A hybrid statistical-machine learning methodology for addressing endogeneity and temporal instability in speeding-crash frequency relationships

Ghalehni, Sajad Asadi; Afghari, Amir Pooyan

DOI
10.1016/j.aap.2025.108284
10.1016/j.aap.2025.108284

Publication date
2026

Document Version
Final published version

Published in
Accident Analysis & Prevention

Citation (APA)
Ghalehni, S. A., & Afghari, A. P. (2026). A hybrid statistical-machine learning methodology for addressing endogeneity and temporal instability in speeding-crash frequency relationships. *Accident Analysis & Prevention*, *224*, Article 108284. https://doi.org/10.1016/j.aap.2025.108284, https://doi.org/10.1016/j.aap.2025.108284

Accident Analysis and Prevention 224 (2026) 108284

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A hybrid statistical-machine learning methodology for addressing endogeneity and temporal instability in speeding-crash frequency relationships

Sajad Asadi Ghalehni [a], Amir Pooyan Afghari [b,*]

[a] Road and Transportation Section, Faculty of Civil Engineering, Tarbiat Modares University, Jalal-e-Al Ahmad, Tehran, Iran
[b] Safety and Security Science Section, Faculty of Technology, Policy and Management, Delft University of Technology, 2628BX Delft, Netherlands

ARTICLE INFO

Keywords:
Speeding
Crash risk
Instrumental variable modelling
Machine learning
Hybrid modelling
Temporal instability

ABSTRACT

Speeding is a key behavioural factor contributing to increased crash frequencies along road segments, especially horizontal curves. Estimating the effect of speeding on crashes is, however, very challenging due to several reasons. Traditional speeding data collection methods often introduce measurement error in the analysis. In addition, there is a complex inter-relationship between driver behaviour, roadway geometry, and crash risk leading to endogeneity between speeding and crash risk. While instrumental variable modelling has been previously used for addressing such endogeneity, the effectiveness of this technique depends on strong instruments that correlate well with speeding but not with crashes. Moreover, the effects of explanatory variables on crashes may vary across locations and time too.

This study aims to address these gaps by developing a new methodology combining improved data collection and a hybrid statistical-machine learning model for better identification of speeding and a more accurate estimation of its effect on crashes. The model, tested on 179 km of horizontal curves along rural roads in Iran, integrates negative binomial regression and gradient boosting with shapley values. The negative binomial model is specified with random parameters and mixed spline indicators accounting for unobserved heterogeneity and temporal instability in the data. Results indicate high predictive power of the machine learning model in predicting speeding from exogenous variables, complemented by intuitive shapley values and feature importance for those variables. A comparison of statistical fit between the proposed model and several state-of-the-art modelling candidates showed that our model is superior to the existing modelling techniques. The results of this model suggest that curve's geometry and traffic characteristics are strong predictors of speeding, while driving more than 20 % over the speed limit substantially contributes to increased crash frequency. The effects of passenger and heavy vehicle traffic on crashes change over time.

1. Introduction

Traffic crashes are among the top ten causes of deaths and severe injuries globally and remain a major public health issue, despite ongoing efforts to improve road safety. These crashes are often complex and are the result of an interaction between drivers, vehicles, and the road environment. Driver behaviour is the primary contributor of these crashes, accounting for nearly 95 % of them in road networks (Abdel-Aty & Radwan, 2000; Petridou & Moustaki, 2000). In particular, speeding behaviour – where the average speed exceeds the + 3 % threshold of the posted speed limit (Chevalier et al., 2016)– has been consistently shown as a primary contributing factor to the frequency (and severity) of traffic crashes (Elvik et al., 2004). It is responsible for approximately 50 % of all crashes worldwide, a figure that is notably higher in low- and middle-income countries (World Health Organization, 2023). Speeding-related crashes are more prevalent on rural roads, horizontal curves, and roads with elevated speed limits (Choudhari & Maji, 2019; Council et al., 2010; Dhungana & Qu, 2005). Rural roads, in particular, are unique due to their distinct roadway geometric features and limited speed enforcement. Sharp horizontal curves, narrow lanes, and decreased visibility are common roadway characteristics in these areas contributing to a higher likelihood of crashes, especially in conjunction with speeding behaviour (Calvi, 2015; Pratt et al., 2019; Wu et al., 2017). Additionally, lower traffic volumes typically observed in rural areas may

* Corresponding author.
E-mail address: a.p.afghari-1@tudelft.nl (A.P. Afghari).

https://doi.org/10.1016/j.aap.2025.108284
Received 15 July 2025; Received in revised form 11 October 2025; Accepted 19 October 2025
Available online 24 October 2025

encourage drivers to drive at a higher speed, thereby increasing the likelihood of crashes (Ali et al., 2007). This is particularly concerning given that rural roads, in many countries, account for a disproportionate share of fatalities (Afukaar et al., 2003).

Accurate estimation of the effect of speeding on crash risk, however, is not straightforward because of several reasons. First and foremost, speeding and crash occurrence may be endogenous because they are inter-dependent and can influence each other. In econometrics and causal inference, endogeneity arises when an explanatory variable is correlated with the error term, often due to simultaneous causality, omitted variables, or measurement error (Nakamura & Nakamura, 1998). While speeding increases the likelihood of crashes due to reduced reaction time / longer stopping distances (Nassiri & Mohammadpour, 2023), crashes can influence speed selection of drivers too because drivers who experience or observe crashes may alter their speed—either by slowing down due to risk-compensating or by speeding in less regulated areas to compensate for the lost time (Oviedo-Trespalacios et al., 2020; Soole et al., 2013; Yasmin et al., 2022). This feedback loop means the two variables affect each other. In addition, unobserved factors may affect both speeding and crash frequency, creating a spurious relationship. For example, risk-taking drivers are more likely to speed and to be involved in crashes (Gheorghiu et al., 2015), but this personality trait is often not measurable particularly in aggregate (segment-specific) crash studies. Finally, speeding data are commonly collected in locations with higher risk of crashes and using speed cameras. As a result, the data may not capture actual driving behaviour of drivers, potentially skewing the observed relationship between speeding behaviour and crash outcomes (Yasmin et al., 2022).

Secondly, while many roadway characteristics and features (such as divided median or shoulder) are designed to mitigate crash frequency (for instance, by separating the opposing traffic or providing recovery space for errant vehicles), they may imply a 'safe opportunity' for drivers to exceed the speed limit and take over other vehicles which in turn may adversely increase crash frequency. Moreover, the effects of the above factors on speeding behaviour and ultimately on crash frequency may vary across locations (Afghari et al., 2018a; Liu and Chen, 2009). These varied effects are referred to as unobserved heterogeneity (F. L. Mannering et al., 2016). These effects may vary across time as well because many external factors such as safety campaigns and enforcement policies result in behavioural changes over time (F. Mannering, 2018). Overlooking either of these three properties (endogeneity, unobserved spatial heterogeneity and temporal instability) may lead to biased parameter estimates and erroneous inferences about the effects of speeding behaviour on crash frequency.

In response to the above issues, advanced methodologies have been developed and used in modelling crash frequency. Instrumental variable models (and latent variable models, both under the umbrella of simultaneous equation modelling technique) have been used to account for endogeneity between driver behaviour and crash frequency (Afghari et al., 2018b, 2019, 2023; Heydari and Forrest, 2024; Yasmin et al., 2022). These models regress the endogenous variable against all exogenous variables and then use its predicted value in the crash frequency model. Random Parameters (with/without heterogeneity in the means and/or variances) (Barua et al., 2016; Chen et al., 2017; Coruh et al., 2015; Heydari, 2018; Heydari et al., 2018; Huo et al., 2020; A. S. M. M. Islam et al., 2023; Shaon et al., 2018) and latent class (or finite mixture) models (Afghari et al., 2016; Kim, 2023; Li et al., 2018; Park & Lord, 2009) have been largely used to address unobserved heterogeneity in the effects of factors on crash frequency. These models allow parameters to vary across observations (or groups of observations), providing a more nuanced presentation of the complex dynamics influencing crash frequency (Anastasopoulos & Mannering, 2016; F. L. Mannering & Bhat, 2014; Washington et al., 2020). Finally, year-specific negative binomial models (Dzinyela et al., 2024; Fu et al., 2022; Mohammadi et al., 2014), year-indicator pooled negative binomial models(Alnawmasi & Mannering, 2022, 2023; Bhowmik et al., 2019b; Pervaz et al., 2024), and most recently, spline-indicator pooled negative binomial models (Marcoux et al., 2024; Phuksuksakul et al., 2025; Shabab et al., 2024) have been used to capture instability of model parameters across time. The spline-indicator pooled model employs spline functions to introduce temporal variations, offering a more flexible and interpretable way to capture changes over time, and has been shown to have superior performance to the other two modelling alternatives (Shabab et al., 2024).

Despite the above methodological advancements, there are still important gaps in understanding the effects of speeding on crash frequency:

(i) The effectiveness of instrumental variable modelling (in mitigating the potential effects of endogeneity bias) highly depends on the presence of strong instruments (exogenous variables) in the data which are well correlated with speeding (the endogenous variable) but are not correlated with crash frequency (the dependent variable). Previous studies have consistently highlighted the challenge of identifying suitable instruments for this purpose (Afghari et al., 2021, 2023). A potential solution for this dilemma is to make use of machine learning algorithms to create the instrumented variable. These algorithms are not bound to any distributional assumptions and use all nuances in the data to predict an outcome and therefore have much higher predictive power than conventional statistical models. A few recent studies have integrated these algorithms into instrumental variable modelling, although in a different context, and found that such an integration enhances the overall accuracy of the models (Afghari et al., 2022; Hussain et al., 2022).

(ii) Most of the conventional data collection methods for assessing speeding behaviour result in measurement error in speeding data (the difference between the measured value of speeding and its unknown - true value). Overt and fixed speed cameras provide censored data on speeding because of behavioural adaptation of drivers at the locations of those cameras (Marciano & Norman, 2015), whereas covert and mobile speed cameras are easily recognisable with the new capabilities of smartphones and in-vehicle technologies. The advent of advanced data collection technologies, such as unmanned aerial vehicles (UAVs), has further improved the precision and granularity of traffic monitoring, offering an innovative solution to the biases associated with traditional, fixed-location sensors (Dronova et al., 2022). UAVs enable capturing detailed data on vehicle trajectories and speed profiles, providing unparalleled insights into the interaction between road geometry, traffic volume, and speeding behaviour (Ghalehni & Boroujerdian, 2023; Karimi & Boroujerdian, 2021; Xing et al., 2019). This level of detail is crucial for enhancing the accuracy of crash frequency models and for developing more targeted interventions to mitigate the risks associated with speeding-related crashes. In addition, UAVs have fewer perspective-view issues or calibration requirements in comparison with fixed cameras (Fig.1). The difficulty in calibrating fixed cameras at long distances, where the perspective becomes more pronounced, limits the coverage area of these cameras. In contrast, UAVs with their orthogonal video recording capabilities, can cover greater lengths with minimum distortion, providing more accurate and expansive data.

2. Analytical framework

The proposed analytical framework in this study consists of two components, operating in two stages subsequently. A machine learning algorithm is used in the first stage, and a statistical model is used in the second stage in which the output of the first stage is used. A schematic of the proposed hybrid framework is presented in Fig. 2. The details of this schematic are presented in the following subsections.

For a better readability of the framework, a few notations are presented in the following and prior to presenting each component. Let i (i

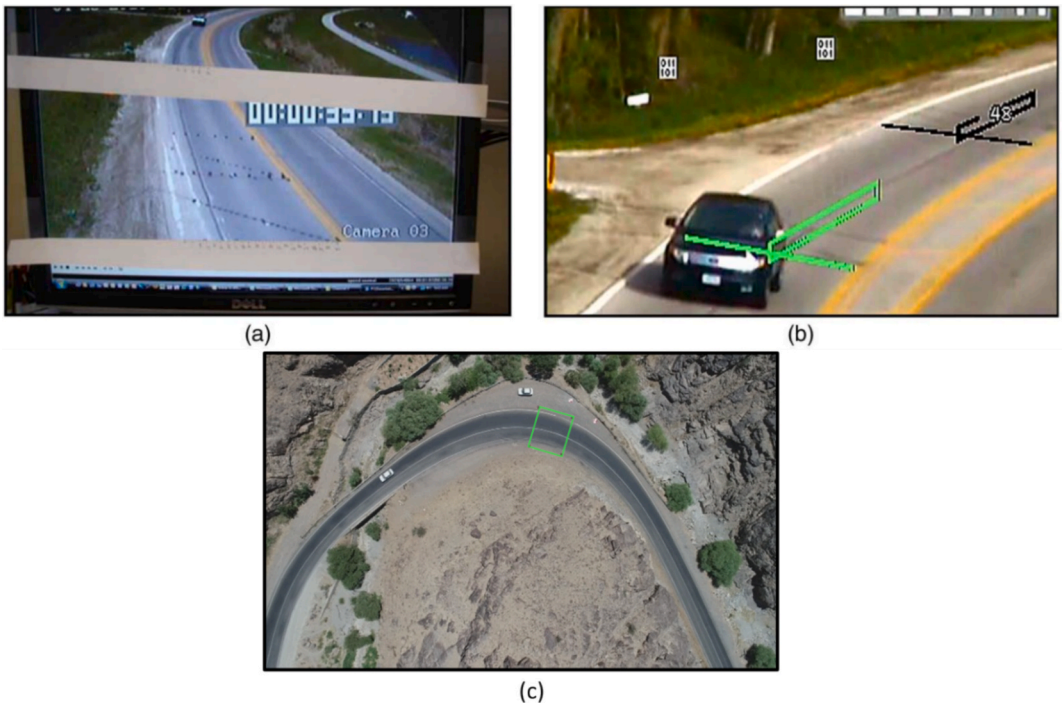

Fig. 1. Comparison of distortion view in fixed cameras (a) and (b) versus an unmanned vehicle (c) (Fitzsimmons et al., 2013).

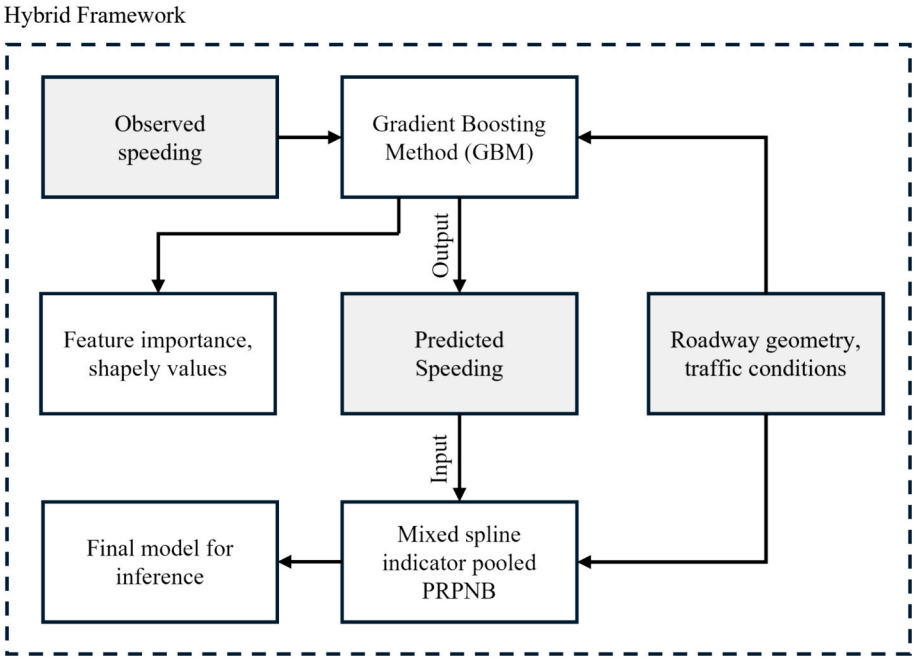

Fig. 2. A schematic of the proposed hybrid framework for estimating the relationship between speeding and crash frequency.

$= 1, 2, \ldots, I$) represent the location of each curve, and t ($t = 1, 2, \ldots, T$) represent the time period, corresponding with different years of data. Let k ($k = 1, 2, 3$) represent the magnitude of speeding categories, with k = 1 representing 'minor speeding' (less than 10 % over the speed limit), k = 2 representing 'moderate speeding' (between 10 % and 20 % over the speed limit), and k = 3 representing 'major speeding' (more than 20 % over the speed limit). Since the proposed analytical framework is segment-specific, the speeding behaviour of drivers is aggregated across a segment and is used in the models as the proportion of vehicles speeding in the above three categories (more on this will be presented in Section 3).

2.1. First Stage: Speeding behaviour component

In the first stage of the framework, the speeding behaviour of drivers (in the form of a proportion) is predicted using a Gradient Boosting Method (GBM) to serve as an instrument for the observed speeding behaviour in the second stage. In other words, observed speeding will be

replaced by predicted speeding in the subsequent crash frequency model which addresses the potential endogeneity between observed speeding behaviour and crash frequency. GBMs are ensemble learning techniques that combine multiple shallow decision trees to capture nonlinear relationships and complex interactions among roadway and traffic characteristics (Natekin & Knoll, 2013; M. Sadeghi et al., 2024; Wen et al., 2021). Compared to alternatives such as Random Forests, which may overemphasize variance reduction, or neural networks, which typically require larger datasets and more intensive calibration, GBMs provide a balanced trade-off between predictive accuracy, interpretability, and computational efficiency—making them particularly suitable for the present dataset.

The GBM is implemented using the XGBoost framework (Boehmke & Greenwell, 2019), leveraging a Gamma distribution to effectively model the skewed nature of the speeding data.[1]

The underlying latent propensity for the speeding behaviour component can be formulated as:

$$\widehat{y}_{kit}^{(r)} = \sum_{m=1}^{M} f_m(x_{it}) = \widehat{y}_{kit}^{(r-1)} + f_r(x_{it}) \tag{1}$$

where $\widehat{y}_{kit}^{(r)}$ denotes the estimated proportion of speeding behaviour for location i and t time period after r iterations, M is the number of additive trees, and x_{it} represents the input features (analogous to independent variables in statistical models) influencing the speeding behaviour.

The objective function for minimizing the loss in the GBM model can be expressed as:

$$\sum_{i=1}^{n} l(y_{kit}, \widehat{y}_{kit}) + \sum_{m=1}^{M} \Omega(f_m) \tag{2}$$

where $l(y_{kit}, \widehat{y}_{kit})$ is the loss function, and $\Omega(f_m)$ represents the regularization term to prevent overfitting and reduce complexity, defined as:

$$\Omega(f_m) = \gamma T_m + \frac{1}{2}\lambda \sum_{n=1}^{T_m} \omega_n^2 \tag{3}$$

In this formulation, T_m denotes the number of leaves in the m^{th} tree, ω_n signifies the L2 norm of the n^{th} leaf scores, and n indicates the total number of speeding instances in the sample data.

Regularization in $\Omega(f_m)$ is applied specifically to manage the model's complexity by penalizing:

Tree Depth: Limiting the maximum depth of each tree reduces the number of splits, making trees shallower and less prone to overfitting.

Number of Leaves: Penalizing the number of leaves in each tree helps prevent the model from memorizing details in the training data, thereby enhancing its generalization ability.

Leaf Weights (Shrinkage): Each leaf in the GBM has an associated weight that contributes to the prediction. By penalizing larger weights, the model ensures that no single tree dominates the prediction, allowing a more balanced contribution from each tree and improving the model's robustness.

This structure, with the combined effects of tree depth, leaf number, and weight penalties, helps the model avoid excessive complexity while providing an accurate estimation of speeding behaviour across iterations.

By employing the GBM model, which effectively identifies key predictors of speeding through its iterative learning process, the analysis captures the complex relationships present in the data without necessitating predefined theoretical assumptions. This capacity is crucial for isolating the instrumental variable, enhancing the subsequent analysis in the Random Parameters Negative Binomial (RPNB) model.

2.2. Second Stage: Crash frequency component

In the second stage, the frequency of crashes across road segments is analyzed using a random parameters negative binomial model (Anastasopoulos & Mannering, 2009) with spline indicator variables originally introduced by Shabab et al., (2024). In this approach, spline indicators provide a piecewise linear representation of time, allowing explanatory variables to capture gradual changes in their effects across years rather than abrupt shifts tied to year-specific dummies. By combining this temporal formulation with random parameters, the model simultaneously accounts for spatial heterogeneity across locations and temporal instability in coefficient estimates.

Let us assume that the frequency of crashes at segment i during time t (c_{it}) follows a negative binomial distribution with the mean (expected value) μ_{it}. The probability density function of c_{it} can be expressed as:

$$(c_{it}) = \frac{\Gamma\left(c_{it} + \frac{1}{\lambda}\right)}{\Gamma(c_{it}+1)\Gamma\left(\frac{1}{\lambda}\right)} \left(\frac{1}{1+\lambda\mu_{it}}\right)^{\frac{1}{\lambda}} \left(1 - \frac{1}{1+\lambda\mu_{it}}\right)^{c_{it}} \tag{4}$$

Where, $P_{(c_{it})}$ is the probability that segment i will experience c_{it} crashes at time t, and λ is the overdispersion parameter in the negative binomial distribution. μ_{it} is structured as a log-link function of exogenous covariates as:

$$\mu_{it} = \exp\left(X_{it}'\beta + Z_i'\theta_i + W_t'\gamma_t + \varepsilon_{it}\right) \tag{5}$$

Where, X_{it} is a vector of exogenous covariates with fixed coefficients (β), Z_i is a vector of exogenous covariates with random coefficients varying at segment (location) level (θ_i) allowing for spatial unobserved heterogeneity, and following a normal distribution with mean $\overline{\theta}$ and standard deviation σ. W_t is a vector of exogenous covariates with coefficients varying at temporal (year) level (γ_t) allowing for temporal instability, and ε_{it} is a random error term; $\exp(\varepsilon_{it})$ follows a Gamma distribution with mean 1 and variance λ.

Borrowing from Shabab et al., (2024), we create a set of time-dependent variables as:

$$Year_1 = Max(Year_{record} - Year_{base}, 0); \tag{6}$$

$$Year_2 = Max(Year_{record} - Year_{base} - 1, 0);$$

$$Year_N = Max(Year_{record} - Year_{base} - (N-1), 0);$$

Where $Year_{record}$ represents the observation year, and $Year_{base}$ is the reference year. A product of the above indicators and the temporally instable variables (changing more than 5 % over time) are then used as covariates in Equation (6). This formulation allows for a piecewise linear representation of temporal effects, enabling a more flexible and efficient evaluation of changes in parameters over time.

Speeding behaviour is now introduced into the model in the form of speeding proportions in the k^{th} category (minor, moderate, and major) denoted by Y_{kit}. The log-link function, incorporating the speeding variable(s), is then expressed as:

$$\mu_{it} = \exp\left(X_{it}'\beta + Z_i'\theta_i + W_t'\gamma_t + Y_{kit}'\omega_{kit} + \varepsilon_{it}\right) \tag{7}$$

where ω_{kit} is a vector of estimable parameters. Y_{kit} is, however, endogenous with observed crash frequencies and thus is replaced by the predicted speeding proportions from the Gradient Boosting Model in the first stage ($\widehat{Y}_{kit}$). The final log-link function of the mean, incorporating the instrumental variable becomes:

$$\mu_{it} = \exp\left(X_{it}'\beta + Z_i'\theta_i + W_t'\gamma_t + \widehat{Y}_{kit}'\omega_{kit} + \varepsilon_{it}\right) \tag{8}$$

[1] Speeding data are usually skewed to the left because minor speeding (less than 10% above the speed limit) is much more common than moderate and major speeding (more than 10% above the speed limit) (Perez et al., 2021). This is the case in the sample of speeding data in this study too.

This hybrid model formulation (stage 1 and 2, together) is referred to as the *mixed spline indicator pooled random parameters negative binomial model with gradient boosting method* (MSIPRPNB-GBM) in this manuscript and accounts for the unobserved heterogeneity across segments and temporal instability while also addressing the endogeneity between speeding behaviour and crash frequency.

2.3. Model performance

The models in the first and second stages are evaluated based on various performance metrics. The predictive accuracy of the GBM model in the first stage is assessed using Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R^2), the Receiver Operating Characteristic (ROC), and Area Under the Curve (AUC), providing a detailed evaluation of predictive performance. These metrics are calculated as follows:

Mean Squared Error (MSE):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_{kit} - \widehat{Y}_{kit})^2 \tag{9}$$

Where N is the total number of observations, Y_{kit} represents the observed speeding proportion, and $\widehat{Y}_{kit}$ denotes the predicted speeding proportion. Lower MSE values indicate higher predictive accuracy.

Mean Absolute Error (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Y_{kit} - \widehat{Y}_{kit}| \tag{10}$$

The MAE provides a measure of prediction error without squaring the deviations, making it less sensitive to outliers than MSE. Lower MAE values signify improved model precision.

R-squared (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_{kit} - \widehat{Y}_{kit})^2}{\sum_{i=1}^{N}(Y_{kit} - \overline{Y}_{kit})^2} \tag{11}$$

where $\overline{Y}_{kit}$ is the mean of observed values. The R^2 metric provides the proportion of variance in speeding proportion explained by the GBM model, with values closer to 1 indicating better explanatory power.

ROC AUC Score:

This score evaluates the model's ability to differentiate between speeding severity levels by measuring the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various thresholds. The AUC (Area Under the Curve) quantifies this performance, with values closer to 1 indicating stronger classification ability. A high AUC demonstrates the GBM model's effectiveness in accurately identifying instances of each speeding category.

The performance of the negative binomial models in the second stage are assessed using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC):

$$AIC = 2P - 2LL \tag{12}$$

$$BIC = PLn(N) - 2LL \tag{13}$$

where P is the total number of estimable parameters, N is the total number of observations, and LL is the log-likelihood value of the model at convergence. The model with the lowest AIC and BIC is preferred as it strikes a balance between goodness of fit and complexity.

By employing many different metrics of performance (MSE, MAE, R^2, ROC AUC, AIC and BIC), the analysis provides a robust evaluation of predictive accuracy (which is needed for the first stage) and statistical fit (which is needed for the second stage).

2.4. Model explanation

To interpret the GBM, both feature importance scores and SHapley Additive exPlanations (SHAP) values are employed. Feature importance in the GBM model is calculated using gain-based importance, which measures the average improvement in model accuracy attributed to each feature across all trees. Each feature's importance score is derived by summing the reduction in prediction error achieved each time that feature is used for a split and averaging it across splits. Mathematically, the importance for feature f_i is given by:

$$Importance_{f_i} = \frac{\sum_{j\in S_{f_i}} Gain_j}{\sum_{K\in F}\sum_{j\in S_{f_k}} Gain_j} \tag{14}$$

where S_{f_i} represents all splits involving feature f_i, $Gain_j$ is the reduction in loss from split j, and F is the set of all features. This allows for a clear, quantitative ranking of features according to their contribution to the model.

In addition, SHAP values offer a comprehensive view of feature impact by calculating the marginal contribution of each feature to the model's prediction for each individual sample (analogous to marginal effects in the statistical models). Let x_i be the i-th sample, $x_{i,j}$ the j −th feature of x_i, and $\bar{y}$ the baseline (expected value) of the model. SHAP values (ϕ_j) for feature j in sample x_i are computed as:

$$y_i = \bar{y} + \sum_{j=1}^{K} f(x_{i,j}) \tag{15}$$

where $f(x_{i,j})$ denotes the contribution of feature j to the prediction for sample x_i. SHAP thus provides an additive approach to decompose each prediction, enabling a detailed analysis of how individual features drive model outcomes.

3. Data and Empirical Design

Speed data in this study were collected using a UAV on rural roads in Khorasan Razavi and Gilan provinces, in Iran. The extent of the network is 179 km comprising segments of various lengths, all with horizontal curves. Speed data were collected over 28 days in May and June 2018. A total of 1,245 vehicles passed through these curves during this period. For vehicle detection and tracking, the YOLOv5 model(Redmon et al., 2016) and a Kalman filter-based Simple Online (Welch & Bishop, 1995) and Realtime Tracking (SORT) algorithm (Bewley et al., 2016) were employed. YOLOv5 is a state-of-the-art deep learning model which is able to detect objects using a convolutional neural network. This algorithm divides each video frame into a grid, assigning bounding boxes to potential objects within each grid cell. Using regression-based probability estimates, YOLOv5 classifies the detected objects—specifically, vehicles—while resolving overlapping bounding boxes through non-max suppression to ensure accurate localization. Following the detection of vehicles, the SORT algorithm tracks them across frames by utilizing Kalman filtering to predict the position and movement of each object. The filter models the bounding box centre, scale, and aspect ratio, alongside their time derivatives. Predicted states are matched to new detections in subsequent frames using the Hungarian algorithm, which minimizes identity switches. This process enables the system to maintain robust tracking, even when objects momentarily disappear from the frame or experience occlusions, ensuring consistent vehicle tracking throughout the video. The final output of the software is vehicle trajectories from which vehicles' speeds were calculated for the purpose of this study.

Since the scope of this study is segment-specific, the speeding behaviour of drivers was aggregated across road segments by taking the average of actual speed for every driver along the horizontal curve and

comparing it with the posted limit. The number of speeding vehicles was then recorded in three categories: (1) *minor speeding*: less than 10 % above speed limit (including no speeding),[2] (2) *moderate speeding*: between 10 % and 20 % above the speed limit and (3) *major speeding*: more than 20 % above the speed limit (Zhou et al., 2024a). Consequently, the speeding variable in this study is expressed as a proportion:

$$\text{Speeding proportion} = \frac{\text{Number of observed vehicles in each speeding category per segment}}{\text{Total number of observed vehicles per segment}}$$

Crash data were collected from police reports for the same road segments for eight years, from 2011 through 2018. The data contained information about the time and location of crashes, severity levels, and crash type. Additionally, vehicle count data were extracted from traffic camera datasets provided by the Provincial Departments of Roads and Urban Development for Khorasan Razavi and Gilan, covering the same period from 2011 to 2018.

To bring the two datasets (speed and crash data) to the same units, a proportional (linear) extrapolation method was used to extrapolate speeding proportions from monthly (28-day observation period) to yearly data from speeding tickets issued to drivers across the same locations over the same 8 years as crash data. In doing so, the ratio of speeding tickets (issued in the same categories as above) across different months and years were used to determine the monthly and yearly trends and then these trends were applied to the observed speeding data to create yearly speeding data for 8 years. The summary statistics of the final speeding and crash data are presented in Table 1 below.

To further illustrate the potential association between crash frequency and speeding behaviour, binned means of crash frequencies across all segments with their 95 % confidence intervals, together with a locally weighted smoothing line (LOWESS) have been plotted in Fig. 3. In this figure, each point represents the average crash frequency within equal-width bins of speeding proportions, while the red line indicates the general trend across the sample. These bins and the line show that crash frequency increases with higher speeding proportions. The increasing slope of the LOWESS curve suggests a positive association between crash frequency and the proportion of speeding, implying that segments with a greater prevalence of speeding are more likely to experience higher crash occurrences.

In addition, roadway geometric characteristics were collected for the same network too. Some of these characteristics, such as lane width, shoulder width, and grade, were directly measured at the site, while other variables, such as deflection angle, curve radius, and length, were calculated from the UAV aerial footage. The radius of curves varied between 30 and 150 m, while the vertical grade ranged between −8% and + 8 %. The summary statistics of road geometric data are presented in Table 2.

To further illustrate the variation in driving behavior across roadway environments, the distribution of vehicle speeds is presented by posted speed limit categories. Table 3 summarizes the average and 85th percentile of speed across road segments with different speed limits, providing a clearer picture of the actual operating speeds relative to the posted speed limit.

[2] Since the scope of this study is to contrast normal/minor speeding behaviour versus high-risk or aggressive speeding, the first speeding category included non-speeding vehicles too, especially because slight speed limit exceedance (e.g., 1–5 km/h over the speed limit) is common and often considered socially acceptable or legally tolerated in Iran.

4. Results and Discussion

4.1. Speeding behaviour component

The GBM model within the first stage of the hybrid model was implemented using the XGBoost library (version 3.0.0) in Python. In developing this model, several hyperparameters were determined through trial and error. The final values included a learning rate of 0.001, a maximum depth of 8 for the trees, a subsample of 0.8, and a total of 1000 trees, with early stopping applied after 100 rounds. A list of all hyperparameters, including both the default values provided by XGBoost and the values adopted in this study, is presented in the Appendix. While trying other values of these hyperparameters may increase the performance of the GBM model, optimizing the hyperparameters is not within the scope of this study. The dataset was divided into 80 % training data and 20 % testing data. Two GBM models were developed, one for minor and one for major speeding categories (the proportions of the three speeding categories must sum to unity and so developing three separate models could introduce inconsistencies in the second stage where their combined predictions might exceed 1). The final GBM models demonstrated strong predictive performance, with R-squared (R^2) values of 0.842 and 0.834, alongside ROC AUC scores of 0.904 and 0.941, indicating high classification accuracy. Table 4 presents the predictive performance metrics for these two GBM models.

Results of the SHAP analysis for the two speeding prediction models

Table 1
Summary statistics of dependent variables in the study.

Dependent Variables	Definitions	Mean	Standard Deviation	Minimum	Maximum
Proportion of Minor speeding	Number of vehicles driving less than 10 % above the posted speed limit / Total number of observed vehicles	0.40	0.16	0.08	0.74
Proportion of Moderate speeding	Number of vehicles driving between 10 and 20 % above the posted speed limit / Total number of observed vehicles	0.41	0.05	0.21	0.48
Proportion of Major speeding	Number of vehicles driving more than 20 % above the posted speed limit / Total number of observed vehicles	0.17	0.24	0.00	0.97
Crash frequency	Total number of crashes recorded per year	3.09	4.60	0.00	22.00

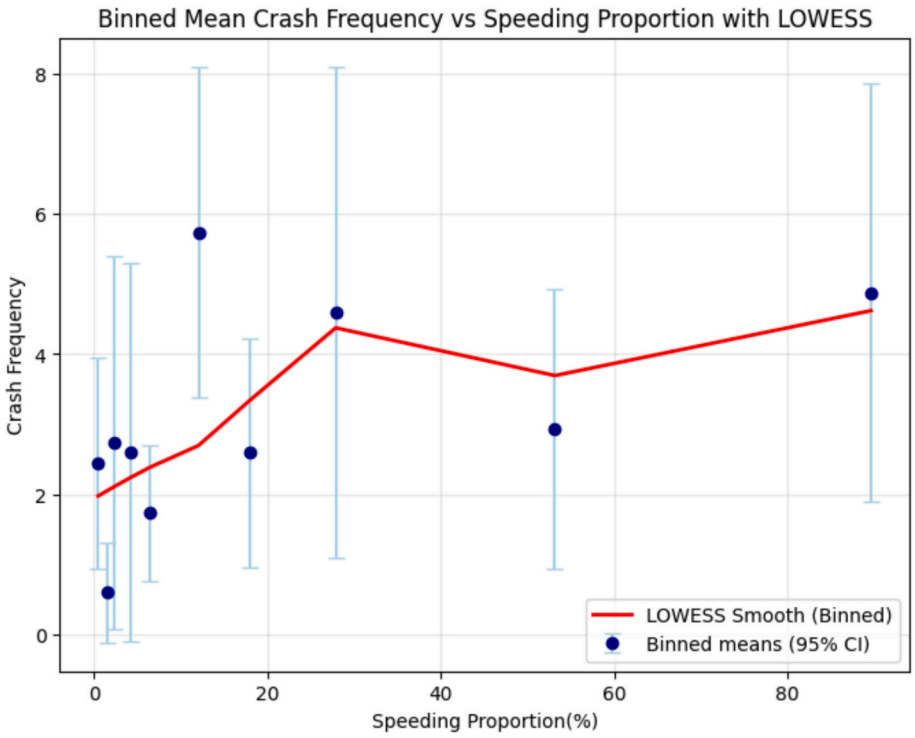

Fig. 3. Binned mean crash frequency versus speeding proportion with 95% confidence intervals and locally weighted smoothing line (LOWESS).

are illustrated in Fig. 4. Each point in these figures represents one observation. Features are ranked by their average predictive power on the outcome variable (speeding proportions), with the location of each point along the horizontal axis representing this power on the specific observation. The colour gradient indicates the feature value (low to high), and the SHAP values indicate the magnitude of the effect. Please note that visual differences between the minor and major speeding plots arise from overlapping points with SHAP values near zero—particularly for low-importance features—not from unequal sample sizes. Features with wider SHAP value ranges appear denser because their points are more spread, while features with narrower SHAP value ranges cluster around zero.

Radius is the most influential feature in predicting minor and major speeding proportions. Higher values of radius (red dots) contribute to decreased predicted speeding proportions, particularly for major speeding category (Fig. 4 (b)). The degree of curve and deflection angle are among predictors of the speeding behavior too: flatter curves and lower deflection angles contribute to increased predicted minor and major speeding proportions.

Pavement Condition Index (PCI) contributes positively to both predictions, with higher PCI (better pavement quality) linked with increased predicted speeding proportions. In contrast, curve length contributes negatively to speeding predictions, where longer curves contribute to slightly lower predicted proportions of minor and major speeding. Outer shoulder width has a negative effect on speeding predictions too, more noticeably for minor speeding.

Widening, MADT, and the percentage of heavy vehicles have mixed effects. Widening appears to result in decreased minor speeding proportions, while higher MADT (red dots in Fig. 4)b() results in increased major speeding proportions. Heavy vehicle percentage plays a stronger role in predicting the minor speeding proportions, where higher values of this feature reduce the predicted proportion of minor speeding.

Additional features such as edge and centerline marking quality, grade, and curve type have smaller but consistent effects. Better marking quality and lower grade values generally contribute to the prediction of reduced speeding proportions.

To have a better understanding of the contributions of the above features to the prediction of speeding proportions, their feature importance (ranked from the highest to the lowest) are presented in Figs. 5 and 6 for the minor and major speeding categories, respectively.

A comparison of the SHAP analysis with the feature importance results reveals differences in the ranking and predictive power of specific features. Nonetheless, these features are analogous to instruments in conventional instrumental variable modelling, and the predicted speeding proportions from this GBM model are postulated to be less subject to endogeneity in comparison with the observed speeding proportions. However, an important conundrum is whether this postulation is valid after all because when the instruments are constructed in a way that maximize prediction, they might capture the same unobserved factors ('noise' in the language of machine learning) that are correlated with crash frequency. While proving this exogeneity is very difficult (as the true effect of speeding on crash frequency in the population is not known), the shapley values together with the feature importance analysis in our study show that the GBM's high accuracy is due to meaningful features rather than unobserved errors. This explanation is even more reinforced noting that most of the important features in predicting the speeding proportions (radius, degree and length of horizontal curves, type of curves, deflection angle, pavement condition index, quality of centerline and edge lane markings, and shoulder width) have intuitive causal link with speeding but are not statistically significant when used directly in the mean function of the negative binomial model in the second stage (more on this will be presented in the next section). In other words, the effect of these features on crashes is only through speeding.

4.2. *Crash frequency component*

In the second stage of the hybrid framework, the negative binomial crash frequency model was estimated using the predicted speeding proportions from the first stage as well as other explanatory variables, using STATA 17.0 statistical software package. In estimating the crash frequency model in the second stage, explanatory variables were selected using a stepwise variable selection criterion. They were tested

Table 2
Summary statistics of independent variables in the study.

Continuous variables	Mean	Standard Deviation	Minimum	Maximum	Temporal Variation
Monthly Average Daily Traffic (MADT) (vehicles/day)	1319.89	220.42	859	1742	28 %
Percentage of heavy vehicle traffic (%)	4.38	1.64	1.97	8.32	25.7 %
Curve length (m)	74.98	47.16	24.7	205.4	<5%
Widening (m)	1.18	1.73	0	7.5	<5%
Inner lane-shoulder-width (m)	2.04	2.23	0	9.5	<5%
Outer lane-shoulder-width (m)	3.02	2.15	0	7.8	<5%
Curve radius (m)	64.89	37.68	15	153	<5%
Deflection angle (°)	107.58	39.72	32	170	<5%
Radius	64.89	37.805	15	153	<5%
Degree of curve*	67.66	99.89	3.82	360	<5%
Centre line marking quality (of tenth)	5.83	2.20	1	8	<5%
Edge line marking quality (of tenth)	3.21	2.16	1	7	<5%
Road vertical slope (absolute value)	4.86	2.42	0	8	<5%
Pavement condition index	69.88	14.44	37	95	<5%
Categorical variables			Sample frequency	Sample share	Temporal Variation
Posted speed limit					
20 km/h			16	0.10	<5%
30 km/h			32	0.21	<5%
40 km/h			40	0.26	<5%
50 km/h			16	0.10	<5%
60 km/h			48	0.33	<5%
Sufficient stopping sight distance**					
Yes			83	0.54	<5%
No			69	0.46	<5%
Curve type					
S-curve			43	0.28	<5%
Normal			109	0.72	<5%
Before Curve Existence					
Yes			61	0.40	<5%
No			91	0.60	<5%
Curve direction					
Right turn			101	0.66	<5%
Left turn			51	0.34	<5%
Outer lane shoulder-type					
Paved			131	0.86	<5%
Unpaved			21	0.14	<5%

* The central angle is created by two radii extending from the centre of a circle to its ends. It is 100 m (feet in imperial units) long.

** Thresholds of 50 m and 80 m has been used for this variable according to (American Association of State and Highway Transportation Officials, 2018).

Table 3
Speed distribution by posted speed limit.

Speed Limit (Km/h)	Average Speed (Km/h)	85th percentile speed (Km/h)
20	28.83	37.35
30	44.86	59.54
40	47.69	53.73
50	52.18	64.03
60	47.06	58.76

Table 4
Performance metrics for Gradient Boosting Machine (GBM) models predicting major and minor speeding proportions.

	MSE	MAE	R-Squared (R^2)	ROC AUC Score
GBM Model for minor speeding	0.007	0.069	0.834	0.941
GBM Model for major speeding	0.009	0.042	0.842	0.904

for multicollinearity by computing the Pearson or Spearman correlation coefficients, and the variables with unacceptably high (>0.7) correlation coefficients were not simultaneously introduced into the model. The parameters of all variables were tested for random parameters specification (across location) and normal distribution was used as the distribution for all of the random parameters. The parameters were considered random only if their standard deviations are statistically significant. The parameters of all variables with more than ± 5 % variation during the study period were tested for temporal instability. The models were estimated using the maximum simulated likelihood approach with 500 Halton draws. The required number of Halton draws was selected so that further increasing the number of draws does not change the estimates significantly. The final specification of the model was based on statistical significance guided by a 95 % confidence level. The results of this model are presented in Table 5 and 6 below. In Table 5, we present a comprehensive examination of temporal fluctuations of each variable's impact on crash frequency while in Table 6 we present the net effects of these variables on crash frequency across the years.

Monthly Average Daily Traffic[3] (MADT) exhibited a statistically significant and temporally dynamic effect on crash frequency, consistent with the patterns observed in the literature (Shabab et al., 2024). As shown in Table 5, the coefficient for the natural logarithm of MADT is zero prior to 2014 and becomes increasingly positive in the subsequent years, peaking in 2015 and gradually declining thereafter. This shift underscores the importance of accounting for temporal instability in modeling traffic safety outcomes. The non-significant role of MADT during the early years might be due to the dominance of single-vehicle (such as run-off-road) crashes in those years as illustrated in Fig. 7.

[3] While Annual Average Daily Traffic (AADT) has been long used as a measure of exposure in road safety modelling (Elvik et al., 2009), the choice between AADT and MADT depends on the purpose of study and the level of accuracy that is needed. AADT is usually widely available and standardized across road agencies. It is useful for long-term planning, network-level safety models, and comparisons between road segments. However, it masks seasonal, monthly, and temporal fluctuations. It may lead to under- or over-estimation of risk if crashes correlate with specific seasonal traffic surges. In contrast, MADT captures temporal variability and is a more precise exposure metric when crashes show seasonal patterns (Jessberger et al., 2016)– which is the case in our study. Therefore, we used MADT for our analysis to capture temporal nuances in the effects of exposure on crash frequency.

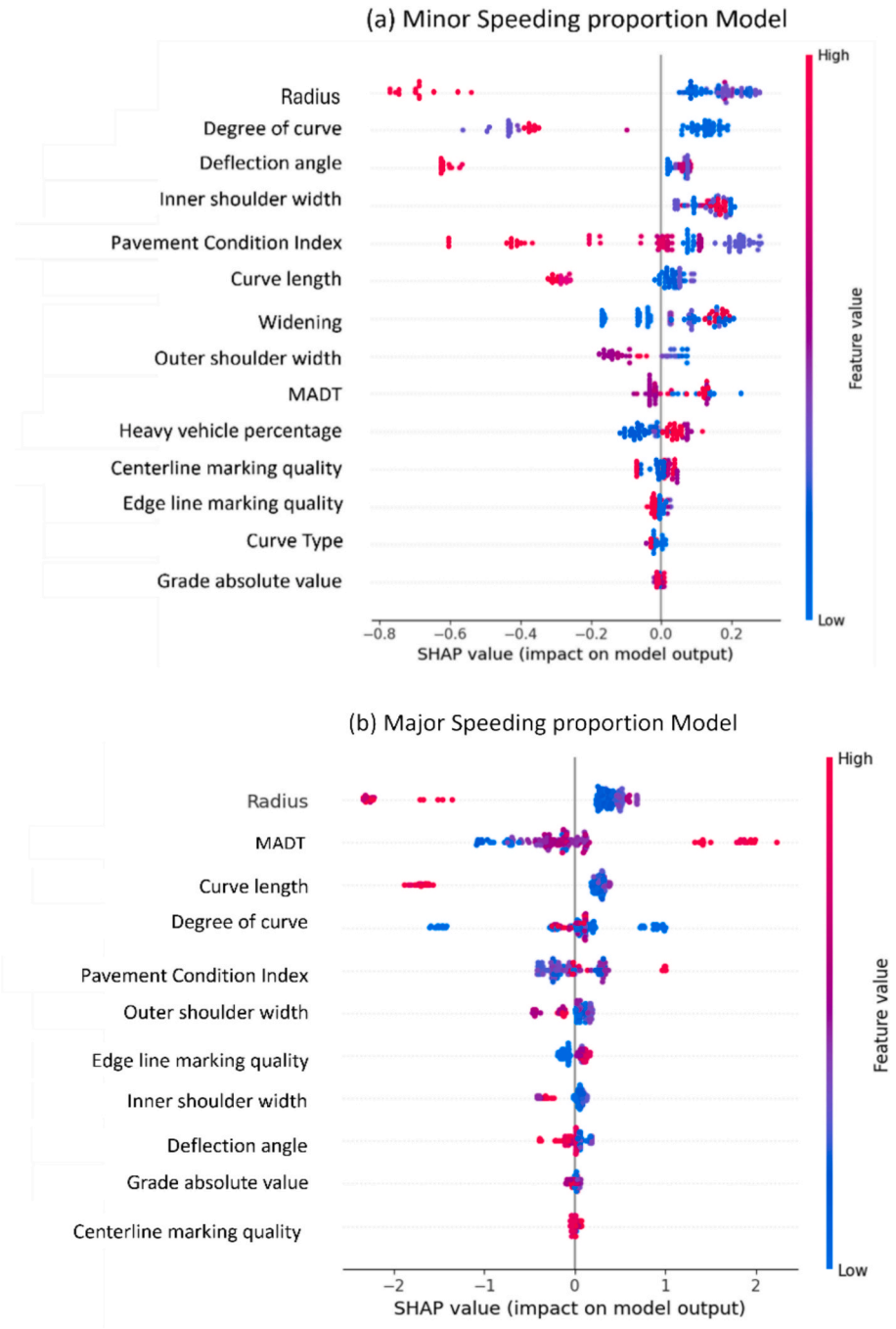

Fig. 4. SHAP summary plot illustrating the impact of road and traffic features on (a) minor speeding predictions, (b) major speeding predictions.

Traffic volume might not be a major risk factor for these crashes. After 2014, however, multi-vehicle crashes become dominant and the effects of MADT become positive and increasing over the next years. This interpretation supported by changes in the distribution of crash types across the years is in line with the findings from the literature noting that increased exposure in dense traffic conditions often leads to a higher likelihood of interaction-based crashes (Geedipally & Lord, 2010). The observed pattern reinforces the benefit of the temporally segmented modelling and the spline-based specification in uncovering evolving risk within crash data, thereby offering a more nuanced understanding of the traffic volume–crash relationship over time.

These findings further reinforce the nonlinear relationship between traffic volume and crash risk, highlighting the dynamic influence of traffic flow on roadway safety, especially with respect to different crash types. Moreover, this finding aligns with those from the existing literature, suggesting that dynamic traffic conditions play a crucial role in shaping road safety outcomes (Alnawmasi & Mannering, 2022; Council et al., 2010).

The percentage of heavy vehicles demonstrated a temporally unstable coefficient in its relationship with crash frequency across the study period. Between 2011 and 2013, the coefficient is positive, suggesting that higher percentage of heavy vehicles increases crash likelihood — a finding consistent with traditional safety concerns of heavy vehicles because of their large mass, limited maneuverability, and longer stopping distances (Zhu & Srinivasan, 2011; Zubaidi et al., 2022). However, between 2014 and 2016, the coefficient becomes negative,

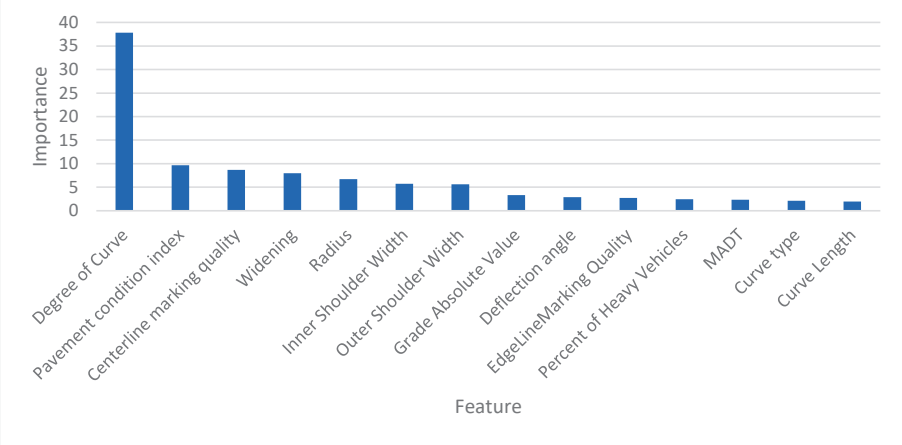

Fig. 5. Feature importance for the predictors used in the minor speeding model.

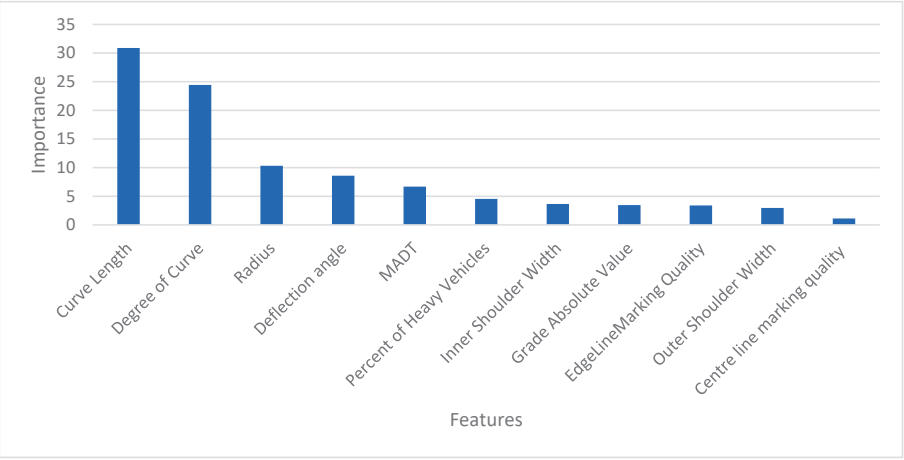

Fig. 6. Feature importance ranking for the predictors used in the major speeding model.

Table 5
Estimation results of mixed spline indicator pooled random parameters negative binomial model with gradient boosting method – base and deviation effect of variables.

Variable	Year							
	2011	2012	2013	2014	2015	2016	2017	2018
Time variant variables								
Constant	−1.756 (−5.61)*	–	–	–	–	–	–	−1.231 (−2.27)
Ln(MADT)	–	–	–	1.058 (7.35)	–	−1.373 (−5.93)	–	–
Percent of heavy vehicles	0.215 (5.13)	–	–	−1.440 (−6.68)	–	2.327 (5.79)	–	–
Time invariant variables								
Inverse absolute grade**	33.660 (4.44)	–	–	–	–	–	–	–
Pavement Condition Index	−0.015 (−2.18)	–	–	–	–	–	–	–
Inner Shoulder type	−0.387 (−2.10)	–	–	–	–	–	–	–
Major speeding (instrumented)								
Mean	2.058 (2.13)	–	–	–	–	–	–	–
Variance over locations	0.006 (16.62)	–	–	–	–	–	–	–
Logarithm of dispersion parameter	−0.876 (−1.97)	–	–	–	–	–	–	–

* Numbers in brackets present the corresponding Z-values.
** The absolute grade was offset by 10 values to compensate instances where the original grade was zero.

and sharply increases again in 2017 and 2018. This non-monotonic trend underscores the temporal volatility of the variable's effect on crash risk.

At first glance, the negative coefficient during the mid-period (between 2014 and 2016) may appear counterintuitive. However, the presence of heavy vehicles could induce more cautious or conservative driving behavior among other road users, thereby offsetting some of their inherent risk (Shabab et al., 2024). This behavioral adaptation — where drivers self-regulate speed, increase following distance, or reduce overtaking around large trucks — may temporarily lower the overall

Table 6
Net effects of variables in mixed spline indicator pooled random parameters negative binomial model with gradient boosting method.

	2011	2012	2013	2014	2015	2016	2017	2018
Constant	−1.756	−3.512	−5.268	−7.024	−8.780	−10.53	−12.29	−15.27
Ln(MADT)	0	0	0	1.058	2.116	1.801	1.486	1.171
Percent of heavy vehicles	0.215	0.43	0.645	−0.58	−1.805	−0.703	0.399	1.501
Inverse absolute grade	33.660	33.660	33.660	33.660	33.660	33.660	33.660	33.660
Pavement Condition Index	−0.015	−0.015	−0.015	−0.015	−0.015	−0.015	−0.015	−0.015
Inner Shoulder type	−0.387	−0.387	−0.387	−0.387	−0.387	−0.387	−0.387	−0.387
Major speeding (instrumented)								
Mean	2.058	2.058	2.058	2.058	2.058	2.058	2.058	2.058
Variance over locations	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006
Logarithm of dispersion parameter	−0.876							

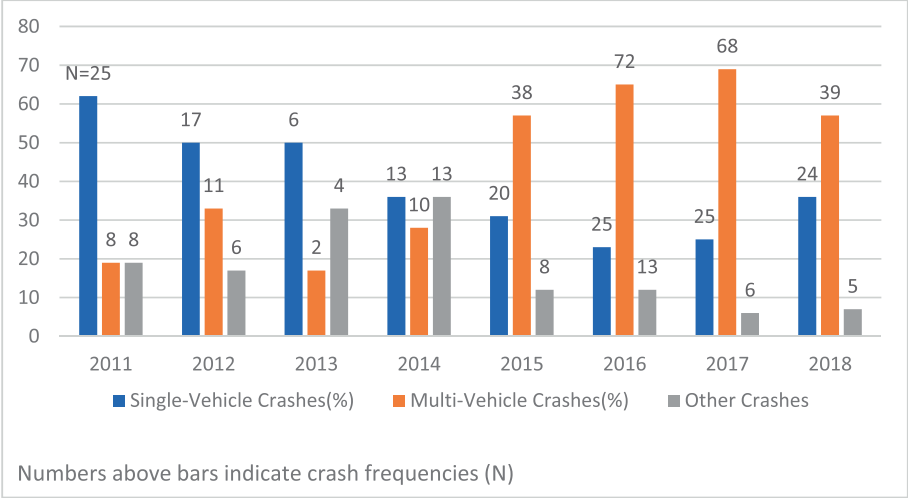

Fig. 7. Sample share and frequency of various crash types across different years in the study area (the numbers above bars indicate crash frequencies).

crash rates. Alternatively, changes in fleet composition, enhanced training or enforcement targeting commercial drivers, or changes in crash type distributions could have moderated their contribution during those years.

The inverse of the absolute grade in the negative binomial model has a statistically significant and positive parameter (33.660 with p-value < 0.001), indicating that as the inverse of the absolute grade increases, the expected number of crashes increases too. This inverse relationship suggests that as the absolute grade increases, the gradient of the crash likelihood decreases, implying that while increasing the steepness of the road increases crash likelihood, this relationship is not linear, and the effect becomes less pronounced at higher grades. This finding indicates that steeper roads experience a higher frequency of crashes compared to flat segments. One plausible reason for this finding is that increased road steepness generally leads to reduced vehicle speeds. This finding aligns with the results from the GBM model too, which highlighted a positive effect of absolute grade on predicting speeding. As the grade becomes steeper, vehicles naturally reduce their speed to navigate the incline, thereby lowering the likelihood of major speeding. Consequently, on steep grades, the resulting lower speeds contribute to a reduced overall crash frequency compared with the segments with lower absolute grades. This finding underscores the importance of considering both the geometric characteristics of the road and driver behaviour in assessing crash frequency. While steeper grades might intuitively suggest higher crash frequency due to increased difficulty in navigation, the reduction in vehicle speed on inclined segments can mitigate this risk.

Paved inner shoulder was found to have a consistent and statistically significant negative association with crash frequency across all years, as evidenced by its stable coefficient (− 0.387) in the model. This negative coefficient suggests that paved inner shoulders contribute to lower crash occurrences compared to unpaved ones, likely due to their capacity to enhance roadway recovery zones and allow greater maneuverability in emergency situations. Paved shoulders may also improve driver perception of lane boundaries and increase operational space, particularly in high-speed environments. This finding is consistent with prior studies which suggest that geometric enhancements — such as paved shoulders — provide critical margins for error that can prevent both run-off-road and sideswipe collisions (Bisht & Tiwari, 2022; Hallmark et al., 2009).

Pavement condition index is negatively associated with crash frequency, indicating that improved pavement conditions is associated with fewer crash occurrences. Poorer pavement quality, associated with lower PCI, could potentially compromise vehicular stability, thereby increasing crash frequency. This association highlights the importance of maintaining high-quality road surfaces to promote safety and aligns with the results from prior studies (Elghriany, 2016; P. Sadeghi & Goli, 2024).

Finally, the parameter estimate for major speeding is 2.058, indicating a positive and statistically significant association with crash frequency. This result implies that higher rates of major speeding lead to an increased number of crashes. Speeding is a well-established risk factor in crash occurrence, as it reduces reaction time, increases stopping distances, and heightens crash severity upon impact. Furthermore, the variance of the parameter across locations suggests that unobserved location-specific factors influence the degree to which speeding contributes to crash risk. These factors may include environmental conditions such as vegetation density, which affects drivers' speed choices by altering visibility, or weather conditions that influence road friction and stopping distances. This finding aligns well with extensive evidence showing that speeding, particularly at major levels, exacerbates both the frequency and severity of crashes (Alnawmasi & Mannering, 2022; Yasmin et al., 2022) and underscores the importance of major speeding

which is often linked with aggressive driving behaviours. It highlights the need for context-specific speed management strategies that account for localized conditions to effectively reduce crash likelihood.

4.3. Statistical evidence for endogeneity between speeding and crash occurrence

In the **Introduction** section, we presented a theoretical argument behind endogeneity between speeding and crash occurrence. To provide statistical evidence for this endogeneity (and hence the need for the 2-stage hybrid model), we estimated the mixed spline indicator pooled random parameters negative binomial model with observed speeding proportions used directly in the mean function of the negative binomial distribution. The parameter of these speeding proportions was then compared with those of the hybrid model using the Durbin-Wu-Hausman statistical test. The result of test is used to demonstrate that speeding is correlated with the error term of the crash frequency model. The Durbin-Wu-Hausman test is the most widely used formal test for showing endogeneity (Patrick, 2021). It compares the difference between estimates from an instrumental variable model and a regular model (the hybrid model and the negative binomial model, in our study). The full estimation results of the negative binomial model with observed speeding are presented in the Appendix. However, the parameter of the speeding variable (major speeding proportion) in the negative binomial model is 1.46 with standard error 1.01. Using these estimates as well as those of the hybrid model (2.058 with standard error 0.966) in the Durbin-Wu-Hausman test, the p-value of the test was found to be 0.042, rejecting the null hypothesis on lack of endogeneity. Therefore, and in addition to the theoretical hypothesis behind the endogeneity between speeding and crash occurrence, there is statistical evidence too for this endogeneity which motivated us to develop the hybrid model.

4.4. Model comparison

To better highlight the significance of our hybrid model, we also estimated a state-of-the-art instrumental variable negative binomial model as well as its extension with spline indicator variables and compared their parameter estimates and statistical fit with those of our hybrid model. The final model candidates for comparison are:

Model #1: a mixed spline indicator pooled random parameters negative binomial model with observed speeding proportions directly in the model (MSIPRPNB–Observed); the model that was used in the Durbin-Wu-Hausman test;

Model #2: an instrumental variable random parameters negative binomial ordered probit fractional split model (PRPNB–OPFS); an instrumental variable model in which we regress the speeding variables (major, moderate, and minor speeding proportions) against all exogenous variables and then use their predicted values in the crash frequency model. Ordered probit factional split model (Bhowmik et al., 2019a) is used in the first stage of this model for predicting speeding proportions;

Model #3: an instrumental variable mixed spline indicator pooled random parameters negative binomial ordered probit fractional split model (MSIPRPNB–OPFS); an extension of model #1 including the spline indicator variables for capturing temporal instability; and.

Model #4: a mixed spline indicator pooled random parameters negative binomial model with gradient boosting method (MSIPRPNB–GBM); the full proposed hybrid model including spline indicator variables for capturing temporal instability.

Model #1 serves as the baseline for comparing the parameter of speeding variable. The comparison between models #2 and #3 determines the suitability of the spline indicator variables in capturing temporal instability whereas the comparison between models #3 and #4 determines the suitability of the proposed hybrid framework in capturing endogeneity.

Full estimation results of models #2 and #3 are presented in the Appendix (Tables A3-A5) too. In model #2, the ordered probit fractional split model in the first stage was used for modelling proportions of the discrete ordered speeding variable (Bhowmik et al., 2019a) aligning well with the hypothesis in this study. The statistically significant variables in this model (Table A3) were mostly the same as the features in the GBM within the hybrid model, and with intuitive parameter estimates. However, its predictive accuracy is notably lower ($R^2 = 0.27$) than the GBM ($R^2 = 0.842$), highlighting a significant limitation for its use in a two-stage instrumental variable crash frequency modelling. The parameter estimates of the random parameters negaive binomial model in the second stage of this model (Table A4) are intuitive too and similar to those of the hybrid model.

In model #3, the same ordered fractional model as in model #2 was used in the first stage but the random parameters model in the second stage was extended and included the spline indicator variables for capturing temporal instability. The parameter estimates of the negative binomial model in the second stage of model #3 (Table A5) reveal a similar pattern to those of model #2 and the hybrid model regarding the statistically significant variables and their effects on speeding and crash frequency. However, the key distinction between them lies in the different estimates for the parameter of the major speeding proportion variable.

The parameter of major speeding proportions in models #2, #3 and #4 are 3.321, 3.94 and 2.058, respectively. The same parameter in model #1 is 1.46 – which is believed to be biased due to endogeneity. To better understand the differences between these parameters, we simulated their distribution and compared their kernel densities. Under the Central Limit Theorem (Kwak and Kim, 2017), we can assume:

$$\overline{\omega}\ N(\overline{\omega}, SE(\overline{\omega}))$$

where $\overline{\omega}$ is the parameter estimate for the speeding variable from any of the models, and $SE(\overline{\omega})$ is its standard error. The corresponding kernel densities of these parameters (Fig. 8) demonstrate that the variance of the distribution from the hybrid model is much lower than that of the other models indicating that this model provides a more efficient parameter estimate for speeding. Moreover, the distribution from the model with observed speeding (red curve) is centered furthest to the left, followed closely by the hybrid model (green curve), both of which exhibit relatively sharp peaks and narrower spreads, suggesting more precise estimates. The instrumental variable model without temporal instability (orange curve) appears more dispersed and centered further to the right, whereas the instrumental variable model with temporal instability (blue curve) has the widest distribution and is centered at the highest values among all models. This finding indicates that the two modelling methodologies (instrumental variable modelling and hybrid modelling) may mitigate the endogeneity bias in the parameter for speeding and yield a more accurate estimate for that parameter. Since the true value for this parameter is not known, it is not directly possible to select the best model. However, a comparison of the statistical fit between the models may shed more light on these findings.

The statistical fit measures between the four model candidates (Table 7) show that the hybrid model has a lower AIC and BIC (618.90 and 647.25, respectively) in comparison with the rest of the models

Table 7
Goodness of fit measures between the hybrid and the conventional models.

	AIC	BIC	log-likelihood
Model #1*	624.62	666.95	−298.30
Model #2**	650.06	668.20	−318.10
Model #3***	632.64	668.92	−304.31
Model #4****	618.90	647.25	−295.44

* Negative binomial model with temporal instability and observed speeding.
** instrumental variable model without temporal instability.
*** instrumental variable model with temporal instability.
**** hybrid model with temporal instability.

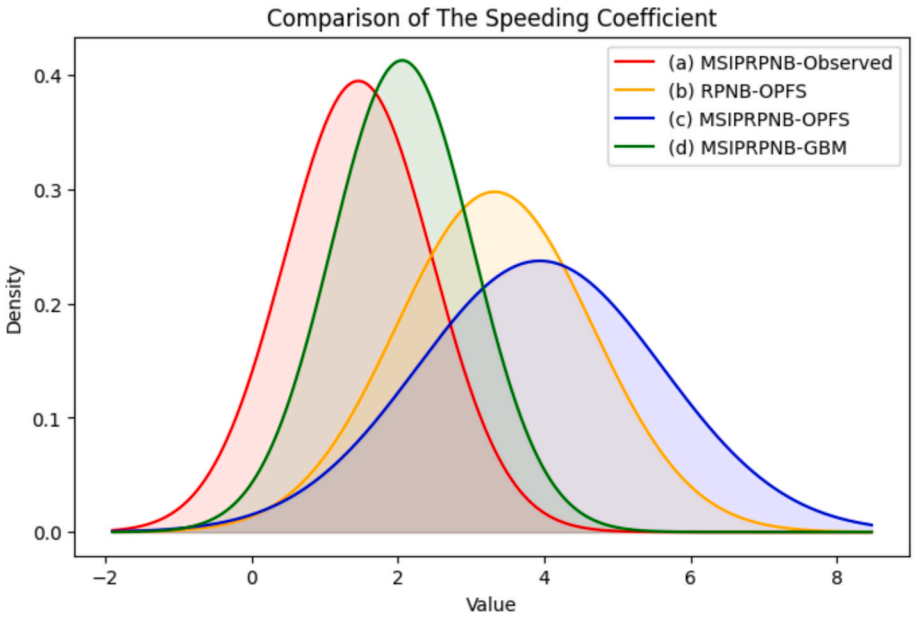

Fig. 8. Kernel densities of the parameter for major speeding proportions in four different models: (a) mixed spline indicator pooled random parameters negative binomial model with observed speeding proportions directly in the model (MSIPRPNB–Observed) (red), (b) instrumental variable random parameters negative binomial ordered probit fractional split model (PRPNB–OPFS) (orange), (c) an instrumental variable mixed spline indicator pooled random parameters negative binomial ordered probit fractional split model (MSIPRPNB–OPFS) (blue) and (d) mixed spline indicator pooled random parameters negative binomial model with gradient boosting method (MSIPRPNB–GBM) (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

demonstrating a superior goodness of fit.

The above findings all together show that the hybrid framework with GBM in the first stage and mixed spline indicator specification of the negative binomial model in the second stage outperforms the existing models for the data in this study. These results further highlight the strength of the machine learning approach in addressing endogeneity and support the growing body of literature that advocates for the use of machine learning techniques combined with econometric methods to overcome the traditional limitations of the later methods in transport safety modeling (Afghari et al., 2022; Jin et al., 2023; Zhou et al., 2024b).

5. Conclusions

Crash risk along horizontal curves, measured by the frequency of crashes at these locations, is influenced by a complex interplay of roadway geometry, traffic characteristics, and driver behaviour, particularly speeding. Addressing endogeneity between speeding behaviour and crash risk remains a key challenge in crash modeling. In addition, traditional data collection methods for collecting speeding data often suffer from measurement error because of driver adaptation with speed cameras. This study aimed to address these gaps by introducing a novel framework, consisting of speeding data collection by unmanned vehicles and using these data in a hybrid mixed spline indicator pooled random parameters negative binomial model with gradient boosting method, for determining the effect of speeding on crash frequency.

Empirical testing of the proposed hybrid model and comparing it with the existing instrumental variable models demonstrated its substantial benefit, in terms of finding suitable instrumental variables for addressing endogeneity and enhancing prediction accuracy. Shapely values of the predictors in the gradient boosting model as well as their feature importance painted an explainable picture of the variables and their impact on predicting speeding. These values were consistent with statistically significant variables in the conventional fractional split model and aligned closely with the findings of (M. Islam et al., 2024), where shapely values were employed for variable selection. This consistency reinforces the robustness of SHAP-based feature importance analysis, highlighting its reliability in determining critical predictors within complex traffic safety models. While machine learning models have high predictive power, they may be prone to overfitting too, capturing noise rather than true causal variation, which can lead to misleading estimates. A highly accurate model may replicate observed values without adequately addressing endogeneity. We demonstrated that further investigations (such as SHAP analysis, feature importance and Durbin-Wu-Hausman test) should be conducted to determine an appropriate balance between accuracy and endogeneity.

Incorporation of random parameters varying across locations into the hybrid model further enhanced the model's ability to capture localized variability. The variance associated with speeding, for example, illustrated the importance of considering site-specific factors in crash prediction models. Similarly, the temporal instability observed in the variability of the vehicular and heavy vehicle traffic over time emphasized the dynamic nature of traffic patterns and their influence on crash likelihoods. The incorporation of temporal instability and heterogeneity not only improves model accuracy but also provides a more nuanced understanding of the factors influencing crash frequency, reinforcing the need for adaptive and context-specific models in transport safety research.

Overall, our study showed that collecting speed via unmanned vehicles and modelling it using a hybrid statistical-machine learning approach allows for a more reliable and robust estimation of the complex relationship between speeding and crash frequency and may ultimately lead to more effective road safety interventions and policy recommendations.

Despite the above benefits, our study has limitations too. The extrapolation method that was used to create yearly speeding data assumes that the trend (rate of change) of such data remains constant beyond the observed 28 days in the study. However, this assumption may not hold, especially over extended time periods. As such, the effects of speeding on crash risk should be interpreted with caution. In addition, in modelling speeding and crash risk, we primarily focused on roadway

characteristics and traffic-related factors; however, incorporating human factors, such as visual perception, and cognitive workload, and linking them with road segments could further enhance predictive accuracy of the models. Furthermore, while our model specification allowed for temporal instability in the effects of all factors on crash risk, our sample data showed such instability only in the effects of traffic characteristics. Future research should collect more detailed (behavioral) data over extended periods of time to be able to show temporal instability in other factors too. Future research may also extend this modeling framework to different roadway environments to provide a more comprehensive understanding of crash risk.

Finally, the reliability of crash data in our study area (Iran) may have been affected by under-reporting (of minor crashes or crashes which may not have been documented because there was no severe injuries or substantial property damage), data accuracy and inconsistencies in data collection. These issues can lead to an incomplete understanding of road safety problems and hinder the development of effective policy recommendations. Furthermore, variations in reporting protocols between the law enforcement agency and healthcare facilities in Iran may have created discrepancies in crash records. Future research should address these challenges using improved data integration across agencies and implementing more comprehensive crash surveillance systems.

CRediT authorship contribution statement

Sajad Asadi Ghalehni: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Amir Pooyan Afghari:** Writing – review & editing, Supervision, Methodology, Conceptualization.

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.

Acknowledgement

None

Appendix

Table A1
Hyperparameters of the XGBoost model used in this study.

Hyperparameter	Description	Default Value	Value in this study
objective	Loss function to optimize	reg: squarederror	reg: gamma
eval_metric	Evaluation criterion	None (training loss)	rmse
max_depth	Maximum depth of trees	6	32
learning_rate (eta)	Step size shrinkage	0.3	0.1
subsample	Fraction of samples per tree	1.0	0.8
colsample_bytree	Fraction of features per tree	1.0	0.8
num_boost_round	Number of boosting iterations	10 (if not specified)	1000
early_stopping_rounds	Stop if no improvement after N rounds	Not enabled	100
min_child_weight	Minimum sum of instance weight in a child	1	1 (default)
gamma	Minimum loss reduction for a split	0	0 (default)
lambda (L2 reg)	L2 regularization term	1	1 (default)
alpha (L1 reg)	L1 regularization term	0	0 (default)
scale_pos_weight	Weight balance for imbalanced classes	1	1 (default)
tree_method	Tree construction algorithm	auto	auto (default)
nthread	Number of parallel threads	All cores	All cores (default)
seed (random_state)	Random seed for reproducibility	0	42 (via train_test_split)

Table A2
Estimation results of the mixed spline indicator pooled random parameters negative binomial model with observed speeding (endogenous variable) directly in the mean function of the negative binomial distribution.

	2011	2012	2013	2014	2015	2016	2017	2018
Constant	−1.768 (−5.63)	–	–	–	–	–		−1.167 (−2.12)
Ln(MADT)	–	–	–	1.075 (7.24)	–	−1.412 (−5.61)	–	–
Percent of heavy vehicles	0.216 (5.15)	–	–	−1.466 (−6.57)	–	2.387 (5.56)		–
Inverse absolute grade**	32.918 (4.35)	–	–	–	–	–	–	–
Pavement Condition Index	−0.012 (−1.84)	–	–	–	–	–	–	–
Inner shoulder type	−0.416 (−2.28)	–	–	–	–	–	–	–
Major speeding (observed)								
mean	1.465 (1.450)	–	–	–	–	–	–	–
variance	0.006 (16.53)							
Logarithm of dispersion parameter	−0.880 (−2.00)	–	–	–	–	–	–	–

*Numbers in brackets present the corresponding Z-values.
**The absolute grade was offset by 10 values to compensate the instances where the original grade was zero.

Table A3
Estimation results of the ordered probit fractional split model in the first stage of the instrumental variable model for speeding and crash frequency.

	Coefficient	Standard error	t-value	P>|z|
Threshold minor speeding | moderate speeding	−49.822	21.916	−2.273	0.023
Threshold moderate speeding | major speeding	−48.565	21.926	−2.215	0.026
Sufficient stopping sight distance	−421.345	14.042	−30.005	0.000
Before curve existence	−218.514	7.527	−29.028	0.000
Deflection angle	−6.281	0.223	−28.129	0.000
Outer shoulder width	−75.781	2.713	−27.933	0.000
Radius	171.123	6.401	26.732	0.000
Widening	−10.902	0.408	−26.712	0.000
Curve type	−148.386	5.609	−26.452	0.000
Outer shoulder type	−4.450	0.176	−25.218	0.000
Center line marking quality	320.249	13.239	24.190	0.000
MADT	1.730	0.071	24.242	0.000
Curve length	3.616	0.161	22.381	0.000
Percent of heavy vehicles	153.404	6.850	22.393	0.000
Degree of curve	9.643	0.421	−22.878	0.000
Inner shoulder width	−67.961	3.302	−20.581	0.000
Curve direction	−48.282	2.330	−20.717	0.000
Pavement condition index	25.085	3.463	7.243	0.000
Predictive performance				
R^2	0.27			

Table A4
Estimation results of the random parameters negaive binomial model in the second stage of the instrumental variable model for speeding and crash frequency.

	Coefficient	Standard error	z-value	P>|z|	95 % confidence interval	
Constant	−15.614	5.304	−2.94	0.003	−26.010	−5.218
Ln(MADT)						
Mean	1.995	0.705	2.83	0.005	0.612	3.378
Variance over locations	0.002	0.001	2.00	0.046	0.000	0.011
Inverse absolute grade*	27.461	9.371	2.93	0.003	9.093	45.829
Pavement condition index	−0.018	0.008	−2.14	0.033	−0.035	−0.001
Major speeding (instrumented)	3.321	1.339	2.48	0.013	0.694	5.947
Logarithm of dispersion parameter	0.592	0.171	3.46	0.000	0.256	0.927

*The absolute grade was offset by 10 values to compensate the instances where the original grade was zero.

Table A5
Estimation results of the mixed spline indicator pooled negative binomial random parameters negative binomial model with base and deviation effect of variables for speeding and crash frequency.

	2011	2012	2013	2014	2015	2016	2017	2018
Time variant variables								
Constant	0.208 (0.23)*	–	–	–	–	–	–	–
Ln(MADT)	−0.228 (−3.96)		–	0.674 (3.99)	−0.336 (−3.03)	–	0.450 (2.86)	–
Percent of heavy vehicles	0.159 (3.23)	–	–	−0.349 (−2.45)	–	–	0.759 (2.82)	–
Time invariant variables								
Inverse absolute grade**	24.598 (2.95)	–	–	–	–	–	–	–
Major speeding (instrumented)								
Mean	3.946 (2.34)							
Variance over locations	0.001 (3.01)							
Logarithm of dispersion parameter	0.240 (2.30)	–	–	–	–	–	–	–

*Numbers in brackets present the corresponding Z-values.
**The absolute grade was offset by 10 values to compensate the instances where the original grade was zero.

Data availability

The data that has been used is confidential.

References

Abdel-Aty, M.A., Radwan, A.E., 2000. Modeling traffic accident occurrence and involvement. Accid. Anal. Prev. 32 (5), 633–642.

Afghari, A.P., Haque, M.M., Washington, S., 2018a. Applying fractional split model to examine the effects of roadway geometric and traffic characteristics on speeding behavior. Traffic Inj. Prev. 19 (8), 860–866.

Afghari, A.P., Haque, M.M., Washington, S., Smyth, T., 2016. Bayesian latent class safety performance function for identifying motor vehicle crash black spots. Transp. Res. Rec. 2601 (1), 90–98.

Afghari, A.P., Imani, A.F., Papadimitriou, E., van Gelder, P., Hezaveh, A.M., 2021. Disentangling the effects of unobserved factors on seatbelt use choices in multi-occupant vehicles. Journal of Choice Modelling 41, 100324.

Afghari, A.P., Papadimitriou, E., Pilkington-Cheney, F., Filtness, A., Brijs, T., Brijs, K., Cuenen, A., De Vos, B., Dirix, H., Ross, V., 2022. Investigating the effects of sleepiness in truck drivers on their headway: an instrumental variable model with grouped random parameters and heterogeneity in their means. Anal. Methods Accid. Res 36, 100241.

Afghari, A.P., Vos, J., Farah, H., Papadimitriou, E., 2023. "I did not see that coming": a latent variable structural equation model for understanding the effect of road predictability on crashes along horizontal curves. Accid. Anal. Prev. 187, 107075.

Afghari, A.P., Washington, S., Haque, M.M., Li, Z., 2018b. A comprehensive joint econometric model of motor vehicle crashes arising from multiple sources of risk. Anal. Methods Accid. Res 18, 1–14.

Afghari, A.P., Washington, S., Prato, C., Haque, M.M., 2019. Contrasting case-wise deletion with multiple imputation and latent variable approaches to dealing with missing observations in count regression models. Anal. Methods Accid. Res 24, 100104.

Afukaar, F.K., Antwi, P., Ofosu-Amaah, S., 2003. Pattern of road traffic injuries in Ghana: implications for control. Inj. Control Saf. Promot. 10 (1–2), 69–76.
Ali, A. T., Flannery, A., & Venigalla, M. M. (2007). *Prediction models for free flow speed on urban streets*.
Alnawmasi, N., Mannering, F., 2022. The impact of higher speed limits on the frequency and severity of freeway crashes: Accounting for temporal shifts and unobserved heterogeneity. Anal. Methods Accid. Res 34. https://doi.org/10.1016/j.amar.2021.100205.
Alnawmasi, N., Mannering, F., 2023. An analysis of day and night bicyclist injury severities in vehicle/bicycle crashes: a comparison of unconstrained and partially constrained temporal modeling approaches. Anal. Methods Accid. Res 40, 100301.
American Association of State and Highway Transportation Officials. (2018). *A Policy on Geometric Design of Highways and Streets*. Publication Code: GDHS-7 ISBN: 978-1-56051-676-7.
Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accid. Anal. Prev. 41 (1), 153–159.
Anastasopoulos, P.C., Mannering, F.L., 2016. The effect of speed limits on drivers' choice of speed: a random parameters seemingly unrelated equations approach. Anal. Methods Accid. Res 10, 1–11.
Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Anal. Methods Accid. Res 9, 1–15.
Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016. Simple online and realtime tracking. IEEE International Conference on Image Processing (ICIP) 2016, 3464–3468.
Bhowmik, T., Yasmin, S., Eluru, N., 2019a. A multilevel generalized ordered probit fractional split model for analyzing vehicle speed. Anal. Methods Accid. Res 21, 13–31.
Bhowmik, T., Yasmin, S., Eluru, N., 2019b. Do we need multivariate modeling approaches to model crash frequency by crash types? a panel mixed approach to modeling crash frequency by crash types. Anal. Methods Accid. Res 24, 100107.
Bisht, L.S., Tiwari, G., 2022. Safety effects of paved shoulder width on a four-lane divided rural highway in India: a matched case-control study. Saf. Sci. 147, 105606.
Boehmke, B., Greenwell, B.M., 2019. *Hands-on machine learning with R*. Chapman and Hall/CRC.
Calvi, A., 2015. A study on driving performance along horizontal curves of rural roads. Journal of Transportation Safety & Security 7 (3), 243–267.
Chen, S., Saeed, T.U., Labi, S., 2017. Impact of road-surface condition on rural highway safety: a multivariate random parameters negative binomial approach. Anal. Methods Accid. Res 16, 75–89.
Chevalier, A., Coxon, K., Chevalier, A.J., Wall, J., Brown, J., Clarke, E., Ivers, R., Keay, L., 2016. Exploration of older drivers' speeding behaviour. Transport. Res. F: Traffic Psychol. Behav. 42, 532–543.
Choudhari, T., Maji, A., 2019. Effect of horizontal curve geometry on the maximum speed reduction: a driving simulator-based study. Transp. Dev. Econ. 5 (2), 1–8.
Coruh, E., Bilgic, A., Tortum, A., 2015. Accident analysis with aggregated data: the random parameters negative binomial panel count data model. Anal. Methods Accid. Res 7, 37–49.
Council, F.M., Reurings, M., Srinivasan, R., Masten, S., Carter, D., 2010. *Development of a Speeding-Related Crash Typology*. Turner-Fairbank Highway Research Center.
Dhungana, P., Qu, M., 2005. The risks of driving on roadways with 50 miles per hour posted speed limit. J. Saf. Res. 36 (5), 501–504.
Dronova, O., Parinov, D., Soloviev, B., Kasumova, D., Kochetkov, E., Medvedeva, O., Sergeeva, I., 2022. Unmanned aerial vehicles as element of road traffic safety monitoring. Transp. Res. Procedia 63, 2308–2314.
Dzinyela, R., Shirazi, M., Das, S., Lord, D., 2024. The negative Binomial-Lindley model with Time-Dependent Parameters: Accounting for temporal variations and excess zero observations in crash data. Accid. Anal. Prev. 207, 107711.
Elghriany, A.F., 2016. *Investigating correlations of pavement conditions with crash rates on in-Service US highways*. University of Akron.
Elvik, R., Christensen, P., Amundsen, A., 2004. Speed and road accidents. An evaluation of the Power Model. Speed and Road Accidents, An Evaluation of the Power Model, pp. 3–7.
Elvik, R., Erke, A., Christensen, P., 2009. Elementary units of exposure. Transp. Res. Rec. 2103 (1), 25–31.
Fitzsimmons, E. J., Asce, A. M., Souleyrette, R. R., Asce, M., Nambisan, S. S., & Asce, M. (2013). *Measuring Horizontal Curve Vehicle Trajectories and Speed Profiles : Pneumatic Road Tube and Video Methods*. *March*, 255–265. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000501.
Fu, X., Liu, J., Jones, S., Barnett, T., Khattak, A.J., 2022. From the past to the future: Modeling the temporal instability of safety performance functions. Accid. Anal. Prev. 167, 106592.
Geedipally, S.R., Lord, D., 2010. Investigating the effect of modeling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson–gamma models. Accid. Anal. Prev. 42 (4), 1273–1282.
Ghalehni, S.A., Boroujerdian, A.M., 2023. Model of encroachment into opposite lanes in horizontal curves of rural roads. IATSS Res.
Gheorghiu, A., Delhomme, P., Felonneau, M.L., 2015. Peer pressure and risk taking in young drivers' speeding behavior. Transport. Res. F: Traffic Psychol. Behav. 35, 101–111.
Hallmark, S. L., McDonald, T. J., Tian, Y., & Andersen, D. J. (2009). *Safety benefits of paved shoulders*.
Heydari, S., 2018. A flexible discrete density random parameters model for count data: Embracing unobserved heterogeneity in highway safety analysis. Anal. Methods Accid. Res 20, 68–80.
Heydari, S., Forrest, M., 2024. Estimating the effect of proximity to school on cyclist safety using a simultaneous-equations model with heterogeneity in covariance to address potential endogeneity. Anal. Methods Accid. Res 41, 100318.
Heydari, S., Fu, L., Thakali, L., Joseph, L., 2018. Benchmarking regions using a heteroskedastic grouped random parameters model with heterogeneity in mean and variance: applications to grade crossing safety analysis. Anal. Methods Accid. Res 19, 33–48.
Huo, X., Leng, J., Hou, Q., Zheng, L., Zhao, L., 2020. Assessing the explanatory and predictive performance of a random parameters count model with heterogeneity in means and variances. Accid. Anal. Prev. 147, 105759.
Hussain, F., Li, Y., Arun, A., Haque, M.M., 2022. A hybrid modelling framework of machine learning and extreme value theory for crash risk estimation using traffic conflicts. Anal. Methods Accid. Res 36, 100248.
Islam, A.S.M.M., Shirazi, M., Lord, D., 2023. Grouped Random Parameters negative Binomial-Lindley for accounting unobserved heterogeneity in crash data with preponderant zero observations. Anal. Methods Accid. Res 37, 100255.
Islam, M., Hosseini, P., Kakhani, A., Jalayer, M., Patel, D., 2024. Unveiling the risks of speeding behavior by investigating the dynamics of driver injury severity through advanced analytics. Sci. Rep. 14 (1), 22431.
Jessberger, S., Krile, R., Schroeder, J., Todt, F., Feng, J., 2016. Improved annual average daily traffic estimation processes. Transp. Res. Rec. 2593 (1), 103–109.
Jin, J., Huang, H., Yuan, C., Li, Y., Zou, G., Xue, H., 2023. Real-time crash risk prediction in freeway tunnels considering features interaction and unobserved heterogeneity: a two-stage deep learning modeling framework. Anal. Methods Accid. Res 40, 100306.
Karimi, A., Boroujerdian, A.M., 2021. Explanatory analysis of the safety of short passing zones on two-lane rural highways. Transp. Res. Rec. 2675 (4), 320–330. https://doi.org/10.1177/0361198120980436.
Kim, S.H., 2023. How heterogeneity has been examined in transportation safety analysis: a review of latent class modeling applications. Anal. Methods Accid. Res 100292.
Kwak, S.G., Kim, J.H., 2017. Central limit theorem: the cornerstone of modern statistics. Korean journal of anesthesiology 70 (2), 144.
Li, Z., Chen, C., Wu, Q., Zhang, G., Liu, C., Prevedouros, P.D., Ma, D.T., 2018. Exploring driver injury severity patterns and causes in low visibility related single-vehicle crashes using a finite mixture random parameters model. Anal. Methods Accid. Res 20, 1–14.
Liu, C., & Chen, C.-L. (2009). *An analysis of speeding-related crashes: definitions and the effects of road environments*.
Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Anal. Methods Accid. Res 17, 1–13.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Anal. Methods Accid. Res 1, 1–22.
Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res 11, 1–16.
Marciano, H., Norman, J., 2015. Overt vs. covert speed cameras in combination with delayed vs. immediate feedback to the offender. Accid. Anal. Prev. 79, 231–240.
Marcoux, R., Pervaz, S., Eluru, N., 2024. Assessing non-motorist safety in motor vehicle crashes–a copula-based approach to jointly estimate crash location type and injury severity. Anal. Methods Accid. Res 42, 100322.
Mohammadi, M.A., Samaranayake, V.A., Bham, G.H., 2014. Crash frequency modeling using negative binomial models: an application of generalized estimating equation to longitudinal data. Anal. Methods Accid. Res 2, 52–69.
Nakamura, A., Nakamura, M., 1998. Model specification and endogeneity. J. Econ. 83 (1–2), 213–237.
Nassiri, H., Mohammadpour, S.I., 2023. Investigating speed-safety association: considering the unobserved heterogeneity and human factors mediation effects. PLoS One 18 (2), e0281951.
Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Front. Neurorob. 7, 21.
Oviedo-Trespalacios, O., Afghari, A.P., Haque, M.M., 2020. A hierarchical Bayesian multivariate ordered model of distracted drivers' decision to initiate risk-compensating behaviour. Anal. Methods Accid. Res 26, 100121.
Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accid. Anal. Prev. 41 (4), 683–691.
Patrick, R.H., 2021. Durbin–wu–hausman specification tests. In *Handbook of financial econometrics, mathematics, statistics, and machine learning*. World Scientific, pp. 1075–1108.
Perez, M.A., Sears, E., Valente, J.T., Huang, W., Sudweeks, J., 2021. Factors modifying the likelihood of speeding behaviors based on naturalistic driving data. Accident Analysis & Prevention 159, 106267.
Pervaz, S., Bhowmik, T., Eluru, N., 2024. An integrated multi-resolution framework for jointly estimating crash type and crash severity. Anal. Methods Accid. Res 42, 100321.
Petridou, E., Moustaki, M., 2000. Human factors in the causation of road traffic crashes. Eur. J. Epidemiol. 16, 819–826.
Phuksuksakul, N., Eluru, N., Haque, M.M., Yasmin, S., 2025. Econometric approaches to examine the onset and duration of temporal variations in pedestrian and bicyclist injury severity analysis. Anal. Methods Accid. Res 45, 100362.
Pratt, M.P., Geedipally, S.R., Dadashova, B., Wu, L., Shirazi, M., 2019. Familiar versus unfamiliar drivers on curves: Naturalistic data study. Transp. Res. Rec. 2673 (6), 225–235.
Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788.
Sadeghi, M., Aghabayk, K., Quddus, M., 2024. A hybrid machine learning and statistical modeling approach for analyzing the crash severity of mobility scooter users considering temporal instability. Accid. Anal. Prev. 206, 107696.

Sadeghi, P., Goli, A., 2024. Investigating the impact of pavement condition and weather characteristics on road accidents. Int. J. Crashworthiness 1–17.
Shabab, K.R., Bhowmik, T., Zaki, M.H., Eluru, N., 2024. A systematic unified approach for addressing temporal instability in road safety analysis. Anal. Methods Accid. Res 43, 100335.
Shaon, M.R.R., Qin, X., Shirazi, M., Lord, D., Geedipally, S.R., 2018. Developing a Random Parameters negative Binomial-Lindley Model to analyze highly over-dispersed crash count data. Anal. Methods Accid. Res 18, 33–44.
Soole, D.W., Watson, B.C., Fleiter, J.J., 2013. Effects of average speed enforcement on speed compliance and crashes: a review of the literature. Accid. Anal. Prev. 54, 46–56.
Washington, S., Karlaftis, M.G., Mannering, F., Anastasopoulos, P., 2020. *Statistical and econometric methods for transportation data analysis*. CRC Press.
Welch, G., & Bishop, G. (1995). *An introduction to the Kalman filter*.
Wen, X., Xie, Y., Wu, L., Jiang, L., 2021. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. Accid. Anal. Prev. 159, 106261.
World Health Organization, 2023. Global status report on road safety 2023: summary. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO.
Wu, L., Lord, D., Geedipally, S.R., 2017. Developing crash modification factors for horizontal curves on rural two-lane undivided highways using a cross-sectional study. Transp. Res. Rec. 2636 (1), 53–61.
Xing, L., He, J., Abdel-Aty, M., Cai, Q., Li, Y., Zheng, O., 2019. Examining traffic conflicts of up stream toll plaza area using vehicles' trajectory data. Accid. Anal. Prev. 125, 174–187.
Yasmin, S., Eluru, N., Haque, M.M., 2022. Addressing endogeneity in modeling speed enforcement, crash risk and crash severity simultaneously. Anal. Methods Accid. Res 36, 100242.
Zhou, Y., Fu, C., Jiang, X., 2024a. Multi-dimensional unobserved heterogeneities: Modeling likelihood of speeding behaviors in different patterns for taxi speeders with mixed distributions, multivariate errors, and jointly correlated random parameters. Anal. Methods Accid. Res 41, 100316.
Zhou, Y., Fu, C., Jiang, X., Liu, H., 2024b. Analyzing the heterogenous effects of factors on high-range speeding likelihood of taxi speeders: does explainable deep learning provides more insights than random parameter approach? Accid. Anal. Prev. 207, 107752.
Zhu, X., Srinivasan, S., 2011. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Accid. Anal. Prev. 43 (1), 49–57.
Zubaidi, H., Alnedawi, A., Obaid, I., Abadi, M.G., 2022. Injury severities from heavy vehicle accidents: an exploratory empirical analysis. Journal of Traffic and Transportation Engineering (english Edition) 9 (6), 991–1002.