Accident Analysis and Prevention 195 (2024) 107407

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Advancing proactive crash prediction: A discretized duration approach for predicting crashes and severity

Diwas Thapa [a], Sabyasachee Mishra [a,*], Nagendra R. Velaga [b], Gopal R. Patil [b]

[a] Department of Civil Engineering, University of Memphis, Memphis, TN 38152, United States
[b] Department of Civil Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India

ARTICLE INFO

Keywords:
Crash likelihood
Crash severity likelihood
Survival model
Choice model
Proactive safety performance function
Predictor stability
Real time prediction validation

ABSTRACT

Driven by advancements in data-driven methods, recent developments in proactive crash prediction models have primarily focused on implementing machine learning and artificial intelligence. However, from a causal perspective, statistical models are preferred for their ability to estimate effect sizes using variable coefficients and elasticity effects. Most statistical framework-based crash prediction models adopt a case-control approach, matching crashes to non-crash events. However, accurately defining the crash-to-non-crash ratio and incorporating crash severities pose challenges. Few studies have ventured beyond the case-control approach to develop proactive crash prediction models, such as the duration-based framework. This study extends the duration-based modeling framework to create a novel framework for predicting crashes and their severity. Addressing the increased computational complexity resulting from incorporating crash severities, we explore a tradeoff between model performance and estimation time. Results indicate that a 15 % sample drawn at the epoch level achieves a balanced approach, reducing data size while maintaining reasonable predictive accuracy. Furthermore, stability analysis of predictor variables across different samples reveals that variables such as *Time of day (Early afternoon), Weather condition (Clear), Lighting condition (Daytime), Illumination (Illuminated)*, and *Volume* require larger samples for more accurate coefficient estimation. Conversely, *Daytime (Early morning, Late morning, Late afternoon), Lighting condition (Dark lighted), Terrain (Flat), Land use (Commercial, Rural), Number of lanes,* and *Speed* converge towards true estimates with small incremental increases in sample size. The validation reveals that the model performs better in highway segments experiencing more frequent crashes (segments where the duration between crashes is less than 100 h, or approximately 4 days).

1. Introduction

Crash prediction models can be categorized into two main types: diagnostic crash prediction models, also known as reactive crash prediction models, and proactive or real-time crash prediction models. These two types of prediction models differ in their application and the variables they incorporate. Reactive crash prediction models rely on historical crash data, as well as static covariates (variables that do not change over time) and dynamic covariates (variables that do change over time), aggregated over a specific period. Examples of such dynamic covariates include Average Annual Daily Traffic and average speed. These models are valuable for developing safety performance functions, which help identify the precursors of crashes and evaluate the impact of safety interventions and policies on highway safety (Yasmin et al., 2018). On the other hand, proactive crash prediction models refer to real-time crash prediction models that utilize historical crash data and static covariates, such as roadway condition and roadway geometry, along with disaggregated dynamic covariates that vary with time. These dynamic covariates can include traffic volume, speed, and weather conditions collected in near real-time. By incorporating dynamic predictors, these models can account for changing traffic and weather conditions, allowing for the forecasting of the likelihood of future crashes in real time. This, in turn, enables the implementation of crash mitigation strategies.

Proactive crash prediction models have garnered significant attention from researchers in recent years due to their potential to forecast and prevent future crashes. The availability of granular traffic flow data, such as near real-time traffic flow data collected at small time intervals, from Intelligent Transportation System infrastructure, coupled with the computational performance of modern computers, has played a crucial

* Corresponding author.
E-mail addresses: dthapa@memphis.edu (D. Thapa), smishra3@memphis.edu (S. Mishra), n.r.velaga@iitb.ac.in (N.R. Velaga), gpatil@iitb.ac.in (G.R. Patil).

https://doi.org/10.1016/j.aap.2023.107407
Received 7 August 2023; Received in revised form 4 November 2023; Accepted 24 November 2023
Available online 5 December 2023

role in increasing the popularity of these models. Modern data-driven methods, such as Machine Learning (ML), have gained popularity as they replace traditional statistical models which are often relatively more difficult to fit (Mannering et al., 2020). Data-driven methods have demonstrated superior data fit and predictive capabilities as they are not constrained by assumptions inherent to traditional econometric frameworks, such as statistical distribution and variable correlation. However, data-driven methods have their own limitations too. They struggle with problems related to model transferability, generalization, and the inability to quantify variable effects. In this context, statistical econometric frameworks, through variable coefficients and elasticities, can quantify variable effects and provide model transferability and generalization. In these respects, statistical models can be considered superior to data-driven methods.

Due to the benefits offered by statistical econometric frameworks, there are ongoing efforts to enhance and refine traditional statistical approaches to address their limitations and apply them to proactive modeling. For instance, researchers have extended standard econometric frameworks by incorporating flexible structures to develop mixed and generalized models. These models can account for unobserved heterogeneity and hierarchical structures for variable correlations and dependencies. More recently, researchers developed and implemented a new crash prediction framework (Thapa et al., 2022). In their study, researchers developed a duration-based crash prediction model that combines elements of the survival model and Multinomial Logit model (MNL). In this modeling approach, the time duration between crashes is divided into 1-hour epochs, which are further subdivided into 4 15-minute time intervals. Each epoch between two consecutive crashes is treated as a separate observation, with the time intervals serving as choice alternatives. By adopting this approach, the framework can forecast the likelihood of future crashes by considering two types of covariates. Firstly, static covariates associated with crashes, such as highway geometry and environmental conditions, are repeated over each epoch. Secondly, dynamic covariates, such as traffic flow and speed, change across epochs and within the 15-minute time intervals. The authors of the study discovered that the duration-based model could generate reasonably accurate estimates even when dealing with small sample sizes.

The current study builds upon the duration-based model by incorporating crash severities. While prediction of crash occurrence has already been addressed in previous research, forecasting likelihood of different crash severities is crucial from multiple perspectives, including safety, economic, and planning considerations. The costs associated with crashes vary significantly depending on their severity. For instance, the comprehensive unit cost of a Property Damage Only (PDO) crash in the US was estimated to be around $12,000 in 2016, whereas a fatal crash was estimated to exceed $11 million (Harmon et al., 2018). Additionally, crash severities are linked to road user costs. Studies have indicated that more severe crashes require more time to clear, resulting in higher road user costs (Golob et al., 1987; Lee and Fazio, 2005). Therefore, prioritizing the identification and addressing of factors contributing to more severe crashes is crucial from both safety and economic perspectives. Furthermore, from a planning standpoint, the ability to forecast crash severities provides transportation agencies with valuable insights. Agencies are often constrained with limited resources and personnel, making it necessary to identify critical segments in advance and proactively address adverse traffic flow conditions. By forecasting crash severities, agencies can prioritize the allocation and deployment of resources and personnel to prevent severe crashes and mitigate their impacts, contributing to more efficient and effective traffic operations and planning.

2. Literature review

Research in crash prediction has focused on forecasting both crash occurrences and severities. In the following sections, we provide a literature review of prediction models based on the specific outcomes they forecast. While we will discuss both proactive and reactive crash prediction models, this review will place greater emphasis on proactive crash prediction models, as they align with the scope of our study.

2.1. Crash prediction models

The first group of studies focuses on real-time forecasting of future crashes, employing both data-driven and statistical methods. Researchers have utilized various approaches to develop these models. Data-driven methods have gained popularity in the literature, with several notable examples including Support Vector Machines (Sun and Sun, 2016; Yu and Abdel-Aty, 2013), decision trees and random forests (Beshah et al., 2011; Pham et al., 2010), neural networks (Li et al., 2020), and Bayesian statistics (Hossain and Muromachi, 2012; Zheng and Sayed, 2020). These data-driven methods have proven effective in capturing complex relationships and patterns in crash data, allowing for real-time forecasting of future crash occurrences.

On the statistical side, the case-control design approach has been the most popular method for developing proactive crash prediction models (Hossain et al., 2019). In this approach, crashes are matched with non-crash events based on specific variables such as location and time of the crash (Abdel-Aty et al., 2004). The resulting dataset, with binary outcomes indicating crash or non-crash events, is well-suited for binary logistic regression. However, researchers have also explored the use of data-driven methods and Bayesian statistics to enhance the modeling capabilities of this approach (Hossain et al., 2019). In addition to the traditional case-control approach, alternative methodologies have been proposed. For example, (Yasmin et al., 2018) developed a MNL that considered 5-minute intervals for the next 30 days as choice alternatives, representing the occurrence of crashes in future time intervals. Given the substantial number of choice alternatives, the authors employed sampling techniques (selecting 29 randomly sampled time intervals and 1 interval with a crash) from the 30-day period.

More recently, researchers implemented a real-time crash prediction model by combining survival model with the MNL model. Survival models or duration models have been employed to model traffic crashes using static data (e.g., (Jovanis and Chang, 1989; Thapa and Mishra, 2021), however, they are incapable of incorporating time-varying covariates. The researchers developed a new method to restructure the crash data by creating forecasting epochs and time-intervals that can be associated with the dynamic covariates (Thapa et al., 2022).

2.2. Crash severity prediction models

The second group of studies focuses on predicting crash severity. Data-driven methods have been used more often to forecast crash severities, with various approaches utilized in different studies. Deep learning methods have been applied in crash severity prediction (Rahim and Hassan, 2021), while Support Vector Machines have been utilized in studies by (Chen et al., 2016; Iranitalab and Khattak, 2017). Random forests have also been used as a predictive technique for crash severity forecasting (Iranitalab and Khattak, 2017). Other methods such as neural networks and decision trees have been explored in some studies (Lee et al., 2019; Ospina-Mateus et al., 2021; Zhang et al., 2020). In recent years, a significant focus has been placed on comparing the performance of these algorithms in crash severity prediction (Santos et al., 2022). It is important to note that most prediction models within this group are reactive in nature, aiming to predict crash severity based on historical data and established patterns.

The most common statistical approach for developing crash severity prediction models is applying discrete choice models, specifically multinomial and ordered response logit/probit models. However, more advanced statistical models such as random parameter mixed models have gained popularity among researchers in recent years, as they offer solutions to the fixed parameter restriction imposed by choice models.

Uncorrelated random parameter models (Fountas and Anastasopoulos, 2017) correlated random parameter models (Ahmed et al., 2021; Fountas and Anastasopoulos, 2017), and generalized ordered response models (Osman et al., 2019; Osman et al., 2018a,b; Yasmin et al., 2014) are some of the examples of these advanced statistical models. These models enable researchers to account for parameter variations across different observations, providing more flexibility in capturing the complexity of crash severity prediction. Another approach for crash severity prediction involves the use of sequential models that can account for the dependency between various levels of crash severities. Studies have explored the application of sequential models in crash severity prediction, allowing for the consideration of dependencies between crash severities (Dissanayake and Lu, 2002; Jung et al., 2010).

With the advent of advanced models, researchers have conducted studies to examine and compare their predictive performance. For instance, Yasmin and Eluru, 2013) compared different generalized and mixed models within the frameworks of ordered and unordered choice modeling. Their findings indicated that mixed generalized ordered logit and mixed MNL models showed promise in predicting crash injury severity. In a study by (J. Zhang et al., 2018), various statistical and machine learning methods were compared, and it was found that machine learning algorithms exhibited better performance. This improvement could be attributed to factors such as the linear utility function and parametric assumptions regarding the error term. (Cerwick et al., 2014) conducted a comparison between mixed MNL and latent class MNL models. Their analysis revealed that the former model provided better average predictions across different severity levels.

2.3. Models predicting crash frequency and severity

The final group of studies focuses on forecasting both crashes and their severity. However, it is important to note that most of these models are primarily designed to forecast crash frequencies rather than the presence or absence of crashes.

Multivariate count data models are commonly employed in these studies, as seen in the works of (Jonathan et al., 2016; Ma and Kockelman, 2006; Park and Lord, 2007). Additionally, random parameter count data models have been used to account for spatial and temporal heterogeneity, as demonstrated by (Barua et al., 2016; Cheng et al., 2017; Dong et al., 2014). Other studies have implemented joint models with two components: (i) a crash prediction component utilizing count data models, and (ii) a crash severity component employing discrete choice models to predict crash counts by severity. This approach has been employed by (Afghari et al., 2020; Pei et al., 2011; Yasmin and Eluru, 2018).

The sequential logit model has also been used to predict the likelihood and severity of crashes. (Xu et al., 2013) developed a model using sequential binary logit models, where crashes were modeled in three stages: Stage 1 (crash vs. non-crash), Stage 2 (property damage only vs. higher severities), and Stage 3 (non-capacitating vs. higher severities). However, a significant drawback of the sequential logit model in the context of proactive crash prediction is that the estimation of multiple models can be computationally demanding and time-consuming, making it impractical for large datasets.

3. Study contributions

Only a limited number of statistical approaches have been developed to date for proactive crash prediction, apart from the commonly used case-control approach. This study introduces a duration-based prediction model for both crash occurrence and crash severity. The model framework involves dividing the time duration between historical crashes into distinct time periods to create forecasting epochs and time intervals. This allows the model to incorporate dynamic covariates and ascertain the probability of crashes occurring in future epochs and time intervals (Thapa et al., 2022). While this modeling approach has previously been demonstrated for crash prediction, the current study extends the framework to incorporate crash severities. The major contributions of this paper can be summarized as follows.

1. We expand upon the duration-based proactive crash prediction model by introducing a novel modeling approach that can forecast both crash occurrence and severity. Our model framework is one of handful statistical approaches for proactive crash prediction that does not rely on the case-control approach (Thapa et al., 2022). Unlike the original model, which solely predicts the likelihood of crashes for discrete future time intervals, our proposed model can also predict the corresponding crash severities.

Furthermore, the proposed model is implemented using a larger dataset. Specifically, the model is applied to crash data collected from interstates in two cities in Tennessee, thereby achieving a broader geographical coverage in comparison to the previous study that focused on a single city. This expanded geographical scope enhances the generalizability of the crash predictors, as it ensures adequate representation of diverse roadway conditions and traffic patterns across the study areas.

2. The proposed modeling framework demands discretizing the time duration between crashes to create forecasting epochs (more on this in this in the next section). Consequently, the size of the initial crash data expands significantly. Prior studies have indicated that appropriate sampling techniques can address estimation complexities arising from large data size, thereby allowing for parameter estimation with a reasonable degree of accuracy (Thapa et al., 2022). However, the incorporation of crash severities adds an additional layer of complexity to the model estimation process.

Therefore, this study aims to investigate the influence of sample size on variable coefficients and identify variables that are sensitive to changes in sample size. Understanding the variables that are particularly impacted by sample size variations is crucial for the implementation of the model. Additionally, this information will play a pivotal role in assessing the reliability of the model and guiding future data collection efforts.

4. Methodology

In this section, we present the methodology under three distinct subsections: the duration-based prediction framework, the nested logit model, and the estimation of the nested logit model. First, we describe the duration-based prediction framework and the process of creating forecasting epochs. This section is followed by the introduction of the two-level nested logit model and its relationship with the duration-based crash prediction framework. Finally, we discuss the estimation processes used in this study to estimate the parameters of the models.

4.1. Duration based prediction framework

In the duration-based crash prediction model, the occurrence of a crash at any time interval dt can be modeled using the MNL framework with alternatives, n and the hazard rate, h given by $U_n = -h(n-1)dt$ (Thapa et al., 2022). By utilizing this relationship, the latent propensity function for each time interval can be expressed as a function of static and dynamic covariates (time-varying factors). The application of this concept is illustrated in the following example.

Example:

Consider the duration between crashes in a highway segment, denoted as s, which is discretized into epochs, denoted as e, each with time intervals, denoted as i, and each interval has a duration of dt. Using these indices, we can examine historical crash data for a roadway segment, $s = 1$, where three consecutive crashes, denoted as A1, A2, and A3, were observed with durations of 2.5 h and 0.5 h apart (see Table 1 (a)). Additionally, available are dynamic covariates, speed and volume for the segment and the crash year at a temporal resolution of dt, as shown in Table 1(b). These covariates, as depicted, exhibit time-varying

Table 1a
Historical crash data with static covariates.

s	Crash	Date of crash	Time of crash	Severity	Terrain (Flat = 1, Rolling = 0)
1	A1	1/1/2023	00:00	Fatal	1
1	A2	1/1/2023	02:30	PDO	1
1	A3	1/1/2023	03:00	Injury	1

Table 1b
Dynamic covariates averaged for 15-min intervals: Vehicle speed (in mph).

Date and time	**1/1/2023 00:00**	**1/1/2023 00:15**	**1/1/2023 00:30**	**1/1/2023 01:00**
Speed	49	51	50	49
Date and time	**1/1/2023 01:15**	**1/1/2023 01:30**	**1/1/2023 01:45**	**1/1/2023 02:00**
Speed	47	50	48	49
Date and time	**1/1/2023 02:15**	**1/1/2023 02:30**	**1/1/2023 02:45**	**1/1/2023 03:00**
Speed	51	50	50	51
Date and time	**1/1/2023 03:15**	**1/1/2023 03:30**	**1/1/2023 03:45**	**1/1/2023 04:00**
Speed	49	48	47	48

characteristics.

For discretization, let us choose $e = 1$ h and $dt = 0.25$ h. Therefore, the number of time intervals in an epoch, denoted by $C = 4$, each identified by the index $i = (1, 2, 3, 4)$. After discretization, the forecasting epochs are created as shown in Table 1(c). Each epoch consists of four 15-minute intervals, and an additional $C + 1th$ column called "*Next epoch*" is added, indicating whether the next crash occurred in the current or future epoch (0 if in the current epoch, 1 if in future epochs). Based on the table, we can express the time elapsed since the previous crash using the equation $t_{e,i} = (e-1)Cdt + (i-1)dt$. For example, the time between crashes A1 and A2 can be determined as $t_{3,2} = (3-1)1 + (2-1)0.25 = 2.25$ hours. As shown in the table, the dynamic covariate *Speed* varies across different time periods. The static covariate *Terrain*, in this example, does not repeat across the time intervals of a crash. However, to account for the effect of time, the variable is multiplied by $t_{e,i}$. For instance, the *Terrain* variable for the first time-interval is 0.25 multiplied by 1, and for the second time interval, it is 0.5 multiplied by 1, and so on. Therefore, all variables vary across epochs and time-intervals. The final data obtained after the creation of forecasting epochs takes the form of panel data with repeated observations for each crash corresponding to the forecasting epochs.

A few observations can be made from Table 1(c), particularly regarding the increase in data size after the creation of forecasting epochs. The final data size is influenced by three factors. The first factor is the size of the original crash data. The more crashes are observed, the larger the data size will be after creating forecasting epochs. The second factor is the choice of discretization. When a smaller time discretization is chosen, more detailed information regarding traffic flow can be obtained. However, this also leads to a considerable increase in data size. The third factor is the distribution of inter-crash duration. If the inter-crash durations are longer, more forecasting epochs will be created, resulting in a larger data size. Considering these factors, implementing a model for a wide geographical area with small discretization can become computationally demanding. Even a slight reduction in time discretization significantly increases computational complexity. To reduce computational complexity, it is suggested to use a smaller sample of the expanded data drawn at the epoch level for model training (Thapa et al., 2022).

Now, based on the example provided, the latent propensity function for crash severities, k observed at a particular time interval, i can be represented as a function of time since crash, static, and dynamic covariates using the utility function, $U_{k,i}$ in Eq. (1).

$$U_{k,i} = \beta_t t_{e,i} + \rho' X_{e,i} \tag{1}$$

In Eq. (1), the coefficient β_t represents the impact of duration on crash severity. The vector of covariates, $X_{e,i}$, captures the effect of covariates, with its values varying across epochs and time intervals. The corresponding vector of coefficients is denoted by ρ'. Similarly, if we assume that the latent propensity function for crash occurrences at any time interval, i consist of only an intercept term, the utility equations for each alternative can be formulated using Eq. (2).

$$V_{e,i} = \beta_i \tag{2}$$

It is worth noting here that as shown in Table 1(c), occurrence of a crash at a specific time interval is dependent on crashes not occurring on previous time intervals. This conditional probability of observing a crash in a particular time interval within an epoch can be expressed using a random variable T_s as follows.

$$P\left(T_s = t_{e,i} \middle| T_s > (e-1)Cdt\right) = \frac{exp\left(V_{e,i}\right)}{\sum_{c=1}^{C} exp\left(V_{e,c}\right) + exp\left(V_{s,e,c+1}\right)} \tag{3}$$

The resulting unconditional probability of a crash at any time interval can be obtained by multiplying the conditional probability in Eq. (3) with the cumulative product of all probabilities for the C + 1*th* intervals preceding the epoch e as represented by Eq. (4).

$$P\left(T_s = t_{e,i}\right) = \frac{exp\left(V_{e,i}\right)}{\sum_{c=1}^{C} exp\left(V_{e,c}\right) + exp\left(V_{e,C+1}\right)} \times \prod_{e^*=1}^{e-1} \frac{exp\left(V_{e^*,C+1}\right)}{\sum_{c=1}^{C} exp\left(V_{e^*,c}\right) + exp\left(V_{e^*,C+1}\right)} \tag{4}$$

4.2. Nested logit model

As discussed prior, the crash outcomes in the example are characterized by: (i) occurrence of crashes or the time interval when a crash happens, and (ii) the severity of the crash that happened at a certain interval. These outcomes can be effectively modeled using a two-level nested logit model, as depicted in Fig. 1. In this model, the time intervals, i and an additional alternative $(C + 1)$ serve as nodes representing the upper-level choice alternatives, while the crash severities correspond to the lower-level alternatives. It is important to note that the crash severities at each time interval are conditional upon the occurrence of a crash within that interval. For simplicity, assume the severity levels are comprised of two categories, denoted by k = (F/I, PDO), where F/I represents Fatal or Injury crashes, and PDO represents

Table 1c
Final crash data after creating forecasting epochs.

s	ID	Time to crash (hr)	Epoch	15-min intervals				Next epoch	Speed (mph)				Severity	Terrain(Flat=1, Rolling=0)			
				1	2	3	4		1	2	3	4		1	2	3	4
1	A1	2.5	1	0	0	0	0	1	49	51	50	49	Fatal	0.25	0.5	0.75	1
1	A1	2.5	2	0	0	0	0	1	47	50	48	49	Fatal	1.25	1.50	1.75	2
1	A1	2.5	3	0	1	0	0	0	51	50	50	51	Fatal	2.25	2.50	2.75	3
1	A2	0.5	1	0	1	0	0	0	49	48	47	48	PDO	0.25	0.5	0.75	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

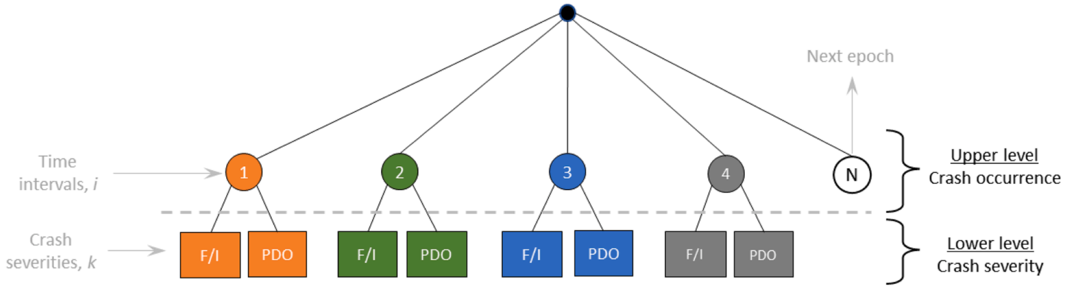

Fig. 1. Two-level nested structure of crash occurrence and severity.

Property Damage Only crashes. The conditional choice probability of the lower-level alternatives, *k* given the upper-level alternatives, *i* can be expressed as follows.

$$P_k = P(k|i)^*P(i) \tag{5}$$

$$P(k|i) = \frac{\exp\left(U_{k,i}/\theta_i\right)}{\sum_k \exp\left(U_{k,i}/\theta_i\right)} \tag{6}$$

$$P(i) = \frac{\exp\left(V_{e,i} + \Gamma_i \times \theta_i\right)}{\sum_i \exp\left(V_{e,i} + \Gamma_i \times \theta_i\right) + \exp\left(V_{e,Next\ epoch}\right)} \tag{7}$$

$$\Gamma_i = \log\left[\sum_k \exp\left(U_{k,i}/\theta_i\right)\right] \tag{8}$$

The parameter θ_i in Eqs. (6), (7), and (8) represents the logsum parameter or nesting coefficient, which captures the underlying correlations for alternatives within a nest. Γ_i in Eq. (8) is the inclusive value for nodes in the upper level. However, the C + 1th alternative, *Next epoch*, lacks the logsum parameter due to its degenerate branch. Consequently, the probability of this alternative can be determined using the following equation.

$$P(Next\ epoch) = \frac{exp\left(V_{Next\ epoch}\right)}{\sum_i exp\left(V_i + \Gamma_i \times \theta_i\right) + exp\left(V_{Next\ epoch}\right)} \tag{9}$$

The probability of F/I crashes in Eq. (6) can be obtained by substituting the value of $U_{k,i}$ from Eq. (1) assuming PDO crashes as the reference case. Similarly, Eq. (7) gives the probability of upper-level alternatives, which is equivalent to Eq. (3) and can be rewritten using Eq. (10).

$$P(i) = \frac{exp\left(V_{e,i} + \Gamma_i \times \theta_i\right)}{\sum_i exp\left(V_{e,i} + \Gamma_i \times \theta_i\right) + exp\left(V_{e,Next\ epoch}\right)} \times \prod_{e^*=1}^{e-1} \frac{exp\left(V_{e^*,C+1}\right)}{\sum_i exp\left(V_{e^*,i} + \Gamma_i \times \theta_i\right) + exp\left(V_{e^*,Next\ epoch}\right)} \tag{10}$$

Assuming each row in the crash data after creation of forecasting epochs is represented using the superscript *n,* the log-likelihood function for the two-level nested logit model can be expressed as the sum of two components using Eq. (11). The first and second components of the equation are associated with the lower and upper-level alternatives, respectively (Brownstone and Small, 1989). The parameters in the two-level nested logit model is estimated by maximizing this equation.

$$L = \sum_n logP^n\left(k^n|i^n\right) + \sum_n logP^n\left(i^n\right) \tag{11}$$

4.3. 4.3. Estimation of the nested logit model

There are several methods available for estimating parameters in nested logit models, with sequential estimation and simultaneous estimation being the most cited approaches. In sequential estimation, the first component of the log-likelihood function (Eq. (11)) is maximized to estimate the parameters in the lower-level. This step provides estimates of the coefficients scaled by their respective nesting parameter θ_i. To simplify the process, the nesting parameters can be assumed to be constant for all nodes, represented as $\theta_i = \theta$. In the next step, inclusive values are calculated for each node using the scaled estimates obtained from the lower level. These inclusive values are then used in the second component of the log-likelihood function to maximize and obtain the values of θ and intercepts β_i for the upper level. It is important to note that while sequential estimation allows for the maximization and estimation of parameters in a stepwise manner, the estimates obtained are not consistent because the scaled parameters from the lower level are substituted to find parameters in the upper level. An alternative approach is simultaneous estimation, where parameters in both levels are estimated simultaneously using a non-linear maximization algorithm. This method is more rigorous compared to sequential estimation, and the estimates obtained are consistent.

5. Data

5.1. Data source and preparation

The estimation and validation of the two-level nested logit model were carried out using data gathered from two main sources. First, historical crash data for the year 2019 was obtained from the Enhanced Tennessee Roadway Information Management System (ETRIMS). This dataset provided information on various crash characteristics such as the date, time, severity, and coordinates of the crash location, as well as details on static covariates such as highway geometry, weather conditions, lighting conditions, land use, and terrain characteristics. The dynamic covariates for the study, namely traffic flow and speed, were obtained from the Radar Data System (RDS) stations located along the highway segments from which the historical crash data was collected. Since our study aimed to implement a practical time discretization with 15-minute intervals, the RDS data was collected specifically for these 15-minute intervals. To match the RDS data with the corresponding crashes, a geospatial mapping approach was employed, aligning the RDS stations with their respective highway segments.

It is important to note that RDS coverage in Tennessee is limited to its major cities, including Memphis, Nashville, Chattanooga, and Knoxville. Therefore, for the purposes of this study, the segments of interstates within the city limits of Memphis and Chattanooga were considered. Specifically, the selected segments included I-40 and I-55 in Memphis, and I-24 and I-75 in Chattanooga.

For this study, the interstates were divided into segments based on four criteria including the direction of traffic, number of lanes, posted speed limit, and terrain type. The segmentation details of the interstates are provided in Table 2. The table includes information on the total number of segments, their lengths in both directions, and the frequency of crashes observed within each segment. In total, the dataset consisted of 2,375 crashes. Table 3 presents a breakdown of the crash frequencies based on various categorical variables. Additionally, the table includes descriptive statistics for the continuous variables in the dataset. The

Table 2
Summary of interstate segmentation.

Interstate	City	Number of segments	Length (mi)	Number of crashes
I-40	Memphis	146	21.51	905
I-55		94	12.28	268
I-24	Chattanooga	48	14.71	675
I-75		70	13.29	527
Total		358	61.79	2,375

table provides a comprehensive overview of the data, highlighting the distribution of crashes across different segments and variable categories.

In this study, the 15-minute traffic volumes were scaled to a range between 0 (minimum value) and 1 (maximum value). This scaling process was applied to avoid the potential influence of larger volumes on the model training process. The duration between crashes exhibited a right-skewed distribution, as indicated by the mean of 516.67 h (about 3 weeks) being greater than the median of 230.46 h (about 1 and a half weeks). This suggests that there is a longer average time period between crashes, with occasional instances of shorter durations. A visual representation of the distribution of inter-crash duration for the four interstates is presented by a density plot in Fig. 2. The density plot provides a graphical representation of the distribution, highlighting the shape and spread of the duration between crashes for each interstate.

From the plot, it can be observed that I-40 has the highest peak, indicating a higher concentration of crashes compared to the other interstates. Furthermore, the density plot reveals that the distribution of crashes on I-40 is less spread out compared to the other interstates. This means that the duration between crashes on I-40 is shorter, indicating a higher frequency of crashes occurring within a shorter period. In terms of increasing spread, the interstates can be ranked as follows: I-40, I-55, I-24, and I-75. This implies that the duration between crashes is longer and more spread out on I-75 compared to the other interstates.

5.2. Data sampling

In this study, the models were calibrated using training data and evaluated on testing data. The process of creating training and testing data involved splitting the historical crash data in a 9:1 ratio, where 90 % of the data was allocated for training and the remaining 10 % for testing. To create forecasting epochs, both the training and testing crashes were expanded. The training data was further sampled at 5 % increments up to 25 % to investigate whether any sample size below 25 % would provide accurate parameter estimates. Thus, the samples used for parameter estimation were 5 %, 10 %, 15 %, and 25 % of the training data. This sampling approach is called epoch level sampling (Thapa

Table 3
Descriptive statistics of crash characteristics.

Categorical variables	Frequency of crashes				Relative abundance		
Time of day							
Early morning (6 a.m. to 9 a.m.)	447				18.82 %		
Late morning (9 a.m. to 12p.m.)	262				11.03 %		
Early afternoon (12p.m. to 3p.m.)	351				14.78 %		
Late afternoon (3p.m. to 6p.m.)	586				24.67 %		
Evening (6p.m. to 12 a.m.)	392				16.51 %		
Night (12 a.m. to 6 a.m.)	337				14.19 %		
Weather condition							
Clear	1,733				72.97 %		
Others (Cloudy, rain, fog, or snow)	642				27.03 %		
Lighting condition							
Daylight	1,612				67.87 %		
Dark lighted	463				19.49 %		
Dark, not lighted	300				12.63 %		
Illumination type							
Illuminated	1,780				74.95 %		
Not illuminated	595				25.05 %		
Terrain							
Flat	715				30.11 %		
Rolling	1,660				69.89 %		
Land use							
Commercial	1,187				49.98 %		
Rural	765				32.21 %		
Mixed	423				17.81 %		
Crash severities							
Fatal or injury	451				18.99 %		
Property Damage Only	1,924				81.01 %		
Continuous variables	Min	Q1	Median	Q3	Max	Mean	SD
Traffic flow characteristics							
Speed (mph)	1.00	59.06	63.47	66.96	91.00	61.41	10.46
Volume (scaled between 0-minimum, and 1-maximum)	0.0002	0.12	0.28	0.46	1.00	0.31	0.22
Highway geometry							
Number of lanes (both directions)	3	6	8	8	12	7.18	1.78
Inter-crash duration (hours)	0.00	68.05	230.46	627.17	7683.03	516.67	783.18

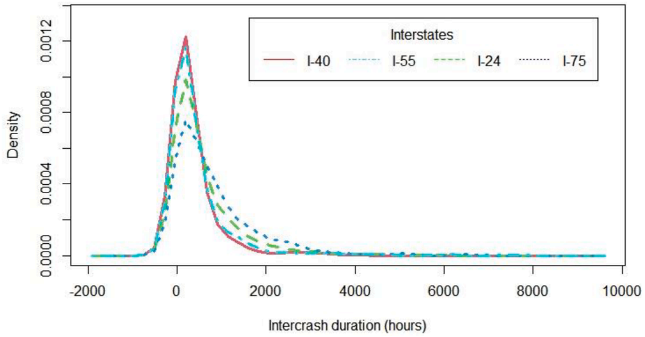

Fig. 2. Distribution of inter-crash duration for the interstates.

et al., 2022). The sampled training data, along with the complete training data, were used to estimate the parameters for the models. For comparison purposes, the parameter estimates obtained from the complete training data (100 % training data) were considered as the "true" estimates.

To evaluate the performance of the trained models, the predicted log-likelihood values were calculated on the training data. In this context, predicted log-likelihood provided a basis for comparing how well the models captured the characteristics of the training data.

6. Results

All model computations, including estimation and validation, in this study were conducted using R version 4.2.3 on a computer equipped with Intel Core i7-11,700 K processor and 16 GB of memory. We initially estimated the model parameters using the complete training data, employing both simultaneous and sequential estimation techniques. The objective of estimating with the complete training data was to obtain "true" parameter estimates and compare the results obtained from different estimation techniques. The estimation results are presented in Table 4. In Table 4, the first column displays the variable groups in the model, along with the corresponding variable categories considered as the base in the models. The second column lists the variables included in the model. The estimation results are then presented, showing the parameter estimates and their respective *t*-statistics for both simultaneous and sequential estimation. The parameter estimates obtained from both estimation methods are comparable, indicating consistency in the results. Additionally, the average values of predicted log-likelihood are also similar between the two methods. When considering estimation complexity, which refers to the time taken for the model to converge from a null model, it was found that sequential estimation offers a considerable advantage. Specifically, using simultaneous estimation, the model took 51.09 h (about 2 days) to converge, which was approximately six times the time taken by sequential estimation, which was 8.63 h. Therefore, sequential estimation may provide consistent estimates with a significant reduction in computational complexity.

The parameters obtained from simultaneous estimation, as shown in the table, can be utilized to express the propensity function for F/I crashes in any time interval using the following utility equation.

$U_{F/I,e,i} = V_i - 6.46 \times t_{e,i} + 1.96 \times \textit{Early morning} + 3.50 \times \textit{Late morning} \cdots - 1.27 \times \textit{Volume}$

For example, the utility equation for the first time-interval can be expressed as follows.

$U_{F/I,e,1} = -8.71 - 6.46 \times t_{e,1} + 1.96 \times \textit{Early morning} + 3.50 \times \textit{Late morning} \cdots - 1.27 \times \textit{Volume}$

Table 4
Results from estimation of model using complete training data.

Variable groups	Variables	Simultaneous estimation		Sequential estimation	
		Estimate	*t*-stat	Estimate	*t*-stat
Upper level					
Duration dynamics	Time since previous crash	−6.46	−22.60	−7.22	−91.82
Time of day (Evening 6p. m. to 12 a.m., Night 12 a.m. to 6 a.m.)	Early morning (6 a.m. to 9 a. m.)	1.96	20.75	2.19	44.87
	Late morning (9 a.m. to 12p. m.)	3.50	22.27	3.91	70.82
	Early afternoon (12p.m. to 3p.m.)	3.48	22.39	3.89	75.27
	Late afternoon (3p.m. to 6p. m.)	2.47	21.74	2.76	58.19
Weather conditions (Others)	Clear	1.92	22.22	2.15	72.37
Lighting condition Dark, not lighted)	Daytime	0.52	9.79	0.58	10.80
	Dark lighted	3.47	22.11	3.88	67.84
Illumination type (Not illuminated)	Illuminated	−0.88	−19.04	−0.99	−32.94
Terrain type (Rolling)	Flat	0.28	9.58	0.32	10.50
Land use (Mixed)	Commercial	−1.06	−18.82	−1.18	−31.78
	Rural	−2.01	−21.60	−2.24	−56.49
Highway geometry	Number of lanes	0.88	22.23	0.98	72.09
Traffic flow characteristics	Speed	−0.08	−23.20	−0.09	−506.52
	Volume	−1.27	−21.70	−1.42	−55.39
Lower level					
Intercepts (Next epoch)	First 15-min interval	−8.71	−131.18	−8.85	−128.09
	Second 15-min interval	−8.74	−130.92	−8.88	−127.94
	Third 15-min interval	−8.77	−130.53	−8.90	−127.68
	Fourth 15-min interval	−8.71	−131.35	−8.85	−128.26
Nesting coefficient	θ	4.36	23.22	4.87	24.80
Goodness of fit					
Number of observations (Training)		1,103,104			
Average initial LL		−213.98			
Average LL at convergence		−2.052			
Number of observations (Testing)		140,591			
Predicted LL		−1.879			
Estimation complexity	Time (hours)	51.09		8.63	

The analysis reveals interesting findings regarding the factors influencing F/I crashes. The duration dynamics coefficient indicates that as the duration between crashes increases, the likelihood of F/I crashes decreases. Moreover, F/I crashes are more likely to occur between 9 am and 3 pm. Clear weather conditions are associated with a higher likelihood of F/I crashes compared to adverse weather conditions such as clouds, rain, fog, or snow. Dark lighted conditions result in more severe crashes, followed by daytime and dark unlighted conditions. Non-illuminated locations are more prone to F/I crashes compared to illuminated locations. Additionally, locations with flat terrain have a higher likelihood of F/I crashes compared to those with rolling terrain. Higher traffic volume leads to a decrease in F/I crashes, due to stop-andgo conditions during congested conditions. Similarly, higher speeds are associated with a lower likelihood of F/I crashes, although the effect size is small. The coefficients for the upper-level nodes, V_i, have similar magnitudes. The nesting parameter has a value of 4.36, indicating cross nesting of alternatives. It is worth nothing that the training data increased significantly after the creation of forecasting epochs, with the original 2,137 crashes expanding to 1,103,104 observations.

Next, we proceeded to estimate parameters using sampled data to explore the tradeoff between model performance and estimation complexity. The results of this estimation can be found in Table 5, which presents the obtained parameter values along with their respective *t*-statistics. Upon visual inspection, it is apparent that the parameter values obtained using the 25 % sample are much closer to the true values compared to the 5 % sample. This finding aligns with a previous study conducted by Thapa et al. (2022). However, it is also crucial to investigate the impact of sample size within the range of 5 % to 25 % to determine the sample that offers the optimal balance between model performance and estimation complexity. To address this, we estimated parameters at 5 % increments, ranging from 5 % to 25 %. Fig. 3 presents a graphical representation of estimation complexity and predicted log-likelihood for the various samples. Notably, the figure indicates a significant improvement in prediction performance beyond the 10 % sample. Furthermore, the models demonstrate similar performance for the 15 %, 20 %, and 25 % samples.

As expected, estimation complexity increases linearly with the sample size. For instance, the model required 2.48 h to train on the 5 % sample, while it took approximately 20 times or 51.09 h (about 2 days) for the full 100 % dataset. Based on the findings depicted in the figure, it is evident that using a 15 % sample can yield comparable estimates and predictive performance to the 25 % sample, while reducing the estimation complexity to 60 % of that offered by the 25 % sample. This suggests that the 15 % sample size strikes a favorable balance between model performance and estimation complexity.

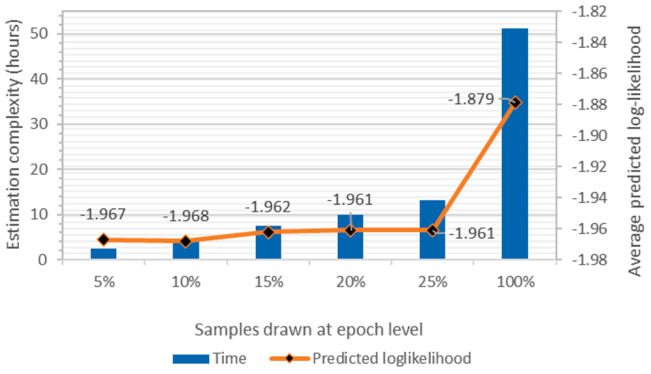

Fig. 3. Improvement in model performance with increase in data size/estimation complexity.

6.1. *Effect of sampling on coefficients*

Based on the parameter estimates, it is evident that certain predictor variables are particularly sensitive to sampling. A notable example is the *Volume* variable, where the coefficients exhibit significant differences between the sampled data and the complete data (refer to Fig. 4). This discrepancy can be attributed to the sampling approach and the scaling of traffic volumes. Since the volumes are scaled between 0 and 1, random sampling can lead to the exclusion of several observations, resulting in considerable variations in the parameter estimates for this variable. On the other hand, coefficients for the *Speed* variable demonstrate consistency. This consistency may be attributed to the fact that the values of the variable do not fluctuate significantly, as indicated by its

Table 5
Results from simultaneous model estimation using samples drawn at the epoch level.

Variable groups	Variables	5 % sample	10 % sample	15 % sample	20 % sample	25 % sample
Upper level						
Duration dynamics	Time since previous crash	−5.05 (−4.21)	−6.55 (−7.15)	−6.60 (−8.65)	−6.99 (−10.10)	−6.94 (−11.36)
Time of day	Early morning (6 a.m. to 9 a.m.)	1.58 (3.93)	1.81 (6.39)	1.77 (7.73)	1.78 (8.97)	1.84 (10.13)
(Evening 6p.m. to 12 a.m.,	Late morning (9 a.m. to 12p.m.)	2.56 (4.13)	3.20 (6.97)	3.43 (8.51)	3.40 (9.87)	3.45 (11.03)
Night 12 a.m. to 6 a.m.)	Early afternoon (12p.m. to 3p.m.)	2.46 (4.14)	3.18 (7.00)	3.60 (8.58)	3.76 (10.03)	3.70 (11.20)
	Late afternoon (3p.m. to 6p.m.)	1.71 (4.02)	2.40 (6.83)	2.48 (8.32)	2.50 (9.66)	2.52 (10.81)
Weather conditions (Others)	Clear	1.43 (4.12)	1.87 (7.00)	1.97 (8.51)	1.98 (9.91)	1.99 (11.08)
Lighting condition	Daytime	0.49 (2.27)	0.42 (2.50)	0.31 (2.37)	0.39 (3.33)	0.36 (3.53)
(Dark, not lighted)	Dark lighted	2.97 (4.15)	3.31 (6.92)	3.36 (8.41)	3.46 (9.81)	3.52 (11.06)
Illumination type	Illuminated	−0.66 (−3.60)	−0.90 (−5.98)	−0.80 (−6.99)	−0.72 (−7.74)	−0.75 (−8.81)
(Not illuminated)						
Terrain type	Flat	0.49 (3.22)	0.85 (5.89)	0.80 (7.05)	0.70 (7.65)	0.63 (8.15)
(Rolling)						
Land use	Commercial	−1.22 (−3.94)	−1.28 (−6.27)	−1.16 (−7.40)	−0.98 (−8.05)	−0.99 (−9.01)
(Mixed)	Rural	−2.03 (−4.14)	−2.15 (−6.87)	−2.08 (−8.28)	−1.98 (−9.54)	−1.96 (−10.63)
Highway geometry	Number of lanes	0.77 (4.19)	0.96 (7.09)	0.86 (8.48)	0.89 (9.89)	0.89 (11.11)
Traffic flow	Speed	−0.07 (−4.29)	−0.08 (−7.34)	−0.08 (−8.86)	−0.08 (−10.32)	−0.08 (−11.55)
characteristics	Volume	−14.94 (−4.26)	−18.77 (−7.26)	−17.74 (−8.77)	−17.98 (−10.23)	−18.02 (−11.47)
Lower level						
Intercepts	First 15-min interval	−8.68 (−28.18)	−8.94 (−39.92)	−8.89 (−49.03)	−8.90 (−56.71)	−8.85 (−63.84)
(Next epoch)	Second 15-min interval	−9.01 (−27.42)	−8.61 (−42.27)	−8.56 (−51.90)	−8.59 (−59.88)	−8.62 (−66.90)
	Third 15-min interval	−8.56 (−28.57)	−8.83 (−40.44)	−8.77 (−49.69)	−8.75 (−57.60)	−8.83 (−64.04)
	Fourth 15-min interval	−8.60 (−28.37)	−8.73 (−40.63)	−8.77 (−49.53)	−8.80 (−57.19)	−8.70 (−64.55)
Nesting coefficient	θ	3.68 (4.30)	4.49 (7.35)	4.42 (8.87)	4.46 (10.34)	4.45 (11.58)
Goodness of fit						
Number of observations (Training)		55,062	110,187	165,378	220,662	275,787
Average initial LL		−212.70	−212.87	−212.94	−213.01	−213.08
Average LL at convergence		−2.052	−2.053	−2.054	−2.053	−2.051
Number of observations (Testing)		140,591				
Predicted log-likelihood		−1.967	−1.968	−1.962	−1.961	−1.961
Estimation complexity	Time (hours)	2.48	4.22	7.49	9.86	13.25

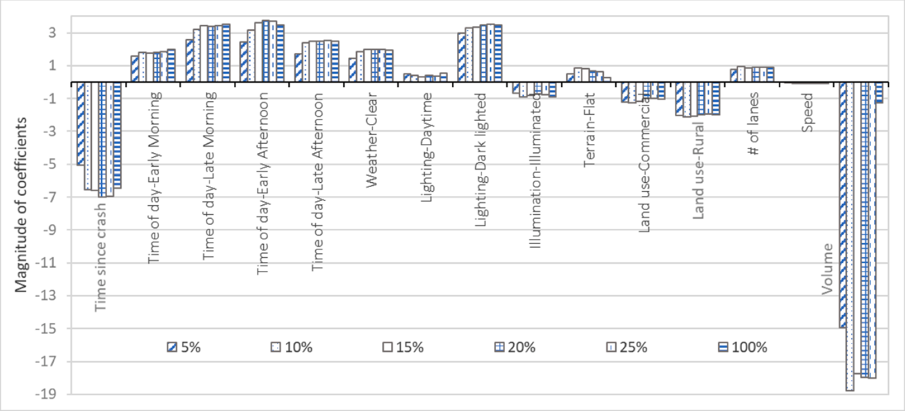

Fig. 4. Coefficient of variables for different training samples.

descriptive statistics, and are less affected by sampling.

Considering these observations, we aim to identify and report variables that are sensitive to sampling. To visualize this, a bar plot in Fig. 4 presents the variable coefficients obtained from the sampled and complete data. From the plot, it can be observed that smaller samples are more likely to overestimate the effect of some variables, for example, *Time since crash*, *Time of day-Early afternoon*, *Terrain-Flat*, *Land Use-Commercial*, and *Volume*. Conversely, variables such as *Time of day-Early Morning* and *Late Morning*, and *Lighting-Daytime* are more likely to be underestimated when smaller samples are used. Overall, these findings emphasize the importance of considering the impact of sampling on parameter estimates, particularly for variables that exhibit sensitivity to sampling.

Considering the impact of sampling, we identified variables that are unlikely to converge toward the true value when small samples are used and those that are more likely to do so. Identification of these variables is crucial from a practical standpoint, especially when analysts and planners seek greater accuracy for specific variables. In the following figures, we present two groups of predictors. The first group consists of variables which are less likely to converge to actual values with small increments in sample size. These variables would require larger samples to achieve more accurate estimation. It is important to recognize the limitations in estimating the coefficients for these variables with smaller sample sizes. The second group comprises variables whose coefficients converge closer to the actual values as the sample size increases. This group includes variables whose coefficients can be obtained with reasonable accuracy, even with small increments in sample size. The findings will be useful in identifying variables that becomes more stable and reliable as the sample size grows. These findings serve as valuable insights for researchers and practitioners, allowing them to prioritize their data collection efforts and allocate resources effectively based on the sensitivity of different predictors to sample size.

The variables which are less likely to converge to true values despite an increase in sample size, ranging from 5 % to 25 %, compared to the full data are *Time since crash*, *Time of day-Early afternoon*, *Weather Condition-Clear*, *Lighting Condition-Daytime*, *Illumination-Illuminated*, and *Volume*. These variables are presented in Fig. 5, indicating the percentage difference of the coefficients from the complete training data. On the other hand, coefficients for *Time of day-Early morning*, *Time of day-Late morning*, *Time of day-Late afternoon*, *Lighting-Dark lighted*, *Terrain-Flat*, *Land Use-Commercial*, *Land Use-Rural*, *Number of lanes*, and *Speed* converge quicker to the actual values as the sample size increases. These variables are displayed in Fig. 6, illustrating the percentage difference compared to the complete training data. These findings highlight the sensitivity of different variables to sample size and provide valuable insights into the accuracy and stability of their coefficient estimates.

7. Validation

The validation of the proposed nested logit model was carried out to assess its predicted capabilities. All validations were conducted using the simultaneous model trained on 15 % data drawn at the epoch level since our analysis suggested that it provided the best tradeoff between accuracy and estimation complexity. As discussed previously, 10 % of the sample was held out for testing. The test sample consisted of 236 crashes, including 39 F/I crashes and 197 PDO crashes. This test sample was used for validation. Similar to the two-step model, validation was conducted to assess predictive abilities for the outcomes considered at the lower and upper levels. These results are discussed in the following subsections.

7.1. *Upper level: crashes at epoch level*

One of the primary objectives of the proposed framework is to

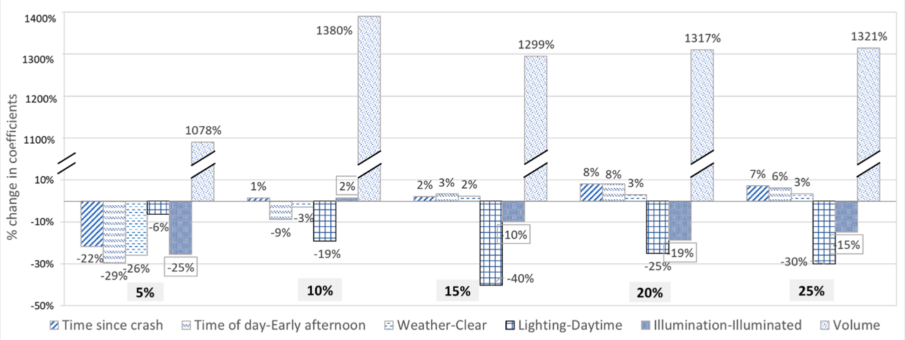

Fig. 5. Variables unlikely to converge to their actual values despite of incremental increase in sample size.

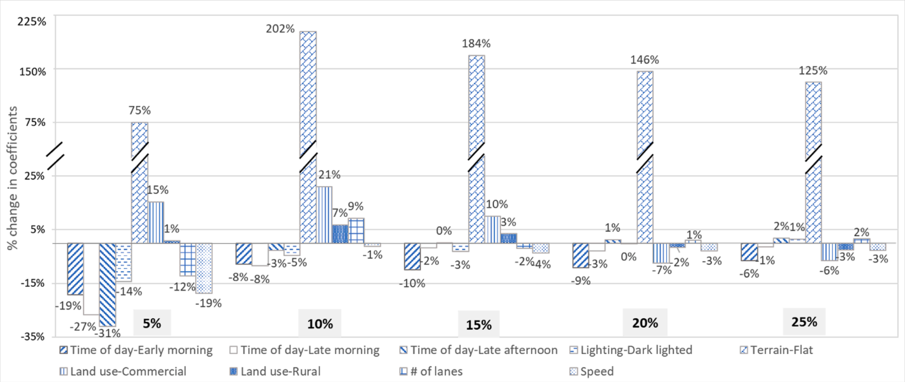

Fig. 6. Variables likely to converge to their actual values with incremental increase in sample size.

predict the occurrence of future crashes. Therefore, it is crucial to evaluate the temporal accuracy of the predicted crashes. To evaluate this, we measured the proximity between the predicted crash epoch and the actual epoch at which crashes were observed, by introducing a metric called Predicted Temporal Proximity (PTP), represented by Eq. (12). This metric quantifies how closely the predicted crash epochs align with the observed epochs.

Furthermore, we also investigated whether the number of epochs impacted the model's performance in terms of PTP. To accomplish this, we calculated the PTP for different subsets of the testing data by excluding crashes with a substantial number of epochs. This was accomplished by creating subsets of the test data to include crashes with fewer than 100 to 1000 epochs, with intervals of 100 epochs. The average values of PTP for these subsets of testing data are depicted in Fig. 7.

$$PTP = \left|\frac{Predicted\ crash\ epoch - Actual\ crash\ epoch}{Actual\ crash\ epoch}\right| \times 100\% \tag{12}$$

It is important to note that, according to the definition of PTP, a smaller value is desired as it indicates that the predicted crash epoch is closer to the observed epoch. The results depicted in Fig. 7 indicate that when there is a substantial number of epochs (i.e., a large inter-crash duration), the value of PTP increases. This suggests that epoch-level prediction is more accurate when the duration between crashes is smaller. In other words, the prediction of crash epochs is more reliable for highway segments that experience crashes more frequently. For example, based on the figure, for crashes with inter-crash durations less than 100 h (approximately 4 days), the predicted crash epoch is within 60 % of the actual epoch, compared to 74 % for durations exceeding 1,000 h.

7.2. *Upper level: crashes in predicted time-intervals*

The accuracy of predicting crash occurrences at specific time intervals can be assessed from two perspectives: i) the accuracy of predicting crashes (true positives), and ii) the accuracy of predicting 'no crashes' (true negatives). Therefore, we relied on the metrics of Specificity and Sensitivity to evaluate the model's predictions. Specificity measures the model's ability to correctly predict 'no crashes' (true negatives) and is defined by Eq. (13). On the other hand, Sensitivity measures the model's ability to correctly predict crashes (true positives) and is defined by Eq. (14). It quantifies the proportion of correctly identified positive cases in relation to the actual positive cases. It quantifies the proportion of correctly identified negative cases in relation to the actual negative cases.

The model's prediction accuracy for crash and severity were evaluated using these metrics. The results are summarized in Table 6 and described as follows.

$$Specificity = \frac{True\ Negatives(TN)}{True\ Negatives(TN) + False\ Positives(FP)} \tag{13}$$

Table 6
Values of Specificity and Sensitivity from the model predictions.

Predictions	TN	TP	FP	FN	Specificity	Sensitivity
Crash occurrence	539	63	173	169	0.76	0.27
Crash severity						
F/I crashes	887	9	20	28	0.97	0.24
PDO crashes	557	41	192	154	0.74	0.21

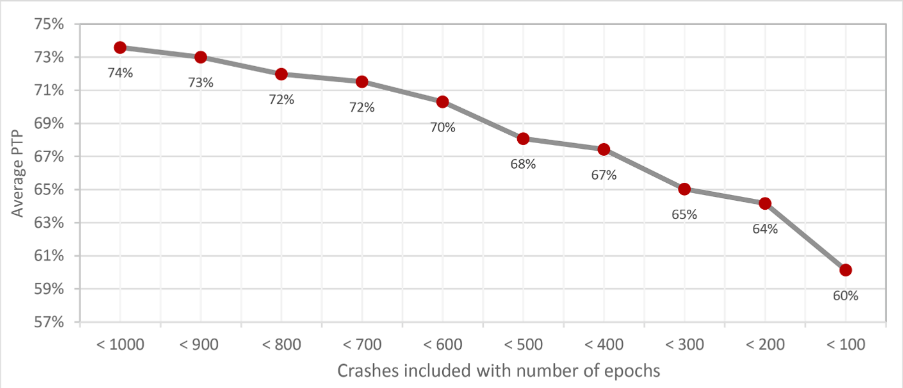

Fig. 7. Average PTP for different subsets of test samples.

$$Sensitivity = \frac{True\ Positives(TP)}{True\ Positives(TP) + False\ Negatives(FN)} \tag{14}$$

The model predictions for the time intervals resulted in the following counts: True Negatives (TN) = 539, True Positives (TP) = 63, False Positives (FP) = 173, and False Negatives (FN) = 169. The Specificity is calculated to be 0.76, indicating a high value. This high value suggests a low rate of false positive predictions. Therefore, the model demonstrates reliability in predicting crashes. In other words, the likelihood of classifying a time interval without a crash as a time interval experiencing a crash is low. On the other hand, the Sensitivity is calculated to be 0.27, indicating a low value. This low value suggests a high rate of false negatives, or in other words, the chances of classifying true crash intervals as having no crash is high.

7.3. Lower level: crash severity for crashes in predicted time-intervals

The Specificity and Sensitivity measures were also utilized to evaluate the model's ability to predict crash severities at each time interval. For F/I crashes, the following results were obtained: TN = 887, TP = 9, FP = 20, FN = 28, resulting in a Specificity of 0.97 and a Sensitivity of 0.24. Similarly, for PDO crashes, the values obtained were TN = 557, TP = 41, FP = 192, and FN = 154, with a Specificity of 0.74 and a Sensitivity of 0.21.

The results indicate that for both severity types, the Specificity values are high. This suggests that the model is capable of reliably predicting both F/I and PDO crashes with a lower chance of false positive predictions. However, it should be noted that the model also exhibits low Sensitivity values, indicating that the model may not always accurately classify the severity types with a high degree of certainty, leading to a higher occurrence of false negative predictions. This outcome is the result of exceptionally higher prevalence of time-intervals without crashes (0 s) in comparison to those with crashes (1 s). Future research can improve upon the model by addressing this imbalance in the frequency of outcomes (e.g., see Morris and Yang, 2021).

8. Conclusion

This study developed a duration-based model to predict crash occurrence and severity using historical crash and traffic flow data from four interstates in Tennessee. The framework involved the reformulation of crash data to create forecasting epochs and time-intervals, which were used to calculate crash and severity likelihoods. The creation of forecasting epochs significantly increased the data size and estimation complexity. Additionally, the adoption of a nested structure further contributed to the complexity of model estimation. To address the computational challenges, we suggested sampling the data at the epoch level to reduce estimation complexity. We aimed to find the optimal sampling strategy by considering the tradeoff between model performance and estimation complexity. After evaluating various samples, we determined that a 15 % sample drawn at the epoch level provided the best balance in reducing data size. Furthermore, we investigated the impact of sampling on the coefficients of predictor variables to identify those most sensitive to changes in sample sizes. Variables such as *Time since crash, Time of day-Early afternoon, Late afternoon, Terrain-Flat, Land Use-Commercial, Number of lanes,* and *Volume* were found to be more likely to be overestimated by smaller samples. Conversely, variables including *Time of day-Early Morning, Late Morning, Lighting-Daytime* and *Dark lighted* were more likely to be underestimated.

When investigating the stability of coefficients for the predictors, it was found that *Time since crash, Time of day-Early afternoon, Weather Condition-Clear, Lighting Condition-Daytime, Illumination-Illuminated,* and *Volume* exhibited a higher degree of instability. Consistent estimation of these coefficients required larger sample sizes. On the other hand, coefficients for *Time of day-Early morning, Late morning, Late afternoon, Lighting-Dark lighted, Terrain-Flat, Land Use-Commercial* and *Rural, Number of lanes,* and *Speed* demonstrated a tendency to converge towards true estimates with incremental increases in sample size. These findings are crucial for obtaining consistent and reliable estimates when utilizing samples for model estimation and clarify the challenges and considerations associated with implementing the duration-based model, including the impact of data sampling on estimation outcomes and the sensitivity of certain variables to changes in sample sizes.

The proposed framework's validation provided satisfactory results. The measure, Predicted Temporal Proximity (PTP), suggests that the model performs better when implemented on segments where crashes are more frequent. For context, the model, trained on a 15 % epoch-level sample, was able to predict crashes within 60 % (i.e., average PTP = 60 %) of the actual epoch for crashes occurring within 100 epochs, or approximately 4 days of each other. On the contrary, the average value of PTP was 74 % for crashes occurring within 1,000 epochs of each other. This finding also sheds light on the practical implications of the model, as it is often impractical to predict crashes too far into the future due to potential changes in traffic, weather, and driving conditions. Similarly, the estimated model displayed a satisfactory value of Specificity, indicating a low rate of false positives. In other words, the model is less likely to falsely predict time intervals without crashes as having experienced crashes. This is particularly important as a reasonable degree of certainty is desired to ensure effective allocation of limited safety resources to critical segments. The value of Sensitivity was comparatively smaller, implying a higher rate of false negatives or missed detections. However, it should also be noted that the frequency of time intervals without crashes is several multiples larger than the frequency of time intervals with crashes (preponderance of 0 s compared to 1 s). Therefore, the low value of Sensitivity is expected in this case.

9. Study limitations and future research

Future research offers opportunities for notable improvements to the proposed model. Firstly, it would be valuable to investigate alternative nesting structures to determine if they provide a better fit, especially considering that the nesting parameter suggests the presence of alternative nests. More complex nesting structures based on distinct categories such as time of day, weather conditions, and other relevant factors could be explored. Additionally, in this study, the upper-level is assumed to be a MNL model without considering the effect of time. Future research should consider addressing this when investigating alternative structures. Secondly, the model estimates could be enhanced by incorporating random effects. Since the reformulated data, after the creation of forecasting epochs, takes the form of panel data with repeated observations for crashes and road segments, accounting for segment and crash-specific heterogeneity could lead to more accurate model estimates. Furthermore, data balancing techniques such as Synthetic Minority Over-sampling Technique can be used to balance the frequency of outcomes and study its impact on model estimates. Finally, alternative estimation techniques leveraging parallel and distributed computing can be implemented to reduce estimation time while still retaining information from complete training dataset. Addressing these limitations would contribute to a more comprehensive understanding of crash prediction and severity estimation and improve the accuracy and applicability of the model in real-world scenarios.

CRediT authorship contribution statement

Diwas Thapa: Conceptualization, Methodology, Formal analysis, Software, Validation, Writing – original draft. **Sabyasachee Mishra:** Conceptualization, Supervision, Project administration. **Nagendra R. Velaga:** Conceptualization, Writing – review & editing. **Gopal R. Patil:** Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

This research was partially supported by Fulbright Fellowship to second author at Indian Institute of Technology (IIT) Bombay, National Science Foundation award # 2222699 and the Center for Transportation Innovations in Education and Research (C-TIER) at the University of Memphis. Any findings and opinions expressed in this paper are those of the authors and do not necessarily reflect the view of the aforementioned agencies.

Appendix

Stepwise process for the application of the crash prediction framework.

Step 1: Collect roadway inventory [Roadway characteristics]
Step 1.1: Select study area and road segments.
Step 1.2: Segment roadways based on attributes, e.g., speed limit, number of lanes, highway terrain, and travel direction.
Step 1.3: Extract highway characteristics for all segments.

Step 2: Extract and merge historical crash data [Roadway characteristics + Crash data]
Step 2.1: Select crashes for the study period on the road segments being studied.
Step 2.2: Extract available crash attributes, e.g., date and time, GPS coordinates, direction of travel, injury severity with two levels.
Step 2.3: Merge crash attributes with highway characteristics obtained from Step 1.3 based on GPS proximity and travel direction.

Step 3: Discretization and creation of forecasting epochs
Step 3.1: Select temporal resolution for discretization: Values of e and dt.
Step 3.2: Discretize time interval between crashes in the same segment (data obtained from Step 2.3) and create forecasting epochs.

Step 4: Extract and merge RDS data [Roadway characteristics + Crash data + RDS data]
Step 4.1: Identify RDS stations on the roadway segments.
Step 4.2: Extract data from RDS stations, e.g., GPS coordinates, travel direction, traffic flow data at dt intervals along with date and time.
Step 4.3: Merge reformulated data obtained after creation of forecasting epochs in Step 3.2 with RDS data based on GPS coordinates, date and time, and travel direction.

Step 5: Model estimation
Step 5: Estimate crash probabilities as a two-level nested logit model using data obtained from Step 4.3.

References

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M.F., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transp. Res. Rec. 1897 (1), 88–95. https://doi.org/10.3141/1897-12.

Afghari, A.P., Haque, M.M., Washington, S., 2020. Applying a joint model of crash count and crash severity to identify road segments with high risk of fatal and serious injury crashes. Accid. Anal. Prev. 144, 105615 https://doi.org/10.1016/j.aap.2020.105615.

Ahmed, S.S., Cohen, J., Anastasopoulos, P.C., 2021. A correlated random parameters with heterogeneity in means approach of deer-vehicle collisions and resulting injury-severities. Anal. Meth. Accid. Res. 30, 100160 https://doi.org/10.1016/j.amar.2021.100160.

Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Anal. Meth. Accid. Res. 9, 1–15. https://doi.org/10.1016/j.amar.2015.11.002.

Beshah, T., Ejigu, D., Abraham, A., Snasel, V., Kromer, P., 2011. Pattern recognition and knowledge discovery from road traffic accident data in Ethiopia: Implications for improving road safety. World Cong. Inf. Commun. Technol. 2011, 1241–1246. https://doi.org/10.1109/WICT.2011.6141426.

Brownstone, D., Small, K.A., 1989. Efficient Estimation of Nested Logit models. J. Bus. Econ. Stat. 7 (1), 67–74. https://doi.org/10.1080/07350015.1989.10509714.

Cerwick, D.M., Gkritza, K., Shaheed, M.S., Hans, Z., 2014. A comparison of the mixed logit and latent class methods for crash severity analysis. Anal. Meth. Accid. Res. 3–4, 11–27. https://doi.org/10.1016/j.amar.2014.09.002.

Chen, C., Zhang, G., Qian, Z., Tarefder, R.A., Tian, Z., 2016. Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accid. Anal. Prev. 90, 128–139. https://doi.org/10.1016/j.aap.2016.02.011.

Cheng, W., Gill, G.S., Dasu, R., Xie, M., Jia, X., Zhou, J., 2017. Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. Accid. Anal. Prev. 99, 330–341. https://doi.org/10.1016/j.aap.2016.11.022.

Dissanayake, S., Lu, J., 2002. Analysis of severity of young driver crashes: sequential binary logistic regression modeling. Transp. Res. Rec. 1784 (1), 108–114. https://doi.org/10.3141/1784-14.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. Accid. Anal. Prev. 70, 320–329. https://doi.org/10.1016/j.aap.2014.04.018.

Fountas, G., Anastasopoulos, P.C., 2017. A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities. Anal. Meth. Accid. Res. 15, 1–16. https://doi.org/10.1016/j.amar.2017.03.002.

Golob, T.F., Recker, W.W., Leonard, J.D., 1987. An analysis of the severity and incident duration of truck-involved freeway accidents. Accid. Anal. Prev. 19 (5), 375–395. https://doi.org/10.1016/0001-4575(87)90023-6.

Harmon, T., Bahar, G., Gross, F., 2018. Crash Costs for Highway Safety Analysis.

Hossain, M., Abdel-Aty, M., Quddus, M.A., Muromachi, Y., Sadeek, S.N., 2019. Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. Accid. Anal. Prev. 124, 66–84. https://doi.org/10.1016/j.aap.2018.12.022.

Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accid. Anal. Prev. 45, 373–381. https://doi.org/10.1016/j.aap.2011.08.004.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prev. 108, 27–36. https://doi.org/10.1016/j.aap.2017.08.008.

Jonathan, A.-V., Wu (Ken), K.-F., Donnell, E.T., 2016. A multivariate spatial crash frequency model for identifying sites with promise based on crash types. Accid. Anal. Prev. 87, 8–16. https://doi.org/10.1016/j.aap.2015.11.006.

Jovanis, P.P., Chang, H.L., 1989. Disaggregate model of highway accident occurrence using survival theory. Accid. Anal. Prev. 21 (5), 445–458. https://doi.org/10.1016/0001-4575(89)90005-5.

Jung, S., Qin, X., Noyce, D.A., 2010. Rainfall effect on single-vehicle crash severities using polychotomous response models. Accid. Anal. Prev. 42 (1), 213–224. https://doi.org/10.1016/j.aap.2009.07.020.

Lee, J.-T., Fazio, J., 2005. Influential factors in freeway crash response and clearance times by emergency management services in peak periods. Traffic Inj. Prev. 6 (4), 331–339. https://doi.org/10.1080/15389580500255773.

Lee, J., Yoon, T., Kwon, S., Lee, J., 2019. Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study. Appl. Sci. 10 (1), 129. https://doi.org/10.3390/app10010129.

Li, P., Abdel-Aty, M., Yuan, J., 2020. Real-time crash risk prediction on arterials based on LSTM-CNN. Accid. Anal. Prev. 135, 105371 https://doi.org/10.1016/j.aap.2019.105371.

Ma, J., Kockelman, K.M., 2006. Poisson regression for models of injury count, by severity. Transp. Res. Rec. 1950, 24–34.

Mannering, F., Bhat, C.R., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Anal. Meth. Accid. Res. 25, 100113 https://doi.org/10.1016/j.amar.2020.100113.

Morris, C., Yang, J.J., 2021. Effectiveness of resampling methods in coping with imbalanced crash data: Crash type analysis and predictive modeling. Accid. Anal. Prev. 159, 106240 https://doi.org/10.1016/j.aap.2021.106240.

Osman, M., Mishra, S., Paleti, R., 2018a. Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. Accid. Anal. Prev. 118, 289–300. https://doi.org/10.1016/j.aap.2018.05.004.

Osman, M., Paleti, R., Mishra, S., 2018b. Analysis of passenger-car crash injury severity in different work zone configurations. Accid. Anal. Prevent. 111 (May 2017), 161–172. https://doi.org/10.1016/j.aap.2017.11.026.

Osman, M., Mishra, S., Paleti, R., Golias, M., 2019. Impacts of work zone component areas on driver injury severity. J. Transp. Eng., Part A: Syst. 145 (8), 04019032. https://doi.org/10.1061/jtepbs.0000253.

Ospina-Mateus, H., Quintana Jiménez, L.A., Lopez-Valdes, F.J., Berrio Garcia, S., Barrero, L.H., Sana, S.S., 2021. Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists. J. Ambient Intell. Hum. Comput. 12 (11), 10051–10072. https://doi.org/10.1007/s12652-020-02759-5.

Park, E.S., Lord, D., 2007. Multivariate poisson-lognormal models for jointly modeling crash frequency by severity. Transp. Res. Rec. 2019, 1–6. https://doi.org/10.3141/2019-01.

Pei, X., Wong, S.C., Sze, N.N., 2011. A joint-probability approach to crash prediction models. Accid. Anal. Prev. 43 (3), 1160–1166. https://doi.org/10.1016/j.aap.2010.12.026.

Pham, M.-H., Bhaskar, A., Chung, E., Dumont, A.-G., 2010. Random forest models for identifying motorway Rear-End Crash Risks using disaggregate data. In: 13th International IEEE Conference on Intelligent Transportation Systems, pp. 468–473. https://doi.org/10.1109/ITSC.2010.5625003.

Rahim, M.A., Hassan, H.M., 2021. A deep learning based traffic crash severity prediction framework. Accid. Anal. Prev. 154, 106090 https://doi.org/10.1016/j.aap.2021.106090.

Santos, K., Dias, J.P., Amado, C., 2022. A literature review of machine learning algorithms for crash injury severity prediction. J. Saf. Res. 80, 254–269. https://doi.org/10.1016/j.jsr.2021.12.007.

Sun, J., Sun, J., 2016. Real-time crash prediction on urban expressways: Identification of key variables and a hybrid support vector machine model. IET Intel. Transport Syst. 10 (5), 331–337. https://doi.org/10.1049/iet-its.2014.0288.

Thapa, D., Mishra, S., 2021. Using worker's naturalistic response to determine and analyze work zone crashes in the presence of work zone intrusion alert systems. Accid. Anal. Prev. 156 https://doi.org/10.1016/j.aap.2021.106125.

Thapa, D., Paleti, R., Mishra, S., 2022. Overcoming challenges in crash prediction modeling using discretized duration approach: An investigation of sampling approaches. Accid. Anal. Prev. 169, 106639 https://doi.org/10.1016/j.aap.2022.106639.

Xu, C., Tarko, A., Wang, W., Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accid. Anal. Prev. 57, 30–39. https://doi.org/10.1016/j.aap.2013.03.035.

Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accid. Anal. Prev. 59, 506–521. https://doi.org/10.1016/j.aap.2013.06.040.

Yasmin, S., Eluru, N., 2018. A joint econometric framework for modeling crash counts by severity. Transport. A: Transp. Sci. 14 (3), 230–255. https://doi.org/10.1080/23249935.2017.1369469.

Yasmin, S., Eluru, N., Bhat, C.R., Tay, R., 2014. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. Anal. Meth. Accid. Res. 1, 23–38. https://doi.org/10.1016/j.amar.2013.10.002.

Yasmin, S., Eluru, N., Wang, L., Abdel-Aty, M.A., 2018. A joint framework for static and real-time crash risk analysis. Anal. Meth. Accid. Res. 18, 45–56. https://doi.org/10.1016/j.amar.2018.04.001.

Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259. https://doi.org/10.1016/j.aap.2012.11.027.

Zhang, C., He, J., Wang, Y., Yan, X., Zhang, C., Chen, Y., Liu, Z., Zhou, B., 2020. A crash severity prediction method based on improved neural network and factor analysis. Discret. Dyn. Nat. Soc. https://doi.org/10.1155/2020/4013185.

Zhang, J., Li, Z., Pu, Z., Xu, C., 2018. Comparing prediction performance for crash injury severity among various machine learning and statistical methods. IEEE Access 6, 60079–60087. https://doi.org/10.1109/ACCESS.2018.2874979.

Zheng, L., Sayed, T., 2020. A novel approach for real time crash prediction at signalized intersections. Transp. Res. Part C 117, 102683. https://doi.org/10.1016/j.trc.2020.102683.