Implementation of a Realistic Artificial Data Generator for Crash Data Generation

Lauren Hoover
Doctoral Student
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Email: spychalskylauren@knights.ucf.edu

Md. Istiak Jahan
Doctoral Student
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Email: mdistiakjahan@knights.ucf.edu

Tanmoy Bhowmik*
Assistant Professor
Dept. of Civil and Environmental Engineering
Portland State University
Email: tbhowmik@pdx.edu
ORCiD number: 0000-0002-0258-1692

Sudipta Dey Tirtha
Postdoctoral Scholar
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 407-543-7521
Email: sudiptadeytirtha2018@knights.ucf.edu

Karthik C. Konduri
Principal Research Scientist
Department of Civil and Environmental Engineering
University of Connecticut
Email: karthik.konduri@uconn.edu

John Ivan
Professor
Department of Civil and Environmental Engineering
University of Connecticut
Email: john.ivan@uconn.edu

Kai Wang
Statistician & Transportation Safety Engineer
Connecticut Transportation Safety Research Center
Connecticut Transportation Institute
University of Connecticut
Email: kai.wang@uconn.edu

Shanshan Zhao
Research Scientist-Project Manager
Connecticut Transportation Safety Research Center
Connecticut Transportation Institute
University of Connecticut
Email: shanshan.h.zhao@uconn.edu

Joshua Auld
Manager, Transportation Systems & Mobility
Argonne National Laboratory
Email: jauld@anl.gov

Naveen Eluru
Professor
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Email: naveen.eluru@ucf.edu
ORCiD number: 0000-0003-1221-4113

Submitted December 10, 2023

---

*Corresponding author

ABSTRACT

In this paper, a framework is outlined to generate realistic artificial data (RAD) as a tool for comparing different models developed for safety analysis. The primary focus of transportation safety analysis is on identifying and quantifying the influence of factors contributing to traffic crash occurrence and its consequences. The current framework of comparing model structures using only observed data has limitations. With observed data, it is not possible to know how well the models mimic the true relationship between the dependent and independent variables. Further, real datasets do not allow researchers to evaluate the model performance for different levels of complexity of the dataset. RAD offers an innovative framework to address these limitations. Hence, we propose a RAD generation framework embedded with heterogeneous causal structures that generates crash data by considering crash occurrence as a trip level event impacted by trip level factors, demographics, roadway and vehicle attributes. Within our RAD generator we employ three specific modules: (a) disaggregate trip information generation, (b) crash data generation and (c) crash data aggregation. For disaggregate trip information generation, we employ a daily activity-travel realization for an urban region generated from an established activity-based model for the Chicago region. We use this data of more than 2 million daily trips to generate a subset of trips with crash data. For trips with crashes crash location, crash type, driver/vehicle characteristics, and crash severity. The daily RAD generation process is repeated for generating crash records at yearly or multi-year resolution. The crash databases generated can be employed to compare frequency models, severity models, crash type and various other dimensions by facility type – possibly establishing a universal benchmarking system for alternative model frameworks in safety literature.

Keywords: realistic artificial data generation, crash data generation

1. INTRODUCTION

Transportation safety modeling has broadly evolved along two streams. The first stream, labeled as crash frequency models, examine the factors affecting the occurrence of crashes on transportation facilities. The second stream, referred to as crash severity models, examine factors affecting crash consequences (usually severity) at the disaggregate level (such as driver, vehicle or crash record). The primary focus of these two streams of safety analysis is on identifying and quantifying the influence of factors contributing to traffic crash occurrence and its consequences. In transportation (and other domains), observed data are generally employed to evaluate the performance of statistical or machine learning methods. The traditional analysis paradigm of model development employs the following steps. A statistical model structure is proposed for a selected empirical dataset. The proposed model and various comparable models are estimated using the empirical dataset. The model fit of the proposed model and the competitive models are compared using various performance measures. Finally, the preferred model for the empirical context is identified.

The application of observed data in such performance evaluation has several drawbacks. First, the observed data only enables researchers to compare the performance of alternative models based on selected statistical measures. But it is impossible to know how well the models mimic the true relationship between the dependent and independent variables which is of utmost interest to researchers (Scott & Wilkins, 1999). For example, crash risk has an explicit relationship with roadway geometric characteristics such as lane width, shoulder with, and median width. With real datasets, it is only possible to find the best model based on how well the models fit the dataset. But we cannot identify the model which most successfully captures the true relationship between crash risk and roadway geometry. Second, real datasets do not allow researchers to evaluate the model performance for different levels of complexity of the dataset. For example, some models may perform reasonably well on datasets without complex data generation processes but perform poorly on datasets with complex data generation processes. Often, it is not possible to compare the performance of alternative approaches on multiple datasets. Finally, some analysis methods demand comprehensive datasets that are resource intensive and scarce.

An effective approach to address these limitations is to consider the development of artificial data (or simulated data) with complete knowledge of the underlying crash generation process (as suggested by Dr. Ezra Hauer; Bonneson & Ivan, 2013). Such a simulated dataset, referred to as Realistic Artificial Data (RAD), can then be used to investigate different questions related to safety modeling analyses. In the RAD generation process, the true relationship is predefined but remains unknown to the analysts. Thus, it is possible for researchers to ideally examine the alternative methods in a more comprehensive manner. RAD data will allow objective evaluation of the methods used by comparing the inferences about the crashes and contributing factors to the assumptions that underlie the synthetic data generation process. In the RAD generation process, it is also possible to impose different degrees of complexity in the dataset which may enable researchers to more closely evaluate the performance of alternative models in handling complexity. Further, these artificially generated crashes can be aggregated at any spatial or temporal resolution to mimic data from the real world and carry out systematic safety analysis methods evaluation. With the RAD generated datasets, researchers can test their model framework on the RAD data and establish a benchmark. The approach is analogous to sample networks used by operation research and transportation researchers to compare runtimes of different algorithms. The RAD generated datasets, if employed as a benchmark, can serve as a guide for model selection.

In this research, we document the development of a RAD for transportation safety crash record generation. The proposed RAD generator recognizes that crashes are a result of travel decisions made by individuals. Hence, to mimic the true crash generation process, we examine crash occurrence as a trip level decision. The generator considers a set of daily trips from a travel demand model framework as input to RAD. Each trip contains information on trip start time, end time, origin, destination, travel mode, and travel route details. Employing this rich set of information, for each trip, we evaluate crash risk. For trips identified to be involved in a crash, detailed crash characteristics are generated. The RAD generator employs a suite of models to process trips with crashes including crash type, crash severity, crash location, driver and vehicle characteristics. The RAD generator is developed employing multiple datasets including Strategic Highway Research Program 2 (SHRP2) Naturalistic Driving Study (NDS) data and Crash Report Sampling System data from National Highway Traffic Safety Administration. The RAD generator produces crashes at a daily resolution with detailed spatio-temporal information. These crashes are generated multiple times to obtain yearly or multi-year datasets. Further, the datasets can be aggregated at any spatial resolution (such as intersection, segment, zone) or temporal resolution (such as morning, evening, seasonal) for frequency and severity analysis. with the embedded randomness, multiple realizations of RAD will generate distinct crash samples. The implementation results from the RAD generator are presented in the paper.

2. EARLIER LITERATURE

The concept of RAD has been applied in a number of disciplines including statistics, econometrics, computer science, ecology, medicine and psychology. In all these disciplines, the primary goal is to assess the ability of the methods to draw inferences about the underlying assumptions and assertions that generated the data. The research team conducted a comprehensive review of research efforts on RAD approaches across various domains. The criterion for selection of a study to be included in our review followed a simple core principle of RAD generation. The data generated in the research effort must be based on a framework that is built on research assumptions (as opposed to entirely real observed data-based simulation efforts). The criterion eliminates two major sets of transportation studies that generate simulated data. First, several travel demand modeling forecast systems such as activity-based models and synthetic population generators generate individual level synthetic data (Eluru et al., 2008; Konduri et al., 2016). However, the generation is entirely based on models estimated using observed data. Second, artificial data is generated in micro-simulation frameworks for traffic flow modeling. In these studies, the simulated data is generated based on well calibrated traffic flow models (Ranade et al., 2007; Asano et al., 2010; Yu & Abdel-Aty, 2014; Mamun et al., 2018). Hence, these studies are also not appropriate for our review.

In our review process, based on the realistic data generation criterion, we have identified several research studies that employed artificial data generation in their analysis. These studies span transportation (including transportation safety and travel behavior), medical science, data science, and information analytics. As opposed to providing a study-by-study summary of earlier research, we provide insights on the important elements of RAD framework that can be observed from earlier research efforts.

2.1. Review Findings

A concise summary of earlier research efforts on RAD generation is presented in Table 1. In this table, we provide information on study objectives, dataset adopted and study region, software/procedure followed for generating RAD, and field of the study (for example transportation safety). For the ease of presentation, the studies presented in Table 1 are categorized along two streams based on the discipline of the study: 1) studies related to transportation and 2) studies related to other disciplines including statistics, economics, ecology and computer science.

Several important observations can be made from Table 1. First, earlier research has explored RAD applications for wide ranging topics including statistical/econometric model performance and comparison, travel demand forecasting, route choice behavior, and data mining. Second, RAD applications have been developed using several software packages or platforms such as R, SAS, GAUSS, Python, and COMPAS. Third, employing RAD datasets, performance of several model structures was considered including ordered logit (OL), multinomial logit (MNL), generalized ordered logit (GOL), mixed multinomial logit (MMNL) and probit models (and their cross-sectional and panel variants), multiple discrete–continuous (MDC) frameworks with probit and extreme value formulations, and recurrent neural networks (RNN). Fourth, it is interesting to note that studies within transportation domain traditionally adopt RAD approaches for econometric models. However, non-transportation domain research typically is more focused on machine learning and data mining approaches. Finally, the number of alternatives in the RAD variable is related to the problem context. The number of alternatives could range from a small number (say 2 for a binary variable based RAD) to a very large number (theoretically infinity for crash counts).

Table 1: Summary of Existing Literature on RAD generation

Study	Study Objectives	Dataset Adopted (Study Region)	Software/Procedure for RAD generation	Field
Transportation Domain				
Bhat et al., 2010	Propose a Composite Marginal Likelihood (CML) approach to estimate ordered response discrete choice models with flexible copula based spatial correlation structures	Simulated and observed data (San Francisco Bay area)	Three independent variables are considered, and the values are drawn from univariate normal distribution. Fixed coefficients are assumed. Error terms are generated using correlation structure. 25 different datasets are generated with 500 observations	Travel behavior
Bhat & Sidharthan, 2010	Investigate the ability of Maximum Approximate Composite Marginal Likelihood (MACML) estimator to recover parameters from finite samples	Simulated dataset	Five independent variables are considered, and the values are drawn from univariate normal distribution. Random coefficients are assumed. Error terms are generated from univariate normal distribution with 0.5 variance. 20 datasets with 5000 observations are generated	Travel behavior
Pinjari & Bhat, 2010	To investigate non-worker out-of-home discretionary activity time-use and activity timing decisions on weekdays using multiple discrete-continuous nested extreme value (MDCNEV) model	Simulated and observed data (San Francisco Bay area)	Independent variable values are assumed to be uniformly distributed. Coefficients are assumed to be nested extreme values. Generate the data for 2500 hypothetical individuals with an assumption that each individual chose the value to maximize the total random utility	Travel behavior
Ferdous et al., 2010	Model the interactions in non-work activity episode decisions across household and non-household members at the level of activity generation using multivariate ordered-response system framework	Simulated and observed dataset (2007 American Time Use survey data)	Values for the independent variables are drawn from univariate normal distribution. A fixed coefficient is assumed and using that, the utility for each individual is computed using a linear combination. The error term is generated with predefined correlation structure. The process is repeated at least 50 times.	Travel behavior
Ye & Lord, 2011	Examining the effects of underreporting crash data using multinomial logit (MNL), ordered probit (OP), and mixed logit (ML) models	Simulated and observed data (Texas)	Weighted exogenous sample maximum likelihood estimator (WESMLE). Computer code was developed for daily travel pattern generation	Safety
Geedipally et al., 2012	Application of a negative binomial (NB) generalized linear model with Lindley mixed effects for analyzing traffic crash data	Simulated and observed data (road segment, Indiana, Michigan)	Coefficients are selected in a way that they seem logical and comparable with existing literature. Crash mean was computed and then crashes are simulated	Safety
Lord & Kuo, 2012	Examining the effects of site selection criteria	Simulated Data	The software R was used to generate sites with crash counts with a predefined overall mean for different dispersion parameters.	Safety
Cummings et al., 2013	Reviews three methods for estimating relative risks in matched-pair crash data	Simulated and observed data	Employing Stata Statistical Software the study generated crash data with an assumed probability of fatality as a function of speed and seatbelt use	Safety

Study	Study Objectives	Dataset Adopted (Study Region)	Software/Procedure for RAD generation	Field
Eluru, 2013	Investigating the performance of the ordered (OL, GOL) and unordered (MNL) injury severity response frameworks	Simulated dataset	Three independent variables are considered. Assume parameters that provides the same aggregate shares. 50 realizations of the data with 5000 observations each are generated for each proportion value. Total 6 aggregate sample shares are generated	Safety
Paleti & Bhat, 2013	Comparison between the maximum-simulated likelihood inference (MSL) and composite marginal likelihood (CML) approach	Simulated dataset	Independent variables are drawn from univariate normal distribution while coefficients are assumed and drawn from multivariate normal distribution. Consider both independent and correlated realizations. Data is generated at least 50 times	Travel behavior
Wu et al., 2015	Generating crash modification factors (CMFs) using NB regression model and compared with assumed true values	Simulated data	CMF values for lane width, curve density, and pavement friction were assumed and used to generate simulated crash counts	Safety
Highway safety and information system, 2017	Use of RAD to assess performance of cross-sectional analysis methods	Artificial realistic data (Rural two-lane highways, Washington)	Data generation was implemented by SAS programs based on an assumed model structure for AADT and roadway geometry factors	Safety
Berke et al., 2022	Generating synthetic mobility data using recurrent neural networks (RNN)	Synthetic data and LBS data from more than 22,700 mobile devices	Population distribution is the input and mobility traces for a synthetic population is generated	Transportation planning and epidemic modeling
Non-Transportation Domain				
Zimmermann, 2012	Generation of diverse data sets reflecting realistic data characteristics	Artificial Data	Data generator was implemented in JAVA	Data science
Devroye et al., 2012	Estimation of a density using real and artificial data	Observed and Artificial data	Data generator was implemented in R. The artificial data is generated from a regression analysis of observed data	Data science
Hazwani et al., 2016	Developing the automatic artificial data generator for generating artificial data set based on the real data	Artificial and real data	Random permutation algorithm was used to generate different sets of artificial data that represent realistic data	Information and Communication Technology
Dahmen & Cook, 2019	Introducing a synthetic data generation method	Simulated and real data	SynSys, a machine learning-based synthetic data generation method	Medical science
Chatterjee et al., 2022	Generating synthetic multiuser datasets for multiuser activity recognition	Simulated and real data	A strategy to generate a multiuser dataset from the existing single-user dataset	Information and Communication Technology
Charalambidis et al., 2022	Developing dataset generator for large-scale electric vehicles charging management	Simulated and anonymized real datasets	Flask—a Python micro web framework, pure HTML and JavaScript	Data science

From our review of earlier literature, the embedded RAD frameworks are consistently single level frameworks, i.e., the underlying decision process consists of only one layer of decisions. For example, in modeling crash occurrence, earlier research has related the crash occurrence to roadway geometry and traffic volume under pre-specified assumptions of what variables will influence crash occurrence (say AADT and lane width). While the approach is useful, it inherently disregards the nature of crash occurrence. The process of crash occurrence is a multi-layered decision process that is dependent on travel decisions (such as mode, travel route, departure time), transportation infrastructure (roadway characteristics, speed limits, facility types) and network interactions (congestion, presence of pedestrians). Hence, in our study, we develop and implement a multi-layered RAD that is more appropriate to represent the underlying crash generation process.

2.2. Current Study in Context

The current research builds on Hauer's earlier work on building RAD framework for crash data. The current paper develops a multi-layered RAD recognizing that crashes occur at an individual trip level. The approach introduces significant realism in the data generation process while also incorporating significant stochasticity in the data generation process. As is evident from the literature review, earlier efforts across different fields have focused on single layer RAD generation and our current study is the first effort to conceptualize and develop a RAD platform with multiple connected layers. The RAD framework identifies trips with crashes and for these selected trips builds crash type, crash severity, crash location, driver and vehicle characteristics.

To operationalize the RAD platform the paper employs data from three data sources including (a) Travel demand model outputs for the Chicago region developed by Argonne National Laboratory, (b) Strategic Highway Research Program 2 (SHRP2) Naturalistic Driving Study (NDS) data and

(c) Crash Report Sampling System data from National Highway Traffic Safety Administration. The RAD platform and the various datasets generated are analyzed to illustrate how they represent current crash data realistically.

The rest of the paper is organized as follows: We present the conceptual framework for crash data generation and present a discussion of data processing steps. Subsequently, the module specific results for RAD components are described. Next, we present an overview of overall RAD datasets generated and outline how the RAD datasets can be used for alternative model comparison. Finally, we provide some concluding thoughts and future directions of research.

3. RAD CONCEPTUAL FRAMEWORK

In this section, we describe the conceptual framework for a high resolution Disaggregate Realistic Artificial Data (RAD) generation. Specifically, we propose a framework of RAD generation embedded with heterogeneous causal structures that generates crash data by considering crash occurrence as a trip level event impacted by trip level factors, demographic characteristics, roadway facility and vehicle attributes. The proposed framework will be general enough to generate crashes for all roadway facility types and also be able to generate data for different combinations of inputs including modeling methods, model formulation, input specification, and unobserved heterogeneity. Employing daily trip level travel information, we will generate crash characteristics including crash occurrence, crash type, and crash severity. Trip level attributes to be considered include driver and other occupant characteristics, vehicle characteristics, and roadway attributes. Toward generating the proposed framework, we employ three specific modules: (a) disaggregate trip information generation, (b) crash data generation and (c) crash data aggregation (see Figure 1). The red tables represent the input data for the step and the green tables represent the data outputs in the step.

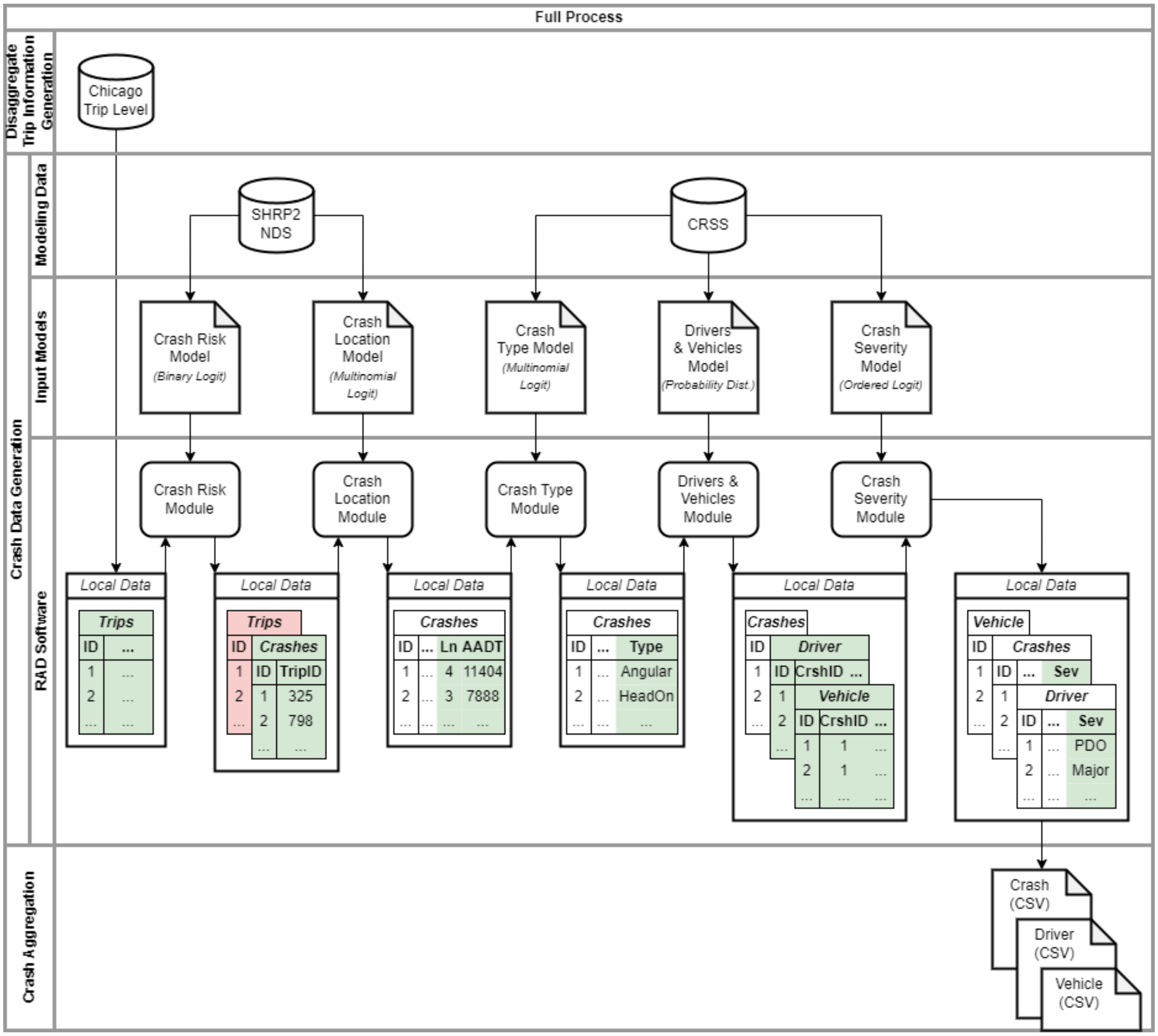

Figure 1: RAD Conceptual Framework

3.1. Disaggregate Trip Information Generation

The travel demand modeling paradigm has undergone a transformation from an aggregate zonal level statistical framework (such as a four step or trip-based model) to a disaggregate individual level framework (tour level and/or activity based models) (Kamel et al., 2019; Pinjari et al., 2008). The disaggregate frameworks accommodating for the influence of socio-demographic characteristics (such as income, age, household structure, education, car ownership), employment characteristics (such as employment industry and location), transportation network characteristics (such as access to travel mode and travel time by mode) and built environment measures (such as population density, land-use mix, public transit density), provide a representation of an individual's travel in continuous time and space. From these travel patterns, high resolution information for trips can be retrieved including trip start and end time, trip start and end location, trip characteristics (such as alone/group trip), vehicle used for the trip and precise route considered. In this research, we will employ a daily activity-travel realization for an urban region generated from an established travel demand model for the Chicago region developed by Argonne National Laboratory (Auld et al., 2016).

3.2. Crash Data Generation

The objective of the Crash Data Generation module is to generate crashes on the transportation system. The framework would utilize the detailed trip information from the previous module to generate crashes. The crash generation would involve identifying the vehicles involved in the crash, crash location, severity of drivers (such as fatal, capacitating injury, non-incapacitating injury and no injury), and crash type (such as head-on, rear-end, vehicle-pedestrian). The framework to be employed for crash generation is described below.

In the *first step* of the framework, the research team will classify the trips on the transportation system into two categories: (a) No Crash and (b) Crash. In urban regions, trips in a typical day amount to several million and are likely to take up large storage space with high resolution details on routing characteristics with geographical information system (GIS) coordinates. The proposed classification process allows us to reduce the number of trips to be used for crash data generation. Given the relatively small proportion of crash involved trips, the classification approach provides an elegant solution to computational and data burdens. The classification problem will be modelled using a binary classification model (such as binary logit or probit model).

The "crash" tagged trips will be processed in the *second step* of the framework to determine crash type, crash location and injury severity. It is important to note that while crash type and crash severity have fixed and well-defined alternatives, crash location alternatives are more complicated. Thus, depending on when crash location is examined, alternative structures for crash variable generation become possible. For example, one sequence can be as follows. For the crash tagged trips, a trip level model is estimated to identify the type of crash (such as head-on, rear-end and vehicle-pedestrian). Using the crash type, a subsequent model for crash severity follows. Finally, conditional on crash type and severity a crash location model is developed (see Figure 2). As we move toward the latter models in the sequence, the reader would recognize that more information is available i.e., additional independent variables can be included in the model estimation. For example, if crash severity follows crash type model, it will be possible to include crash type as an independent variable in the model. The crash location model that follows can have crash type and crash severity as independent variables. The attributes of other drivers and vehicles (for multi-vehicle crashes) involved in crashes will also be generated based on the driver and vehicle characteristics of the crash trips. Alternatively, the sequence of the variables can be altered to crash location followed by crash type and crash severity. In this sequence, crash location model estimation will be based on trip level characteristics and crash type and crash severity variable will have access to location variables in the model (see Figure 3 for a potential model structure).

The final step of crash data generation framework would involve determining the econometric model framework. Given that crash type and crash location are categorical variables a multinomial logit model framework would be appropriate. For the severity variable, given the inherent ordered nature of the variable, an ordered logit model structure would be employed.

3.3.Crash Aggregation

The crash data generation module will provide as outputs, the crash data including crash type, crash severity and crash location along with time and number of vehicles in the crash for a typical day in the year. However, for crash datasets it might be necessary to aggregate data temporally by facility type (such as crashes on a segment or intersection in a 6-month period or multiple years), and spatially (such as crashes in a zone, county). We can run the framework developed for a typical day multiple times with different random seeds (to ensure we don't just duplicate the same crashes in each run) to aggregate the data.

4. DATA SECTION

Three datasets were used in the development of this project: (1) Strategic Highway Research Program 2 (SHRP 2) Naturalistic Driving Study (NDS) data from Virginia Tech Transportation Institute (VTTI), (2) Crash Report Sampling System (CRSS) data from the National Highway Traffic Safety Administration (NHTSA) and (3) Chicago trip level data from Argonne National Laboratory.

4.1.SHRP2 NDS Data

The SHRP2 NDS data was used to develop the models for crash risk and crash location. This data was collected through a naturalistic driving study where cameras and sensors were placed in participants' cars to track their driving over an extended period of time. The data that we obtained information on 1,951 trips resulting in a crash, and 1,000,000 trips that did not result in a crash. These 1,000,000 trips were randomly selected from a full sample of 5,512,900 trips (Hankey et al., 2016). The data included information on trip data (such as start time, end time, day of week, facility locations, and facility speeds), driver demographics (such as age, gender, education, and income), crash event details (such as collision type, crash severity, driver impairments, and weather), and roadway segments and intersections (such as number of lanes, roadway classification, and AADT). Of the 1,951 trips where a crash occurred, 814 of those crashes were categorized as a "low risk tire strike", and were therefore removed from the list of crashes, leaving 1,137 crashes.

4.2.CRSS Data

The CRSS data was used to develop the models for crash type, drivers and vehicles, and crash severity. This data is a sampling of police reported crashes from across the United States. The data was from 2016 through 2019 and contained records for 200,682 crashes in all 50 states and the District of Columbia. The data included crash information (such as hour, day, location, lighting, weather, vehicle type, vehicle age, number of lanes, and speed limit) and driver information (such as age and gender). Of the original set of crashes, those with missing values were removed from analysis, leaving 113,983 crashes.

Table 2: SHRP2 Descriptive Statistics

Categorical Variables	
Variable Name	**Share of Category**
Age Distribution	
Age: 16-19	0.023
Age: 20-24	0.064
Age: 25-29	0.081
Age: 30-74	0.758
Age: > 74	0.074
Mileage Distribution	
Avg. annual miles: < 10,000	0.229
Avg. annual miles: 10,000 to 25,000	0.637
Avg. annual miles: > 25,000	0.134
Employment Status	
Worker: Full-time	0.48
Worker: Part-time	0.19
Worker: Not working outside the home	0.33
Gender Distribution	
Gender: Male	0.49
Gender: Female	0.51
Crash History	
Previous Crash (within 3 years): Yes	0.26
Previous Crash (within 3 years): No	0.74

Table 3: CRSS Descriptive Statistics

Categorical Variables	
Variable Name	**Share of Category**
Time of Day	
Hour: AM Peak	0.15
Hour: PM Peak	0.23
Hour: Off-Peak	0.62
Day of the Week	
Day: Weekday	0.77
Day: Weekend	0.23
Location of Crash	
Location: Urban	0.74

Location: Rural	0.26
Roadway Classification	
Highway: Yes	0.11
Highway: No	0.89
Lighting Condition	
Light: Day	0.69
Light: Dark, no light	0.12
Light: Dark, with light	0.18
Weather	
Weather: Clear	0.72
Weather: Adverse	0.28

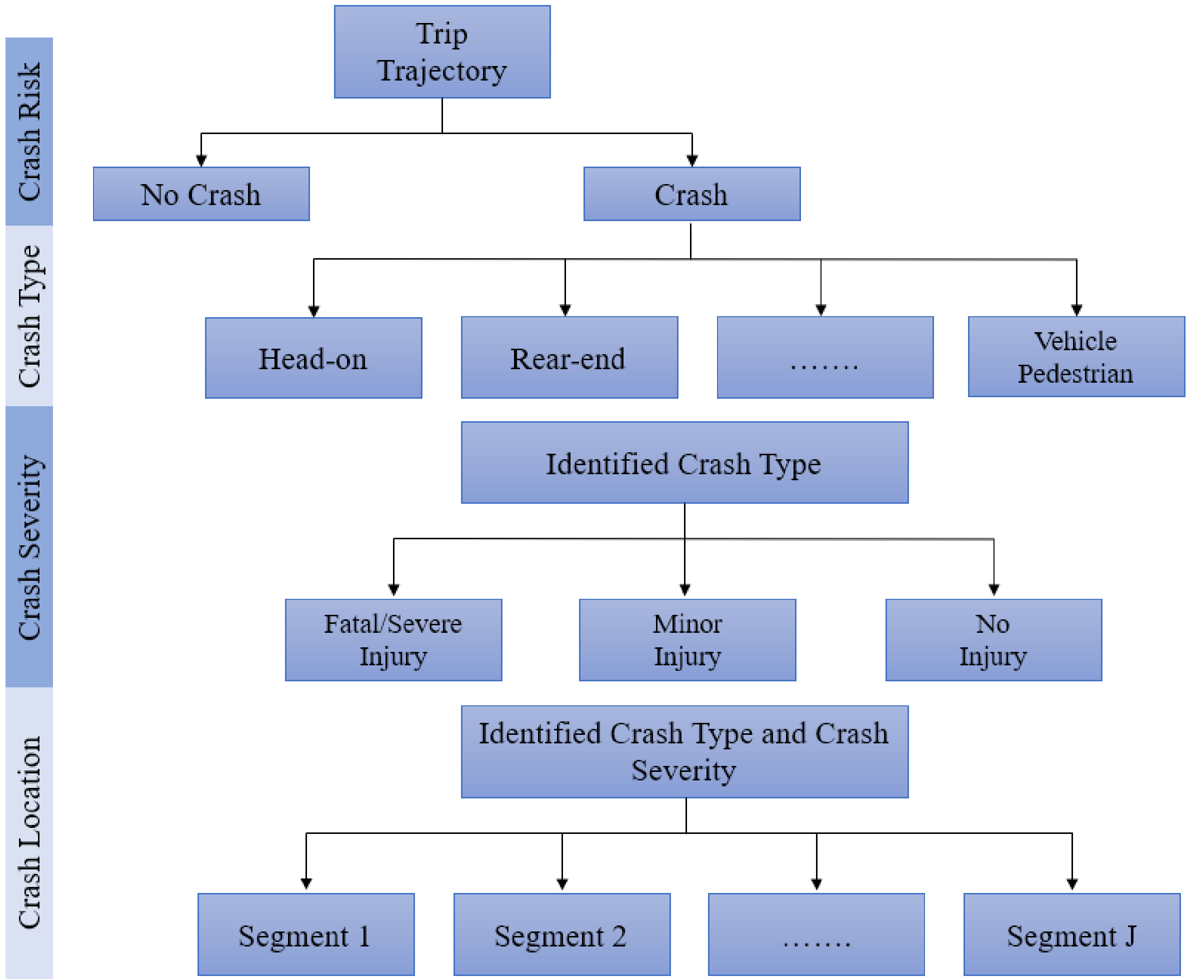

Figure 2: Sequential Approach I: Crash Risk → Crash Type → Crash Severity → Crash Location

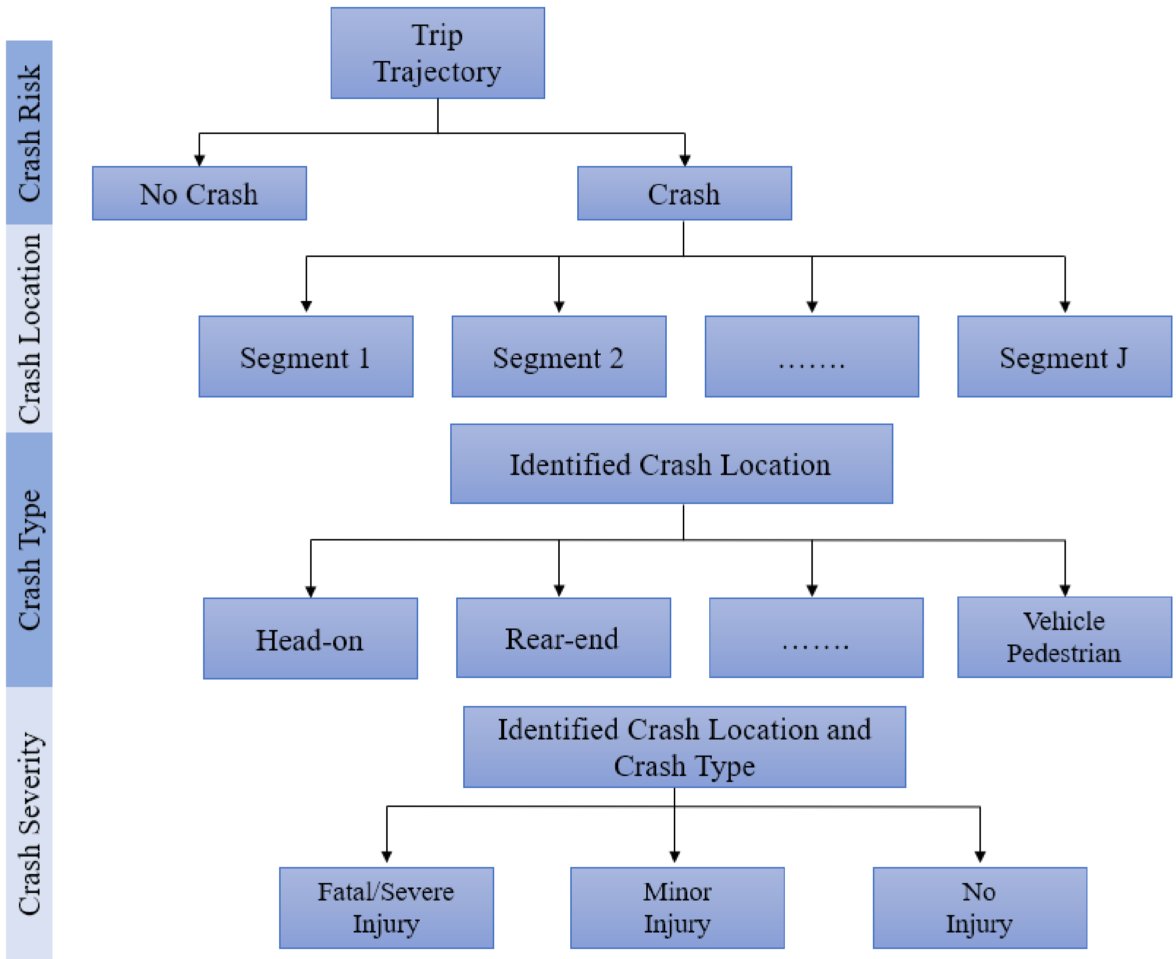

Figure 3: Sequential Approach II: Crash Risk → Crash Location → Crash Type → Crash Severity

4.3.Chicago Trip Level Data

The Chicago trip level data was used as an input for implementation of the RAD generator. The data contained 2,256,502 trips, with information on trip data (such as start time and duration), driver demographics (such as age and education), and roadway segments (such as AADT, number of lanes, and roadway type).

5. RAD MODULE DEVELOPMENT

In our analysis, based on data availability the first sequence presented in Figure 3 were employed. The module development included the estimation of five models described in this section.

5.1. Crash Risk

The goal of the crash risk module is to evaluate each trip and determine stochastically if a crash will occur during that trip. To develop the crash risk model, we used the SHRP2 NDS dataset. In the dataset there were 1,137 trips resulting in a crash and 1,000,000 trips that did not result in a crash. For this model we removed any trips that were missing relevant trip or driver information. This left 1,004 trips resulting in a crash and 774,873 trips that did not result in a crash. We had to further filter the data because crashes accounted for only 0.13% of trips, making them very difficult to model. Therefore, we under sampled the trips not resulting in a crash, randomly selecting 10% to be used for analysis. The final dataset that was used for model development contained 78,336 trips, 1,004 resulting in a crash and 77,332 that did not result in a crash. The reader is encouraged to review Hoover et al., 2022 for methods employed to minimize the impact of sampling.

For modeling crash risk, a binary logit model was used. The results of the model estimation are presented in Table 4. In this model, the only variable that had a statistically significant parameter at the 90% confidence level was age. Drivers less than 30 years old (with teenage drivers being the most likely) and greater than 74 years old were found to be more likely to be in a crash relative to other drivers.

Table 4: Crash Risk Model

Parameters	**Coefficients**	**T-value**
Constant	-5.4234	-86.994
Age (Base: 30-74 years)		
16-19 years	3.4055	36.915
20-24 years	2.5765	29.576
25-29 years	1.0682	8.252
Greater than 74 years	1.6611	15.553
N = 78,336 *LL = -4,593.91* *Note: All coefficients in the model are significant at the 95% confidence level (p-value <0.05).*		

Due to the under sampling of non-crash trips, the constant in the binary logit model is skewed towards a high crash risk. The constant was calibrated to match the true population crash shares. At the project sponsor's request, the calibrated parameter is not reported to avoid replication of our RAD software.

5.2. Crash Location

For developing the crash location model, we used the SHRP2 NDS dataset crash records with location information (about 857 crashes). As a significant amount of information was missing from the roadway data, the missing data was imputed based on the existing distributions observed in the data instead. For modeling the crash segments, the outcomes in the model could be very large for longer trips. To avoid computational complexity due to a large number of alternatives, a sampling of segments was considered for large trips spanning a large number of segments. The sampling process included the crash segment alternative and 29 additional segments randomly sampled from the trip segments (see Faghih-Imani and Eluru, 2015 for similar sampling approaches in literature). The results of the multinomial logit model estimated for crash location is presented in Table 5. For each trip, the longer segments tend to have a higher risk of a crash occurring. Additionally for each trip, segments with more lanes, those with a higher AADT, and collector roads tend to have a lower risk of a crash occurring.

Table 5: Crash Location Model

Parameters	Coefficients	T-value
Link length (x100)	0.7932	14.869
Number of lanes	-0.078	-1.910
AADT (/10,000)	-0.0387	-3.191
Collector Road	-0.5801	-4.611
N = 19,891 *LL = -2,378.94* *Note: All coefficients in the model are significant at 90% confidence level (p-value <0.1).*		

5.3. Crash Type

The goal of this module is to generate the type of crash that will occur based on trip and roadway variables. For developing the crash type model, we used the CRSS dataset. In the dataset there were 113,983 crashes. Of the 113,983 crashes in the CRSS dataset, 25,000 were randomly selected to be used for developing the crash type model. The alternatives considered for crash type were rear end crash, head on crash, angular crash, sideswipe crash, crash with fixed objects, crash with non-fixed objects, and non-motorized crash. Since different datasets were used for modeling and implementation, only those variables that were present in both datasets were considered when developing the model. The results of the multinomial logit model estimation can are presented in Table 6.

For this model, rear end crashes are used as the base alternative, with angular crashes and crashes with fixed and non-fixed objects having a higher probability of occurrence and head on crashes, sideswipe crashes, and non-motorized crashes having a lower probability of occurrence. Also, as the number of lanes increases, the probability of any crash, other than a rear end crash, decreases. Crashes on freeways have a higher likelihood of sideswipe crashes, and a lower probability of head on crashes, angular crashes, crashes with non-fixed objects, and non-motorized crashes. On weekdays, the probability of rear end crashes increases and the probability of head on crashes and crashes with fixed and non-fixed objects decreases. During the morning peak (7AM to 10AM), the probability of crashes with fixed and non-fixed objects and non-motorized crashes decreases. During the evening peak (4PM to 7PM), the probability of any crash, other than a rear end crash, decreases.

Table 6: Crash Type Model

Parameters	Rear end	Head on	Angular	Sideswipe	Crash with fixed objects	Crash with non-fixed objects	Non-motorized crash
Intercept	-	-1.49 (-13.74)	0.65 (12.37)	-0.82 (-13.16)	1.27 (22.01)	1.01 (15.06)	-0.24 (-3.39)
Roadway variables							
Number of lanes	-	-0.1 (-3.86)	-0.24 (-18.49)	-0.02 (-1.61)	-0.44 (-28.05)	-0.55 (-28.11)	-0.35 (-16.46)
Freeway	-	-2.02 (-8.81)	-2.3 (-21.71)	0.3 (5.51)	-	-0.24 (-3.41)	-2.31 (-11.76)
Temporal variables							
Weekdays	0.11 (2.87)	-0.19 (-2.16)	-	-	-0.54 (-12.68)	-0.47 (-9.23)	-
Morning peak	-	-	-	-	-0.33 (-6.1)	-0.48 (-6.89)	-0.18 (-2.32)
Evening peak	-	-0.25 (-2.95)	-0.19 (-4.72)	-0.34 (-6.53)	-0.84 (-16.23)	-0.74 (-12.16)	-0.34 (-5.17)
Format: Coefficient (t-statistic) *N = 25,000* *LL = -41,976.75* *Note: All coefficients in the model are significant at the 95% confidence level (p-value <0.05).*							

5.4. Drivers and Vehicles

The goal of this module is to generate data for each driver and vehicle involved in a crash. For the drivers and vehicles module we used Illinois crashes from the CRSS dataset. From this data we developed a probability distribution of different driver demographics (such as age, gender, and seatbelt use) and vehicle characteristics (such as type and age), which were used to generate driver and vehicle information for the generated crashes. The first step in generating the driver and vehicle information is determining the number of vehicles involved in the crash. This is partially based on the crash type generated in the previous module. If the crash type was defined as crash with fixed objects, crash with non-fixed objects, or non-motorized crash then it was considered a single vehicle crash. Otherwise, the number of cars was generated as 2 or 3 cars. The probabilities from the Illinois dataset for multivehicle crashes were 88.1% two vehicles and 11.9% three vehicles. This number was generated using a cumulative probability table as described in previous modules. Once the number of vehicles was determined, data was generated for each driver and vehicle involved in a crash. The first driver would have the same age as the primary driver in the trip data, but subsequently, all other information was generated.

5.5. Crash Severity

The goal of this module is to generate the severity of the crash for each driver based on trip data, roadway information, driver demographics, vehicle information, and crash type. Of the 25,000 crashes that were used in the crash type model, driver information was available for 24,351 crashes, resulting in 42,039 drivers that were used in developing the crash severity model.

For modeling crash severity an ordered logit model was used. In this case, the alternatives were property damage only (PDO), minor, major, and severe. The results of the model estimation are presented in Table 7. In this model, drivers that are less than 25 years old are less likely to experience high severity. Crashes that occur on freeways and those with a higher number of lanes are more likely to result in high severity. Crashes that occur on weekdays or during peak hours are likely to be less severe. Using rear end crashes, crashes with non-fixed objects, and non-motorized crashes as a base, sideswipe crashes are less likely to result in severe crashes, while head on crashes, angular crashes, and crashes with fixed objects are more likely to result in severe crashes. Using automobiles, motorcycles, and buses as a base, drivers in utility vehicles and trucks are less likely to sustain severe injuries. The reader would note that while motorcycles and buses are very different from automobiles, these three vehicle types were grouped together due to small sample sizes for motorcycles and buses.

Table 7: Crash Severity Model

Parameters	**Estimates**	**T-Value**
Thresholds		
a1	-0.1563	-4.787
a2	0.9296	28.171
a3	1.9466	56.983
Demographics		
Age (Base: 25 years and more)		
Less than 25 years	-0.1525	-6.68
Roadway variables		
Base: Other roadways		
Freeway	0.2432	7.882
Number of lanes	0.0286	4.513
Temporal Variable		
Base: Weekend		
Weekday	-0.1879	-8.015
Base: Off-peak		
Morning peak (7AM-10AM)	-0.1493	-5.563
Evening peak (4PM-7PM)	-0.1126	-5.1
Crash type		
Base: Rear end, crash with non-fixed objects, and non-motorized crash		
Head on	1.5195	30.1
Side swipe	-0.8238	-24.143
Angular crash	0.3918	17.54
Crash with fixed objects	0.6821	18.206
Vehicle type		
Base: Automobiles, motorcycle and bus		
Utility vehicles	-0.1199	-5.055
Light truck	-0.0827	-3.259
Medium and heavy truck	-0.3874	-6.62
N = 41,132 *LL = -49,819.08* *Note: All coefficients in the model are significant at the 95% confidence level (p-value <0.05).*		

6. RAD IMPLEMENTATION

The implementation of RAD modules involved the process of employing Monte Carlo simulation for each module discussed above. Typically, the simulation process involves generating the cumulative probability function (CPF) for all alternatives using the module specific model. Then, by generating a uniform random number between 0 and 1 and comparing it with the CPF, the chosen alternative is identified. Across different modules, different CPF formulae are employed. The rest of the process remains stable across all modules. The implemented routines are validated and checked to ensure the model outcomes follow expected distributions.

For testing, the RAD generator was used to generate a full year of data. For one year data, RAD generator is employed 365 times using the 2 million daily trip data records. Across each day, a different sample of crash records are generated and processed to generate crash location, crash type, driver/vehicle characteristics, and crash severity. The one-year RAD generated data on crash type, driver and vehicle characteristics, and crash severity are compared to the CRSS dataset (see Figures 4 through 7). In these figures the generated data is comparable to the CRSS dataset. In Figure 4, the biggest differences are the rear end crashes, which are slightly underestimated, and crashes with fixed and non-fixed objects, which are slightly overestimated. In Figure 5 and Figure 6, the main differences can be found in the number of vehicles and the driver age, which are partially affected by inputs from preceding modules. Crash severity results are well-aligned with the input data. The reader will note that the objective of RAD generation is not to match the observed data exactly. The emphasis is on ensuring that RAD generated crash data aligns with the observed crash data. Further, the embedded models can be carefully tweaked to produce different data distributions that might not be possible in empirical datasets.

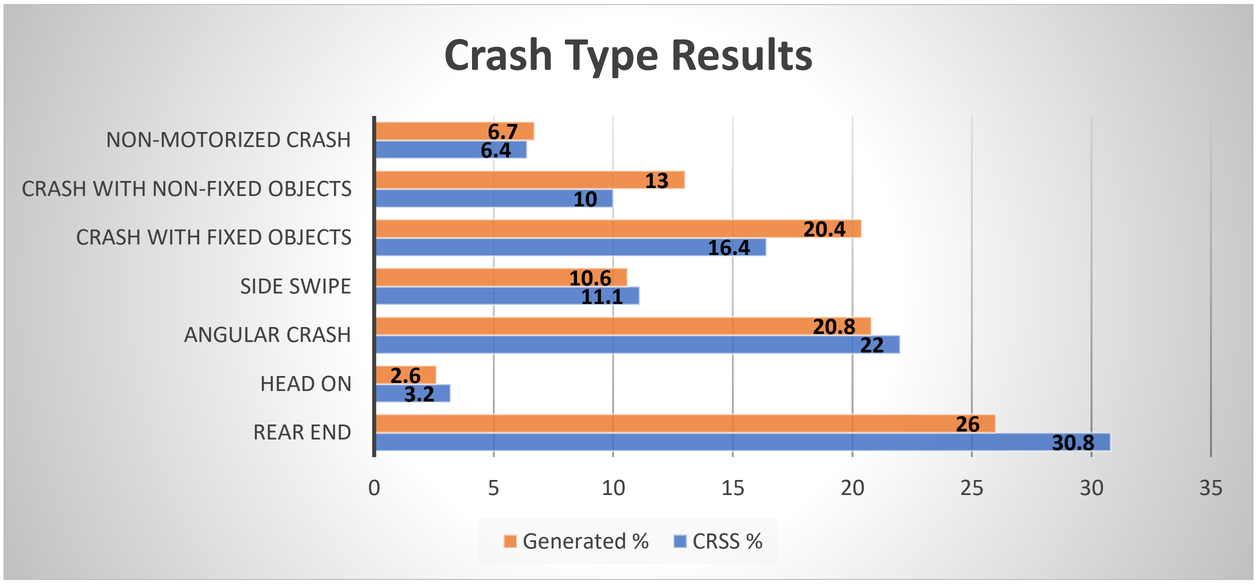

Figure 4: Crash Type Results

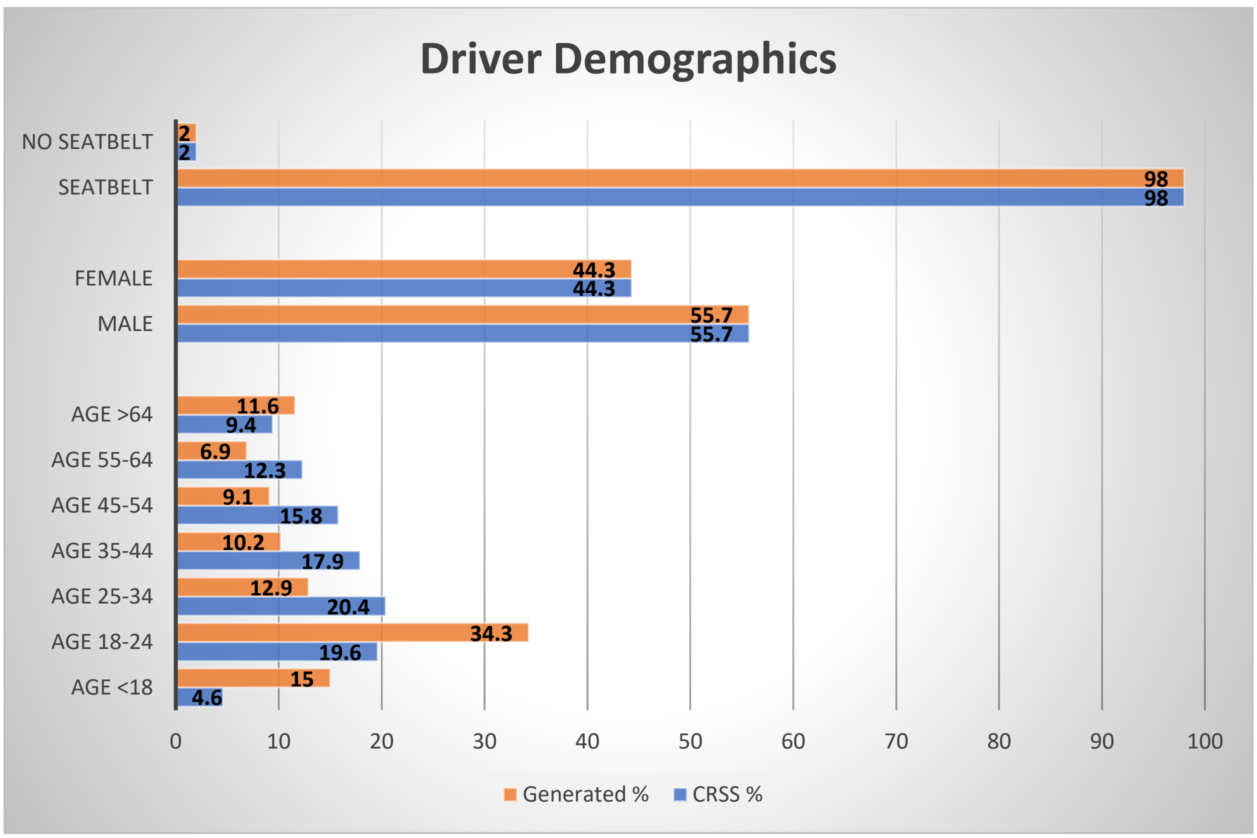

Figure 5: Driver Demographics Results

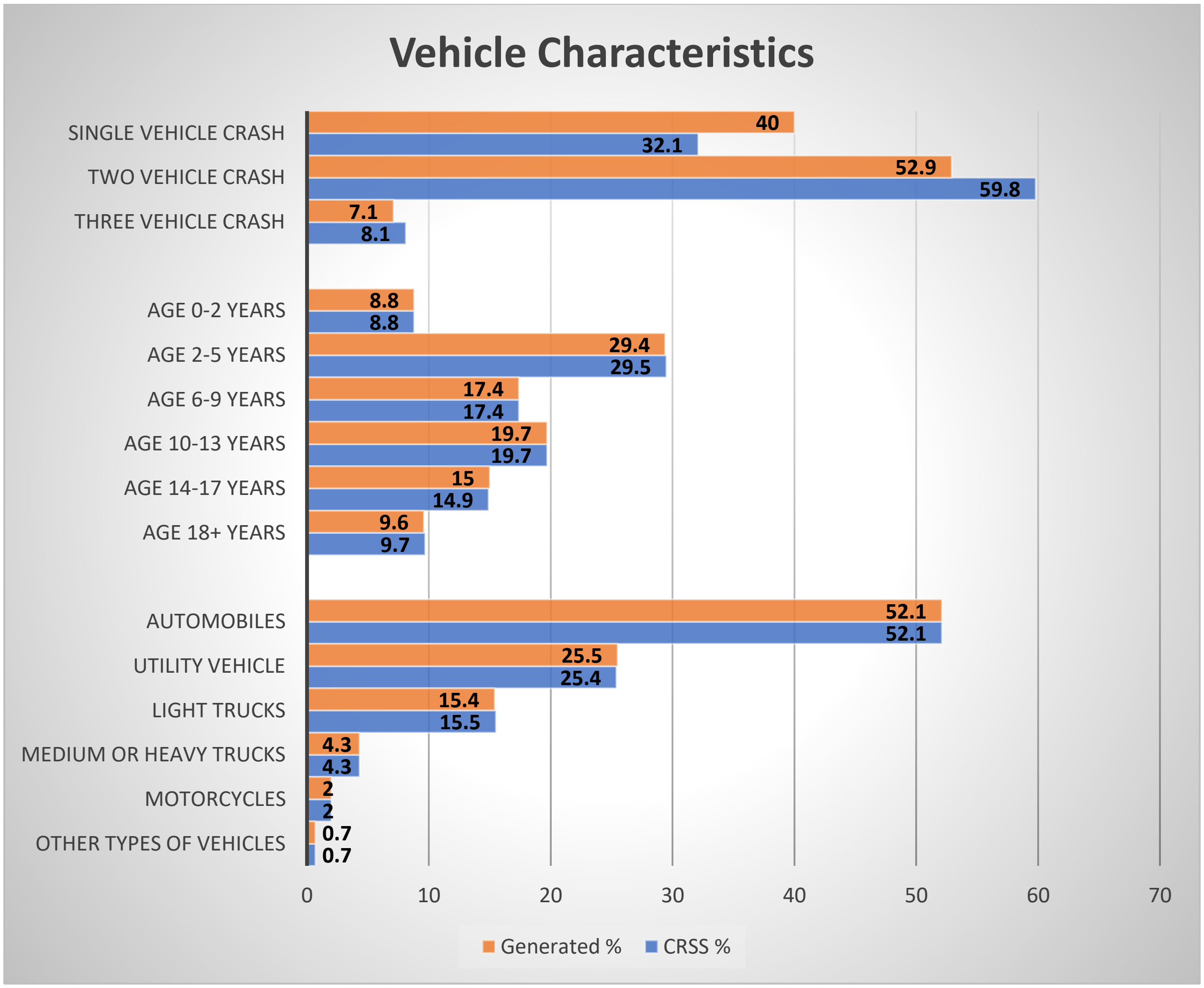

Figure 6: Vehicle Characteristics Results

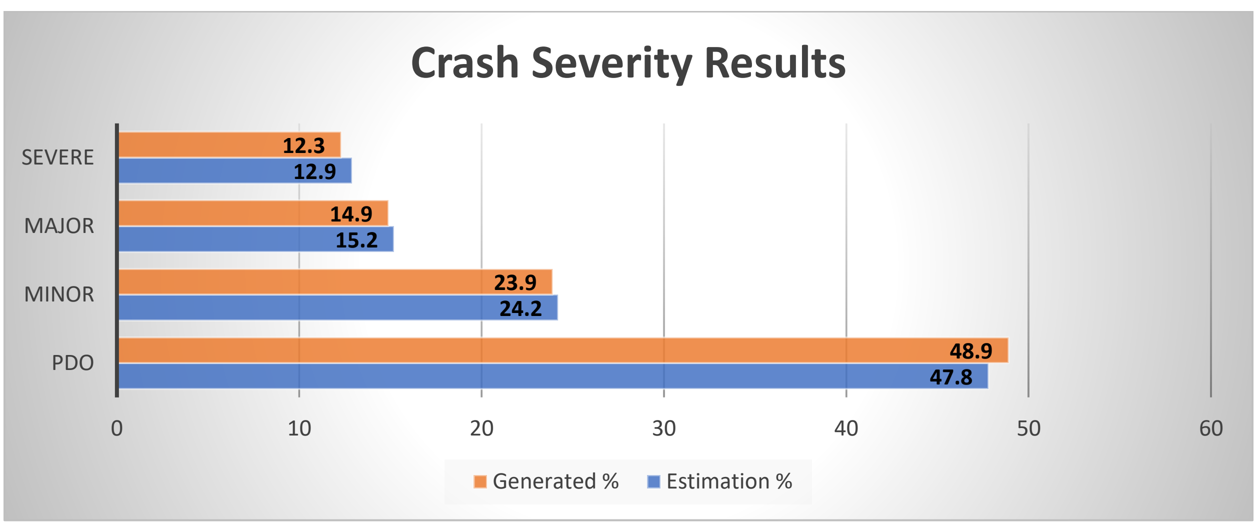

Figure 7: Crash Severity Results

7. RAD DATASETS

The RAD generator is implemented to produces 3 data files: the crash file (containing information on crash details such as location, type, and severity), the driver file (containing information on each driver involved in a crash and their individual injury severity), and the vehicle file (containing information on each vehicle involved in a crash). The three files generated are cross-linked and columns from one dataset can be readily merged into the other two files as needed. The user can specify the number of years of crash data to be produced by the RAD generator, as well as the number of instances of data for that number of years. For example, a user can specify that they want two sets of three years of crash data. When the RAD generator is run two different crash files, driver files, and vehicle files will be produced, each containing three years of data.

The crash dataset that is produced by the RAD generator can be used in a variety of ways. To analyze the crash data produced by the RAD generator, it can be aggregated by facility type (such as crashes on a segment in a 6-month period or multiple years) and spatially (such as crashes in a zone or county). There are also multiple variables that can be used for analysis. A selection of the variables (and their distribution) that could be used for analysis are shown in Figure 8. A user could analyze the data for roadway characteristics such as number of lanes, type of roadway, or AADT. A user could also analyze the data by crash characteristics such as time of crash, type of crash, or severity of crash. The crash databases generated can be employed to compare frequency models, severity models, crash type and various other dimensions by facility type. The development of the disaggregate RAD can serve as a universal benchmarking system for alternative model frameworks in safety literature.

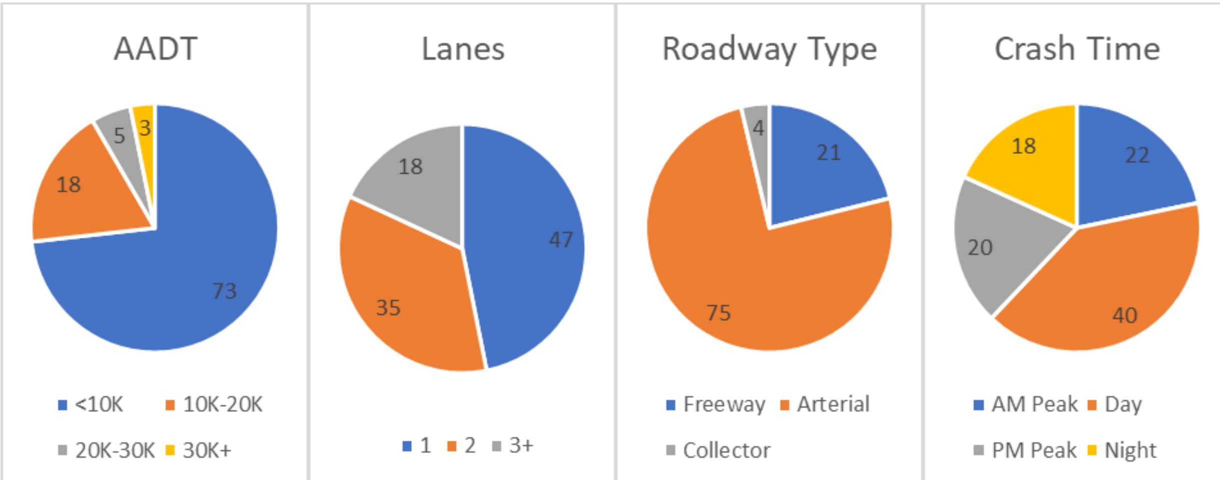

Figure 8: Sample Variable Distribution

8. CONCLUSION

Transportation safety modeling has broadly evolved along two streams: crash frequency models and crash severity models. The primary focus of these two streams of safety analysis is on identifying and quantifying the influence of factors contributing to traffic crash occurrence and its consequences. Traditionally approaches to model evaluation have relied on observed data and have multiple drawbacks. Realistic artificial data (RAD) is a potential innovative approach to address the over-reliance on observed datasets. In this paper, we implemented this solution by proposing a RAD generation framework which will generate a compilation of traffic crashes and characteristics to be used for safety analysis.

The proposed approach builds on Hauer's earlier work on generating crash data. In our study, we build on previous single level data generation process by employing a multi-level crash data generation process using trip level data for crash generation. Specifically, we generate crash data by considering crash occurrence as a trip level event impacted by trip level factors, demographic characteristics, roadway facility and vehicle attributes. This conceptual framework has five stages of crash data generation that are described in this paper. The first stage of data generation is the crash risk stage, which evaluates a series of trips using a binary logit model to classify each trip as “crash” or “no crash”. The second stage of data generation is the crash location stage, where the location of each “crash” trip is determined using a multinomial logit model. The third stage of data generation is the crash type stage, where the type of each crash is determined using a multinomial logit model. The fourth stage of data generation is the drivers and vehicles stage, where data on the driver(s) and vehicle(s) associated with each crash are generated using a probability distribution table. The fifth and final stage of data generation is the crash severity stage, where the severity of the crash is generated for each driver involved in a crash using an ordered logit model. Each of these modules is implemented sequentially in the RAD generator using the Python programming language. After Monte Carlo implementation of the RAD generator, the software will provide crash data in three interconnected files including (a) crash file, (b) driver file and (c) vehicle file. In future work, the crash databases generated can be employed to compare frequency models, severity models, crash type and various other dimensions by facility type. The development of the disaggregate RAD can serve as a universal benchmarking system for alternative model frameworks in safety literature. The approach can be enhanced further by employing trip level data from multiple urban regions.

It is important to note that the crash frequency variables generated in our RAD originate from a multi-level aggregation of crashes on a single day. Hence, the crash frequency models developed with this data might not always be aligned with the current state of the art crash frequency models that assume a count over an aggregated timeframe. It will be an interesting future exercise to test how model specifications will vary between RAD datasets and traditional aggregated observed datasets. The RAD generator was developed based on trip data from one jurisdiction. It would be beneficial to update the RAD generator with data from multiple jurisdictions to enhance wider applicability. The RAD framework developed in the current study should serve as starting point for future efforts that can establish benchmarks for safety modeling selection in the future.

ACKNOWLEDGMENT

This study was supported by the U.S. DOT Federal Highway Administration through the grant No.693JJ31950017 titled as “Disaggregate Realistic Artificial Data (Rad) Generator - Design, Development and Application for Crash Safety Analysis”. However, the authors are solely responsible for the facts and accuracy of the information presented in the paper.

AUTHOR CONTRIBUTIONS

The authors’ confirmed contributions are as follows; study conception and design: Eluru, Konduri, Ivan, Zhao, and Wang; literature review: Tirtha, Bhowmik, Eluru; data collection: Hoover, Jahan, Tirtha, Bhowmik, Auld, Eluru; Model Estimation: Hoover, Jahan, Bhowmik; Analysis and Interpretation: Hoover, Jahan, Bhowmik; Eluru; Draft Manuscript: Hoover, Jahan, Bhowmik, Review: All Authors

REFERENCES

Asano, M., Iryo, T., and Kuwahara, M. (2010) 'Microscopic pedestrian simulation model combined with a tactical model for route choice behaviour', *Transportation Research Part C: Emerging Technologies*, 18(6), pp. 842-855.

Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., and Zhang, K. (2016) 'POLARIS: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations', *Transportation Research Part C: Emerging Technologies*, 64, pp. 101-116.

Berke, A., Doorley, R., Larson, K., and Moro, E. (2022) 'Generating synthetic mobility data for a realistic population with RNNs to improve utility and privacy', *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. April. pp. 964-967.

Bhat, C.R., Sener, I.N., and Eluru, N. (2010) 'A Flexible Spatially Dependent Discrete Choice Model: Formulation and Application to Teenagers' Weekday Recreational Activity Participation', *Transportation Research Part B*, 44(8-9), pp. 903-921.

Bhat, C.R. and Sidharthan, R. (2011) 'A Simulation Evaluation of the Maximum Approximate Composite Marginal Likelihood (MACML) Estimator for Mixed Multinomial Probit Models', *Transportation Research Part B*, 45(7), pp. 940-953.

Bonneson, J. and Ivan, J. (2013) 'Theory, Explanation, and Prediction in Road Safety: Promising Directions', Transportation Research Circular, E-C179

Charalambidis, G., Akasiadis, C., Rigas, E.S., and Chalkiadakis, G. (2022) 'A realistic dataset generator for smart grid ecosystems with electric vehicles', *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems*. June. pp. 432-433.

Chatterjee, S., Singh, A., Mitra, B., and Chakraborty, S. (2022) 'Realistic Multiuser, Multimodal (IMU, Acoustic) HAR Data Generation through Single User Data Augmentation', *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, IEEE. May. pp. 533-534.

Cummings, P., McKnight, B., and Weiss, N.S. (2003) 'Matched-pair cohort methods in traffic crash research', Accident Analysis & Prevention, 35(1), pp. 131-141.

Dahmen, J. and Cook, D. (2019) 'SynSys: A Synthetic Data Generation System for Healthcare Applications', *Sensors*, 19(5), pp. 1181.

Devroye, L., Felber, T., and Kohler, M. (2012) 'Estimation of a density using real and artificial data', *IEEE Transactions on Information Theory*, 59(3), pp. 1917-1928.

Eluru, N., Pinjari, A.R., Guo, J.Y., Sener, I.N., Srinivasan, S., Copperman, R.B., and Bhat, C.R. (2008) 'Population updating system structures and models embedded in the comprehensive econometric microsimulator for urban systems', *Transportation Research Record: Journal of the Transportation Research Board*, 2076, pp. 171-182.

Eluru, N. (2013) 'Evaluating Alternate Discrete Choice Frameworks for Modeling Ordinal Discrete Variables', *Accident Analysis & Prevention*, 55(1), pp. 1-11.

Faghih-Imani, A. and Eluru, N. (2015), ‘Analyzing Bicycle Sharing System User Destination Choice Preferences: An Investigation of Chicago’s Divvy System’, *Journal of Transport Geography*, 44, pp. 53-64.

Ferdous, N., Eluru, N., Bhat, C.R., and Meloni, I. (2010) 'A Multivariate Ordered Response Model System for Adults' Weekday Activity Episode Generation by Activity Purpose and Social Context', *Transportation Research Part B*, 44(8-9), pp. 922-943.

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

Hankey, J.M., Perez, M.A., and McClafferty, J.A. (2016) *Description of the SHRP 2 Naturalistic Database and the Crash, Near-Crash, and Baseline Data Sets*. Blacksburg, VA: Virginia Tech Transportation Institute.

Hazwani, R.A., Wahida, N., Shafikah, S.I., and Ellyza, P.N. (2016) 'Automatic artificial data generator: Framework and implementation', *2016 International Conference on Information and Communication Technology (ICICTM), IEEE*. May. pp. 56-60.

Highway Safety Information System (HSIS). (2017) *Use of “Artificial Realistic Data” (ARD) to Assess the Performance of Cross-Sectional Analysis Methods in Capturing Causal Relationships between Individual Roadway Attributes and Safety*.

Hoover, L., Bhowmik, T., Yasmin, S., and Eluru N., (2022) “Understanding Crash Risk using a Multi-Level Random Parameter Binary Logit Model: Application to Naturalistic Driving Study Data”, *Transportation Research Record* 2676(10), 737–745

Kamel, J., Vosooghi, R., Puchinger, J., Ksontini, F., and Sirin, G. (2019) 'Exploring the impact of user preferences on shared autonomous vehicle modal split: A multi-agent simulation approach', *Transportation Research Procedia*, 37, pp. 115-122.

Konduri, K.C., You, D., Garikapati, V.M., and Pendyala, R.M. (2016) 'Enhanced synthetic population generator that accommodates control variables at multiple geographic resolutions', *Transportation Research Record: Journal of the Transportation Research Board*, 2563, pp. 40-50.

Lord, D. and Kuo, P.F. (2012) 'Examining the effects of site selection criteria for evaluating the effectiveness of traffic safety countermeasures', *Accident Analysis & Prevention*, 47, pp. 52-63.

Mamun, S., Caraballo, F.J., Ivan, J.N., Ravishanker, N., Townsend, R.M., and Zhang, Y. (2018) 'Identifying association between pedestrian safety interventions and street-crossing behavior considering demographics and traffic context', *Journal of Transportation Safety & Security*, pp. 1-22.

Paleti, R. and Bhat, C.R. (2013) 'The Composite Marginal Likelihood (CML) Estimation of Panel Ordered-Response Models', Journal of Choice Modelling, 7, pp. 24-43.

Pinjari, A., Eluru, N., Srinivasan, S., Guo, J.Y., Copperman, R., Sener, I.N., and Bhat, C.R. (2008) 'Cemdap: Modeling and microsimulation frameworks, software development, and verification', *Proceedings of the transportation research board 87th annual meeting*. January.

Pinjari, A.R. and Bhat, C.R. (2010) 'A Multiple Discrete-Continuous Nested Extreme Value (MDCNEV) Model: Formulation and Application to Non-Worker Activity Time-Use and Timing Behavior on Weekdays', *Transportation Research Part B*, 44(4), pp. 562-583.

Ranade, S., Sadek, A.W., and Ivan, J.N. (2007) 'Decision support system for predicting benefits of left-turn lanes at unsignalized intersections', *Transportation Research Record: Journal of the Transportation Research Board*, 2023, pp. 28-36.

Scott, P.D. and Wilkins, E. (1999) 'Evaluating data mining procedures: techniques for generating artificial data sets', Information and software technology, 41(9), pp. 579-587.

Wu, L., Lord, D., and Zou, Y. (2015) 'Validation of crash modification factors derived from cross-sectional studies with regression models', *Transportation Research Record: Journal of the Transportation Research Board*, 2514, pp. 88-96.

Ye, F. and Lord, D. (2011) 'Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: multinomial logit, ordered probit, and mixed logit', *Transportation Research Record: Journal of the Transportation Research Board*, 2241, pp. 51-58.

Yu, R. and Abdel-Aty, M. (2014) 'An optimal variable speed limits system to ameliorate traffic safety risk', *Transportation research part C: emerging technologies*, 46, pp. 235-246.

Zimmermann, A. (2012) 'Generating Diverse Realistic Data Sets for Episode Mining', *IEEE 12th International Conference on Data Mining Workshops*.