommended. Note that the tree can only be used for data with 0.1 < mean < 20 and 1 < VMR < 100. Also, it is assumed that the sample size is large.

Figure 3. Heuristic to select a model between the NB and NB-L distributions.

The classification between the NB and NB-L distributions can be seen in a binary-classification fashion. The confusion matrix for the results of the classification problem can be structured as shown in Table 6. The overall misclassification error (FP + FN) is equal to 5.90%. The sensitivity and specificity of the classification are equal to 89.96% and 99.21%, respectively.

Table 6. NB vs. NB-L: Confusion Matrix Based on the Results of the Decision Tree Classifier

Predicted	NB-L True	NB True
NB-L	49.64% (TP)	5.54% (FN)
NB	0.36% (FP)	44.46% (TN)

Finite Sample Bias Adjustment

This section presents the bias-correction procedure based on the approximated bias correction formulation provided by McCullagh and Nelder (8), and a case study on infrastructure safety evaluation.

Methodology

The Poisson regression assumes that the frequency of events Y_i (e.g., the crash count) follows a Poisson distribution,

$$Y_i \sim \text{Poisson}(\lambda_i \cdot E_i),\ i = 1,2,\cdots,n, \tag{10}$$

where λ_i is the expected crash rate for the i^{th} road segment or the i^{th} driver, and E_i is its corresponding exposure, which could be the length of the observation period or the total vehicle miles traveled. A log link function is used to link the expected event rate λ_i with a linear transformation of p explanatory variables, $X_{i1}, X_{i2}, \cdots, X_{ip}$,

$$\log(\lambda_i) = X_i'\beta, \tag{11}$$

where β is a vector of regression coefficients, $\beta = (\beta_0, \beta_1, \cdots, \beta_p)'$; X_i is the covariates vector for entity i, $X_i = (1, X_{i1}, \cdots, X_{ip})$. The coefficient β_j indicates the impact of the j^{th} variable on crash risk, $j = 1, \cdots, p$. The estimation of β is the focus of safety evaluation.

Denote the MLE as $\hat{\beta}$. Based on the approximated bias provided by McCullagh and Nelder (8), the bias-corrected coefficient estimate $\tilde{\beta}$ can be calculated as

$$\tilde{\beta} = \hat{\beta} - \widehat{\text{bias}(\hat{\beta})} = \hat{\beta} - \left(\text{X}'\widehat{\text{W}}\text{X}\right)^{-1}\text{X}'\widehat{\text{W}}\hat{\xi}, \tag{12}$$

where $\mathrm{X} = (X_1, X_2, \cdots, X_n)'$, $\mathrm{W} = \mathrm{cov}(\mathrm{Y})$, and ξ is an n-dimensonal vector with the i^{th} element being $\xi_i = -\frac{1}{2}Q_{ii}$. Q_{ii} is the i^{th}diagonal element of the matrix $\mathrm{Q} = \mathrm{X}(\mathrm{X}'\mathrm{W}\mathrm{X})^{-1}\mathrm{X}$.

The bias-correction procedure is applicable to both the Poisson and NB regressions. The only difference is that the $\widehat{\mathrm{W}}$ of an NB model involves the estimated dispersion parameter.

Results

We applied the bias correction to a case study of infrastructure safety evaluation. The dataset includes information from 5,238 short road segments collected from 2012 to 2015. The total number of crashes was 32,298 for a total of 10,894,920 passing vehicles, resulting in the average crash rate being 2.96×10^{-3} crashes/passing vehicle. There are 59.9% zero responses in the dataset. Table B-1, in Appendix B, lists the seven covariates used in the analysis, along with the number of observations and percentage in each stratum (the second column). The crash frequencies of all strata for one covariate add up to the total of 32,298 crashes.

The NB regression was implemented because of the existence of overdispersion. The estimated dispersion parameter is 2.17. Table B1 also provides the difference between bias-corrected coefficient estimates $\tilde{\beta}$ and the regular MLE $\hat{\beta}$, as well as the percentage change $\frac{\tilde{\beta}-\hat{\beta}}{\hat{\beta}} \times 100\%$. The results indicate that the bias correction is generally larger for a stratum having a smaller number of events.

To test if the number of crashes affects the magnitude of bias, we also conducted a bias correction for two hypothetical pavement datasets: a "half-year" dataset and a "quarter-year" dataset. The crash count for each road segment is reduced to only 1/6 and 1/12, respectively, of the original pavement dataset. The exposure (vehicle-miles traveled) is also reduced by the corresponding fraction in the two hypothetical pavement datasets, while the covariates are the same as the original dataset. After reducing the number of crashes, the percentage of zero responses is 76.7% and 82.5% in the "half-year" dataset and the "quarter-year" dataset, respectively.

By comparing the results from the original dataset and two hypothetical datasets, the results show that the magnitude of the correction increases with the decrease in crash frequency. This testifies that the number of crashes is the factor that influences the magnitude of bias rather than the number of observations. The balance of event counts in one stratum compared to the reference stratum also impacts the magnitude of bias correction.

Decision-Adjusted Modeling Framework

Traditional statistical model selection methods are typically based on an overall generic statistical metric, such as likelihood ratio test or the AUC value of the ROC. The resulting model is optimized according to the generic metric, which might not be optimal for the specific goal of a study. For example, in predicting high-risk drivers based on logistic regression, a model selected by AUC might perform poorly when the goal is to identify a very small percentage of the riskiest drivers. While the AUC criterion selects a model with respect to the entire spectrum of possible decision points, the prespecified small percentage of riskiest drivers concerns only that particular decision point. In this study, we propose a decision-adjusted modeling framework that directly links the study goal with a decision-based objective function in the model selection/optimization process.

This framework will ensure that the output model is optimized with respect to the specific study objective. The framework is illustrated in Figure 4.

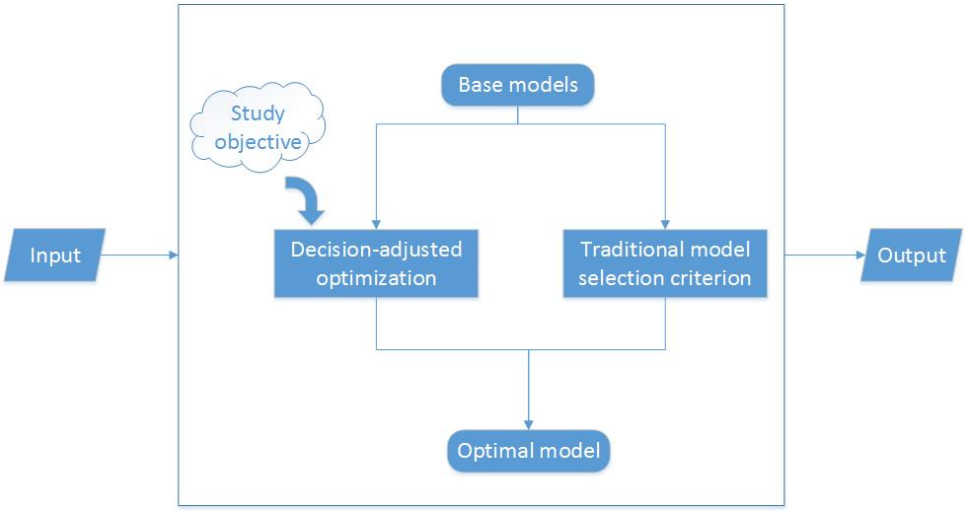

Figure 4. Decision-adjusted modeling framework.

Based on the objective function, the model selection/optimization process includes model form determination, variable selection, and parameter tuning. For binary response data, the model form could be a logistic regression, prediction tree, neural network, or so forth. Variable selection determines which covariates set should be incorporated into the optimal model. As to the parameter tuning, it refers not only to the hyperparameter tuning for the selected model form but also to the critical value adjustment in building certain predictor variables. We demonstrated this framework using the SHRP 2 NDS data to identify optimal prediction models for high-risk drivers by kinematic signatures.

Results

We applied the decision-adjusted framework to predict a small percentage of high-risk drivers using the SHRP 2 NDS data. The models identify the optimal threshold values for elevated longitudinal acceleration (ACC), deceleration (DEC), and lateral acceleration (LAT). We compared the decision-adjusted model $\mathcal{M}_2$ with two benchmark models $\mathcal{M}_0$ and $\mathcal{M}_1$. These models are specified as:

- $\mathcal{M}_0$: Traditional driver assessment model, without the kinematic predictors;
- $\mathcal{M}_1$: Kinematic-event-based driver risk assessment model, optimized by AUC;
- $\mathcal{M}_2$: Kinematic-event-based driver risk assessment model, optimized by the decision rules.

In three models, a regularized logistic regression (elastic-net) is used to assess the driver risk. In addition to kinematic event rates, age, gender, and crash/violation history, eight other traditional risk predictors are included in the model. The models' performance was evaluated by "prediction precision," the percentage of correct identifications among identified high-risk drivers.

Generally, our proposed method $\mathcal{M}_2$ performed the best among the three models, as shown in Figure 5. The decision-adjusted model improves the prediction precision by 7.8%–41.7% compared to baseline model $\mathcal{M}_0$. It is also superior to $\mathcal{M}_1$ by 6.5%–41.7%. The improvement is more prominent when identifying a small percentage of the riskiest drivers (e.g., <5%). $\mathcal{M}_1$ is also

better than $\mathcal{M}_0$ in the most decision spectrum. The benefit is credited to the inclusion of high g-force event rates. The results confirm that using kinematic information can improve individual driver risk prediction, and the improvement is more significant when a decision-adjusted modeling approach is applied. Furthermore, the prominent improvement of $\mathcal{M}_2$ when targeting a small percentage of high-risk drivers indicates that our decision-adjusted modeling is more desirable for the imbalanced data problem.

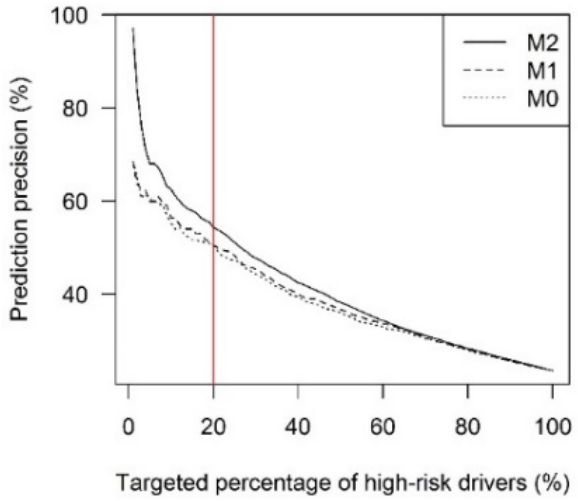

Figure 5. The comparison of three models' prediction precision, the percentage of correct identification among the drivers labeled by the model as high risk.

Cluster Analysis

The most popular method in partitional algorithms is the K-means, which is simple, efficient, and easy to implement (2). A similar algorithm, namely Partitioning Around Medoids (PAM) (8), can also be used. Unlike K-means, which creates centroids as centers of clusters, PAM identifies actual observations for clusters called medoids that represent different clusters. This makes PAM more robust to outliers compared to K-means. The Clustering Large Applications (CLARA) (8) method can be used to deal with larger datasets since it does not require the calculation of the entire distance matrix all at once, whereas PAM stores the distance matrix in central memory (9). Therefore, CLARA can be considered an alternative if the number of observations becomes so large that the memory is insufficient to store the distance matrix. Compared to CLARA and PAM, a more efficient algorithm, Clustering Large Applications based on RANdomized Search (CLARANS), was proposed in Chen and Zhang (10). Similar to CLARA, CLARANS utilizes a randomized search to cluster large data sets. However, unlike CLARA, CLARANS does not restrict the search to a localized area. Instead, it dynamically draws samples in each step to improve the quality of clustering. Through several experiments, it was shown that CLARANS outperformed both CLARA and PAM for small, medium, and large data sets (10).

In hierarchical algorithms, Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) was proposed in Gan, Ma, and Wu (11) and is suitable for performing cluster analysis on very large numerical data sets as it considers the available memory and time constraints. BIRCH's clustering decisions are not based on the entire data set and also consider the natural closeness of data points. BIRCH creates dendrograms, known as clustering feature trees (CF tree), that store

the number of data points, linear sum of the data points, and square sum of the data points in clusters. These CF trees are built using two parameters: a branching factor, B, and a threshold, T. BIRCH allows users to select a desired number of clusters or a desired threshold such as cluster radius or diameter. In some problems, distance-based metrics for partitioning Boolean and categorical data may not be suitable. A robust hierarchical clustering algorithm, ROCK, was proposed by Lord and Geedipally (12) that exploits a link-based approach to measure similarities between data points. It has also shown that ROCK has good scalability properties when dealing with large data sets.

Among many approaches in different clustering algorithm categories, a few methods were selected to develop cluster analysis models: PAM from partitional algorithms, DIANA from hierarchical algorithms, and DBSCAN from density-based algorithms were examined. A sample dataset prepared by the Virginia Tech Transportation Institute (VTTI) was used to test these approaches. The sample data include 3,592 observations (rows) and 39 variables (columns). Seven variables (gender, age group, marital status, annual mileage, years driving, income, and site) were used to develop clustering models. To be consistent with the VTTI analysis, these are the same independent variables investigated by VTTI. Missing data were excluded, resulting in a total of 3,237 observations.

Results

For several clustering algorithms, such as partitional and hierarchical approaches, it is important to determine the best number of clusters. The silhouette, elbow, and/or gap statistics are methods that are typically employed. The silhouette metric (41) shows how similar each point is to other points in a cluster compared to other points in the neighboring cluster. Applying the silhouette method, two clusters were found to be optimum for both PAM and DIANA, as shown in Figure 6 through Figure 8. A higher silhouette value represents better clustering results. Looking at the silhouette plot for the DIANA method, three clusters had almost the same silhouette value as two clusters. Hence, the silhouette plots for three clusters are also presented. The average silhouette values are quite small (highest is 0.22), which suggests that there may not be a clear separation of the data.

In order to implement the DBSCAN approach, two tuning parameters for Eps-neighborhood (Eps) and the minimum number of points (MinPts) need to be selected. The rule of thumb is to use the number of predictors plus one (in our case: 7 + 1 = 8) as MinPts. To select Eps, a common practice is to plot the k-nearest neighbor distances in the data space with k being equal to MinPts. This plot shows the k-distances in an ascending order. The k-distance value (i.e., Eps) that corresponds to a "knee" in the plot is considered as the optimal value of Eps (34). The "knee" should be detected as a sharp change that occurs along the k-distance plot. Looking at this plot (Figure 9), three knees can be detected at 8-NN values of 0.01, 0.16, and 0.29. These values resulted in an average silhouette width of -0.27, 0.022, and 0.088, respectively. In general, these show poor clustering performance as the silhouette values are very small. Even the highest value partitions the data into one large cluster with 3,214 data points and 23 noise points (outliers). Treating the noise points as a cluster, the Eps parameter was tweaked to obtain a more balanced data proportion without a significant decrease in the silhouette value. The Eps value of 0.28 was determined after trial and error, which led to one large (2,880 data points) and one small (357 outliers) cluster, with an

average silhouette width of 0.082. This suggests that the data space may not have varying density, and thus the algorithm only discovers a large cluster.

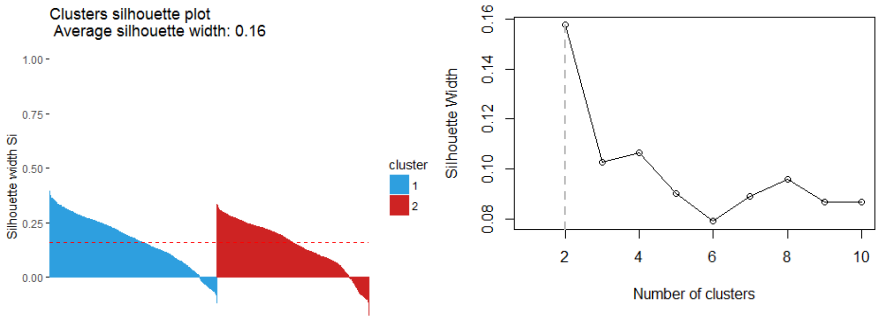

Figure 6. Silhouette visualization (left) and best number of clusters (right) for PAM.

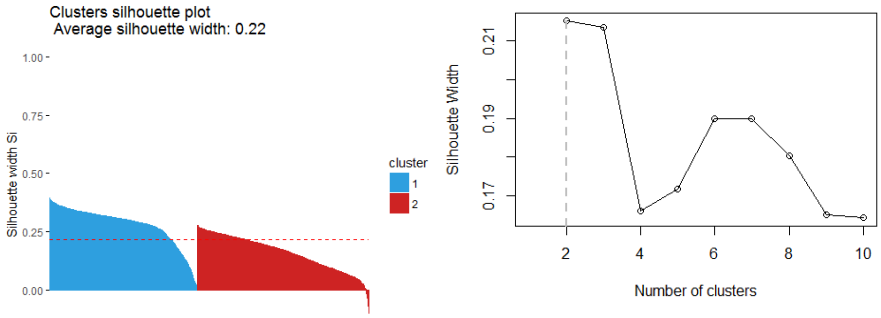

Figure 7. Silhouette visualization (left) and best number of clusters (right) for DIANA, 2 clusters.

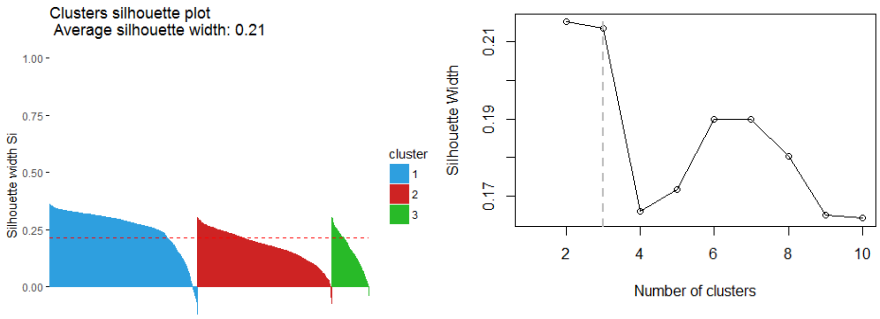

Figure 8. Silhouette visualization (left) and best number of clusters (right) for DIANA, 3 clusters.

DIANA showed the highest clustering performance in terms of the silhouette metric. Using the same metric, the DBSCAN's clustering performance was worse than the PAM and DIANA algorithms. However, it may not be appropriate to favor one approach to the other solely based on the silhouette width value, especially when the silhouette values of different methods are close to each other or when they are not considerably high. In general, although a higher value of the silhouette width shows better clustering performance, the clustering results, serving as new predictors (i.e., categorical variables), can determine the actual benefit of clustering. Also, DBSCAN is more efficient in terms of its run time (42) compared to the other two approaches, and therefore is more suitable when dealing with big data.

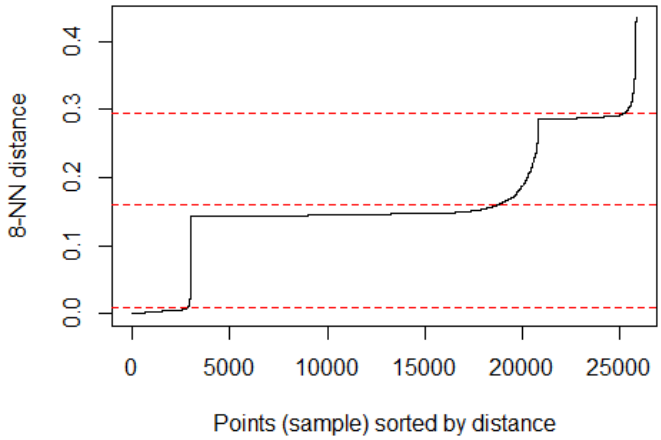

Figure 9. Determining the Eps parameter for DBSCAN method.

Discussion

This section reviews the proposed guidelines that were produced from this research. The guidelines should help researchers and practitioners handle highway-safety data with different characteristics. When datasets have a large percentage of zero responses (50% or above), the recommendations are as follows:

- When the percentage of zeros is higher than 70%, aggregate the data only if the change in CV of all variables when data are aggregated compared to the disaggregated data is less than 7%.
- When the percentage of zeros is less than 70%, aggregate the data only if the change in CV of all variables when data are aggregated compared to the disaggregated data is less than 4%.

When data aggregation is not possible, the guidelines for selecting the appropriate model are as follows:

- Select NB-L over NB when the skewness is greater than 1.92, independent of the number of zero responses.
- The selection of PLN over NB is governed by the percentage of zeros and the kurtosis. The boundaries are presented in Figure 2.

For datasets with a small number of crashes or that are imbalanced (small number of crashes in one category of a covariate), the bias-correction procedure is recommended for reducing errors with the estimation of the coefficients. Typically, bias adjustment should be performed when the number of crashes is less than 50 in any stratum. For rare-event prediction, a decision-adjusted framework is recommended, which will provide better predictive power.

To discover hidden patterns in the data, especially when the dataset becomes large, cluster analysis can be applied to create new predictors to potentially produce insight or reduce data dimension (i.e., number of attributes). In particular, CLARANS from the partitional algorithms, BIRCH and ROCK from hierarchical algorithms, and DBSCAN from density-based algorithms have been identified in past efforts as approaches suitable for conducting cluster analysis when the dataset is fairly large.

Conclusions and Recommendations

This report has documented issues and provided guidelines associated with the analysis of crash data. This kind of data has unique characteristics not found with datasets used in other types of research. Three characteristics were examined: (1) datasets with a large percentage of zeros; (2) datasets with few crashes (may not be necessarily characterized by having a large percentage of zeros); and (3) big data. These unique characteristics can negatively influence analyses of crash or other safety-related data. This report documented how these characteristics can be handled by either manipulating the dataset or providing statistical tools that are specifically tailored for these characteristics. Although the project provided useful guidelines, further research is recommended:

- Examine if statistics other than the change in the coefficient of variation of the independent variables can be used to determine when aggregated data should be used over disaggregated data.
- More statistical models should be compared using the heuristics method.
- For datasets with a small number of crashes or that are imbalanced (i.e., small number of crashes in one category of a covariate), the bias-correction procedure should be tested to determine the boundary conditions when the procedure is no longer needed.
- More cluster analysis methods should be examined. In addition, the predictors resulting from cluster analysis methods should be incorporated into crash-risk models to determine if they can enhance model performance and/or reduce the number of data dimensions.

Additional Products

The Education and Workforce Development (EWD) and Technology Transfer (T2) products created as part of this project can be downloaded from the Safe-D website here. The final project dataset is located on the Safe-D Dataverse.

Education and Workforce Development Products

Undergraduate and graduate courses

- TTI/Texas A&M – CVEN 626 – Highway Safety (Fall 2019): The material for the Fall 2019 graduate course CVEN 626 will be included in the slides and class notes. At the time this report was written, the class notes have not been yet updated. They will be made available on Dr. Lord's website.
- Virginia Tech – STAT4504 – Applied Multivariate Analysis (Fall 2019) and STAT5504G and STAT5594 – Statistical Epidemiology and Observation Study (Spring 2020): The models developed will be used in the multivariate statistics class STAT4504 (Fall 2019); the results will be used in the graduate course STAT5594, which will be offered in Spring 2020.
- Seminars – There were no specific training seminars that have been prepared from this project, but presentations were made for TRB, INFORMS and the University of Michigan. The first two (TRB and INFORMS) are listed on the Safe-D website. All three were presented by Dr. Ali Shirazi. A fourth one was presented by Maggie Mao and it is described below.

- Two UTC presentations/webinars are currently under preparation based on the results of this research. One will focus on VTTI's work, while the other will describe the characteristics of the heuristics method. These presentations are anticipated to be performed in the fall of 2019 and will be available on the on the Safe-D website after completion.

Educational Audience:

- University
- Professional

Student Funding:

- TTI – one graduate student – Ph.D., Mohammadali (Ali) Shirazi: "Advanced Statistical Methods for Analyzing Crash Datasets with Many Zero Observations and a Long Tail: Semiparametric Negative Binomial Dirichlet Process Mixture and Model Selection Heuristics" Status: Completed December 2018
- Virginia Tech – one graduate student – Ph.D., Huiying (Maggie) Mao: "Decision-Adjusted Approach for Driving Risk Evaluation" Status: Anticipated August 2019

Student Enrichment:

- For Ali Shirazi, the project has been very beneficial. In additional to adding knowledge to the science of highway safety research, this allowed Ali to find an academic position at the University of Maine. His work also helped find a postdoc position at the University of Michigan, which started in the fall 2018. The papers published from this work helped Ali secure these positions.
- For Maggie Mao, the project provided the main motivation examples for her dissertation research. Her work on this project help her to secure a postdoc position at the Statistical and Applied Mathematical Sciences Institute (SAMSI), one of the most prestigious statistical research institutions.
- Ali gave a presentation of the work he performed on this project to professors and students at the University of Michigan (An Innovative Method towards Automation of Model Selection using Big Simulated Data and Machine Learning, GG Brown Laboratory – 2029, Thursday, December 6 2018). About 25 people attended the presentation.
- Maggie gave a presentation of the work at the Joint Statistical Meeting (JSM) with the Transportation Statistics Interest Group (Joint Statistical Meeting, Vancouver, Canada, August 1st, 2018). JSM is one of the biggest conferences for the Statistics community.

Technology Transfer Products

The T2 portion of this project was met through the development of the guidelines described in this report and the Project Brief, providing a brief description of the project and summary results, available on the project page of the Safe-D website. In addition, the publications and presentations which have been produced from the research conducted in this project are listed on the project page of the Safe-D website.

Data Products

Descriptions of the Data Products for this project can be found in Appendix C.

References

1. Lord, D., and S. R. Geedipally. Safety Prediction with Datasets Characterised with Excess Zero Responses and Long Tails. In Safe Mobility: Challenges, Methodology and Solutions. Emerald Publishing Limited, 2018, pp. 297-323.

2. Usman, T., L. Fu, and L. F. Miranda-Moreno. (2011). Accident Prediction Models for Winter Road Safety: Does Temporal Aggregation of Data Matter? Transportation Research Record, Vol. 2237, No. 1, 2011, pp. 144-151.

3. Pratt, M. P., S. R. Geedipally, B. Wilson, S. Das, M. Brewer, and D. Lord. Pavement Safety-Based Guidelines for Horizontal Curve Safety. TxDOT 0-6932. Texas A&M Transportation Institute, College Station, TX, 2018.

4. Cafiso, S., C. D'Agostino, and B. Persaud. Investigating the Influence of Segmentation in Estimating Safety Performance Functions for Roadway Sections. Journal of Traffic and Transportation Engineering, Vol. 5, No. 2, 2018, pp. 129-136.

5. Lord, D., and F. Mannering. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. Transportation Research Part A: Policy and Practice, Vol. 44, No. 5, 2010, pp. 291-305.

6. Mannering, F. L., and C. R. Bhat. Analytic Methods in Accident Research: Methodological Frontier and Future Directions. Analytic Methods in Accident Research, Vol. 1, 2014, pp. 1-22.

7. Miaou, S. P., and D. Lord. Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes. In Transportation Research Record, Vol. 1840, 2003, pp. 31-40.

8. McCullagh, P., and J. A. Nelder. Generalized Linear Models (Vol. 37). CRC Press, 1989.

9. Hashem, I. A. T, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan. The Rise of 'Big Data' on Cloud Computing: Review and Open Research Issues. Information Systems, Vol. 47, 2015, pp. 98-115.

10. Chen, C. L. P., and Chun-Yang Zhang. Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. Information Sciences, Vol. 275, 2014, pp. 314-347.

11. Gan, G., Ma, C. and Wu, J., 2007. Data clustering: theory, algorithms, and applications (Vol. 20). Siam

12. Lord, D., and S. R. Geedipally. The Negative Binomial–Lindley Distribution as a Tool for Analyzing Crash Data Characterized by a Large Amount of Zeros. Accident Analysis & Prevention, Vol. 43, No. 5, 2011, pp. 1738-1742.

13. Geedipally, S. R., D. Lord, and S. S. Dhavala. (2012). The Negative Binomial-Lindley Generalized Linear Model: Characteristics and Application using Crash Data. Accident Analysis & Prevention, Vol. 45, 2012, pp. 258-265.

14. Vangala, P., D. Lord, and S. R. Geedipally. Exploring the Application of the Negative Binomial–Generalized Exponential Model for Analyzing Traffic Crash Data with Excess Zeros. Analytic Methods in Accident Research, Vol. 7, 2015, pp. 29-36.

15. Shirazi, M., D. Lord, S. S. Dhavala, and S. R. Geedipally. A Semiparametric Negative Binomial Generalized Linear Model for Modeling Over-dispersed Count Data with a Heavy Tail: Characteristics and Applications to Crash Data. Accident Analysis & Prevention, Vol. 91, 2016, pp. 10-18.

16. Lindley, D. V. 1958. Fiducial Distributions and Bayes' Theorem. Journal of the Royal Statistical Society Series B (Methodological), 1958, pp. 102-107.

17. Zamani, H., and N. Ismail. Negative Binomial-Lindley Distribution and its Application. Journal of Mathematics and Statistics, Vol. 6, No. 1, 2010, pp. 4-9.

18. Shirazi, M., S. S. Dhavala, D. Lord, and S. R. Geedipally. A Methodology to Design Heuristics for Model Selection based on the Characteristics of Data: Application to Investigate when the Negative Binomial Lindley (NB-L) Is Preferred over the Negative Binomial (NB). Accident Analysis & Prevention, Vol. 107, 2017, pp. 186-194.

19. Lord, D., S. P. Washington, and J. N. Ivan. Poisson, Poisson-gamma and Zero-inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. Accident Analysis & Prevention, Vol. 37, No. 1, 2005, pp. 35-46.

20. Da Costa, J. O., E. F. Freitas, and P. A. A. Pereira. Collision Prediction Models with Longitudinal Data: A Note on Modeling Collision Frequency in Road Segments in Portugal. In Transportation Research Board 95th Annual Meeting (No. 16-2131), 2016.

21. Firth, D. Bias Reduction of Maximum Likelihood Estimates. Biometrika, Vol. 80, No. 1, 1993, pp. 27-38.

22. Kosmidis, I., and D. Firth. Bias Reduction in Exponential Family Nonlinear Models. Biometrika, Vol. 96, No. 4, 2009, pp. 793-804. doi:10.1093/biomet/asp055

23. Kosmidis, I., and D. Firth. A Generic Algorithm for Reducing Bias in Parametric Estimation. Electronic Journal of Statistics,Vol. 4, 2010, pp. 1097-1112. doi:10.1214/10-Ejs579

24. Cordeiro, G. M., and P. Mccullagh. Bias Correction in Generalized Linear-Models. Journal of the Royal Statistical Society Series B-Methodological, Vol. 53, No. 3, 1991, pp. 629-643.

25. Lambrecht, B., W. Perraudin, and S. Satchell (1997). Approximating the Finite Sample Bias for Maximum Likelihood Estimators Using the Score–Solution. Econometric Theory, Vol. 13, No. 2, 1997, pp. 310-312.

26. Giles, D. E., and H. Feng. Reducing the Bias of the Maximum Likelihood Estimator for the Poisson Regression Model. Economics Bulletin, Vol. 31, No. 4, 2011, pp. 2933-2943.

27. Saha, K., and S. Paul. Bias-corrected Maximum Likelihood Estimator of the Negative Binomial Dispersion Parameter. Biometrics, Vol. 61, No. 1, 2005, pp. 179-185. doi:DOI 10.1111/j.0006-341X.2005.030833.x

28. Cox, D. R., and E. J. Snell. A General Definition of Residuals. Journal of the Royal Statistical Society Series B-Statistical Methodology, Vol. 30, No. 2, 1968, pp. 248-265.

29. Guo, F., and Y. J. Fang. Individual Driver Risk Assessment using Naturalistic Driving Data. Accident Analysis and Prevention, Vol. 61, 2013, pp. 3-9. doi:10.1016/j.aap.2012.06.014

30. Simons-Morton, B. G., Z. W. Zhang, J. C. Jackson, and P. S. Albert. Do Elevated Gravitational-Force Events While Driving Predict Crashes and Near Crashes? American Journal of Epidemiology, Vol. 175, No. 10, 2012, pp. 1075-1079. doi:10.1093/aje/kwr440

31. Li, K-C., H. Jiang, and A. Y. Zomaya. Big Data Management and Processing. CRC Press, 2017.

32. Zerhari, B., A. A. Lahcen, and S. Mouline. Big Data Clustering: Algorithms and Challenges. In Proc. of Int. Conf. on Big Data, Cloud and Applications (BDCA'15), 2015.

33. Jain, A. K. Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters, Vol. 31, No. 8, 2010, pp. 651-666.

34. Hahsler, M., M. Piekenbrock, and D. Doran. Dbscan: Fast Density-Based Clustering with R. Journal of Statistical Software, Vol. 25, 2017, pp. 409-416.

35. Geedipally, S., M. Martin, R. Wunderlich, D. Lord. Highway Safety Improvement Program Screening Tool. Technical Memorandum. Traffic Operations Division, Texas Department of Transportation, 2017.

36. Pudlo, P., J. M. Marin, A. Estoup, J. M. Cornuet, M. Gautier, and C. P. Robert. Reliable ABC Model Choice via Random Forests. Bioinformatics, 32, No. 6, 2015, pp. 859-866.

37. Hastie, T., R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer, New York, 2001.

38. James, G., D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning (Vol. 6). Springer, New York, 2013.

39. Gonzales-Barron, U., and F. Butler. A Comparison between the Discrete Poisson-gamma and Poisson-lognormal Distributions to Characterise Microbial Counts in Foods. Food Control, Vol. 22, No. 8, 2011, pp. 1279-1286.

40. Khazraee, S. H., V. Johnson, and D. Lord. (2018). Bayesian Poisson Hierarchical Models for Crash Data Analysis: Investigating the Impact of Model Choice on Site-specific Predictions. Accident Analysis & Prevention, Vol. 117, 2018, pp. 181-195.

41. Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, Vol. 20, 1987, pp. 53-65.

42. Ester, M., H. P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Kdd, Vol. 96, No. 34, 1996, pp. 226-231.

Appendices

Appendix A

Appendix A provides the simulation protocol used in this report to study the aggregation of crash datasets. Let us, first, define the parameters and variables as follows:

x_{ij}^{m} : The value of the j-th covariate for i-th site at time period 'm'.

φ^{m} : Inverse dispersion parameter at each time period 'm' calculated from real data.

y_i^{m} : simulated observation for the i-th site at each period 'm'.

μ_i^{m}: mean response of the NB distribution at the i-th site at period 'm'.

β_j: The true parameter for the j-th covariate (derived from a known model)

β_j^{n*}: The estimated parameter for the j-th covariate at iteration 'n' of simulation.

The steps of the simulation protocol can be summarized as follows:

1. Find the mean of crashes at each site 'i' as follows[1]:

$$\mu_i^{m} = e^{\sum_{j=1}^{d} \beta_j x_{ij}^{m}}$$

2. Repeat the following steps for 'N' times:
   2.1 Simulate the observation at each site i=1 to n at the m-th period from the NB distribution as follows:

$$y_i^{m} \sim \mathrm{NB}(\mu_i^{m}, \varphi^{\mathrm{m}})$$

   2.2 Creating the experiment datasets.
      2.2.1 Create the disaggregated dataset (D_1)

[1] For the purpose of simulation, the x_{ij}^{m} with "NA" values are replaced with $\min_{m}(x_{ij}^{m}) + \frac{\max_{m}(x_{ij}^{m}) - \min_{m}(x_{ij}^{m})}{m}$ in Step 1. However, the records 'NA' values eventually are removed in Step 2.2.1.3 and Step 2.2.2.1.

2.2.1.1 Create the datasets D^m at each period 'm', with (y_i^m, x_{ij}^m) elements (where the index 'i' denote a row and 'j' a column of the dataset).

2.2.1.2 Merge the D^m datasets into a single dataset D_1.

2.2.1.3 Remove the records of D_1 that include an 'NA' value.

2.2.1.4 Shuffle the records in D_1.

2.2.2 Create the aggregated dataset (D_2):

2.2.2.1 Find $\bar{x}_{ij} = \underset{m}{\text{mean}}\, x_{ij}^m$ (exclude x_{ij}^m with the "NA" values when $\bar{x}_{ij}$ is calculated).

2.2.2.2 Create the D_2 dataset with $(\sum_m y_i^m, \bar{x}_{ij})$ elements (where the index 'i' denote a row and 'j' a column of the dataset).

2.2.2.3 Shuffle the records in D_2.

2.3 Refitting the simulated datasets

2.3.1 Fit an NB GLM to D_1 and record the estimated coefficients in $\beta_j^{n^*}(D_1)$.

2.3.2 Fit an NB GLM to D_1 and record the estimated coefficients in $\beta_j^{n^*}(D_2)$.

3. Comparison.

3.1 For each j-th covariate, find the standard deviation of the estimated coefficients over 'n' iterations and denote them by $\beta_j^{std}(D_1)$ and $\beta_j^{std}(D_2)$.

3.2 Compare $\beta_j^{std}(D_1)$ and $\beta_j^{std}(D_2)$, the one with a smaller value indicates a more reliable implementation.

Appendix B

Appendix B presents summary statistics for the data used for the Finite Sample Bias Adjustment section.

Table B-1. Descriptive Statistics and Corresponding Bias Magnitude for the Explanatory Variables of Pavement Data

Variables	Freq (pct)	Original Dataset: No. of Crashes	Original Dataset: $\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$ ($\times \mathbf{10^{-3}}$) (pct change)	"Half-Year" Dataset: No. of Crashes	"Half-Year" Dataset: $\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$ ($\times \mathbf{10^{-3}}$) (pct change)	"Quarter-Year" Dataset: No. of Crashes	"Quarter-Year" Dataset: $\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$ ($\times \mathbf{10^{-3}}$) (pct change)
RTE type							
I	2,236 (42.7%)	29,543		4,851		2,377	
SR	1,160 (22.1%)	1,582	0.1 (0.0%)	208	6.2 (-0.7%)	82	29.6 (-1.8%)
US RTE	1,842 (35.2%)	1,173	0.1 (0.0%)	133	6.8 (-0.6%)	43	32.6 (-1.6%)
Entrance/Exit							
0	5,163 (98.6%)	31,740		5,098		2,460	
1	61 (1.2%)	245	0.0 (0.0%)	41	1.2 (0.4%)	15	14.7 (-4.2%)
2	14 (0.3%)	313	0.0 (0.0%)	53	1.0 (0.1%)	27	3.6 (0.3%)
Intersection							
0	4,906 (93.7%)	31,236		5,032		2,425	
1	233 (4.4%)	685	0.0 (0.0%)	103	0.3 (0.0%)	50	-1.4 (-0.1%)
2	61 (1.2%)	254	0.1 (0.0%)	37	4.9 (0.4%)	18	11.5 (0.6%)
3	14 (0.3%)	55	0.3 (0.0%)	8	9.8 (0.4%)	4	34.1 (1.3%)
4	24 (0.5%)	68	0.9 (0.1%)	12	19.3 (0.8%)	5	52.9 (1.7%)
Divided Highway							
0	1,883 (35.9%)	1,913		248		100	
1	3,355 (64.1%)	30,385	0.1 (0.0%)	4,944	3.8 (-0.3%)	2,402	19.8 (-1.3%)
Rural/Urban							
Rural	3,467 (66.2%)	2,975		365		130	
Urban	1,771 (33.8%)	29,323	0.0 (0.0%)	4,827	-1.0 (-0.1%)	2,372	-3.4 (-0.3%)
No. of Lanes							
1	1,752 (33.4%)	1,439		174		64	
2	2,481 (47.4%)	6,439	0.0 (0.0%)	962	-1.4 (0.7%)	441	-5.3 (2.8%)
3	624 (11.9%)	12,930	0.0 (-0.1%)	2,136	-1.7 (-0.9%)	1,038	-6.0 (-5.5%)
4	380 (7.3%)	11,387	0.0 (-0.2%)	1,903	-1.8 (4.3%)	950	-6.1 (3.0%)
5	1 (0%)	103	0.0 (0.0%)	17	0.0 (0.0%)	9	2.1 (0.2%)
Pavement Type							
ACP	3,846 (73.4%)	10,626		1,615		730	
ACP/PCCP	44 (0.8%)	620	0.0 (0.0%)	102	0.8 (0.2%)	52	2.9 (0.5%)
BST	119 (2.3%)	20	16.1 (-2.5%)	0	NA	0	NA
PCCP	1,229 (23.5%)	21,032	0.0 (0.0%)	3,475	-0.1 (0.0%)	1,720	-0.2 (0.0%)

Appendix C

Appendix C describes the characteristics of the data used in this research.

Data used for decision-adjusted modeling

Summary Description of Analysis:
Predicting crash risk and identifying high-risk drivers are critical for developing appropriate safety countermeasures, driver education programs, and user-based insurance. However, predicting driver risk is a challenging task because crashes are rare events and many factors contribute to individual crash risk. As in-vehicle data collection becomes more prevalent and cost-effective, it has become more feasible to improve risk prediction by utilizing kinematics information. Currently, there are several challenges to implementing kinematics-based driver risk prediction models. We focus on two primary issues: (1) the decision rule and (2) the optimal threshold values for kinematics predictors.

Data Scope:
The naturalistic driving data collected from the second Strategic Highway Research Program (SHRP 2) is used to identify optimal prediction models for high-risk drivers by kinematic signatures. The dataset includes 3440 rows of drivers and 200 columns of features. Each row represents one driver, and the columns represent the characteristics of each driver.

Data Specification:
The specific data description is shown in this link: https://doi.org/10.15787/VTT1/QQEZOP.

Data used for the adjust finite-sample bias for traffic safety modeling

Summary Description of Analysis:
The Poisson and NB models are generally estimated using the maximum likelihood method. When the sample size is small and/or when the number of events is limited (e.g., small number of crashes), the maximum likelihood estimators (MLEs) are biased and the bias could be substantial. This finite sample bias could lead to incorrect estimation of the impacts of risk factors and jeopardize traffic safety improvement efforts. This project addresses this gap by studying the finite sample bias for the parameter estimation of Poisson and NB regression models in the context of traffic safety modeling.

Data Scope:
To illustrate the benefit of bias correction and examine the magnitude of bias, we applied the bias-correction procedure to an infrastructure safety evaluation dataset.

This dataset includes information from 5,238 short road segments, which are collected from 2012 to 2014 in the State of Washington. The length for each segment is 0.1 mile. The covariates used in the analysis include route type, whether the road segment is an entrance/exit, whether it is an intersection, whether it is a ramp, whether it is a wye connection, whether it is a divided highway, rural/urban, number of lanes, pavement type, friction, gradient, and horizontal curvature.

Data Specification:
The specific data description are listed in this link: https://doi.org/10.15787/VTT1/QQEZOP.

In order to comply with participant informed consent and IRB requirements, a portion of this data set is governed by a Data Use License (DUL). Additionally, the other portion of the data will not be available for re-use. Please send inquiries to datasharing@vtti.vt.edu.

Link for Data:
https://doi.org/10.15787/VTT1/QQEZOP

Public Link for Project:
https://www.vtti.vt.edu/utc/safe-d/index.php/projects/big-data-methods-for-simplifying-traffic-safety-analyses/