Proposed segment length for safety evaluation studies on rural divided highways in India

Bisht, Laxman Singh; Chand, Sai; Tiwari, Geetam

DOI
10.1016/j.iatssr.2024.05.002

Publication date
2024

Document Version
Final published version

Published in
IATSS Research

Citation (APA)
Bisht, L. S., Chand, S., & Tiwari, G. (2024). Proposed segment length for safety evaluation studies on rural divided highways in India. *IATSS Research*, *48*(3), 331-341. https://doi.org/10.1016/j.iatssr.2024.05.002

IATSS Research 48 (2024) 331–341

Contents lists available at ScienceDirect

IATSS Research

journal homepage: www.sciencedirect.com/journal/iatss-research

Research Article

Proposed segment length for safety evaluation studies on rural divided highways in India

Laxman Singh Bisht [a,*], Sai Chand [b], Geetam Tiwari [b]

[a] Department of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, Delft 2628CN, the Netherlands
[b] Assistant Professor, Transportation Research and Injury Prevention Centre (TRIP Centre), Indian Institute of Technology Delhi, Hauz Khas, New Delhi, Delhi 110016, India

ARTICLE INFO

Keywords:
Accident analysis
Crash prediction models
LMICs
Road safety
Safety performance function
Segmentation

ABSTRACT

Determining an appropriate segment length for highway safety evaluations in low- and middle-income countries (LMICs) poses a significant challenge. This study aims to address this issue by recommending a suitable segment length for such evaluations in India, using a 167 km intercity expressway as a case study. We employed negative binomial (NB) models on datasets segmented from 100 m to 1000 m with 100 m increments. Our findings strongly suggest that segment lengths from 300 m to 700-m suit various safety assessments. However, the study reveals that parameter estimates vary significantly with both segment length and sample size. This highlights the sensitivity of parameters to data aggregation and sample size across different segment lengths, making it difficult to identify a single optimal length. Therefore, we propose selecting the segment length and segmentation approach based on specific local conditions, highway context, data availability and quality. The methodology presented here can guide policymakers in LMICs to make informed choices regarding segment length for safety evaluations, including blackspot identification and treatment on their highways.

1. Introduction

In recent years, several low- and middle-income countries (LMICs) have rapidly expanded their road infrastructure with little focus on safety. The burden of road traffic crashes (RTCs) has increased in such countries, although an opposite trend was observed for many infectious diseases [1]. There is a constant demand to improve the safety status of the existing road infrastructure. The Highway Safety Manual (HSM) suggests that the safety evaluation of a geometric design element or a road safety intervention must be carried out using crash prediction models, also commonly known as Safety Performance Functions (SPFs) [2].

The SPFs are generally developed based on highway segment-specific crash frequency or crash severity data [3]. These models are based on the premise that crash data distribution is not entirely random, though dependent on a highway's geometric and traffic characteristics [4]. Thus, one of the major steps in developing SPFs is to divide the highway stretch or network into discrete segments [5]. Choosing an appropriate segmentation approach and length for SPFs is crucial as they can affect the inference about the effectiveness of a road safety intervention [6,7]. It is also often argued that segmentation is critical to generating unbiased estimates of statistical models and, consequently, to correct inferences [8]. Studies have shown that the selection often depends on data availability status and analysts' choice [8,9]. Past research has highlighted that the SPFs recommended by HSM are not transferable to other conditions and settings due to variations in many factors, such as driver behaviour, road environment, data availability and consistency [10]. Therefore, researchers have tried to build local site-specific SPFs [11]. In addition, LMICs need a distinct approach to build SPFs because of the heterogeneous traffic, highly variable roadside environment, different road user behaviour, and vehicle standards [1]. In addition, the activities along and around the highways in LMICs are comparatively high compared to LMICs. The design standards and road safety interventions in HICs may not be effective or infeasible to apply in LMICs. Hence, the SPFs built for LMICs need a customized approach.

In this context, this study is motivated by the challenges faced by a road safety practitioner in LMICs for selecting the adequate segment length in a highway safety evaluation study. In contrast to HICs, LMICs have no guiding document like the HSM available to select segment lengths for crash analysis for road safety practitioners. Further, in LMICs, the complexities of implementing the HSM suggested segment length approaches increase due to issues related to road safety data, such

* Corresponding author.
E-mail address: laxmanlsbisht@gmail.com (L.S. Bisht).

https://doi.org/10.1016/j.iatssr.2024.05.002
Received 14 December 2023; Received in revised form 23 April 2024; Accepted 16 May 2024
Available online 17 May 2024

as low quality, limited availability, inconsistencies, and absence of a central database and non-linking of databases [8,12].

This study aims to recommend an adequate segment length for road agencies in India. As a case study, we collected data on fatal crashes, geometrics, speed, AADT, and roadside environment for an Indian six-lane rural expressway.

2. Literature review

2.1. Overview of segmentation approaches

Roadway segmentation aims to establish a unit of analysis for the safety evaluation [13]. The unit of analysis can be of fixed or variable lengths based on the homogenous geometric, crash data, and traffic characteristics. The roadway segmentation approaches can be divided into three categories as follows [9]:

i. Fixed length-based approach: The highway network is divided into segments with pre-specified fixed length [13].
ii. Homogenous traffic and roadway characteristics-based approach: The highway network is divided into segments based on changes in one or more traffic and roadway characteristics [2,14].
iii. Spatial clustering-based approach: In this method, the segments are divided based on the spatial distribution of the RTCs on the highway network. Techniques such as K-means clustering [9,15,16], and Fisher's method have been used for segmentation [5,17,18]. These methods divide roadway segments into homogeneous groups by grouping segments with similar crash distributions.

The HSM recommends a homogenous segmentation approach with a minimum segment length of approximately 0.10 miles (161 m) and the physical characteristics of the highway [2,12]. The HSM segmentation approaches have been suggested and developed when the network-level data is extensively available [14,19]. However, in LMICs, data availability is limited, and data collection is arduous and time-consuming without a centralized database [30].

2.2. Segmentation approaches in the crash prediction models

Ghadi and Török [9] argue that the accuracy of the Crash Prediction Models (CPMs) and highway safety evaluation methods depends on the accuracy of the crash data distribution on the segmented highways; the more accurate the crash data is segmented, the more accurate the SPF. This will also affect the performance of some safety evaluation methods that rely on CPMs in their criterion [4,20] – [6]. One of the early studies by Thomas [21] suggested that the statistical form of the crash distribution varies as the length of the road segment increases. Fitzpatrick et al. [19] employed a two-step segmentation approach, somewhat like the approach suggested by the HSM. First, they divided the roadway based on the horizontal curve and tangent segments and then predicted the number of midblock crashes for each segment. Second, they identified the segments with intersections. At the same time, a study by Koorey [14] concluded that fixed-length segments are computationally easier to create from constant-interval raw crash data. Besides, the study also recommended minimizing the short segments by not creating new segments when the length is less than 50 m.

Recent studies have extensively used the HSM-suggested segmentation approach. For example, Agostino [22] analyzed the performance of two different methods of segmentation using homogeneous segments with varying lengths based on the HSM approach on a sample of Italian four-lane rural motorways. Cafiso et al. [12] suggested that the homogenous segmentation approach is too complicated and impractical when the available variables are too high. Therefore, the fixed segmentation approach is most feasible in such cases. M. Ghadi & Török [9] opined that the effectiveness of the segmentation approaches determines the accuracy of the developed SPFs. The study concluded that the performance of SPFs changed as the segmentation approach changed. Elagamy et al. [20,23] concluded that the segmentation approach based on the presence of curvatures was best for total crashes, considering the time trend. However, the fixed-length segmentation approach was the best for fatal and injury crashes and property damage-only crashes without considering the time trend.

A Brazilian study by Silva et al. [6] emphasized that the longer segments minimize the effects related to the improper and uncertain georeferencing of the crash data. Tahir et al. [5] concluded that the HSM and fixed segmentation (1 km) were the best methods for predicting crash modification factors. Bartin et al. [10] developed the SPF based on the HSM segmentation approach with a minimum of 160 m (0.1 mile) segment length for the undivided two-lane urban and suburban arterial segments in New Jersey, US. Table 1 summarizes the studies using the three common segmentation approaches based on the regions.

2.3. Segmentation approaches in India

Limited literature is available in the context of LMICs compared to the HICs. Vayalamkuzhi and Amirthalingam [34] employed the fixed-length segmentation approach to develop CPMs for a four-lane divided rural highway in India. Singh et al. [35] also adopted a fixed-length segmentation approach, dividing the selected rural highway stretches of 250 km into 68 uniform segments of varying lengths. Similarly, Dhankute and Parida [31] divided the selected 115 km of four-lane rural highway stretch into 23 segments of a fixed length of 5 km each. A recent study by Bisht and Tiwari [30] implemented a fixed-length segmentation approach of 100 m to assess the safety effectiveness of the paved shoulder on a four-lane divided highway in India. Bisht and Tiwari [27] also used a fixed segmentation approach to study the effect of risk factors on pedestrian fatal crashes on a rural six-lane highway in India. Similarly, another study by Bisht and Tiwari [28] implemented the fixed length (100 m) segmentation approach for a six-lane intercity expressway in India.

A few studies also applied the HSM-based segmentation approach in India. Dinu and Veeraragavan [38] employed the homogenous segmentation approach to divide India's 200 km of selected two-lane rural highways. Mitra et al. [37] implemented the HSM-based homogenous segmentation to divide the selected four-lane highway stretch into fifty-

Table 1
Summary of the studies based on adopted segmentation approaches and the study region.

Adopted segmentation approach	Study region	
	HICs	LMICs
Fixed length	Thomas [21], Cafiso et al. [13], Lu et al. [17], Ma et al. [24], Cafiso et al. [12], Tahir et al. [5]	Bhavsar et al. [26], Bisht and Tiwari [27], Bisht and Tiwari [28], Bisht & Tiwari [29], Bisht and Tiwari [30], Silva et al. [6], Dhankute and Parida [31], ChikkaKrishna et al. [32], ChikkaKrishna et al. [33], Vayalamkuzhi and Amirthalingam [34], Singh et al. [35]
Homogeneous characteristics	Koorey [14], Cafiso et al. [13], Agostino [22], Cafiso et al. [12], Tahir et al. [5], Bartin et al. [10]	Elagamy et al. [20], Elagamy et al. [23], Nair and Bhavathrathan [36], Mitra et al. [37], Dinu and Veeraragavan [38]
Spatial clustering	Thomas [21], Anderson [16], M. Ghadi et al. [25], M. Ghadi and Török [9], Shen et al. [18]	Nair and Bhavathrathan [36], Bisht et al. [30]

Table 2
Summary of segmentation methods and adopted segment length based on regions.

Authors/Country	Highway context	Number of lanes	Segmentation method	Adopted segment length (m)
HICs				
Koorey [14]/New Zealand	rural	all	Fixed length	250 m
Cafiso et al. [13]/Italy	rural	2	The HSM	500 m to 4290 m
Lu et al. [14]/US	both	minimum 4 lanes	All three	The HSM and Spatial clustering methods: variable segment length Fixed length method: 800 m
Agostino [22]/Italy	rural	4	The HSM	76 m to 10,195 m
Ma et al. [24]/China	rural	4 & 6	Fixed length and the HSM	Fixed length: 1000 m The HSM method: 400 m to 1400 m
Cafiso et al. [12] /Italy	rural	4	Fixed length and the HSM	Fixed length: 650 m The HSM method: varying segment length
M. Ghadi and Török [9]/ Hungary	NA	4	Fixed length and the HSM	Fixed length: 750 The HSM method and Spatial clustering based: Varying segment length Recommended method: Clustering based method with 6760 m of segment length
Shen et al. [18]/China	rural	NA	Clustering	minimum 200 m
Tahir et al. [5]/Australia	rural	2	All methods	Fixed length method: 1000 m Other methods: minimum 200 m to maximum 17,300 m
Bartin et al. [10]/US	urban and suburban	2	The HSM	minimum 161 m (0.1 mile)
LMICs				
Dinu & Veeraragavan [38]/ India	rural	2	The HSM	1000 m to 22,000 m
Boroujerdian et al. [39]/Iran	rural	NA	Dynamic	Varying segment length
ChikkaKrishna et al. [32,33]/ India	rural	4	Fixed length	Minimum 160 m (0.1 mile)
Mitra et al. [37]/India	rural	4	The HSM	300 m to 1600 m
Elagamy et al. [20,23]/Egypt	rural	2 & 4	Fixed length and the HSM	Fixed method: 1000 m The HSM method: AADT, number of lanes, shoulder widths, median width, curvature, U-turns
Bhavsar et al. [26]/India	rural	4	Fixed length	1000 m
Silva et al. [6]/Brazil	rural	4	Fixed length	Tested segment length: 500 to 5000, with increments of 500 m recommended: 4500 m to 5000 m
Bisht & Tiwari [30]/ India	rural	4	Fixed length	100 m
Nair & Bhavathrathan [36]/ India	both	2 & 4	The HSM and Spatial clustering	NA
Bisht & Tiwari [28]/ India	rural (expressway)	6	Fixed length	100 m
Bisht & Tiwari [29] /India	rural	6	Fixed length	100 m
Bisht & Tiwari [27]/India	rural	6	Fixed length	100 m

two homogeneous road segments, varying segment lengths between 300 m and 1600 m. A study by Nair and Bhavathrathan [36] proposed a hybrid segmentation methodology, combining the HSM and clustering-based segmentation approaches. However, as evident from Table 2, very few studies implemented the clustering-based segmentation approach for highways in India [30].

The reviewed studies indicated that the choice of the segmentation method and segment length varied between the studies for various reasons, one of the major reasons being data availability. The methodological framework of the reviewed studies showed that the selection was also influenced by the analyst's choice [40]. The adopted segmentation approaches and segment length also depend on the roadway type, hierarchy, and speed limit [8]. Hence, it can be summarized that the segmentation approach and adopted segment length vary due to analyst choice influenced by data availability roadway type and the study region [8,40].

In the case of LMICs, the studies suggested that researchers preferred a fixed-length segmentation approach. In these countries, guidance on the segmentation approach is scant for road safety practitioners. Literature also suggests that segmentation approaches based on the per se of adequate segment length for safety evaluation studies have not been explored in LMICs. Therefore, this study will contribute by suggesting adequate segment-length implementation for safety evaluation studies on rural divided highways in India.

3. Methodology

3.1. Study area and data description

In this study, a 167 km long access-controlled six-lane Indian expressway known as the Yamuna expressway was selected, shown in Fig. 1.

The geometry, crash, speed, and traffic volume data were collected from the expressway's governing body, i.e., the Yamuna Expressway Industrial Development Authority (YEIDA) and the concessionaire Jaypee Infratech Limited. The expressway geometric data consists of vertical alignment, including vertical alignment gradient, vertical gradient length, and vertical curve length. On the other hand, horizontal alignment variables include horizontal alignment radius, curve length and straight length of the horizontal alignment.

The crash dataset consists of fatal crashes on the expressway from August 2012 to October 2018. Each fatal crash consists of information such as the date and time of the crash, location, vehicle type and road users involved in a crash, severity and fatalities, causes of the crash, and type of crash. The details related to driver and victim information, weather conditions, and lighting data were unavailable. Traffic volume data were extracted with the help of speed camera data, and toll transaction details data of the toll plazas on the expressway. The share of the cars in the total traffic volume was the highest as it varies between 72 and 80% of the total traffic volume on the expressway. Therefore, the car's 85th percentile speed was considered the representative speed for a

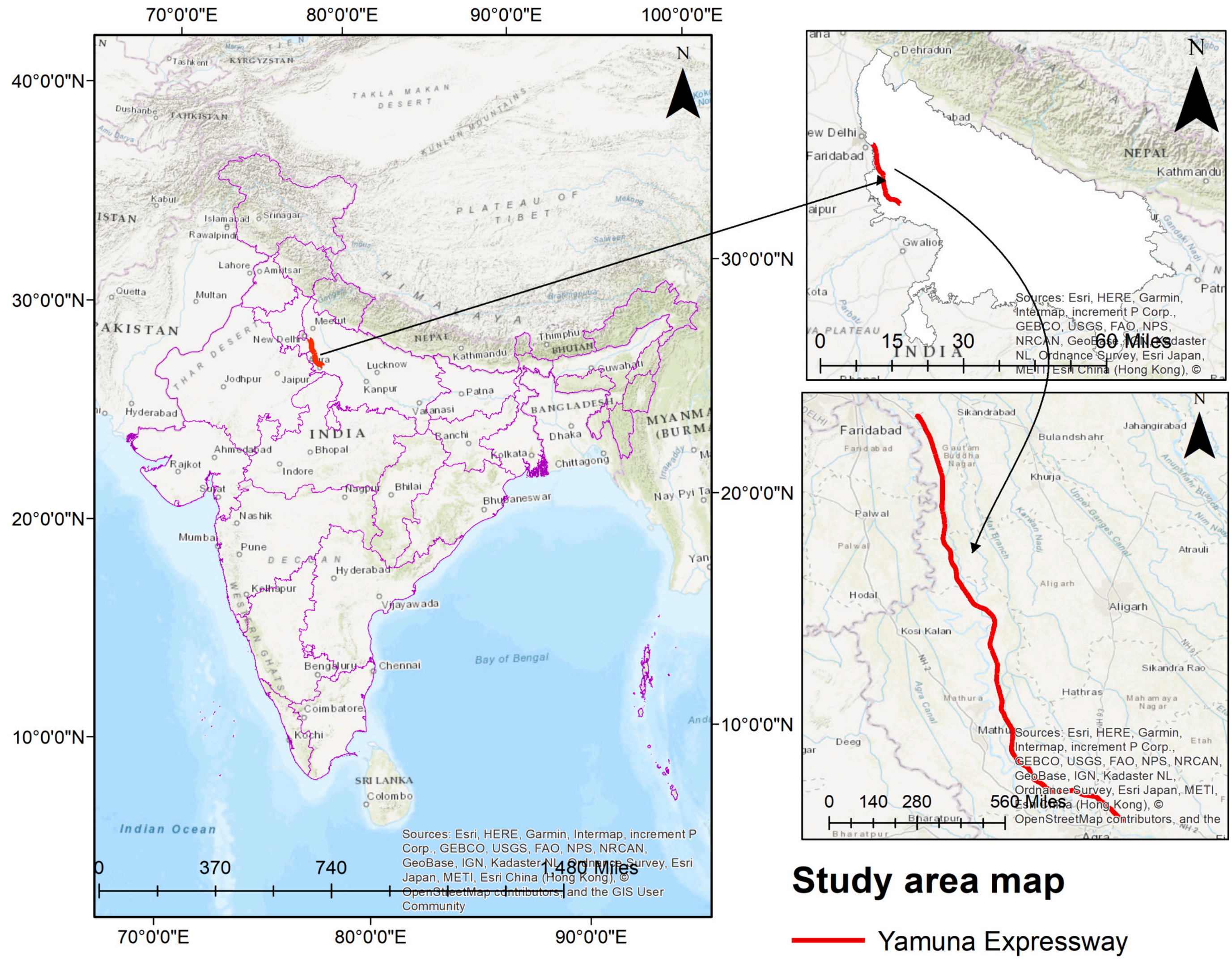


Fig. 1. A typical map of the selected expressway as a case study stretch.

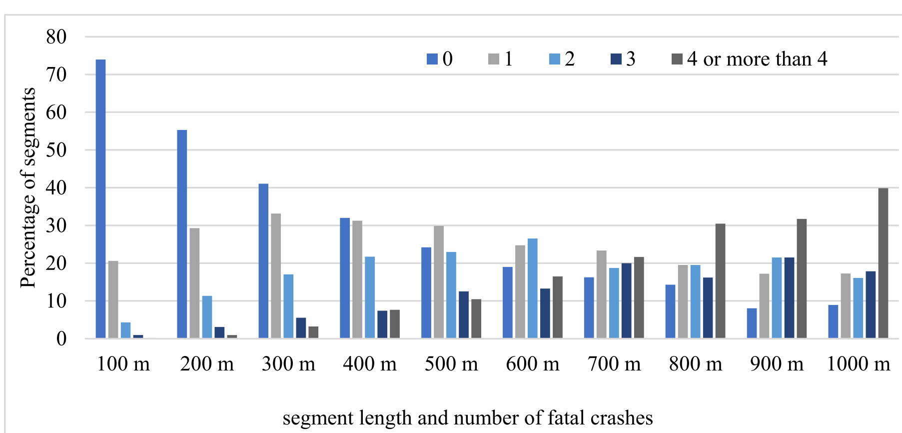


Fig. 2. Distribution of the fatal crashes based on the segment length.

segment. The lane-wise speed data was recorded by speed cameras installed at five locations, each in both directions, staggered along the expressway stretch. Further, the posted speed limit for cars was 100 kmph and 60 kmph for trucks and buses.

In addition, the segments with rest areas and the presence of an underpass were also considered. The 100 m influence length on both sides of the underpass was considered as the pedestrian activity observed. Due to the presence of the nearby village, pedestrian and motorized two-wheeler activities were observed on the expressway. Further, the hazards on the shoulder consist of overhead gantry board

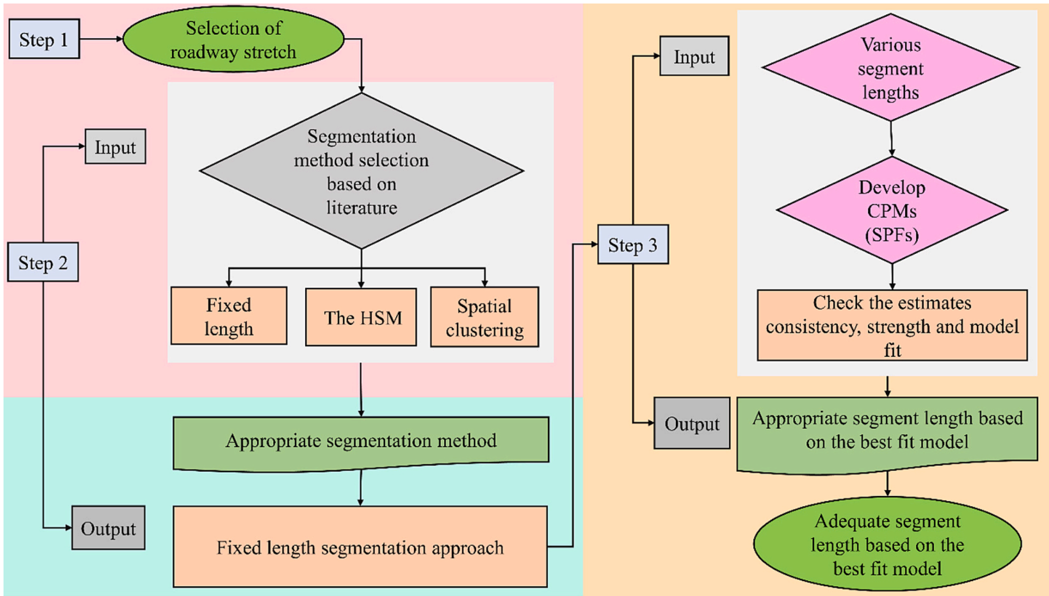


Fig. 3. Methodological framework of the study.

and road signage posts, non-standard signage poles, overhead electric cable posts, full-grown trees, and kilometer stones. The hazards were considered if they were present within the clear zone of the expressway. The next subsection discusses the approach adopted for the segmentation process.

3.2. Adopted segmentation approach

The original geometric dataset was available for the 20 m section of the expressway. Therefore, in this study, consecutive 20 m sections were combined from 100 m to 1000 m segments for the unit of analysis. There were two distinct purposes for combining the sections into segments. First, to create the desired segment-specific datasets for developing explanatory models. For example, the average value of five consecutive 20 m sections of the geometric variable was considered a representative value for the 100 m segment. Thus, based on the conclusion of the literature review, we adopted a fixed-length segmentation approach. A similar approach for combining various sections into segments was utilized for all the independent variables. Since the location inaccuracy with the crash data exists, it is difficult to determine the exact location of the crash [41]. However, based on the location information available, we can determine the geo-location of the crash. Second, entering crash data into the segments will improve the accuracy of fatal crash locations on the expressway [41].

Fig. 2 illustrates the change in the number of segments with fatal crashes as segment length increases from 100 to 1000 m. It shows that the zero crash segments decrease with the increase in the segment length. Compared to smaller segment lengths, the segment with 500 to 700 m length has a uniform distribution of segments per the number of fatal crashes.

The model specifications employed the data for the fatal crash, geometric, speed, volume, and roadside characteristics. The next section discusses the results of the developed explanatory models concerning parameter transferability testing. Fig. 3 demonstrates the analytical framework adopted in this study.

3.3. Explanatory model development

Generally, the Poisson regression model is a starting point in the crash analysis, as crash counts are non-negative integer values [42]. The basic premise of the Poisson distribution is that mean and variance should be equal. However, crash data are generally over-dispersed as the variance exceeds the mean. The Poisson model specification suggests that highway segments with the same independent variable's magnitude are expected to have the same number of crashes, which is hardly true. Moreover, the Poisson specification lacks transferability due to unobserved effects [43]. Researchers have shown that crashes are due to many risk factors, and many are unobservable or omitted in the model specification [44].

The segments with the same values of independent variables (X_i) still may differ in many other omitted variables. Consequently, the λ_i also varies because of omitted variables, leading to variance greater than the mean, called overdispersion. As a result, the negative binomial (NB) is employed to handle the overdispersion issue in the crash data [45,46]. The NB model specification assumes that the Poisson parameter follows a Gamma probability distribution, also known as the Poisson-Gamma model.

In the NB model, overdispersion is accounted for with the help of an overdispersion parameter denoted by α. Therefore, the model selection between Poisson and NB depends on the estimate of α.In the case of the NB model, the value of α is greater than zero. Hence, in the NB model, the parameter λ_i can be well approximated by a Gamma distribution function with a quadratic variance and mean relationship. The Gamma probability density function can take any shapes based on the values of rate (a) and shape (b) parameters, and can be expressed through Eq. (1):

$$f(\lambda_i) = \frac{a^b \lambda_i^{b-1} exp(-a\lambda_i)}{\Gamma(b)} \tag{1}$$

As mentioned earlier, the maximum likelihood method uses the NB density function for parameter estimation. The expectation can be computed as $E(\lambda_i) = \frac{b}{a}$.In contrast, variance is computed as $Var(\lambda_i) = b/a^2$. Hence, the expected number of crashes for the expressway segment (i) with given values of independent variables X_i can be expressed as Eq. (2):

$$\lambda_i = e^{(\beta X_i + \varepsilon_i)} \tag{2}$$

Where ε_i is stochastic and accounts for the unobserved heterogeneity in λ_i. The $exp(\varepsilon_i)$ is a Gamma distributed error term with a mean value equal to one and variance equal to α. The conventional NB model specification provides the marginal distribution as an NB distribution. Therefore, the probability of segment i having n_i the number of crashes can be estimated with the help of Eq. (3) [47]:

Table 3
Summary descriptive statistics of the study variables for 100 m segment length.

Variable description	Mean	Std. Dev.	Min.	Max.
Dependent variable				
Number of fatal crashes	0.33	0.63	0.00	6.00
Independent variables [unit] (abbreviation)				
Vertical gradient [%] (VGRAD)	0.06	1.05	−2.00	33.66
Vertical curve length [km] (VCRUVL)	0.20	0.20	0.00	0.79
Horizontal curve radius [%] (HCURVRD)	0.17	1.98	−10.00	7.50
Horizontal curve length [km] (HCURVLC)	0.52	0.80	0.00	3.20
AADT, Cube root of traffic volume [vehicles/day]	14.44	2.35	11.00	18.00
Proportion of the segment with hazard on the shoulder [m] (HZSHLD)	0.07	0.25	0.00	1.00
Proportion of the segment with underpass [m] (UNDPS)	0.11	0.32	0.00	1.00

$$P(n_i) = \frac{\Gamma\left(n_i + \frac{1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right) n_i!} \left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i}\right)^{\frac{1}{\alpha}} \left(\frac{\lambda_i}{\frac{1}{\alpha} + \lambda_i}\right)^{n_i} \tag{3}$$

Where $\Gamma(*)$ is a Gamma function, and if α is equal to zero, the NB model is reduced to the Poisson model. The model transferability was tested based on McFadden's pseudo (ρ^2) likelihood-based ratio tests. The likelihood ratio test evaluates the difference in log likelihoods between the null model and the full model and can be estimated as per Eq. (4):

$$\chi^2 = -2\left[LL(\beta_A) - LL(\beta_B)\right] \tag{4}$$

Where $LL(\beta_A)$ and $LL(\beta_B)$ are the log-likelihood at convergence for two competing models, A and B, respectively. Whereas chi-squared (χ^2) is a test statistic (chi-square distributed) with a degree of freedom equal to the difference in the number of estimable parameters between two competing models.

In this study, the statistical performance of the developed NB models in terms of model fit was assessed and compared by computing log-likelihood at convergence, Akaike information criteria (AIC), and Bayesian information criteria (BIC). The smaller the AIC and BIC values, the better the model is as the estimates converge close to the true parameters [44].

$$AIC = -2LL(\beta) + 2k \tag{5}$$

$$BIC = -2LL(\beta) + k^*ln(N) \tag{6}$$

Where $LL(\beta)$ is the log-likelihood at convergence, $LL(0)$ is the log-likelihood value for the intercept-only model, N is the number of the segments, k is the number of parameters. To sum up, the NB models were developed to assess the transferability of the parameters. In this study, the 100 m segment model was considered the base model for assessing the parameter transferability across the different tested segment length models.

4. Results and discussion

In this study, we originally collected thirty-two independent variables. As an initial step, the independent variable selection process includes checking univariate distributions, correlation analysis, and multi-collinearity for the independent variables. Therefore, one of the highly correlated variables was removed. Next, the variance inflation factor (VIF) was computed to check the multi-collinearity among the variables. The variables having a VIF of more than ten were excluded from the study. Therefore, thirteen independent variables were retained for the model fitting in the next step.

Next, the backward stepwise selection regression method was used to select the best subset of the thirteen retained independent variables. Therefore, the final seven independent variables were retained and are considered continuous data. The segment-wise number of fatal crashes was considered the dependent variable. The presented results are the parameter estimation of the final NB regression models for all the considered regression models. The NB models were developed for all the segment lengths in the next stage. Table 3 presents the descriptive summary of the final retained independent variables in the study for the 100 m length segment.

Table 4 presents the NB model estimates for all the segment lengths. The significance level was set to 10% for all the models. The bold ones are statistically significant parameter estimates at or less than a 10% significance level. Since none of the roadside variables was statistically significant, therefore, they were dropped during the variable selection process.

As shown in Table 4, for the 100 m segment length model, the statistically significant variables are vertical curve length, AADT, and segment with the presence of an underpass. However, with the increase in the segment length, the sample size decreases, and the parameter estimates change to statistically insignificant, except for the AADT. The intercept magnitude decreases with the segment length increase and varies between −3.016 and − 0.319. However, its parameter estimate was consistent and statistically significant till 600 m. In the case of the vertical gradient variable, the parameter estimates magnitude fluctuates

Table 4
NB model parameter estimates for all developed models for 100 to 1000 m segment length.

Independent variable description	Segment-specific parameter estimates									
	100 m	200 m	300 m	400 m	500 m	600 m	700 m	800 m	900 m	1000 m
Intercept	**−3.016**	**−2.315**	**−1.865**	**−1.684**	**−1.230**	**−1.028**	−0.819	−0.892	−0.319	−0.383
Vertical gradient	−0.069	−0.076	0.008	−0.017	0.029	−0.160	−0.248	−0.099	−0.051	0.089
Vertical curve length	**−0.638**	**−0.655**	−0.443	−0.387	−0.405	−0.496	−0.295	−0.304	−0.265	−0.595
Horizontal curve radius	0.044	0.048	0.049	0.052	0.055	0.055	0.067	0.072	0.057	0.083
Horizontal curve length	0.086	0.101	0.106	0.093	0.113	0.096	0.100	0.132	0.116	0.147
AADT, Cube root of traffic volume	**0.124**	**0.122**	**0.120**	**0.127**	**0.116**	**0.116**	**0.118**	**0.128**	**0.101**	**0.112**
Proportion of the segment with hazard on the shoulder	0.177	0.077	−0.130	*0.206*	−0.394	−0.376	−0.563	−0.168	−1.137	−0.970
Proportion of the segment with the underpass	**0.806**	**1.127**	**0.833**	0.756	0.413	0.515	−0.479	−0.353	−0.410	−0.196

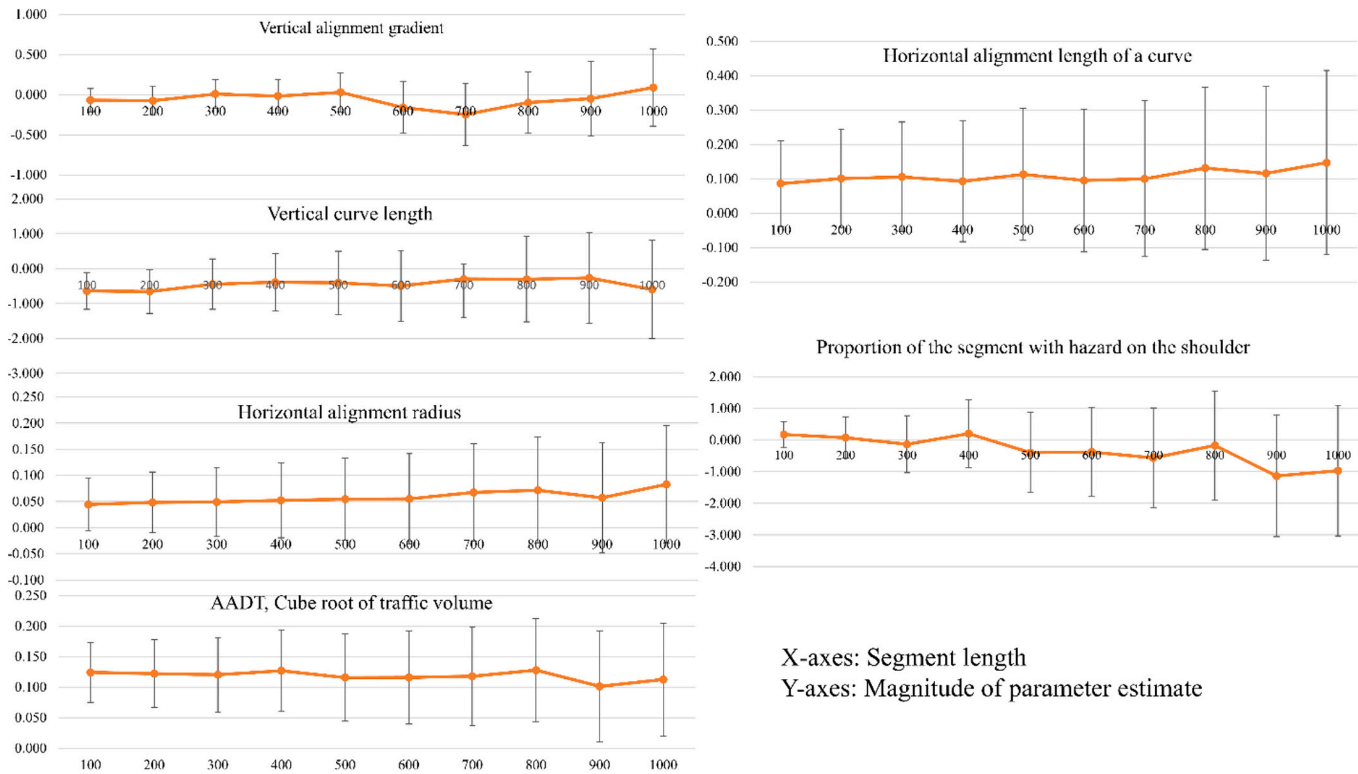
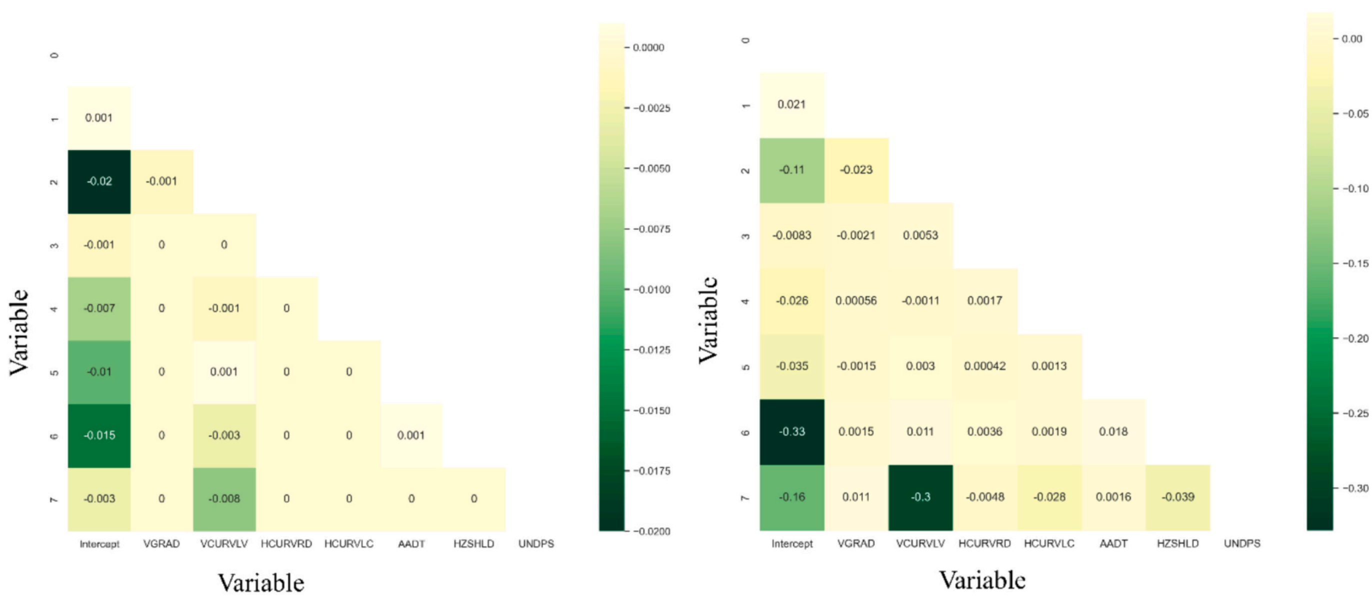


Fig. 4. Parameter estimate's variation with change in segment length.



a. Covariance heatmap for parameters of the 100 m model.

b. Covariance heatmap for parameters of the 1000 m model

Fig. 5. Comparison of the covariance estimates for the estimated parameters.

between −0.248 and 0.089. However, its parameter estimates are statistically insignificant for all models. The direction of the parameter estimates changes from positive to negative, suggesting the change in the association with fatal crashes as the segment length varies; however, the effect is inconsistent.

Further, the effect is statically insignificant for all the models. The parameter's estimate of the vertical curve length variable varies from −0.655 to −0.265, suggesting a negative association with fatal crashes. The parameter's estimates show consistency in effect on the fatal crashes. However, its effect is statistically significant for segment lengths to 200 m. On the other hand, the variable horizontal curve radius has a consistently positive effect and varies between 0.044 and 0.083. In addition, the direction of the parameter estimate is positive and remains unchanged as the segment length increases. Again, its effect is statistically insignificant for all segments. However, the strength of the effect is negligible as the standard deviation and variance are low; refer to Table 4. Similarly, consistency (i.e., 0.086 to 0.147) and low effect were observed in the case of the variable horizontal curve length. However,

Table 5
Model fit statistics for all the segments.

Model for segment length	Sample size	The goodness of fit statistics			Overall model fit#	
		Log-likelihood	AIC	BIC	Chi-square value	*p*-value
100	1674	−1214.337	2444.673	2488.057	**64.907**	0.000
200	837	−911.365	1838.731	1876.569	**43.938**	0.000
300	558	−757.472	1530.943	1565.538	**23.811**	0.001
400	419	−654.445	1324.891	1357.194	**18.927**	0.008
500	335	−579.920	1175.839	1206.352	**16.173**	0.024
600	279	−522.859	1061.717	1090.767	**14.920**	0.037
700	240	−478.261	972.521	1000.366	**15.120**	0.034
800	210	−442.032	900.063	926.840	13.177	0.068
900	186	−411.115	838.229	864.035	11.057	0.136
1000	168	−384.513	785.026	810.017	13.287	0.065

Note: degree of freedom = 7; Critical chi-square at 90% significance level is 12.017. #Comparison of the fitted model against the intercept-only model.

none of the effects were statistically significant.

The coefficient of the exposure variable, such as AADT, varies between 0.101 and 0.128, and the effect was statistically significant for all models. The parameter estimates of the variable proportion of the segment with hazard on the shoulder are inconsistent as the direction changes as the segment length varies. However, its effect is consistent for segment lengths greater than 500 m. The parameter's magnitude varies between −1.137 and 0.206. Lastly, the parameter estimates of the variable proportion of the segment with underpass vary between −0.479 and 1.127, suggesting inconsistency in the direction of the effect. However, the effect varied as it was positive up to 600 m. However, the effect was statistically significant for segments to 300 m. Further, it was negative for the segment length greater than 600 m. However, a larger dataset shall be tested to consolidate the findings, and other highway facility types should also be considered.

Fig. 4 summarizes the variation in the parameter estimates as segment length changes. In Fig. 4, the y-axis represents the magnitude of the parameter estimate, whereas the x-axis represents the segment length.

Fig. 5 illustrates the change in the covariance between the estimated parameters as segment length varies. The results imply that the covariance between the parameters increases as the segment length increases.

As shown in Table 5, the models are statistically insignificant for segment lengths greater than 700 m at or less than a 10% significance level. For abbreviations definition refer to Table 3. In Table 5, the bold Chi-square values indicate statistical significance at or less than a 10% significance level. As evident, with the increase in the segment length, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) statistics decrease, suggesting a better model fit, though the model fit beyond 700 m segment length was statistically insignificant.

The model results showed that the parameter estimates' consistency, statistical significance, effect, and magnitude depend on the sample size and segment length. Further, the developed models in this study suggest that the models are statistically insignificant as the segment length increases more than 700 m. However, the magnitude of AIC and BIC statistics decreases as segment length increases. Fig. 6 presents the cumulative residual (CURE) plots for all the segments of the expressway.

In the end, the results indicate that additional information is required to improve the model specifications, such as increasing the study period and including more highway stretches. Due to the research funding crunch in LMICs, additional data collection may not always be feasible. Therefore, the methodology presented here can guide practitioners in getting an optimal length of the segments for CPM development and blackspot treatment on the highways.

5. Conclusions

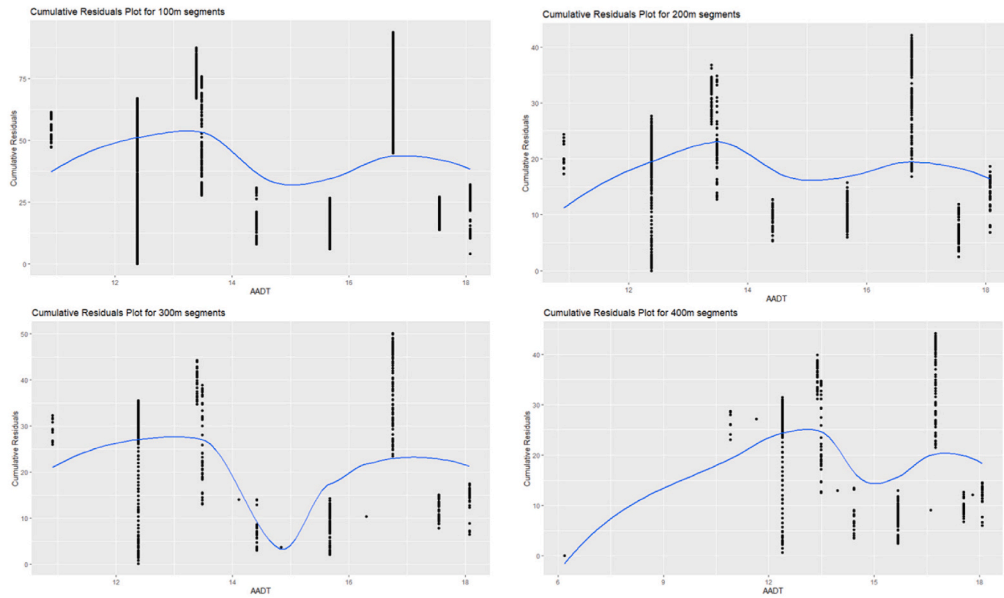
This study sought to answer the question of adequate segment length for highway safety evaluation studies in LMICs. In LMICs, where data availability and accuracy are issues, data collection for very short segments is challenging and very long sections dilute the variability [8,48]. Due to various issues, the segmentation approaches vary between studies and at the regional level. Though we did not primarily aim to present a safety evaluation study, the major motivation was to provide an adequate segment length in the limited data availability scenario. In this direction, we present a case study of an Indian rural six-lane expressway.

This study has shown that the consistency of the parameter estimates varies with the segment length and sample size. The findings suggest that it is difficult to suggest a point estimate for the adequate segment length. However, findings strongly suggest that a segment length between 300 and 700 m can be considered adequate for various safety evaluation studies for multi-lane rural highways. The study highlighted the parameter sensitivity to sample size and data aggregation for different highway segment lengths. Therefore, the methodology presented here can guide practitioners to judicially select an adequate segment length for safety evaluation, such as blackspot identification and treatment, and for fitting the CPMs on the highways.

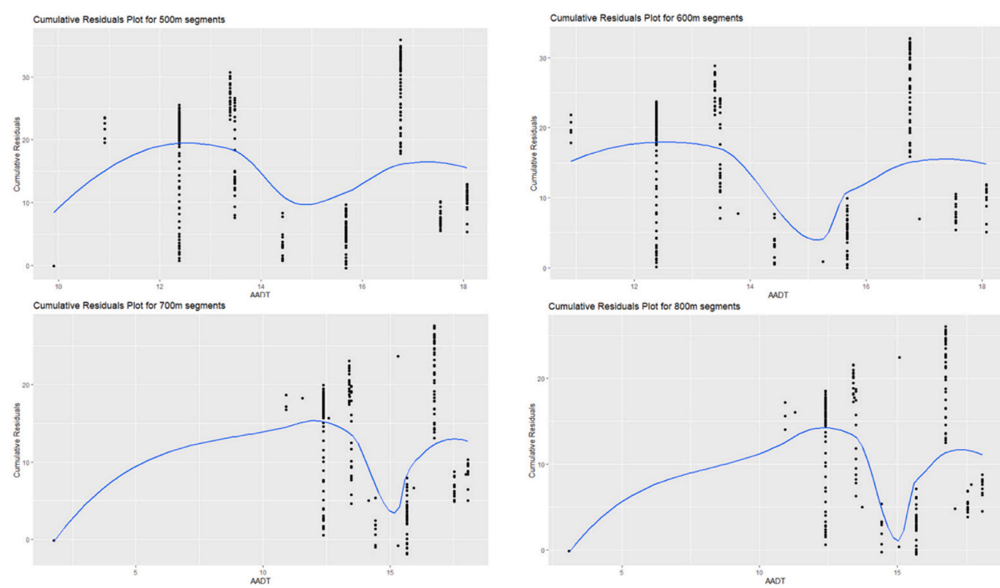
One of the study findings' major indications is that road safety-related data availability and quality shall be improved in India, which also applies to other LMICs [49,50]. In addition, though modelling techniques may not be novel in this study, the segmentation framework employed may provide valuable insight to policymakers, practitioners, and road safety practitioners in LMICs. This study employed the fixed-length segmentation approach, which is pragmatic and straightforward. In addition, considering the data availability and sample size, the fixed-length segmentation approach was deemed fit [5].

The absence of extensive data availability limited this study. The generalisability and transferability of this study's findings are subject to caution and certain limitations. The limitation of the study is that most of the parameter estimates are insignificant, irrespective of the segment length. Thus, an additional study can be conducted with additional data to enhance the sample size. Another limitation is that we have adopted the 100 m incremental increase (fixed length) in the segment length for developing different segment length-related scenarios. Therefore, an alternative segmentation approach with additional data should also be tested for Indian highways in the future. In addition, other modelling techniques can be tested in future studies.

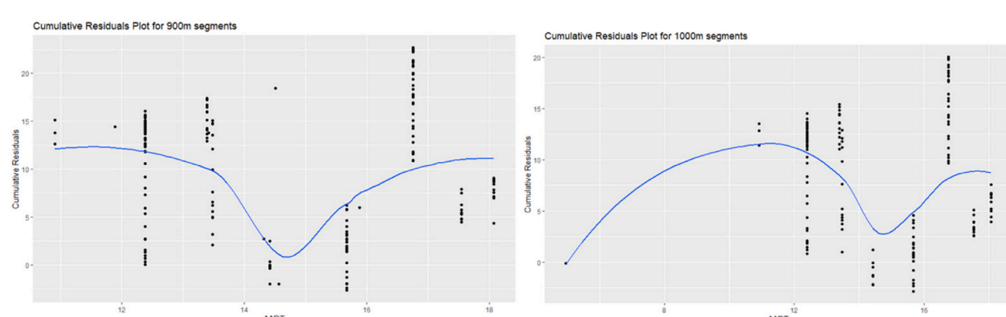
Since data collection is cost-intensive, expressway crash data in LMICs usually have small observations. Additionally, some rural expressways inherently have low crash rates for many segments; however, with high severity, the sample size remains low in such cases. Such crash data has characteristics of low sample mean, excessive zero



(a) CURE plot for 100 m and 400 m segments.



(b) CURE plot for 500 m and 800 m segments.



(c) CURE plot for 900 m and 1000 m segments.

Fig. 6. CURE plots for all ten segments.

preponderance, and negative skewness [8,42]. Moreover, with the available crash data, many risk factors remain unobserved; such an issue is called unobserved heterogeneity [46]. Lord & Mannering [42] have argued that the NB model is susceptible to the above-discussed data issues and unobserved heterogeneity and estimates the parameters with less reliability. Therefore, an alternative approach could be to implement random parameter techniques in future studies [42,46]. The random parameter techniques address the unobserved heterogeneity issues by allowing the parameters to vary across the expressway segments.

Overall, this study strengthens the idea that adequate segment length and segmentation approaches rely on many factors, which is not a straightforward question to answer [51]. Therefore, selecting segment length and segmentation approach should be based on the local conditions, highway context, level of data availability and data quality.

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Laxman Singh Bisht: Conceptualization, Methodology, Writing – original draft, Software, Formal analysis. **Sai Chand:** Conceptualization, Methodology, Writing – review & editing, Supervision, Validation. **Geetam Tiwari:** Conceptualization, Writing – review & editing, Supervision, Validation.

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.

References

[1] D. Mohan, et al., PROTOCOL: effectiveness of road safety interventions: an evidence and gap map, Campbell Syst. Rev. 16 (1) (2020), https://doi.org/10.1002/cl2.1077.
[2] AASHTO, Highway Safety Manual, 2010.
[3] G. Yannis, et al., Road traffic accident prediction modelling: a literature review, Proc. Inst. Civ. Eng. Transp. 170 (2017) 1–10, https://doi.org/10.1680/jtran.16.00067.
[4] X. Zhao, D. Lord, Y. Peng, Examining network segmentation for traffic safety analysis with data-driven spectral analysis, IEEE Access 7 (2019) 120744–120757, https://doi.org/10.1109/ACCESS.2019.2937001.
[5] H. Bin Tahir, S. Washington, S. Yasmin, M. King, M.M. Haque, Influence of segmentation approaches on the before-after evaluation of engineering treatments: a hypothetical treatment approach, Accid. Anal. Prev. 176 (April) (2022) 106795, https://doi.org/10.1016/j.aap.2022.106795.
[6] P.B. Silva, M. Andrade, S. Ferreira, Influence of segment length on the fitness of multivariate crash prediction models applied to a Brazilian multilane highway, IATSS Res. 45 (4) (2021) 493–502, https://doi.org/10.1016/j.iatssr.2021.05.001.
[7] R. Srinivasan, K. Bauer, Safety Performance Function Development Guide: Developing Jurisdiction-Specific SPFs 5. Report Date September 2013 6. Performing Organization Code, North Carolina, 2013.
[8] M. Schlögl, R. Stütz, Methodological considerations with data uncertainty in road safety analysis, Accid. Anal. Prev. 130 (2019) 136–150, https://doi.org/10.1016/j.aap.2017.02.001.
[9] M. Ghadi, Á. Török, A comparative analysis of black spot identification methods and road accident segmentation methods, Accid. Anal. Prev. 128 (February) (2019) 1–7, https://doi.org/10.1016/j.aap.2019.03.002.
[10] B. Bartin, K. Ozbay, C. Xu, Safety performance functions for Two-Lane urban arterial segments, Saf. Sci. 167 (July 2022) (2023) 106284, https://doi.org/10.1016/j.ssci.2023.106284.
[11] K. Ozbay, H. Nassif, B. Bartin, C. Xu, A. Bhattacharyya, Calibration/Development of Safety Performance Functions for New Jersey, New York, 2019.
[12] S. Cafiso, C.D. Agostino, B. Persaud, Investigating the influence of segmentation in estimating safety performance functions for roadway sections, J. Traffic Transp. Eng. (English Ed.) 5 (2) (2018) 129–136, https://doi.org/10.1016/j.jtte.2017.10.001.
[13] S. Cafiso, A. Di Graziano, G. Di Silvestro, G. La Cava, B. Persaud, Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables, Accid. Anal. Prev. 42 (4) (2010) 1072–1079, https://doi.org/10.1016/j.aap.2009.12.015.
[14] G. Koorey, Road data aggregation and sectioning considerations for crash analysis, Transp. Res. Rec. 2103 (2009) 61–68, https://doi.org/10.3141/2103-08.
[15] M.Q. Ghadi, Á. Török, Comparison of different road segmentation methods, Promet - Traffic - Traffico 31 (2) (2019) 163–172, https://doi.org/10.7307/ptt.v31i2.2937.
[16] T.K. Anderson, Kernel density estimation and K-means clustering to profile road accident hotspots, Accid. Anal. Prev. 41 (3) (2009) 359–364 [Online]. Available, http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed11&NEWS=N&AN=354712900.
[17] J. Lu, A. Gan, K. Haleem, W. Wu, Clustering-based roadway segment division for the identification of high-crash locations, J. Transp. Saf. Secur. 5 (3) (2013) 224–239, https://doi.org/10.1080/19439962.2012.730118.
[18] L. Shen, J. Lu, M. Long, T. Chen, Identification of accident blackspots on rural roads using grid clustering and principal component clustering, Math. Probl. Eng. 2019 (2019) 12, https://doi.org/10.1155/2019/2151284.
[19] K. Fitzpatrick, W.S. Iv, J. Carvell, Using the rural two-lane highway draft prototype chapter, Transp. Res. Rec. 1950 (2007) 44–54.
[20] S.R. Elagamy, S.M. El-Badawy, S.A. Shwaly, Z.M. Zidan, U.E. Shahdah, Segmentation effect on developing safety performance functions for rural arterial roads in Egypt, Innov. Infrastruct. Solut. 5 (2) (Aug. 2020), https://doi.org/10.1007/s41062-020-00318-7.
[21] I. Thomas, Spatial data aggregation: exploratory analysis of road accidents, Accid. Anal. Prev. 28 (2) (1996) 251–264, https://doi.org/10.1016/0001-4575(95)00067-4.
[22] C.D. Agostino, Investigating transferability and goodness of fit of two different approaches of segmentation and model form for estimating safety performance of motorways, in: Procedia Engineering, Elsevier B.V, 2014, pp. 613–623, https://doi.org/10.1016/j.proeng.2014.10.478.
[23] S.R. Elagamy, S.M. El-Badawy, S.A. Shwaly, Z.M. Zidan, U.E. Shahdah, Segmentation effect on the transferability of international safety performance functions for rural roads in Egypt, Safety 6 (3) (2020) 1–24, https://doi.org/10.3390/SAFETY6030043.
[24] Z. Ma, H. Zhang, S.I. Chien, J. Wang, C. Dong, Predicting expressway crash frequency using a random effect negative binomial model : a case study in China, Accid. Anal. Prev. 98 (2017) 214–222, https://doi.org/10.1016/j.aap.2016.10.012.
[25] M. Ghadi, Á. Török, Z. Szalay, Analysis of traffic accident black spots: an application of spatial clustering segmentation method, East West Cohesion (2019) 10 [Online]. Available: https://www.researchgate.net/publication/340062435.
[26] R. Bhavsar, A. Amin, L. Zala, Development of model for road crashes and identification of accident spots, Int. J. Intell. Transp. Syst. Res. 19 (1) (2021) 99–111, https://doi.org/10.1007/s13177-020-00228-z.
[27] L.S. Bisht, G. Tiwari, A matched case–control approach to identify the risk factors of fatal pedestrian crashes on a six-lane rural highway in India, Int. J. Inj. Control Saf. Promot. (2023) 1–17, https://doi.org/10.1080/17457300.2023.2242336.
[28] L.S. Bisht, G. Tiwari, Assessment of fatal rear-end crash risk factors of an expressway in India: a random parameter NB modeling approach, J. Transp. Eng. Part A Syst. 149 (1) (2023) 1–19, https://doi.org/10.1061/jtepbs.0000767.
[29] L.S. Bisht, G. Tiwari, Assessing the effect of paved shoulder width on road traffic crash risk on a six-lane rural highway in India, J. Transp. Eng. Part A Syst. 149 (7) (2023) 1–16, https://doi.org/10.1061/jtepbs.teeng-7296.
[30] L.S. Bisht, G. Tiwari, Safety effects of paved shoulder width on a four-lane divided rural highway in India: A matched case-control study, Saf. Sci. 147 (November 2021) (Mar. 2022) 105606, https://doi.org/10.1016/j.ssci.2021.105606.
[31] A. Dhankute, M. Parida, Risk analysis of rural four lane divided highway based on risk index determination by road safety audit, J. East. Asia Soc. Transp. Stud. 13 (2019) 1927–1947.
[32] N.K. ChikkaKrishna, M. Parida, S.S. Jain, Calibration of safety performance function for crashes on inter-city four lane highways in India, Cogent Eng. 2 (1) (2015), https://doi.org/10.1080/23311916.2015.1031929.
[33] N.K. ChikkaKrishna, M. Parida, S.S. Jain, Identifying safety factors associated with crash frequency and severity on nonurban four-lane highway stretch in India, J. Transp. Saf. Secur. 9 (May) (2017) 6–32, https://doi.org/10.1080/19439962.2016.1150927.
[34] P. Vayalamkuzhi, V. Amirthalingam, Influence of geometric design characteristics on safety under heterogeneous traffic flow, J. Traffic Transp. Eng. (English Ed.) 3 (6) (2016) 559–570, https://doi.org/10.1016/j.jtte.2016.05.006.
[35] G. Singh, S.N. Sachdeva, M. Pal, M5 model tree based predictive modeling of road accidents on non-urban sections of highways in India, Accid. Anal. Prev. 96 (2016) 108–117, https://doi.org/10.1016/j.aap.2016.08.004.
[36] S.R. Nair, B.K. Bhavathrathan, Hybrid segmentation approach to identify crash susceptible locations in large road networks, Saf. Sci. 145 (October 2021) (2022) 105515, https://doi.org/10.1016/j.ssci.2021.105515.
[37] S. Mitra, M. Haque, M.J. King, Effects of access, geometric design, and heterogeneous traffic on safety performance of divided multilane highways in India, J. Transp. Saf. Secur. 9 (1) (2017) 216–235, https://doi.org/10.1080/19439962.2016.1237600.
[38] R.R. Dinu, A. Veeraragavan, Random parameter models for accident prediction on two-lane undivided highways in India, J. Saf. Res. 42 (1) (2011) 39–42, https://doi.org/10.1016/j.jsr.2010.11.007.
[39] A.M. Boroujerdian, M. Saffarzadeh, H. Yousefi, H. Ghassemian, A model to identify high crash road segments with the dynamic segmentation method, Accid. Anal. Prev. 73 (2014) 274–287, https://doi.org/10.1016/j.aap.2014.09.014.

[40] E.R. Green, Segmentation Strategies for Road Safety Analysis [Online]. Available: http://ezphost.dur.ac.uk/login?url=https://search.proquest.com/docview/2066657623?accountid=14533%0Ahttp://openurl.ac.uk/ukfed:dur.ac.uk?genre=dissertations+%26+theses&issn=&title=Segmentation+Strategies+for+Road+Safety+Analysis&volume=&issue=&date=2018-0, 2018.
[41] Victoria Gitelman, Etti Doveh, Roby Carmel, Shalom Hakkert, The influence of shoulder characteristics on the safety level of two-lane roads: a case-study, Accid. Anal. Prev. 122 (2019) 108–118.
[42] D. Lord, F. Mannering, The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives, Transp. Res. Part A Policy Pract. 44 (5) (Jun. 2010) 291–305, https://doi.org/10.1016/j.tra.2010.02.001.
[43] N. Venkataraman, V. Shankar, J. Blum, B. Hariharan, J. Hong, Transferability analysis of heterogeneous overdispersion parameter negative binomial crash models, Transp. Res. Rec. 2583 (2016) 99–109, https://doi.org/10.3141/2583-13.
[44] T.U. Saeed, T. Hall, H. Baroud, M.J. Volovski, Analyzing road crash frequencies with uncorrelated and correlated random-parameters count models: an empirical assessment of multilane highways, Anal. Methods Accid. Res. 23 (2019) 100101, https://doi.org/10.1016/j.amar.2019.100101.
[45] P.C. Anastasopoulos, F.L. Mannering, A note on modeling vehicle accident frequencies with random-parameters count models, Accid. Anal. Prev. 41 (1) (2009) 153–159, https://doi.org/10.1016/j.aap.2008.10.005.
[46] F.L. Mannering, V. Shankar, C.R. Bhat, Unobserved heterogeneity and the statistical analysis of highway accident data, Anal. Methods Accid. Res. 11 (2016) 1–16, https://doi.org/10.1016/j.amar.2016.04.001.
[47] W. Greene, Functional forms for the negative binomial model for count data, Econ. Lett. 99 (3) (2008) 585–590, https://doi.org/10.1016/j.econlet.2007.10.015.
[48] M. Shirazi, D. Lord, Use/analysis of crash data and underreporting of crash data, in: International Encyclopedia of Transportation, E. Ltd., Ed, 2021, pp. 726–730, https://doi.org/10.1016/B978-0-08-102671-7.10132-0.
[49] M. Barffour, S. Gupta, G. Gururaj, A.A. Hyder, Evidence-based road safety practice in India : assessment of the adequacy of publicly available data in meeting requirements for comprehensive road safety data systems, Traffic Inj. Prev. ISSN 13 (1) (2012) 17–23, https://doi.org/10.1080/15389588.2011.636780.
[50] F. Wegman, H.Y. Berg, I. Cameron, C. Thompson, S. Siegrist, W. Weijermars, Evidence-based and data-driven road safety management, IATSS Res. 39 (1) (2015) 19–25, https://doi.org/10.1016/j.iatssr.2015.04.001.
[51] M. Schlögl, R. Stütz, Methodological considerations with data uncertainty in road safety analysis, Accid. Anal. Prev. 130 (2019) 136–150, https://doi.org/10.1016/j.aap.2017.02.001.