Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company’s public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Investigating the fatal pedestrian crash occurrence in urban setup in a developing country using multiple-risk source model

Dipanjan Mukherjee\textsuperscript{a,}\textsuperscript{*}, Sudeshna Mitra\textsuperscript{b}

\textsuperscript{a} Department of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur - 721302, West Bengal, India
\textsuperscript{b} Global Road Safety Facility, The World Bank, Washington, DC 20433, USA

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Pedestrian fatality
Risk factors
Developing Country
Behavior and perception

\textbf{ABSTRACT}

Pedestrian fatalities and injuries are a major public health burden in developing countries. In the safety literature, pedestrian crashes have been modelled predominately using single equation regression models, assuming a single underlying source of risk factors. In contrast, the fatal pedestrian crash counts at a site may be an outcome of multiple sources of risk factors, such as poor road infrastructure, land use type, traffic exposures, and operational parameters, site-specific socio-demographic characteristics, as well as pedestrians’ poor risk perception and dangerous crossing behavior, which may be influenced by poor road infrastructure and lack of information, etc. However, these multiple sources are generally overlooked in traditional single equation crash prediction models.

In this background, this study postulates, and demonstrates empirically, that the total fatal pedestrian crash counts at the urban road network level may arise from multiple simultaneous and interdependent sources of risk factors, rather than one. Each of these sources may distinctively contribute to the total observed crash count. Intersection-level crash data obtained from the “Kolkata Police”, India, is utilized to demonstrate the present modelling methodology. The three-components mixture model and a joint econometric model are developed to predict fatal pedestrian crashes. The study outcomes indicate that the multiple-source risk models perform significantly better than the single equation regression model in terms of prediction ability and goodness-of-fit measures. Moreover, while the single equation model predicts total fatal crash counts for individual sites, the multiple risk source model predicts crash count proportions contributed by each source of risk factors and predicts crashes by a particular source.

\section{Introduction}

Pedestrian injury and fatalities resulting from road traffic crashes are a major concern all over the world. There are worthwhile insights into pedestrian safety from developed countries as a good number of research studies have been carried out since the 1970s (Buehler and Pucher, 2021; Esmaili et al., 2021; McIlroy et al., 2019; Zhang et al., 2019; Truong et al., 2019; Buehler and Pucher, 2017; Diogenes and Lindau, 2010; Odero et al., 1997; Jacobs and Hutchinson, 1973). Presently, pedestrian crashes in developed countries indicate a decreasing trend due to the ongoing investment in road safety research programs and countermeasures (Buehler and Pucher, 2021; Rifaa et al., 2017a; Tulu et al., 2013). On the other hand, pedestrian safety remains one of the concerning issues in developing countries, with less understanding of the unique challenges faced by pedestrians in a developing country (Pareekh et al., 2019).

The recent report published by the Ministry of Road Transport and Highways (MoRTH) shows that in the year 2019, in India, 25,858 pedestrians were killed (17% of overall fatalities). Moreover, as per the MoRTH data, every year pedestrian-related death is increasing by 2% from the last few years (MoRTH, 2019). Pedestrian safety status in urban India is more severe as most of the Indian metropolitans have been recording more than 50% pedestrian-involved fatal crashes in recent years (Mohan et al., 2016; Mitra et al., 2016; Mohan et al., 2009; Pathak et al., 2009).

In Indian cities pedestrian volume and density are significantly high (NTDPC, 2012); however, pedestrian-friendly infrastructure is largely absent (Gupta et al., 2009). The basic requirements of the pedestrians in terms of the marked crosswalk, sidewalk, convenient grade-separated crossing facilities, and traffic signal with pedestrian phases are often
missing or inadequate (Priyadarshini and Mitra, 2018; Jha et al., 2017). As a result, pedestrians are forced to share the same road space with motorized traffic and are sometimes bound to cross the roads, even in the presence of cross traffic, leading to conflicts, crashes, and even fatalities. It is also important to highlight that motorization and infrastructure development are rather new in Indian cities (Gupta et al., 2009). Therefore, there are deficiencies in both infrastructure as well as information and understanding on risk from motorized traffic. In addition, in urban India, the focus of road safety improvement projects has mostly been restricted to enhancing the safety and efficiency of vehicular traffic rather than non-motorized road users (Gupta et al., 2009). As a result, pedestrian safety remains a major social concern in urban India.

Effective interventions to protect pedestrians and to encourage safe walking require an understanding of risk factors associated with pedestrian injuries and fatalities (Stoker et al., 2015; Mitra and Mukherjee, 2017). Globally a good number of research studies have pointed out several reasons for pedestrian crashes. Among several risk factors, vehicle and pedestrian volume are identified as the most important factors associated with crashes (Mukherjee and Mitra, 2019a; Mukherjee and Mitra, 2019b; Mukherjee and Mitra, 2019c; Mukherjee and Mitra, 2019d; Moreno et al., 2011; Torbic et al., 2010; Ewing and Dumbaugh, 2009; Lee and Abdel-Aty, 2005). Besides traffic parameters, built environment and road geometry play an important role in pedestrian safety (Mukherjee and Mitra, 2020a; Mukherjee and Mitra, 2020b; Mukherjee and Mitra, 2020c; Priyadarshini and Mitra, 2018; Corazza et al., 2016; Rankavatk and Tiwari, 2015). Inefficient design of road network (Boarnet et al., 2005), the presence of on-road parking (Agran et al., 1996), lack of sight distance (Mukherjee and Mitra, 2020d; Mukherjee and Mitra, 2020e; Mukherjee and Mitra, 2019b), the presence of wider at grade crossing (Mukherjee and Mitra, 2020f; Priyadarshini and Mitra, 2018), and lack of visibility during the night-time (Loukaitou-Sideris et al., 2007) increase the likelihood of pedestrian crash occurrence. The vehicular speed is also recognized as an important factor influencing crash frequency as well as crash severity (Mukherjee and Mitra, 2020a; Kong and Yang, 2010; Rosén and Sander, 2009; Gärder, 2004).

Beyond the road infrastructure and traffic operational parameters, land use type also has a significant role in pedestrian safety. Harwood et al. (2008) identified that the presence of public-school zones significantly affects pedestrian fatality risk. Mukherjee and Mitra (2020a) showed that a high share of the commercial sector without adequate crossing facilities for pedestrians increases the likelihood of pedestrian-vehicular crashes. Furthermore, a few studies showed the impact of vehicle characteristics on pedestrian crash severity (oh et al., 2005; Matsui, 2014).

Pedestrian’s dangerous crossing behavior and the pedestrian’s state of crossing such as signal violation (Schwebel et al., 2012; Ren et al., 2011; Zhuang and Wu, 2011), not using designated pedestrian crosswalk (Zhang and Zhang, 2019), running behavior (Preusser et al., 2002), usage of electronic devices while crossing (Mukherjee and Mitra, 2020b; Mukherjee and Mitra, 2019b), etc. also increase the likelihood of pedestrian-vehicular crashes. Such road use behavior is often linked with poor road planning and design as was demonstrated by several prior research studies. For example, the lack of sight distance at a junction is likely to increase pedestrian signal violation probability by 76% (Mukherjee and Mitra, 2020b). Pedestrians experiencing more than 48 s delay before crossing at an urban midblock crosswalk, become intolerant and start involved in jaywalking (Chandrap et al., 2016). Improper location of a zebra crossing or non-standard marking of a zebra crossing is likely to increase pedestrian’s perceived risk by 34% (Mukherjee and Mitra, 2021). The absence of a dedicated pedestrian phase at a signalized junction is also expected to increase pedestrian signal violation probability by 33% (Mukherjee and Mitra, 2020b) and perceived crossing risk by 22% (Mukherjee and Mitra, 2021). Pedestrian’s gap acceptance, rolling behavior, and running behavior are also significantly affected by road geometries such as the lack of physical median, the presence of uncontrolled junction, absence of refuge island (Zafri et al., 2020; Mukherjee and Mitra, 2020a), and traffic flow characteristics (Ishaque and Noland, 2008). Alternatively, the presence of a refuge island, marked and accessible pedestrian crosswalk, and wider sidewalk facility with pedestrian guardrail significantly recover pedestrian crossing behavior and safety (Xu et al., 2013).

Past studies also documented the influence of location-specific sociodemographic factors on pedestrian crossing behavior as well as safety (Mukherjee and Mitra, 2020a; Mukherjee and Mitra, 2020c; Wang et al., 2016; Chakravarthy et al., 2012; Ukkusuri et al., 2011; Laflamme et al., 2010; Mabunda et al., 2008). A few studies showed the significance of road safety campaigns and awareness programs such as ride safe programs, road courtesy campaigns, and the anti-drink drive campaign to enhance pedestrian safety; even though the return on safety education show week evidence of effectiveness (Zhang et al., 2013; Haque, 2011; Lam, 2001). Therefore, the past pieces of literature on pedestrian safety highlight that there are more than one major source of risk factors, which may be characterized into three broad components (three major groups of pedestrian risk factors) namely (a) factors associated with road infrastructure planning and design, land use, traffic exposures, and operational characteristics, (b) factors associated with pedestrians’ crossing behavior and risk perception, and (c) factors associated with location-specific sociodemographic characteristics.

To evaluate the safety performance of existing traffic facilities, researchers have paid attention to the development of crash prediction models which is also known as Safety Performance Function (SPF). There are several motivations for developing SPFs, including the identification of risk factors, predicting crash frequency, crash severity, and the capability to analyze unsafe sites (Anastasopoulos and Manering, 2009). Researchers have primarily applied single equation crash prediction models to develop SPFs. As road traffic crashes are essentially random discrete events, the Poisson regression analysis is a suitable approach to estimate the likelihood of a crash occurrence. However, the Poisson regression model has some potential issues. The most important constraint of the Poisson model is the assumption of the equality of mean and variance of crashes (Washington et al., 2010; Chin and Quddus, 2003; Jones et al., 1991), which may be overcome by the negative binomial (NB) model specification (Anastasopoulos et al., 2012; Savolainen et al., 2011; Lord and Geedipally, 2011; Lord and Manering, 2010; Lyon and Persaud, 2002). For that reason, a good number of researchers have primarily executed the NB model to develop pedestrian crash prediction models (Mukherjee and Mitra, 2020c; Mukherjee and Mitra, 2020d; Mitra et al., 2019; Mukherjee and Mitra, 2019d; Mukherjee and Mitra, 2018; Pulugurtha et al., 2012; Pulugurtha and Sambhara, 2011). Even though utilizing the more precise NB model as an alternative to the Poisson model, the parameters of these conventional crash prediction models are likely to be fixed while they can vary across the observations; consequently, the heterogeneity problem remains unsolved (Park et al., 2016). To capture the effects of unobserved heterogeneity of several variables across sites researchers have suggested a random parameter NB model instead of a fixed-parameter NB model (Ükkusuri et al., 2011). Further, the Zero Inflated Negative Binomial model has also been utilized by the investigators to model pedestrian crash incidence when crash data are characterized by a preponderance of zero (Pour et al., 2012; Shankar et al., 2003).

The basic assumption behind these traditional single equation crash prediction models (i.e., SPFs) is that there is a single underlying source of risk factors (Schlabach et al., 2012). In contrast, the pedestrian crash count may be an outcome of multiple sources of risk factors, including road infrastructure, traffic volume, speed, and operational parameters, pedestrians’ unsafe crossing behavior and risk perception; and sociodemographic characteristics of a site, etc. However, these multiple sources are generally not captured in traditional single equation crash prediction models. Several studies reported that in Indian cities pedestrian’s unsafe crossing behavior and risk-taking attitude positively increases the likelihood of pedestrian-vehicular crashes.
3. Modelling methodology

The present section explains the concept of the multiple risk source model that accommodates three sources of risk factors (i.e., three risk components). A brief description of the proposed modelling methodology is given first, subsequently, goodness-of-fit measures utilized to evaluate the prediction performance of the proposed model are explained.

2.1. Model development

To associate the crash frequency with the intersection parameters, the Poisson or NB model is usually preferred. To overcome the issue related to over-dispersion, the Negative Binomial models are commonly exploited (Washington et al., 2010; Washington and Haque, 2013).

2.1.1. Single equation NB model

In the single equation crash prediction model Y_i indicates the total crash frequency at the intersection “i”. The total crash count at the intersection “i” (Y_i) follows a Poisson probability distribution with the mean \mu_i (Washington and Haque, 2013; Washington et al., 2010),

\[ Y_i \sim \text{Poisson}(\mu_i) \]

The mean value of the Poisson model is as follows:

\[ \log(\mu_i) = \beta X_i \] (1)

Where \( X_i \) is representing site-specific attributes, \( \beta \) is the regression parameter. To overcome the problem associated with over-dispersion, a gamma-distributed error term (\( \epsilon \)) is included the Poisson model and the expression for NB model is as follows:

\[ Y_i \sim \text{NB}(\mu_i, \phi) \]

\[ \log(\mu_i) = \beta X_i + \epsilon_i \]

\[ \mu_i = \alpha_0 F_{1i} \hat{F}_{2i} \exp \sum \beta X_i \] (2)

Where, \( \alpha_0 \) is the constant term. \( \phi \) is the dispersion parameter. \( F_{1i} \) is the average daily traffic volume at the junction and \( F_{2i} \) is the average daily pedestrian-vehicular volume ratio at the junction. \( \gamma_1 \) and \( \gamma_2 \) are the estimated regression parameter.

2.1.2. Three-Components mixture model

The primary concept of the mixture model was first developed by Washington and Haque (2013). The authors assumed that the total crash count at a site arises from three risk-generating processes. It was claimed that these three processes can be modeled distinctly to gain better insight into crash causes and that the model signifies a more accurate representation of reality compared to the conventional NB models. To determine the relative weights for different risk components the authors applied a simulation-based approach. Further, Afghari et al. (2016) utilized the concept of multiple risk-generating processes to identify the predominant source of risk in treating back spots. A Bayesian latent class model with random parameters was established by the authors to capture the unobserved heterogeneity across the sites. To account for the effects of endogeneity between two risk-generating processes simultaneous equation and instrumental variables models were developed. The authors suggested that the complexity of crash incidence can be well approached using three primary sources of risk factors, that is, (a) engineering, (b) unobserved spatial, and (c) driver behavioral factors. The authors also confirmed that the multiple risk source model significantly outperformed the conventional single risk source model in terms of prediction ability and goodness of fit measures.

Even though the joint econometric model provides better overall estimation, the ordinary “three-components mixture model” was primarily developed in this study as the starting point to model multiple sources of risk factors. Subsequently, a simultaneous equation model was also established to account for the possible overlap of the risk sources due to the correlation of the elements of risk factors included in the model.

The basic idea of the mixture model is that \( Y_i \) (total crashes at a site ‘i’) arises from multiple separate sources of risk factors (three distinct
A latent mixture modelling approach was established to associate multiple risk sources with the mean of the Negative Binomial distribution. The risk associated with road infrastructure, land use, planning, and traffic operational characteristics was assumed as

$$\theta_1 \sim U[0.20, 0.80]$$

Further, the risk associated with pedestrian crossing behavior and risk perception was assumed as

$$\theta_2 \sim U[0.10, 0.70]$$

Therefore, the risk associated with location-specific sociodemographic characteristics was assumed as

$$\theta_3 = \{1 - (\theta_1 + \theta_2)\} \sim U[0.05, 0.33]$$

Eq. (8) indicates an investigation where total crashes are randomly and consistently drawn from the observed “road infrastructure, land use, planning, traffic exposures and operational factor” between 20% and 80%, from the “behavior and risk perception related factors” between 10% and 70%, and “sociodemographic factors” influence the remainder. Subsequently, using these distribution weights ($\theta_1, \theta_2, \theta_3$), observed total crash counts were allocated to three distinct crash frequency counts for each intersection “i”.

Since there is no available literature that defined the range of various sources of pedestrian risk factors in the context of urban India; a wide range for each category of risk source was primarily assumed. Using these distribution weights, observed crash counts were assigned to three frequency counts for each intersection. To assess the consistency and sensitivity of outcomes to assumed weights, 20 different distribution proportions and corresponding models were estimated. The best-fitted model was selected from the 20 sets of crash prediction models.

### 2.1.3. Joint multiple risk source model

Three sources of risk factors described previously may be correlated via observed or unobserved factors. However, the “three-component mixture model” does not consider the intercorrelations or the overlaps between the risk components (Washington and Haque, 2013). Overlooking these correlations may consequence in biased parameter estimates and will yield incorrect implications on the effects of variables on risk. Therefore, to account for possible associations across risk components a simultaneous equation modelling was additionally established in this context.

It was assumed that $Y_i$ is the total number of the observed crash count at intersection $i$, and $k$ is the total number of risk components or sources of risk factors as described previously. The contribution of each source to the total observed crash count is unobserved, or latent. A starting point for estimating these latent probabilities is to first utilize the probability of total observed crash count across intersections $Y_i$, which approximately follows Negative Binomial distribution with mean ($\mu_i$) and dispersion parameter ($\phi$):

$$Y_i \sim \text{NB}(\mu_i, \phi)$$

A latent mixture modelling approach was established to associate multiple risk sources with the mean of the Negative Binomial distribution.
\[
\mu_i = \sum_{k=1}^{K} \mu_k \eta_k^i \tag{9}
\]
\[
\mu_k = \eta_k \mu_i \tag{10}
\]

where \(\sum_{k=1}^{K} \eta_k = 1\) and \(\eta_k\) is the proportion (or weight) of the total predicted crash count at intersection \(i\) attributed to latent risk source \(k\). Assuming exponential functions for the decomposed means of the negative binomial distribution, each of the above-mentioned predicted means is a function of a variety of contributing factors associated with the unique risk source \(\text{Afghari et al., 2018}\):

\[
\hat{Y}_k = g(a_0, \hat{X}_k) + \delta_k \tag{12}
\]
\[
\hat{X}_k\text{ is the predicted value of the endogenous variable, } a_0 \text{ is the set of estimable parameters, } \delta_k \text{ is the random error term, and follows standard Normal distribution. The functional form of the structural equation, } g(\cdot) \text{ is left to the analyst and should be defined based on the characteristics of the original endogenous variable represented by the instruments (e.g. definite positive) (Afghari et al., 2018). The original endogenous variable } X_k\text{ is left to the analyst and should be defined based on the characteristics of the original endogenous variable represented by the instruments (e.g. definite positive) (Afghari et al., 2018).}

2.2. Goodness-of-Fit test

In the present study, two global goodness-of-fit standards, namely, (a) Mean Squared Predictive Error (MSPE), and (b) Predictive Loss Criteria (PLC) were applied to select the best model (Washington and Haque, 2013). Assume, \(\xi_i\) and \(\zeta_i\) are mean and variance of maximum likelihood estimation-based crash prediction for location ‘i’. The expression for MSPE is as follows:

\[
\text{MSPE} = \sum_{k=1}^{3} \left( \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 \right) / N \tag{14}
\]

Where \(k\) denotes the number of sources of risk factors or crash occurrence processes (in the present study, \(k = 3\), as a three-components model has established), \(Y_i\) is the observed crash data and \(N\) is the number of observed sites. The MSPE depends on the mean of likelihoods and does not take into account the variance of predictions. Alternatively, the PLC (Haque et al., 2010) comprises the variance of predictions and therefore PLC might be a more accurate measure for model selection,

\[
\text{PLC} = \sum_{k=1}^{3} \sum_{i=1}^{N} \zeta_i + \left( \frac{w}{w+1} \right) \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 \tag{15}
\]

Where \(w\) is the weight factor. Subsequently, equal weight was placed for variance and mean differences to calculate the PLCs values (Haque et al., 2010). From the statistical point of view, models with comparatively lower MSPE and PLC values are considered better models.

3. Research design

The research methodology includes collection crash data, identification of risky intersections for pedestrian crashes, survey and data collection, the formation of traditional and proposed crash frequency models. The methodological framework of the current study is presented in Fig. 1.

3.1. Crash data collection and selection of study intersections

The crash data (2011–2016) were collected from the headquarter of the Kolkata traffic police department. The crash database comprises the following parameters:

- Location of the fatal crashes only
- Crash severity level
- Time of the crash incidence (i.e., date and time of the crash)
- Road user’s types
- Details of the accused vehicle
- Demographic characteristics of the victim

However, the data lacked a few important information such as the manner of collisions, socio-economic characteristics of the crash victim, weather conditions during the crash, road infrastructure, and geometric condition of the crash site, etc. Further, in a developing nation such as India, underreporting of injury crashes is a significant problem (Singh et al., 2018; Mitra et al., 2016; Dandona et al., 2008); hence, the scope of the current study is restricted to fatal crashes only.

The past study conducted by Mukherjee and Mitra (2019e) found that the risk of fatal pedestrian crashes is not uniform across the city. Therefore, to recognize the high-risk intersections, nine major road corridors in Kolkata were initially selected such that a maximum number of crash-prone intersections could be incorporated for further study (Chakraborty et al., 2019). Subsequently, 110 major intersections were selected from nine critical corridors (Chakraborty et al., 2019). The selected study locations are presented on a GIS-based map (Fig. 2).

3.2. Survey and data Collection:

Data collection for this study involved four distinct groups, namely (a) road inventory survey, (b) speed studies, (c) video-graphic survey, and (d) questionnaire survey.

A road inventory survey was executed to collect information related to road geometry, land use type, sight distance, road making and signage, presence and accessibility of pedestrian crosswalks, pavement condition, and type of junction (signalized/unsignalized), etc. (Mukherjee and Mitra, 2019a). The road inventory survey was conducted by a group of survey experts. The survey experts visited all sites, and collected/recorded the relevant information (Table 1).

Afterward, a spot speed survey was performed to estimate the average vehicle speed at a junction (Chakraborty et al., 2019; Mukherjee and Mitra, 2020c; Mukherjee and Mitra, 2020d). The ‘spot speed survey’ was conducted for morning 10 AM and 11 AM, and afternoon 3 PM and 4 PM, so that both peak and off-peak hour vehicle speed could be captured. At all study locations, a minimum of 50 samples (Chakraborty et al., 2019) for each vehicle category were taken in both peak and off-peak hour vehicle speed could be captured. At all study locations, a minimum of 50 samples (Chakraborty et al., 2019) for each vehicle category were taken in both peak and off-peak hours to determine the average vehicular speed at the site.

To estimate average daily traffic and pedestrian volume, and pedestrian crossing behavior video graphics survey was performed across 110 intersections. The video-graphics survey was conducted for 24 h at each of the intersections. In this study, the entire pedestrian and classified traffic volume with a turning movement were extracted for 24 h at each of the study sites. Besides, pedestrians’ crossing behavioral data was extracted for six hours between morning 10 AM and 1 PM, and
afternoon 3 PM and 6 PM. The time frame for pedestrians’ crossing behavior data extraction was selected in such a way that pedestrian-vehicular interactions could be capture during both peak and off-peak hours.

The video data was upgraded with millisecond time code using ‘AVS video editor 5.1’ software. The pedestrians’ crossing behavioral data were extracted with an accuracy of the millisecond by clicking a step forward option of the video extraction software. Frame-by-frame progress of videotapes was examined by research associates manually, and values were recorded in a pre-designed excel format. A total of 128,854 pedestrians’ crossing behavior was extracted from 110 intersections. Pedestrians’ crossing behavior data includes (a) whether a pedestrian is following zebra crossing, (b) pedestrian’s waiting time before crossing, (c) the overall time to cross the junction, (d) pedestrian-vehicular post-encroachment time (Table 1). Additionally, information related to pedestrian’s demographic characteristics and pedestrian’s state of the crossing were also extracted from the video images (Table 1). To reduce human error in the data extraction process, complete video data was extracted by three/four well-trained members.

To assess pedestrian risk perception a questionnaire survey was also conducted across the study sites (Fig. 3). The questionnaire survey timing also included peak and off-peak hours covering the same periods of pedestrians’ behavioral data extraction. A questionnaire survey instrument was prepared in a way so that much realistic information could be captured within a short duration. Several well-trained interviewers were used to examine the pedestrians’ perceptions towards safety and satisfaction level when using a particular crossing. Meaning and importance of the questions were explained to each pedestrian personally to obtain their responses on (a) crossing difficulty (b) safety, and (c) satisfaction with the overall environment of the junction on a scale of 1 to 6; where 1 represents ‘excellent’ and 6 represents ‘very poor’ condition i.e., ‘most difficult’ or ‘unsafe’ or ‘uncomfortable’ as applicable. Besides, the pedestrian’s age, gender, and the frequency of using that particular crossing were also asked. The survey respondents were randomly selected on the site based on their willingness to participate in the questionnaire survey. From each study site at least 50 pedestrians’ responses were collected, and a total of and 6875 pedestrians’ perception data were collected from 110 intersections. The questionnaire survey results indicate that 73% of the total respondents were male and 27% of the respondents were female. 9% of the overall respondents were below 18 years, 79% of the total respondents were between the age group 18 and 49 years, and the rest of the respondents were the age group above 50 years. 81% of the overall respondents were daily users and 19% of the survey respondents were frequent users.

The process of video-graphic survey, data extraction, and questionnaire survey was also implemented by former researchers (Killi and Vedagiri, 2014; Mukherjee and Mitra, 2019b; Mukherjee and Mitra, 2020a). Finally, the information related to the sociodemographic appearances of an intersection was extracted from Census Database 2011, Govt. of India (Mukherjee and Mitra, 2020b). The variables collected and or extracted from different surveys are summarized in Table 1. A few site-specific observations are presented in Fig. 4.

To assess pedestrian safety status at the intersection level, all the pedestrian level data collected or extracted from the questionnaire survey and video-images were combined across each junction to obtain intersection-specific information such as the average waiting time of the pedestrians at an intersection, the average crossing time, post-encroachment time of a junction, the share (or proportion) of the male or female pedestrians at an intersection, the median value of pedestrians’ perception regarding crossing difficulty at a junction, etc. (Mukherjee and Mitra, 2020b).

4. Result

In the current study, statistical models were estimated using “R-
Programming” and “Stata-13 statistical package” software. Ther significant outcomes obtained from different models (traditional NB model and suggested three-components mixture model, joint multiple risk source negative binomial model) are explained in the subsequent subsections.

4.1. Single equation NB model

To establish a single equation fatal pedestrian crash prediction model, NB regression analysis was performed because the variance is statistically and significantly higher than the mean value of the crashes (Table 2). To present the fatal pedestrian crash occurrence in terms of parameters associated with “road infrastructure, planning, land use, traffic operational characteristics”; “pedestrians’ crossing behavior and risk perception” and “location-specific sociodemographic factors”, initially three separate SPFs were formed. Sequentially, a combined model with several independent variables (i.e., considering three major sources of pedestrian risk factors) was developed. In the case of the combined model, to overcome the issues associated with multicollinearity, only certain independent variables were taken into account for the modelling purpose, even though they were significant when individual effects were checked. Traffic exposures such as variables associated with pedestrian and vehicle volume were considered as common parameters in all models because there are no probable crashes without traffic exposures.

To establish the relationship between fatal pedestrian crash frequency and variables associated with “road infrastructure, land use, planning, and operational parameters” an NB regression model was developed (Table 2, model 1). The model outcomes show that the logarithm of average daily vehicle volume at an intersection, speed, and pedestrian-vehicular volume ratio positively and significantly affect fatal pedestrian crash frequency. It was also observed that the presence of a high share of the commercial hub, inaccessibility of pedestrian crosswalks, and inadequate sight distance considerably increase the possibility of fatal pedestrian crashes. On the other hand, the model result shows that the presence of police personnel is helpful to reduce pedestrian crash incidence in Kolkata city.

Similarly, another NB regression model was formed to identify the association between pedestrians’ crossing behavior, risk perception, and the police-reported fatal pedestrian crashes. The model findings are shown in Table 2. The findings obtained from ‘model 2’ show that pedestrians’ risky crossing behavior such as a high share of pedestrians “not following zebra crossing” at a junction, and longer waiting time

Fig. 2. Study intersections.
before crossing increases the possibility of fatal crashes. Pedestrian-vehicular post encroachment time was also found to be negatively associated with fatal pedestrian crashes at a junction indicating that the risk of fatal pedestrian crashes substantially increases with reducing in post-encroachment time. The model result also indicates a positive association between pedestrians’ perceived crossing difficulty and fatal pedestrian crash occurrence.

Model 3 presents the relationship between fatal pedestrian crash incidence and several factors related to sociodemographic aspects of the site (Table 2). The presence of the slums near the intersection was found to be strongly associated with the fatal pedestrian crash incidence of a junction. Similarly, it was found the presence of a “pedestrian attraction zone” near a junction without an adequate crossing facility increases pedestrian fatality risk.

The combined model with risk factors from three major sources resulted in a single equation NB model (model 4 in Table 2) reveals that logarithm of daily traffic volume or log(ADT) ($\beta = 1.55$; $p < 0.001$), pedestrian-vehicular volume ratio of an intersection ($\beta = 0.30$; $p < 0.001$), speed of the motorized vehicle ($\beta = 0.04$; $p < 0.001$), inaccessibility of pedestrian crosswalk ($\beta = 0.61$; $p < 0.001$), inadequate sight distance ($\beta = 0.43$; $p < 0.001$) and lower value of post-encroachment time ($\beta = -0.96$; $p < 0.001$) significantly increase the probability of

---

### Table 1

| Characteristic                      | Variables                                      | Source of Data                  | Type of Variable Description and Explanation |
|-------------------------------------|------------------------------------------------|---------------------------------|---------------------------------------------|
| **Dependent Variable**              | Fatal pedestrian crashes at an intersection    | Kolkata Police                  | Continuous (integer value)                   |
| **Independent Variable**            | Log (Average Daily Traffic Volume / ADT)       | Video-graphic Survey            | Continuous Average daily pedestrian volume (Mukherjee and Mitra, 2019a) |
|                                     | Pedestrian vehicular ratio                      | Video-graphic Survey            | Continuous The ratio of average daily pedestrian and vehicle volume (Mukherjee and Mitra, 2019a) |
| **Traffic Operational Characteristics** | Speed                                          | Spot speed Study                | Continuous The average vehicular speed of motorized traffic at an intersection (Chakraborty et al., 2019) |
|                                     | Overtaking tendencies of vehicles              | Video-graphic Survey            | Categorical Overtaking behavior of the vehicle driver at an intersection: Presence = 1; Absence = 0 (Mukherjee and Mitra, 2020b) |
|                                     | Traffic police                                  | Site visits                     | Categorical Presence = 1; Absence = 0 (Mukherjee and Mitra, 2020b) |
| **Infrastructure and Roadway Factors** | Width of road                                   | Road Inventory Survey           | Continuous (meter) Major and minor road width (Mukherjee and Mitra, 2019b) |
|                                     | Presence of Zebra Crossing                     | Road Inventory Survey           | Categorical Presence = 1; Absence = 0 (Mukherjee and Mitra, 2020b) |
|                                     | Pedestrian Following Zebra Crossing (i.e., zebra crossing): Yes (1) No (0) | Video-graphic Survey            | Categorical Presence = 1; Absence = 0 (Mukherjee and Mitra, 2020b) |
|                                     | Waiting Time before Crossing                   | Video-graphic Survey            | Continuous (Sec) Waiting time of the pedestrian before crossing the road (Mukherjee and Mitra, 2020b) |
|                                     | Crossing Time                                   | Video-graphic Survey            | Continuous (Sec) Pedestrian’s crossing time at an intersection (Mukherjee and Mitra, 2020b) |
|                                     | Post Encroachement Time (PET)                  | Video-graphic Survey            | Continuous (Sec) Time difference between the end of encroachment of crossing pedestrian and the time that the through vehicle reaches at the possible point of the collision (Prijavardhanini and Mitra, 2019). |
| **Pedestrians’ State of Crossing**  | Pedestrian carrying oversized loads: Yes (1) No (0) | Video-graphic Survey            | Categorical Pedestrian is carrying a load on his/her head that exceeds the standard or ordinary legal size that obstruct visibility of the pedestrian (Mukherjee and Mitra, 2019b) |
|                                     | Distracted pedestrian: Yes (1) No (0)          | Video-graphic Survey            | Categorical Pedestrian is using an electronic device while crossing (i.e., using a cell phone, tablet, etc.) (Mukherjee and Mitra, 2019b). |
| **Pedestrians’ Risk Perception**    | Pedestrian’s perceived satisfaction level       | Questionnaire Survey            | Ordered Based on the pedestrian perception: Excellent = 1 to Very Poor = 6 scale, (Likert scale) (Mukherjee and Mitra, 2021) |
|                                     | Pedestrian’s perceived difficulty               | Questionnaire Survey            | Ordered Not Difficult = 1 to Highly Difficult = 6 (Mukherjee and Mitra, 2021) |
| **Factors Associated with Location-Specific Socio-Demographic Characteristics** | Log (Total Population at the Junction)         | Census India, 2011              | Continuous Overall population near the intersection (i.e., population density) (Mukherjee and Mitra, 2020b) |
|                                     | Slum Population                                 | Road Inventory Survey           | Continuous (in %) The portion of the slum Population near an intersection (Mukherjee and Mitra, 2020b) |
|                                     | Presence of Zone of Attraction: Presence = 1; Absence = 0 | Road Inventory Survey | Categorical Presence of Zone of Attraction near the intersection (Mukherjee and Mitra, 2020b) |
fatal pedestrian crashes. The combined model outcomes with acceptable goodness of fit value ($\rho^2 = 0.300$) and a significant value of dispersion parameter ($\alpha = 0.11, p < 0.100$) indicate that the model presented in this paper is satisfactorily fitted to predict the fatal pedestrian crash occurrence using these independent variables.

4.2. Three-Components mixture regression model

To develop the three-components mixture model, twenty sets of trial models were developed with different randomly drawn weight distributions (i.e., $\theta_1, \theta_2, \theta_3$). Subsequently, the best-fitted model was chosen based on standard goodness-of-fit criteria and logical application of variables. Goodness-of-fit values (MSPE and PLC values) for these twenty trial models are shown in Table 3. Among these twenty trial model sets, the 7th trial shows the best fit compared to the other trials with a Mean Squared Predictive Error of 1.24 and Predictive Loss Criteria of 312.

Estimation results of the best-fitted model (i.e., 7th trial) are presented in Table 4. In this trial, it was assumed that in Kolkata, pedestrians’ unsafe crossing behavior and poor risk perception are implicated in 40% of overall pedestrian fatal crashes; lack of infrastructure and risky traffic operation is implicated in 50% of the total fatal crashes, and 10% of the fatalities are implicated with location-specific sociodemographic aspects (shown in Fig. 5).

To estimate the risk associated with “road infrastructure, planning, land use, traffic exposures, and operational characteristics” the poisson regression model was chosen because the variance of fatal pedestrian crashes was not statistically significantly different than the mean value of the crashes ($\sigma = 1.10, \mu = 1.00$). The outcome of the Poisson model shows that logarithm of ADT, a higher value of pedestrian-vehicular volume ratio, speed, inadequate sight distance, inaccessibility of crosswalk, the presence of overtaking tendency of vehicles, land-use patterns such as a high share of the commercial area or residential area considerably affect the possibility of fatal pedestrian crashes at intersection-level in Kolkata.

To estimate the risk related to pedestrians’ unsafe crossing behavior and risk perception, the Poisson model was preferred because the over-dispersion parameter for the negative binomial model was insignificant (Mitra et al., 2017) ($\sigma = 1.00, \mu = 0.73$). The model outcomes indicate that a lower value of post-encroachment time, higher crossing difficulty, and poor satisfaction level at an intersection significantly increase the odds of fatal pedestrian crashes. Pedestrian’s state of crossing such as a high proportion of “pedestrian carrying overhead load” was found to be significantly associated with fatal pedestrian crash frequency at an intersection.

To study the impact of sociodemographic factors, another Poisson model was developed ($\sigma = 0.21, \mu = 0.20$). The model outcome suggests that pedestrian fatality risk is significantly higher near the slum areas. It was also established that the probability of pedestrian fatal crash occurrence significantly increases with the presence of a “pedestrian zone of attraction” such as an educational institute, heritage building, hospital, shopping mall, bar, and pubs, etc. There is also evidence that the odds of fatal pedestrian crashes are generally higher where population density is too high.

4.3. Comparative study between traditional and proposed mixture models

This section compares the performance of several alternative models (Table 5). The Mean Squared Predictive Error for the traditional single equation combined negative binomial model is 2.27 (model 4, Table 2), while the Mean Squared Predictive Error for the three-components mixture model is about 1.24 (Table 3). The Predictive Loss Criteria for the single equation combined NB model is 526 (model 4, Table 2), whereas the Predictive Loss Criteria value for the three-components mixture model is 312 (Table 3). The study outcomes indicate that the three-components mixture model is statistically and significantly superior to the traditional single-equation models, and the accuracy level of the suggested model is nearly 41% better compared to the traditional single equation model. Fig. 6a and b show a comparative study of PLC and MSPE across four sets of single equation traditional models and the proposed three-components mixture model.

Fig. 7a compares the prediction performance of the traditional single equation combined model (model 4, Table 2), proposed three-components mixture models, and actual crash data across 20 intersections (randomly chosen from the dataset of 110 intersections). The figure certainly indicates that the accuracy level of the three-component mixture model is significantly better than the traditional models.

Fig. 7b presents a bar diagram for the crash components (sources of risk factors) described previously, abbreviated as “road infrastructure, planning, land use, traffic exposures, and operational characteristics”, “pedestrians’ crossing behavior and perception”, and “sociodemographic characteristics”. The figure shows a set of 20 intersections...
randomly chosen from the dataset of 110 intersections) with crash frequencies on the vertical axis. The critical point here is that theoretically these three components of crashes contribute to the total crashes at an intersection (three different sources of pedestrian risk factors contribute to the overall crashes). Interestingly it was found that at several intersections, for example, “intersection 8” and “intersection 9”, the risk associated with sociodemographic characteristics is almost insignificant. However, at “intersection 6”, the risk associated with the sociodemographic characteristics is significantly higher compared to the other intersections. Similarly, it was observed that at “intersection 20”, the risk associated with pedestrians’ crossing behavior and perception is practically negligible; while at several junctions such as “intersection 1”, “intersection 2”, and “intersection 5” risks associated with pedestrians’ crossing behavior and risk perception are relatively higher. Alternatively, traditional single equation NB models are unable to capture such insights (Fig. 7c). Therefore, the comparative study between several alternative models evidently highlights the limitations of traditional count data models over the recommended three-components mixture model.
Table 2
Single equation fatal pedestrian crash prediction model.

| Characteristics                  | Attributes                                                                 | Single Equation NB Model-1 (Road infrastructure, planning, and land use, traffic operational characteristics) | Single Equation NB Model-2 (Pedestrians’ behavior and risk perception) | Single Equation NB Model-3 (Sociodemographic characteristics) | Single Equation NB Model-4 Combined Model |
|----------------------------------|---------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------|-----------------------------------------------------------------|------------------------------------------|
| Model Constant                   |                                                                           | Model Coefficients (t-stat)                                                                                  |                                                                       |                                                                  |                                          |
| Traffic Exposures                |                                                                           | 1.699 (3.72)**                                                                                              | 1.636 (2.48)**                                                         | 0.886 (2.12)**                                                   | 1.556 (3.93)**                            |
| Traffic Operational Parameters   |                                                                           | 0.272 (4.14)**                                                                                              | 0.149 (3.81)**                                                         | 0.118 (1.79)*                                                    | 0.301 (5.27)**                            |
| Land use                         |                                                                           | 1.204 (2.49)**                                                                                              |                                                                       |                                                                  |                                          |
| Road Infrastructure              |                                                                           | -0.560 (1.99)**                                                                                             |                                                                       |                                                                  |                                          |
| Pedestrian Crossing Behaviour    |                                                                           | -0.384 (2.03)**                                                                                             |                                                                       |                                                                  |                                          |
| Pedestrian Risk Perception       |                                                                           | 0.887 (5.73)**                                                                                              |                                                                       |                                                                  |                                          |
| Sociodemographic Factors         |                                                                           | 0.610 (3.29)**                                                                                              |                                                                       |                                                                  |                                          |
| Model Summary                    |                                                                           | Alpha (α)                                                                                                   |                                                                       |                                                                  |                                          |
|                                |                                                                           | 0.213 (7.87, p = 0.002)                                                                                     | 0.100 (2.97, p = 0.045)                                               | 0.181 (6.84, p = 0.003)                                         | 0.110 (1.82, p = 0.088)                   |
| Dispersion Parameter for Count Data Model |                       | Overall Goodness-of-fit                                                                                      |                                                                       |                                                                  |                                          |
|                                |                                                                           | Restricted Log-Likelihood function                                                                          |                                                                       |                                                                  |                                          |
|                                |                                                                           | -234.694                                                                                                    | -234.694                                                              | -234.694                                                        | -234.694                                  |
|                                |                                                                           | Log-Likelihood function                                                                                        |                                                                       |                                                                  |                                          |
|                                |                                                                           | -169.005                                                                                                    | -167.850                                                              | -171.295                                                        | -164.086                                   |
|                                |                                                                           | 𝜒²                                                                                                          |                                                                       |                                                                  |                                          |
|                                |                                                                           | 0.279                                                                                                        | 0.285                                                                 | 0.270                                                           | 0.300                                     |
|                                |                                                                           | Sample Size (i.e., Number of Intersections)                                                                |                                                                       |                                                                  |                                          |
|                                |                                                                           | 110                                                                                                         | 110                                                                   | 110                                                             | 110                                      |

*Significant at 90% Confidence Interval; **Significant at 95% Confidence Interval; ***Significant at 99% Confidence Interval.

Table 3
Compression of different sets of three-components mixture models.

| Trial No. | Share of Risk Categories (θ₁, θ₂, θ₃) | Behavior and Risk Perception (θ₄) | Location Specific sociodemographic factors (θ₅) | Global Goodness of Fit | MSPE PLC |
|-----------|--------------------------------------|----------------------------------|-----------------------------------------------|------------------------|----------|
|           | Road Infrastructure, Land Use, Planning, Traffic Operational Characteristics (θ₁) |                                 |                                               |                        |          |
| Trial 1   | 80                                   | 10                               | 10                                            |                        | 2.871    | 482.000 |
| Trial 2   | 70                                   | 10                               | 10                                            |                        | 2.149    | 421.000 |
| Trial 3   | 60                                   | 10                               | 10                                            |                        | 1.781    | 408.000 |
| Trial 4   | 60                                   | 20                               | 20                                            |                        | 1.852    | 383.000 |
| Trial 5   | 55                                   | 10                               | 10                                            |                        | 1.449    | 341.000 |
| Trial 6   | 50                                   | 20                               | 20                                            |                        | 1.472    | 321.000 |
| Trial 7   | 50                                   | 10                               | 10                                            |                        | 1.240    | 312.000 |
| Trial 8   | 50                                   | 5                                | 5                                              |                        | 1.379    | 329.000 |
| Trial 9   | 50                                   | 15                               | 15                                            |                        | 1.307    | 315.000 |
| Trial 10  | 45                                   | 5                                | 5                                              |                        | 1.366    | 316.000 |
| Trial 11  | 45                                   | 9                                | 9                                              |                        | 1.354    | 317.000 |
| Trial 12  | 45                                   | 10                               | 10                                            |                        | 1.355    | 317.000 |
| Trial 13  | 35                                   | 10                               | 10                                            |                        | 1.572    | 352.000 |
| Trial 14  | 43                                   | 20                               | 20                                            |                        | 1.364    | 317.000 |
| Trial 15  | 48                                   | 9                                | 9                                              |                        | 1.417    | 330.000 |
| Trial 16  | 53                                   | 8                                | 8                                              |                        | 1.295    | 313.000 |
| Trial 17  | 35                                   | 30                               | 30                                            |                        | 1.781    | 362.000 |
| Trial 18  | 33                                   | 33                               | 33                                            |                        | 1.800    | 363.000 |
| Trial 19  | 30                                   | 10                               | 10                                            |                        | 1.866    | 377.000 |
| Trial 20  | 20                                   | 70                               | 10                                            |                        | 2.425    | 442.000 |
The outcomes obtained from the “three-components mixture model” presented in this paper are extremely hopeful, but they are not finally conclusive and suffer from several challenges. Pedestrians’ behavioral data captured in this study were limited to the proxy variables reflecting the aspects of pedestrians’ unsafe crossing behavior that may be endogenous to road infrastructure planning design as well as the land use. Further, pedestrians’ crossing behavior may be influenced by location-specific sociodemographic factors. For example, pedestrian-vehicular PET may be influenced by road infrastructure, land use patterns, and traffic operational characteristics (Kadali and Vedagiri, 2016; Killi and Vedagiri, 2014). Similarly, pedestrians’ perceived crossing difficulty could be correlated with the lack of pedestrian-friendly infrastructure (i.e., absence of zebra crossing, traffic signal, etc.), vehicular operational characteristics (driver’s speeding behavior, yielding behavior, etc.), land use type, and several societal aspects (Mukherjee and Mitra, 2020a; Mukherjee and Mitra, 2019b; Kadali and Vedagiri, 2015). The share of pedestrians carrying overhead loads is generally influenced by the commercial activities of a location (Mukherjee and Mitra, 2020a). Hence, three sources of risk factors described previously may be correlated via observed or unobserved factors.

Further, in the three-components mixture model share of risk components (mixture weights) were assumed and finalized based on a simulation method, which in reality may be different. Thus, there may be a possibility of dissimilarity between the actual scenario and the assumed distribution of risk. In contrast, with the help of the joint multiple risk source negative binomial model, the mixture weights for risk sources (i.e., the contribution of each risk component) were modified based on the ratio of predicted crash count by each risk source to the total predicted crash count. The key observations obtained from the proposed model are summarized below.

To account for the endogeneity between three risk components, selected variables were instrumented by replacing their observed by

\[
\chi^2 (p-Value) 26.41 (0.010) *) \\
\chi^2 (p-Value) 83.58 (0.010) ** \\
\chi^2 (p-Value) 57.854 (0.010) *** \\
\chi^2 (p-Value) 44.648 (0.010) ** \\
\chi^2 (p-Value) 3.85 (0.010) ** \\
\chi^2 (p-Value) 2.31 (0.021) \\
\chi^2 (p-Value) 0.754 (0.060) \\
\chi^2 (p-Value) 1.65 (0.098) * \\
\chi^2 (p-Value) 1.66 (0.096) * \\
\chi^2 (p-Value) 1.44 (0.228) \\
\chi^2 (p-Value) 0.318 (0.574) \\
\chi^2 (p-Value) 83.58 (0.010) ** \\
\chi^2 (p-Value) 57.854 (0.010) *** \\
\chi^2 (p-Value) 44.648 (0.010) ** \\
\chi^2 (p-Value) 3.85 (0.010) ** \\
\chi^2 (p-Value) 2.31 (0.021) \\
\chi^2 (p-Value) 0.754 (0.060) \\
\chi^2 (p-Value) 1.65 (0.098) * \\
\chi^2 (p-Value) 1.66 (0.096) * \\
\chi^2 (p-Value) 1.44 (0.228) \\
\chi^2 (p-Value) 0.318 (0.574) \\
\chi^2 (p-Value) 83.58 (0.010) ** \\
\chi^2 (p-Value) 57.854 (0.010) *** \\
\chi^2 (p-Value) 44.648 (0.010) ** \\
\chi^2 (p-Value) 3.85 (0.010) **
predicted values obtained from structural equation models. Road infrastructure, land use, and traffic operational factors were included as instruments within the structural equations, and negative binomial density functions were used for the factors associated with pedestrians’ crossing behavior and risk perception. Upon estimating the model, mixture weights for risk sources (i.e., the contribution of each risk component) were computed based on the ratio of predicted fatal pedestrian crash count by each risk component (each source of risk factors) to the total predicted fatal pedestrian crash count at an intersection (Eq. (10)). The final mixture weight (proportion of risk components) of each risk component is shown in Table 6.

The average proportion (mixture weight) for “road infrastructure, land use, and traffic operational characteristics”; “pedestrians’ crossing behavior and risk perception” and “location-specific sociodemographic characteristics” instrumented on “pedestrians’ crossing difficulty” related risk factors were statistically significant. Pedestrian-vehicular PET is negatively correlated with the frequency of fatal pedestrian crashes indicating as PET value reduces pedestrian fatality risk increases. Pedestrian fatality risk increases with an increase in the share of pedestrians carrying an overhead load. Subsequently, it was found that pedestrians’ perceived crossing difficulty at an intersection is positively associated with pedestrian fatality risk. Pedestrian-vehicular volume ratio, speed, sight distance, accessibility of pedestrian crosswalk, commercial activities near an intersection, the presence of pedestrian attraction zone, pedestrian-vehicular post encroachment time, the percentage share of pedestrians carrying an overhead load, and pedestrians’ perceived crossing difficulty.

The factors associated with “road infrastructure, land use, and traffic operational characteristics” and “location-specific sociodemographic characteristics” instrumented on “pedestrians’ crossing behavior and risk perception” related risk factors were statistically significant. Pedestrian-vehicular PET is negatively correlated with the frequency of fatal pedestrian crashes indicating as PET value reduces pedestrian fatality risk increases. Pedestrian fatality risk increases with an increase in the share of pedestrians carrying an overhead load. Subsequently, it was found that pedestrians’ perceived crossing difficulty at an intersection is positively associated with pedestrian fatality risk. Pedestrian-vehicular volume ratio, speed, sight distance, accessibility of pedestrian crosswalk, commercial activities near an intersection, the presence of pedestrian attraction zone, pedestrian-vehicular post encroachment time, the percentage share of pedestrians carrying an overhead load, and pedestrians’ perceived crossing difficulty.

Table 5 presents the suggested joint multiple risk source negative binomial model, considering probable endogeneity effects of three sources of risk factors. The multiple risk source model includes average daily traffic volume, pedestrian-vehicular volume ratio, the average speed of the approaching vehicle, sight distance, accessibility of pedestrian crosswalk, commercial activities near an intersection, the presence of pedestrian attraction zone, pedestrian-vehicular post encroachment time, the percentage share of pedestrians carrying an overhead load, and pedestrians’ perceived crossing difficulty.

The factors associated with “road infrastructure, land use, and traffic operational characteristics” and “location-specific sociodemographic characteristics” instrumented on “pedestrians’ crossing behavior and risk perception” related risk factors were statistically significant. Pedestrian-vehicular PET is negatively correlated with the frequency of fatal pedestrian crashes indicating as PET value reduces pedestrian fatality risk increases. Pedestrian fatality risk increases with an increase in the share of pedestrians carrying an overhead load. Subsequently, it was found that pedestrians’ perceived crossing difficulty at an intersection is positively associated with pedestrian fatality risk. Pedestrian-vehicular volume ratio, speed, sight distance, accessibility of pedestrian crosswalk, commercial activities near an intersection, the presence of pedestrian attraction zone, pedestrian-vehicular post encroachment time, the percentage share of pedestrians carrying an overhead load, and pedestrians’ perceived crossing difficulty.

Table 7 presents the suggested joint multiple risk source negative binomial model, considering probable endogeneity effects of three sources of risk factors. The multiple risk source model includes average daily traffic volume, pedestrian-vehicular volume ratio, the average speed of the approaching vehicle, sight distance, accessibility of pedestrian crosswalk, commercial activities near an intersection, the presence of pedestrian attraction zone, pedestrian-vehicular post encroachment time, the percentage share of pedestrians carrying an overhead load, and pedestrians’ perceived crossing difficulty.
The multiple risk source model had lower MSPE (1.054) and PLC (310.4) compared to the single equation model (2.272 and 526.33 respectively) and the three-components mixture model (1.24 and 312.0 respectively), showing considerable improvement in statistical fit. Fig. 8 presents a comparative study between the single equation crash prediction model, the three-components mixture model, and the joint multiple risk source crash prediction model. The overall study outcomes indicate that the multiple-source risk models perform significantly better than the single equation risk model in terms of goodness-of-fit measures (Fig. 8a, Fig. 8b), and crash prediction ability (Fig. 8c).

Table 6
Contribution of Risk Components (Sources of Risk Factors) to Total Crash Count

| Characteristics                                      | Mean  | Standard Deviation | Maximum | Minimum |
|------------------------------------------------------|-------|--------------------|---------|---------|
| Road Infrastructure, Planning, Land Use, Traffic     | 0.528 | 0.399              | 1.000   | 0.000   |
| Operational Parameters                               |       |                    |         |         |
| Pedestrian Behavior and Risk Perception              | 0.385 | 0.336              | 1.000   | 0.000   |
| Location-specific Sociodemographic Factors          | 0.086 | 0.116              | 0.800   | 0.000   |

Fig. 7. Comparative study of single-equation and three-components model.
The present study argues from a hypothetical perspective that observed fatal pedestrian crashes in Kolkata city, India is not caused by a single source of risk factors and instead arise as an outcome of multiple sources of risk factors such as the road infrastructure and traffic operational factors, pedestrians’ dangerous crossing behavior and risk perception, and location-specific sociodemographic factors. Based on this concept, the present study primarily formulates a mathematical model that takes into account these three sources of pedestrian risk factors to explain the fatal pedestrian crash occurrence. Subsequently, a three-component mixture model, as well as a joint multiple risk source negative binomial model, are established to account for possible endogeneity effects between various sources of pedestrian risk factors.

Fatal pedestrian crash data for the years 2011 to 2016 of Kolkata, India, is utilized to demonstrate the modelling methodology. The goodness-of-fit of the proposed three-components mixture model and joint multiple risk source negative binomial model are compared with the traditional single equation count data models. The study outcome shows that both the three-components mixture model and joint multiple risk source negative binomial model presented in this paper perform better than the traditional single equation crash prediction models. The MSPE and the PLC are decreased by at least 1.7 times using the multiple risk source models, promising that from the statistical viewpoint the multiple risk source model offers a superior fit. Several key contributions to the present study are highlighted below:

- While researchers are generally utilized single equation crash prediction models (Sanhosh et al., 2020; Pulugurtha et al., 2012; Pulugurtha and Sambhara, 2011; Dissanyake et al., 2009), the present work demonstrates a unique approach to model fatal pedestrian crashes by combing three major groups of risk factors, namely, (a) road infrastructure, land use, planning, traffic exposures, and operational characteristics, (b) pedestrians’ unsafe crossing behavior as well as poor risk perception, and (c) sociodemographic features of the intersection. The modelling approach, when applied in the context of Kolkata City, is found to produce a model with superior goodness-of-fit as compared to the model developed using the conventional approach of modelling. The MSPE and the PLC are found to decrease by at least 41%, indicating that the proposed models perform better than the traditional crash prediction models.
- the three-components model and joint multiple risk source negative binomial model which are established in this paper are fairly new modelling techniques in the safety literature and previously not explored to evaluate pedestrian safety in particular. Moreover, in previous researches, elements associated with road users’ behavior were not taken into account as a part of the mixture models (Washington and Haque, 2013). Alternatively, in the current study, pedestrians’ behavioral data was collected and extracted based on a systematic approach, and pedestrians’ behavioral issues were considered as a major element of the proposed multiple source risk estimation models.
- Based on the systematic approach demonstrated in this study, it was found that pedestrians’ unsafe crossing behavior and poor risk perception (which is influenced by the lack of road infrastructure, land use pattern, and poor traffic operations) are implicated in 38% of overall fatal pedestrian crashes; while the lack of infrastructure and poor traffic operation and management still holds the highest share with 53% of the risk of the total fatal pedestrian crashes, and a small share of 9% comes from the location-specific sociodemographic aspects. The current finding indicates that the risk associated with lack of road infrastructure, land use pattern, and risky traffic operations are the foremost source of pedestrian risk in Kolkata city. However, the conventional single equation modelling technique is unable to capture the key source of hazard.
- If models were to predict crash counts shown in Fig. 7c as an alternative to Fig. 7c, then pedestrian safety investments could be scientifically targeted to recommend suitable countermeasures based on the major source of risk at a particular location.
- Methods that depend on safety performance functions—including the common concept of single equation modelling technique to identify hot spots—might not be effective since it would fail to shed light on the predominant source of pedestrian risk, thereby failing to target the key source of hazard. In other words, several intersections may be dominated by inefficient planning and design or lack of pedestrian infrastructures or traffic operational issues. The
reasonably large standard deviations for the proportion of risk factors (i.e., 0.39, 0.33, and 0.11, Table 6) indicate that the contribution of the three risk sources could vary across the intersections. However, the traditional single equation crash prediction model is unable to capture that insight.

• It is evidenced from the present study that pedestrians’ behavioral data captured in this study are correlated with road infrastructure planning design as well as land use. Thereby, it is very likely that the improvement in infrastructure, design and planning would improve pedestrians’ behavior as well safety.

Similar to other studies the present study is not without limitations. The limitations of the present study that need to be targeted in the future study are as follows:

• Firstly, in this study, the risk factors are grouped into three comprehensive categories because three types of risk factors were primarily observed in the context of urban India. However, the number of major risk sources may be more than three. Moreover, the grouping of risk factors primarily depends on the researchers.

• Secondly, in the three-components mixture model as the contributions of the risk components should sum to unity; it is mathematically impossible that all weights are probabilistic. Only two of the three weights can be probabilistic at a time.

• Lastly, the modelling approach was examined on one dataset and performed satisfactorily. Further, the application of the present modelling approach to other developing countries would be useful to develop a generalizable inference about the foremost sources of pedestrian risk factors. Moreover, the share of risk components associated with non-fatal pedestrian crashes may be very dissimilar to fatal crashes. Consequently, it will also be valuable for future studies to target non-fatal crashes to capture additional inferences.

Despite several restrictions, findings from the present study can provide a plausible direction in terms of the components of pedestrian risk at urban intersections in the framework of a developing country. Further, it was statistically justified that the total fatal pedestrian crash count of an urban intersection in a developing country is a combination of crashes influenced by multiple causal mechanisms. This understanding is the most important contribution of the present study.

Fig. 8. Comparative study single equation NB model, three-components mixture model, and joint risk source NB model.
Funding

The present study did not receive any funding from external agencies.

CRediT authorship contribution statement

Dipanjan Mukherjee: Conceptualization, Methodology, Data curation, Software, Writing - original draft, Investigation, Formal analysis, Supervision, Writing - review & editing. Sudeshna Mitra: Conceptualization, Methodology, Writing - original draft, Investigation, Supervision, Writing - review & editing.

Declaration of Competing Interest

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

References

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. record 2601 (1), 90–98.
Afghari, Amir Pooyan, Simon Washington, Md Mazharul Haque, Zili Li, 2018. A comprehensive joint econometric model of motor vehicle crashes arising from multiple sources of risk. Accid. Anal. Methods Accid. Res. 18 (2018) 1–14.
Agran, P.F., Winn, D.G., Anderson, C.L., Tran, C., Del Valle, C.P., 1996. The role of the physical and traffic environment in child pedestrian injuries. Pediatrics 98 (6), 1096–1103.
Anastasopoulos, P.C., Manning, 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., Manning, F.L., Shankar, V.N., Haddock, J.E., 2012. A study of factors affecting highway accident rates using the random-parameters tobit model. Accid. Anal. Prev. 45, 621–633.
Boarnet, M.G., Day, K., Anderson, C., McMillan, T., Alfonzo, M., 2005. California’s Safe Routes to School program: impacts on walking, bicycling, and pedestrian safety. J. Am. Plann. Assoc. 71 (3), 301–317.
Buehler, R., Pucher, J., 2021. The growing gap in pedestrian and cyclist fatality rates in urban areas. Transp. Rev. 41 (1), 61–85.
Buehler, R., Pucher, J., 2021. The growing gap in pedestrian and cyclist fatality rates in urban areas. J. Urban. Int. Res. Placemaking Urban Sustain. Journeys, LTA 6, 38–47.?
