IATSS Research 48 (2024) 378–392

Contents lists available at ScienceDirect

IATSS Research

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

Research Article

Developing accident frequency prediction models for urban roads: A case study in São Paulo, Brazil

Cassiano Augusto Isler [a,*], Yue Huang [b], Lucas Eduardo Araújo de Melo [a]

[a] Department of Transportation Engineering, Escola Politécnica, University of São Paulo, Av. Prof. Almeida Prado, Travessa 2, 83, 05508-900 São Paulo, SP, Brazil
[b] Institute for Transport Studies, University of Leeds, 34-40 University Road, Leeds LS2 9JT, UK

ARTICLE INFO

Keywords:
Urban roads
Crash prediction
Stepwise poisson regression
Road safety

ABSTRACT

The growing number of vehicles and the evolving behaviour of road users present new and additional challenges to road safety. Study on the variables that influence the frequency of crash occurrences such as road geometry, junction, speed and land use are needed as they have proven effects on the number and severity of crashes. In this paper, we identify and assess the variables, namely road geometry, vehicle speed, traffic volume, land use and junction type, and develop accident frequency prediction models for a main urban transport corridor in São Paulo, Brazil. Crash data was provided by the traffic management company of the city, other datasets were obtained from a mix of primary and secondary sources including roadside cameras, Geographic Information Systems (GIS) and digital mapping tools. The studied road was segmented and the coefficients associated with variables in the segments were obtained using Poisson regression through a stepwise variable selection procedure. Two models with junctions density per type (access/km, T-junction unsignalised/km, T-junction signalised/km and crossroads/km) and junction density per merged type (signalised/km and unsignalised/km) along with land use per type (commercial and residential) are developed. The junction density and land use are found to be significant and positively correlated with crash frequency. The models were evaluated by statistical means for their accuracy of predicting the crashes, and validated with additional information obtained from field observation.

1. Introduction

Road infrastructures play a crucial role in traffic safety, particularly in developing countries that undergo a rapid increase in motorization. For instance, between 2006 and 2022, vehicle fleet in Brazil grew from 45 million to 115 million, with the number of motorcycles tripling to 25.5 million [1]. However, road deaths in 2017 exceeded 34,000 with 33.6% involving motorcycles. This number is 45.3% higher than the total road deaths in that year (23,392) within the 28 European Union (EU) member states combined [2]. Brazil's road fatality rate of 21.5 per 100 million inhabitants (2009–2011) is four times higher than Europe [3]. The economic impact of traffic accidents is also substantial, costing 1.2% of Brazil's GDP annually [4].

São Paulo is the largest Brazilian city with around 11 million inhabitants and faces escalating road safety concerns due to the growth of motorcycles, which are responsible for the highest number of fatal accidents in the city since 2016. Among 372 such incidents in 2017, 117 died from collisions with fixed objects and 229 from collisions with other vehicles [5]. Vulnerable road users (VRUs) account for 35% of road deaths in São Paulo, exceeding the averages in low- and middle-income countries [6]. This warrants new studies on the impact of infrastructure (e.g. road geometry, junction) on road user behaviour (e.g. speeding) as the high VRU fatality rate cannot be simply explained by exposure [7].

Crash prediction models were developed predominantly between the mid-1990s and mid-2010s in Europe and North America. They can be used to identify locations and designs that are associated with high number of accidents, and to plan for countermeasures. However, only a few were developed for urban roads and very few are using observed traffic data that present a growing share of two wheelers and the evolving behaviours of road users. In this paper, we assess the variables which are found to influence the frequency of road accident occurrence, namely the geometry (lentgh and radius), speed, traffic volume, land use and junction. An accident database was used to develop a crash

* Corresponding author at: University of São Paulo, Department of Transportation Engineering - Escola Politécnica, Av. Prof. Almeida Prado, Travessa 2, 83, 05508-900 São Paulo, Brazil.
E-mail addresses: cassiano.isler@usp.br (C.A. Isler), Y.Huang1@leeds.ac.uk (Y. Huang), lucasmelo@usp.br (L.E.A. de Melo).

https://doi.org/10.1016/j.iatssr.2024.07.002
Received 25 March 2024; Received in revised form 12 June 2024; Accepted 4 July 2024
Available online 23 July 2024

prediction model for a main transport corridor in the city of São Paulo. Roadside cameras provided data on vehicles' speed, and data on the land use were obtained from publicly available database. Field observations were carried out on the road to identify the locations of traffic signalling and to provide geo-referenced images that will help intervention design. We discussed how a crash prediction model developed using this approach may enhance the safety of urban road users.

2. Literature review

Road designs have proven effects on the occurrence and severity of accidents ([8,9]). Road alignment, traffic condition and roadside environment are the primary factors that determine the workload of drivers, who then respond by adjusting the vehicle speed and lane choice. This section reviews the contributing factors to accidents and the evolvement of crash prediction models; some are discussed in the Highway Safety Manual (HSM) [10] as part of practical guide to enhance the safety design of highways.

2.1. Operating speed (V_{85})

Vehicle speed is a major factor contributing to accidents, explaining the use of road geometry by many crash models to predict the operating speed under specific traffic and environmental conditions. These models often focus on the 85th percentile speed (V_{85}) under free flow conditions. Lamm et al. [11] established an empirical relation between V_{85} and degree of curve, lane width, shoulder width, and Annual Average Daily Traffic (AADT). Later on, Lamm et al. [12] simplified this equation by excluding variables related to lane width, shoulder width, and AADT, as their influence explained only 5.5% of the variation in operating speeds. Castro et al. [13] developed a similar V_{85} model in Spain, correlating it with the Radius of Curve (r). However, these speed prediction models are context-specific and require caution when applied to new locations.

In a non-linear approach, Morrall and Talarico [14] linked V_{85} on horizontal curves to the degree of curve using data from two-lane rural Canadian highways. Lamm [15] modified this model for horizontal alignments with transition curves. De Oña and Garach [16] emphasized the importance of adapting accident prediction models to local circumstances to take into account factors such as driver behaviour and traffic regulations.

2.2. Accident prediction models

Lamm et al. [11] proposed ΔV_{85} (difference in 85th percentile speed between consecutive road segments) as a consistency criterion for accident prediction. It was used by Anderson et al. [17] to predict accidents on rural two-lane highways in the USA. Hadi et al. [18] and Martin [19] highlighted that the effect of traffic flow on crash rate increases with AADT on roads with higher levels of traffic, while the effect decreases with AADT on roads with lower traffic volumes. The same study found weaker correlation between traffic volume and fatal accidents. Road length is another form of exposure. The segment length was used in the crash modelling by Silva et al. [20], whose study divided the investigated segment of BR-116, which connects the cities of São Paulo and Rio de Janeiro, into 10 segments of different lengths. The better performance (i.e. smallest errors and highest values of R^2) of the longer segments (4.5 to 5.0 km) was explained as the model aggregated more crashes into one segment. Road length was also used by Fitzpatrick et al. [21] to predict accidents on horizontal curves over a 3-year period, based on a study of 5287 horizontal curves. Two forms of accident prediction are available in literature, presented in Eq. (1) and Eq. (2).

$$Y = e^{-7.1977} AADT^{0.9224} L^{0.8419} e^{0.0662 \times \Delta V_{85}} \quad (1)$$

$$Y = e^{-0.8571} MVKT e^{0.0780 \times \Delta V_{85}} \quad (2)$$

where Y is the number of accidents on horizontal curve in a 3-year period, L is the length of the curve, $MVKT$ denotes million vehicle-kilometres travelled. It can be seen as a variable that combines the length with traffic. Both models indicate that speed reductions (ΔV_{85}) compared to the preceding curve or tangent have statistically significant relationships with accident frequency.

Additional measures investigated by Anderson et al. [17] included the average radius, rate of vertical curvature and ratio of individual curve radius to average radius, while Taylor et al. [22] used average speed ($\overline{V}$) for accident frequency prediction as in Eq. (3).

$$Y = e^{-14.93} AADT^{0.7268} L^{1.000} e^{2.479 \times \overline{V}} \quad (3)$$

De Oña and Garach [16] validated Anderson's model using data from 1748 km of Spanish two-lane rural highways, which included 10,289 horizontal curves over 306 highway sections. The accident prediction model can be presented in a general form, as in Eq. (4) and Eq. (5), in which the variables were determined by different researchers using regression analysis (see Table 1), including Ng and Sayed [23] who studied a total of 319 horizontal curves and 511 tangents from two-way rural highways in Canada.

$$Y = e^{\alpha 1} AADT^{\alpha 2} L^{\alpha 3} e^{\alpha 4 \Delta V_{85}} \quad (4)$$

$$Y = e^{\beta 1} MVKT e^{\beta 2 \Delta V_{85}} \quad (5)$$

Ng and Sayed [23] further developed the model by replacing ΔV_{85} with variables of exposure at critical locations such as percentage of heavy vehicles, geometry (e.g. longitudinal gradient and curvature change rate) and access density (e.g. number of access/km). Tangent-to-curve transitions and successive curves of substantial difference in radius are typical examples of the critical locations. Camacho-Torregrosa et al. [24] selected 65 two-lane rural road segments of 2–5 km each in Spain to develop a new design consistency model using negative binomial regression for road safety evaluation.

Researchers have also incorporated roadside features as independent variables in crash prediction models. For instance, Ben-Bassat and Shinar [25] analysed a 13.4 km long road with experimental scenarios, followed by a questionnaire with 8 participants. They found that the roadside features and drivers' perception of risks and manoeuvres (e.g. actual speed, lane position) are interrelated. For instance, guardrails give drivers a sense of security and they drive faster in their presence. Cafiso et al. [26] calibrated a total of 19 accident models with varying

Table 1
Variables in accident frequency models.

Model	α1	α2	α3	α4	β1	β2
Anderson et al. [17]	−7.1977	0.9224	0.8419	0.0662	−0.8571	0.0780
De Oña [16] (for constant accel/decel)	−3.1008	1.1228	0.9932	0.0128	−1.9596	0.0124
De Oña [16] (for variable accel/decel)	−3.0680	1.1236	1.0002	0.0124	−1.9420	0.0119
Ng and Sayed [23]	−3.796	0.5847	0.8874	0.004828		
Camacho et al. [24]	−4.9462	0.7683*	0.8645	−0.7285		

* AADT is the average of the road segment.

explanatory variables, using a sample of 168 km long two-lane rural roads in Italy. The accident prediction model used in that study has a comprehensive set of variables such as geometry, speed, length and roadside context. Details are presented in Table 2.

Further calibration resulted in several crash models. For example, Cafiso et al. [26] had Eq. (6) that comprises at least one variable in each of the above variable groups.

$$Y = e^{-7.812} AADT^{0.753} L^{1} e^{-1.948 \times CR + 0.0872 \times \Delta V_{10} + 0.067 \times DD + 0.185 \times RSH} \quad (6)$$

In summary, a widely accepted ([23,26,27]) form of accident prediction model can be presented as in Eq. (7).

$$Y = e^{\alpha 1} AADT^{\alpha 2} L^{\alpha 3} e^{\sum_{i=1}^{m} b_i x_i} \quad (7)$$

where Y is the expected accident frequency; L is the section length; $AADT$ is the annual average daily traffic; x_i is any of the m variables in addition to L and $AADT$ (see Table 2); and a_1, a_2, and b_i are the coefficients related to the variables.

Further discussion on the accident frequency models and the estimation procedures were made by Lord and Mannering [28]. Generalized Linear Models (GLM) are widely used to estimate the parameters of the accident frequency models [27]. Negative Binomial and Poisson distributions are appropriate to model non-negative integer values as it is the case of accident frequency [28].

2.2.1. Observed speed

Hauer et al. [29] found a correlation between higher speed limits and fewer crashes, suggesting that higher speeds on flat terrain, associated with gentle curves and gradients, may lead to fewer accidents, especially with a lower number of non-motorized road users. The question arises as to whether actual vehicle speed alone should be used to model crash frequency, and if so, whether traditional variables like radius (r) and traffic volume (AADT) remain necessary in the model. Camacho-Torregrosa et al. [24] employed an innovative in-vehicle GPS-data collection method based on continuous operating speed profiles for more accurate observations of driver behaviour. Guo et al. [30] placed emphasis on the motorway speed influenced by driving behaviour in a study. Hossain [31] also stressed the values for a real-time crash prediction model that can be used to inform design of intervention measures.

Different to previous studies carried out on two-lane rural roads, Caliendo et al. [32] investigated the correlation between accident frequency and traffic (AADT), length (L), curvature (1/r), surface friction and longitudinal gradient on a 4-lane median-divided motorway of 46.6 km length in Italy consisting of circular curves without transition curves and tangents over a 5-year monitoring period. They developed separate equations for predicting accidents on curves (Y^{curve}) and tangents ($Y^{tangent}$), as in Eq. (8) and Eq. (9), respectively.

Table 2
Examples of explanatory variables in accident frequency models.

Variable Group	Abbreviation	Description
Exposure	L (km)	Length of homogeneous section
	AADT (vpd)	Average annual daily traffic
Geometric/ Operational	CCR (°/km)	Curvature change rate
	W (m)	Paved width
	TR (%)	Tangent ratio
	CR (%)	Curve ratio
	V_{avg} (km/h)	Average operating speed
	σ (km/h)	Standard deviation of operating speed
Design consistency	ΔV_{10} (no./km)	Number of speed difference higher than 10 km/h
	ΔV_{20} (no./km)	Number of speed difference higher than 20 km/h
Roadside context	RSH	Roadside hazard
	DD (no/km)	Driveway density

$$Y^{curve} = e^{-0.07130 + 0.80311 \times lnL + 0.27017 \times \frac{1}{r}} + 0.32660 \times AADT \times 10^{-4} \quad (8)$$

$$Y^{tangent} = e^{0.50347} \left(e^{0.85729 \times lnL + 0.23960 \times AADT \times 10^{-4}} + 0.22848\, J\left(AADT \times 10^{-4}\right) \right) \quad (9)$$

In which J = 0 where there is no junction and J = 1 where there is junction (50 m to signalised intersection or 20 m to other types of intersection). While this might be accurate, it creates a difficulty in practice separating them in a given road segment. Besides, 31.1% of all crashes and 33.4% of severe crashes occurred on curves, which represent 29.7% of the total length of the motorway. This means there might be no significant difference in the number of accidents in relation to the tangent/curve ratio. It can also be seen that speed is not a variable in Eq. (8) and Eq. (9) as the model deals with motorway. Study carried out by the UK Transport Research Laboratory [22] used the speed instead of change in speed in a crash model to predict accident frequency (Y) as in Eq. (10), where DS is the density (number per km) of sharp bend and DX is the density of crossroad. Note that Eq. (10) by Taylor et al. [22] is a more general model than the ones proposed by Caliendo et al. [32] for curves and tangents in Eq. (8) and (9), respectively.

$$Y = 3.152 \times 10^{-7}\, Q^{0.728} L^{1.039} V^{2.431} e^{0.121 DS + 0.286 DX} \quad (10)$$

Road design such as the taper length, and roadside environment such as availability for parking, are important variables that influence drivers' workload. A study by Garach et al. [27] of 972 km long two-lane rural highway over a flat terrain in Spain found that longitudinal grade has little effect on crash frequency. In comparison, the bendiness and percentage of HGV have the most significant effects on crash occurrence, e.g. a 10% increase of HGV in the vehicle mix is associated with 22.7% increase of crash incidents in road segments that have high traffic volume (> 4000 vehicles per day). Variables such as shoulder width and access density, which are regulated for high traffic roads, have a relative high influence on crashes on low traffic roads. Interestingly, this is different to what is found on mountainous highways in China. In a study by Fu et al. [33], the accident rates for a 85 km long highway with an average longitudinal gradient of up to 6.5% were found to increase exponentially with the gradient, and in particular on segments after 2–3 km of continuous descent. However, no prediction model was provided by the study and the types of accidents were not stated.

2.2.2. Junctions and land use

Modelling accidents on urban roads is more complicated mainly due to the frequent presence of junctions, non-motorized road users and increased variance in collision patterns. Unsignalised intersections, which may have stop control, yield control or no traffic control, are more subject to drivers' discretion than signalised ones. In a study of 142 km urban road links including 1036 junctions but excluding roundabouts (due to insufficient data) in Demark, an accident model was developed by Greibe [34]. Generalized linear modelling techniques were used to relate accident frequency to explanatory variables and simple models containing only traffic variables are found almost as good as more complicated ones, and variables describing junction design improve the models by only a small extent. The reason can be attributed to behaviour such as explained by Savolainen and Mannering [35], that motorcycle riders are more reckless in ideal riding conditions and by Morency et al. [36], that lenient geometry may increase operating speed. Moore et al. [37] also suggested that good geometry may be offset by reckless behaviour, and campaign for behaviour such as helmet use and driver awareness of cyclists should be encouraged.

Another finding by Greibe [34] is that the total number of accidents is very similar for signalised and unsignalised junctions with the same traffic flow. The same view is held by Haleem and Abdel-Aty [38] in a study in the USA. Greibe [34] further developed the model into a more general form to predict accidents on urban roads, as in Eq. (11), in which

Y is the accident frequency per km per year. Besides, specific values are given for speed, road geometry and roadside environment j^{th} variables as in Table 3.

$$Y = \left(6.09 \times 10^{-4}\right) AADT^{0.8} e^{\sum_{i=1}^{m} b_j \times x_j} \quad (11)$$

Table 3 shows that narrow road links with high number of accesses, and with shops and chances for roadside parking, are prone to accidents. The effects of speed are complicated. For instance, high speed roads tend to have few vulnerable road users (VRUs), which may explain the low accident frequency compared to roads of low speed limits. Haleem and Abdel-Aty [38] indicated that 25–30% of fatal crashes occur at or near an intersection. In the study using 1547 three-legged and 496 four-legged unsignalised intersections in the USA, aggregated binary probit model was proved effective in analysing crash severity. Results indicated that in urban areas, young drivers (demographic factors), grade-separated ramps (geometry factors), low AADT on the major approach and low speed limit (traffic factors) contributed to a decrease in severe crash injuries.

3. Development of a crash model

Broadly, there are two types of crash prediction models. One is represented by Nilsson [39] and Cameron and Elvik [40] that relate casualties and/or crashes to changes of speed limit. These time-series studies require data from a relatively long period of time, and exclude the effects of traffic, road geometry and roadside environment, unless there is a change of them in addition to change of speed limit. The other type is cross-section studies, which hold the time constant and relate crashes and/or casualties to different road sections. In this paper, the effects of road geometry, traffic and roadside environment are studied using a cross-sectional approach with the following steps illustrated in Fig. 1.

In STEP 1, variables and the functional form of the regression model for accident prediction are identified based on literature. STEP 2 involves collecting crash data from secondary sources, primary data on traffic volumes and speeds, and data for road geometry using digital mapping tools. STEP 3 focuses on model development, when the studied road is segmented, the dataset is analysed with the support of Geographic Information Systems (GIS), and the coefficients for the model variables are estimated through a stepwise variable selection procedure. STEP 4 assesses the model performance and suggests adjustments to the model that best estimates accident frequency based on a set of independent variables. Finally, STEP 5 discusses the model, incorporating information from field visit, and proposes countermeasures for traffic management, capital investment and strategic planning to reduce accident frequency.

Table 3
Explanatory Variables for Crash Prediction on Urban Roads.

Geometry/Roadside	Scenario	Estimation [34]
Speed limit	50 km/h	2.25
	60 km/h	2.85
	70 km/h	1.00
Road width	5.0–7.5 m	0.83
	8.0–8.5 m	0.68
	9.0–15.0 m	0.80
Number access road	0	0.72
	0–5	0.75
	5–10	1.00
	>10	1.25
Parking	Prohibited	1.19
	Rarely	1.00
	Permitted	1.77
Land use	Scattered housing	1.00
	Residential blocks	1.56
	Industrial blocks	1.58
	Shops	2.44

We developed a functional form of the regression model in order to take into account the traffic related attributes (*AADT* and ΔV_{85}), the road geometry attributes (radius) and other related attributes (junction and land use) to predict accident frequency. The first model derived from STEP 3 as depicted in Fig. 1 is presented in Eq. (12):

$$Y = e^{\alpha 1} AADT^{\alpha 2} L^{\alpha 3} e^{\alpha 4 \Delta V_{85}} e^{\sum_{i=1}^{m} b_i x_i} \quad (12)$$

where Y is the expected accident frequency per section of the studied road; *AADT* is the annual average daily traffic per section; L is the section length; x_i refers to the explanatory variables, e.g. speed, average radius, land use, number of lanes and number of junctions per type; and a_0, a_1, a_2,α_3, α_4 and b_i are model coefficients.

The following variables were initially considered to be included in the crash frequency model: length (meters); average radius (meters); average speed (kilometres per hour – kph); 85th percentile speed (kph); difference in 85th percentile speed between successive sections (kph); AADT (vehicles per day – vpd); type of land use (residential or commercial) at the roadside; average number of lanes; and density of intersections per type (crossroad/km, T-junction signalised/km, T-junction unsignalised/km and access/km). Two models have been considered after an assessment of the correlation between independent variables; Model (1) comprised the intersections per type and Model (2) comprised the intersections merged (un-signalised junctions including the number of access and unsignalised T-junction, and signalised junctions including the number of crossroad and signalised T-junction) along with other variables after an assessment of the correlation between dependent and independent variables.

A stepwise Poisson regression model was developed considering variables after an assessment of the correlation between them. Poisson model assumes the dependent variable Y_i (crash frequency in the i^{th} segment) have the following distribution as represented in Eq. (13) [41].

$$P(Y_i;\mu_i) = \frac{\mu^{Y_i} exp(-\mu_i)}{Y_i!}, Y_i = 0,1,2\ldots \quad (13)$$

and $\mu_i = \mu_i(x_{ij}) = exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)$ where x_{ij} ($j = 0,1,\ldots,k$ and $x_{i0} = 1$) are independent variables and β_j ($j = 0,1,\ldots,k$) are regression parameters. In this paper, $\mu_i = \mu_i(x_{ij}) = exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right) = exp\left(e^{\alpha 1} AADT^{\alpha 2} L^{\alpha 3} e^{\alpha 4 \Delta V_{85}} e^{\sum_{i=1}^{m} b_i x_i}\right)$.

Frome et al. [42] described the maximum likelihood (ML) method that is used to estimate the coefficients β_j in the Poisson regression model while several measures of goodness-of-fit have been proposed in the literature. The Akaike information criterion (AIC) [43] has been extensively used for these purposes, where $AIC = -log(LL) + (k+1)$ with *LL* equals the final log-likelihood of the estimation and lower AIC values indicate better fit of the model [41]. Given the Poisson model to estimate the coefficients that correlate the independent variables, a variable selection procedure can be applied to develop the model containing the independent variables that best estimate the values for the dependent variables. Several methods have been proposed for variable selection to identify the best subset of independent variables that are relevant for regression modelling. Stepwise methods consider a goodness-of-fit measure (for example, the AIC) in an iterative procedure to select and remove the independent variables in order to obtain the best model fit. In summary, the model starts without input variables. Next, variable x_{ij} is selected for the model with a unique significant independent variable and highest goodness-of-fit measure. A selection method is performed to compare the criterion values of all models that include the first x_{ij} (or x_{ij}'s) and one additional $x_{i\bar{j}}$ variable. Next, a

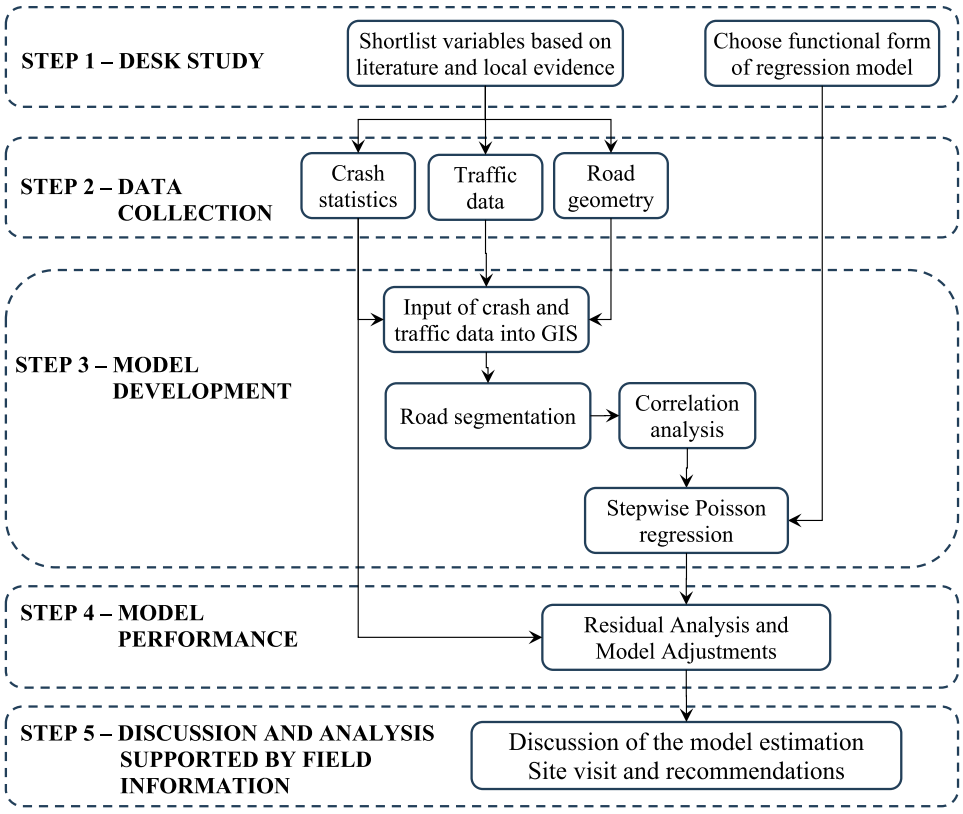


Fig. 1. Flow diagram of the crash model development.

deletion is performed to compare the criterion value of all models that include the x_{ij}'s without one x_{ij}. The method stops when there is no inclusion or deletion to be performed. Yamashita et al. [44] provided further details on the AIC stepwise method for variable selection in linear regression models.

After obtaining the independent variables that provided the bet model fit, we analysed its final measures of performance, and we discussed the values and significance of the coefficients related to each independent variable. We also performed an analysis of correlation between the independent variables after a model estimation through the Variance Inflation Factor (VIF) and we discussed the errors through Residual and Cumulative Residual (CURE) plots. VIF is used to detect multi-collinearity between the independent variables of a regression and measures the magnitude at which the variance of the estimated regression coefficients is inflated, compared to when these variables are not linearly related [45]. CURE plots are used to assess how well a model fits the data used to estimate the model coefficients. They may present poor fit when the cumulative residual line plotted in the y-axis of a chart against the ascending values of each independent variable in the x-axis increases or decreases significantly at a certain point [46]. Further information on CURE plots can be found in studies by Hauer [47] and Lin et al. [48].

Finally, the models were evaluated using a cross-validation procedure that involved splitting the dataset by year from 2017 to 2021. In each iteration, a four-year period was used as the calibration dataset to derive the coefficients of the model. The remaining year was set aside as the validation dataset that played a crucial role in assessing the predictive performance of the model. This process was applied again by removing another year of data and assigning it as the validation dataset. It is repeated until all the years are used as validation dataset.

The Root Mean Square Error (RMSE) presented in Eq. (14) was utilized to obtain the disparity between the predicted accident frequencies and the actual frequencies observed.

$$RMSE = \frac{\sqrt{\sum_i (E_i - Oi)^2}}{n} \tag{14}$$

where E_i is the i-th estimated accident frequency of the validation dataset, O_i is the observed accident frequency of the validation dataset, and n is the size of the dataset.

4. Case study

4.1. Desk study and data collection

The city of São Paulo is located in the southeast region of Brazil and is the largest city of the country, and Radial Leste is one of its major roads connecting the eastern highly populated neighbourhoods to the city centre where most of job opportunities are (see Fig. 2). In this study, Avenida Radial Leste was divided into 11 sections in the eastbound direction and 9 sections in the westbound direction as shown in Fig. 3(a)-(b). This division was based on the locations of roadside cameras that record the vehicle speeds, which were used to derive the average speed of traffic. We opted for this segmentation instead of a homogeneous or fixed-length segmentation because the former would result in the same average speed for both straight and curved segments, thus introducing bias to the model, and the latter would lead to uniform average speed and vehicle flow across several successive segments, resulting in poor model fit.

The State of São Paulo use a georeferenced InfoSiga[1] database which is updated monthly since 2016. InfoSiga contains information on traffic accidents, including fatalities and injuries, in every city of the State. The database includes details on vehicles, age and gender of the casualties,

[1] https://www.infosiga.sp.gov.br/

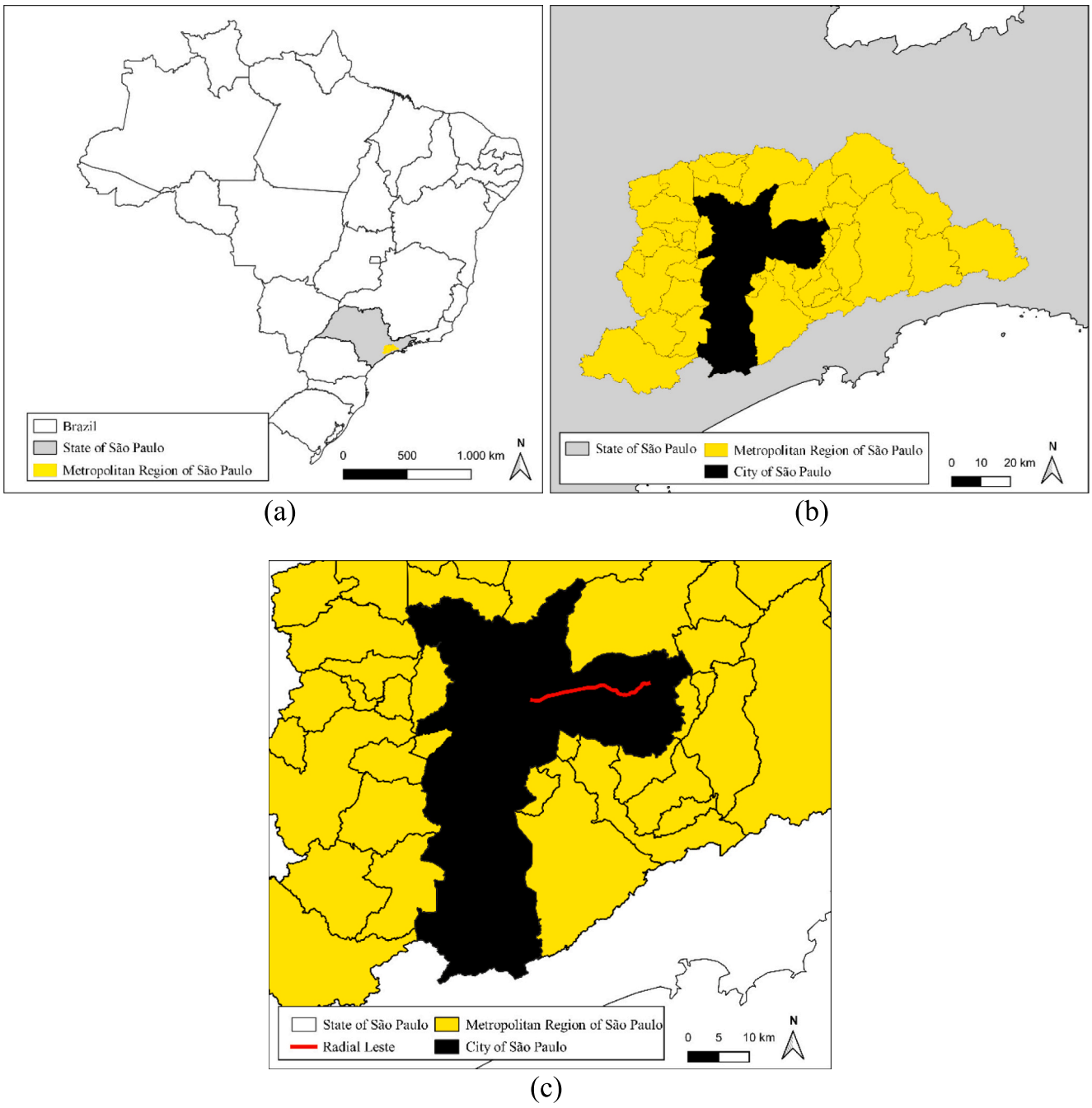


Fig. 2. Location of the (a) State of São Paulo and its Metropolitan Region, (b) city of São Paulo in the Metropolitan Region of São Paulo (MRSP) and (c) Radial Leste in the city.

and collision types. Additionally, the traffic management company CET (Companhia de Engenharia de Tráfego) in São Paulo has access to the Digital Registry of Civil Police Occurrences and data from the Medical Legal Institute, collecting around 227 million records annually from 13 traffic cameras. CET publishes annual data on traffic collisions, injuries and victim characteristics. Records from 2017 to 2021 were used in the model development. Fig. 4(a)-(b) illustrates accident locations on sections of Radial Leste in the eastbound and westbound directions.

A total of 556 accidents occurred over the 20 studied segments of Radial Leste between 2017 and 2021. The majority of accidents (84.8%) occurred in the eastbound direction, with 75.7% of them taking place between 2017 and 2019. Fig. 5 depicts the distribution of accidents in both directions between 2017 and 2021. Half of the segments had between 0 (zero) and 2 accidents in this period, while the remaining segments had between 3 and 22 accidents. Note that the occurrence rate of accidents adheres to a Poisson distribution that aligns with the functional form and estimation approach used in this study.

Following the variables presented in Section 2.2, the length and average radius were measured using GIS. We obtained the average speed, 85th percentile speed (V_{85}) and difference in 85th percentile speed between successive sections (ΔV_{85}) from observed (camera) data. A large dataset was made available by CET containing the records of each vehicle that passed through the studied sections of Radial Leste between 2017 and 2021. The average speeds were calculated by dividing the travelled distance between cameras by time. However, only 23% of the total records were employed because the rest contained information about the vehicle's license plate captured only at the entrance, or at the exit, and not both of a section. Data on AADT was also obtained

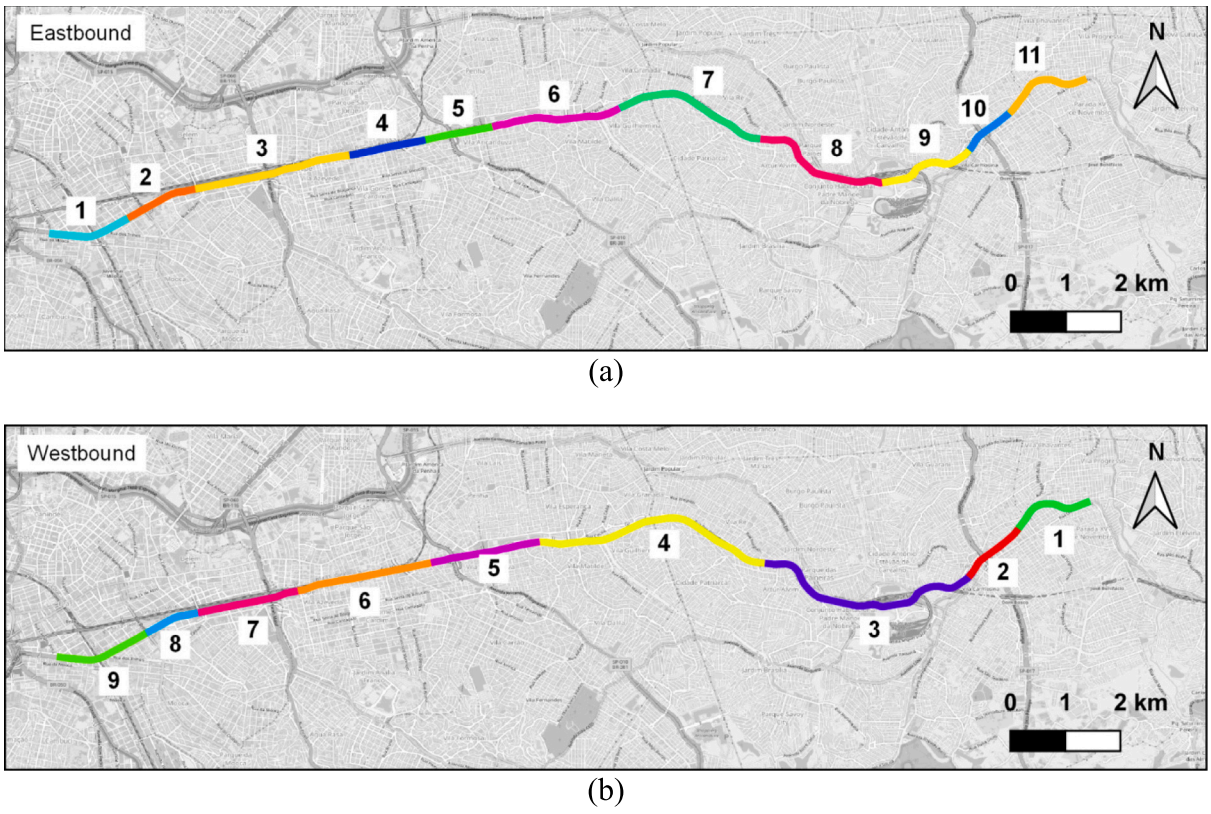


Fig. 3. Location of the segments in Radial Leste in the eastbound (a) and westbound (b) directions.

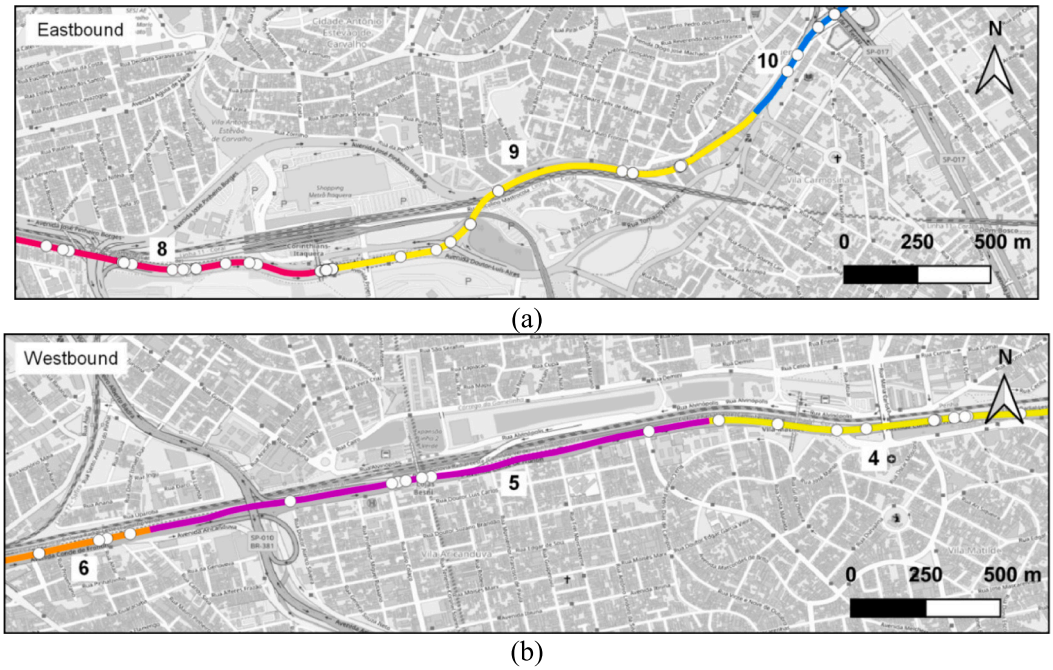


Fig. 4. Example of accident locations in the sections of Radial Leste in the eastbound (a) and westbound (b) directions.

from the traffic cameras at the entrance/exit of each section containing information about vehicles and speeds. The percentage of land use (residential or commercial) was calculated using GIS with the support from a dataset made available by Geosampa[2] that contains information about the land use over São Paulo. The average number of lanes was

[2] https://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx

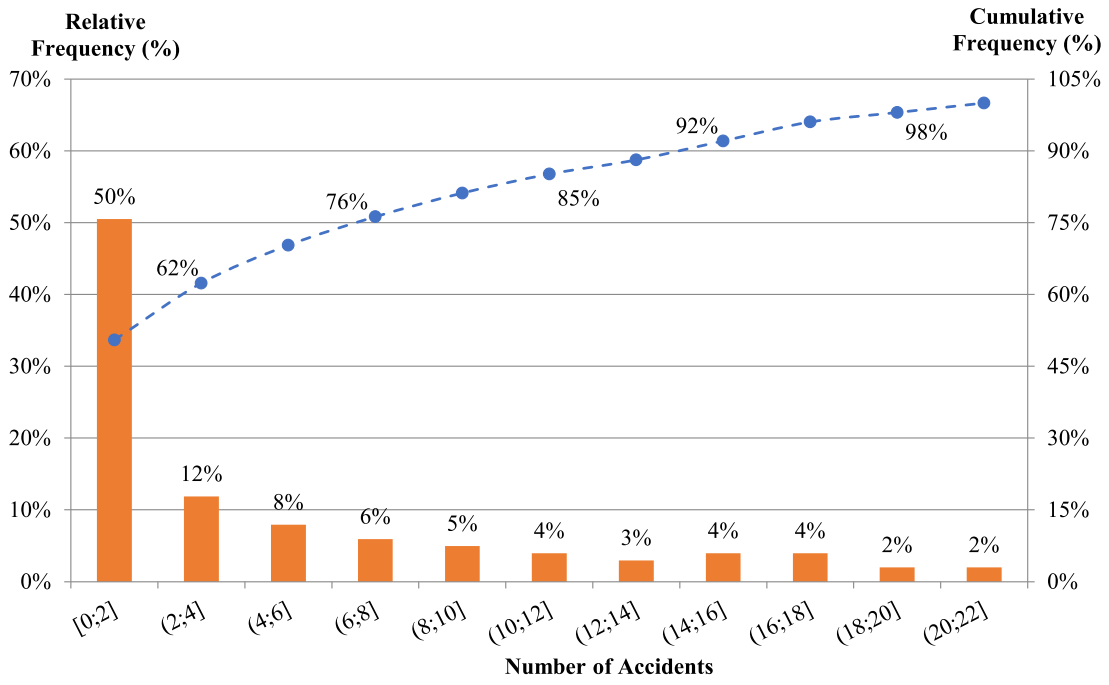


Fig. 5. Relative and cumulative frequency of accidents in the studied sections of Radial Leste between 2017 and 2021 in the eastbound and westbound directions.

obtained from the OpenStreetMaps[3] database also with the support of GIS. Finally, the number of intersections per type was obtained from the OpenStreetMap database.

Appendix A and B of the supplementary material present the values for the predictive variables (number of accidents) and the exploratory variables (road geometry, AADT, land use, number of lanes and number of intersections per type) in the eastbound and westbound directions, respectively. These values were used in regression analysis to estimate the coefficients of the exploratory variables in the model. Table 4 presents the descriptive statistics of accidents and of the variables.

The correlation between the crash frequency and the independent variables were calculated and presented in Table 5. Fig. 6 presents the correlation plot for the independent variables.

Fig. 6 highlights a significant negative correlation between the average radius and various factors associated with junction types (T-junction unsignalised per km, access per km, and unsignalised junctions per km), AADT and the number of lanes. Similarly, the length and residential land use exhibit negative correlations with variables related to junctions, AADT and the number of lanes, so does commercial land use. Conversely, the number of lanes shows a strong positive correlation with types of junctions (T-junction unsignalised per km, access per km, and unsignalised junctions per km), and with AADT.

Table 4
Average statistics of accident and of the explanatory variables per direction.

		Eastbound	Westbound
Total number of accidents		472	84
Accidents per section		8.6	1.9
Length (m)		1890.9	2318.2
Radius (m)		154.3	153.9
Speed (kph)		33.9	34.0
V_{85} (kph)		44.5	45.2
ΔV_{85} (kph)		0.01	−0.96
AADT		14166	16842
Land Use (%)	**Residential**	28	23
	Commercial	53	55
	Other	19	22
Number of Lanes		3	3
Number of Crossroad		2	1
Number of T-junction	**Unsignalised**	2	1
	Signalised	1	2
Number of Access		8	5
Number of Intersection merged by type	**Signalised (crossroad and T-junction)**	3	3
	Unsignalised (access and T-junctions)	9	5

4.2. Model development

We applied the stepwise variable selection procedure to model the accident frequency as a function of the independent variables in Model (1) comprising the junctions per type (access, unsignalised T-junction, crossroad and signalised T-junction) and Model (2) comprising the junctions merged by signalised and unsignalised along with other variables discussed in section 4.1. Fig. 7(a) and 7(b) presents the AIC and the variables included in each step of the selection procedure, respectively, for Model (1) and Model (2).

Based on the final AIC of Model (1) and Model (2), we calculated the final log-likelihood of the models and compared them through a likeli-

Table 5
Correlation between accident frequency and the independent variables.

ln (AADT)	ln (length)	ΔV85 (kph)	Average radius	Land use: residential	Land use: commercial	Number lanes	Crossroads/km	T-junctions unsignalised/km	T-junctions signalised/km	Access/km	Signalised junctions/km	Unsignalised junctions/km
0.1	0.14	0.08	0.13	0.02	0.09	-0.02	0.43	0.24	-0.12	0.44	0.44	0.42

[3] https://www.openstreetmap.org/

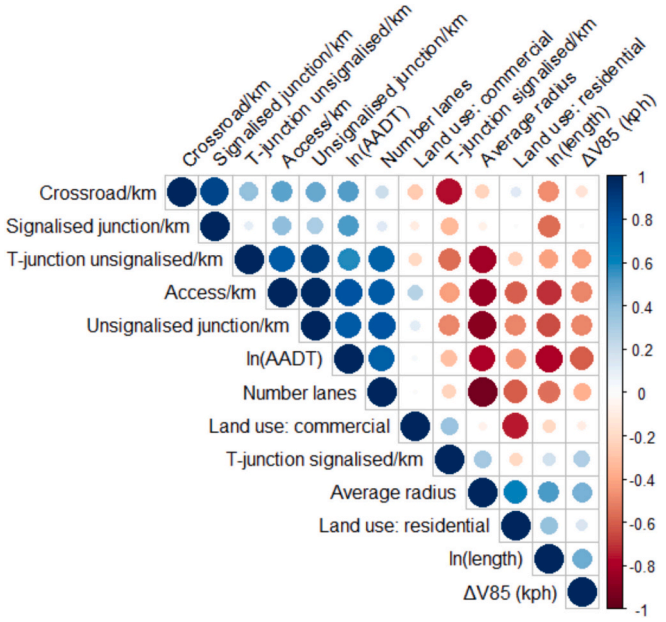


Fig. 6. Correlation between all the independent variables considered in the models.

hood ratio test [49] with null model (H_0) equals Model (2) and alternative model (H_1) given by $LR = -2.\left[LL_{H_0:Model\ (2)} - LL_{H_1:Model\ (1)}\right]$. Table 6 summarizes the measures of performance and shows that the models exhibit similar performance ($LR = 0.6 > \chi^2_{2;95\%} = 5.991$) and with a preference for Model (1) that uses junctions per type and other variables in estimating the accident frequency in the studied case.

The values of the coefficients in both models are presented in Table 7. As expected for the stepwise estimation, all the variables are significant at 95% confidence level ($|\text{z-value}| > 1.96$). In Model (1), length, commercial and residential land use, and all types of junctions (crossroad/km, T-junction unsignalised/km, T-junction signalised/km and access/km) are found to be significant, while in Model (2), the junctions merged are found to be significant. Note that the values of the coefficients for both models have the expected signs and magnitude similar to those found in the literature (see Table 1 and Table 7). It is worth noting that number of lanes, average radius, AADT and difference of 85th percentile speed between successive segments have been dropped from the models in accordance with the correlation observed between variables as shown in Fig. 6. It indicates that the frequency of accidents is primarily associated with land use and traffic control of the road rather than the geometry (except by the segment length) and traffic characteristics.

The signs and values of the parameters in both models agreed with our expectation. The constants reveal an expected accident frequency value regardless of other variables in the model when the Poisson distribution is inferred. The accident frequency increase when the length segment increases exponentially by a factor of 1.087 and 1.089 in Model (1) and Model (2), respectively. The variables related to land use are significant in all models and range between 1.980 and 2.598 in Model (1) and 2.042 and 2.532 in Model (2), contributing to an increase of accidents. It can be explained by the fact that drivers (or riders) usually reduce their speed to access the facilities and buildings along the road and interfere with the traffic when leaving such places. Additionally, commercial land use has higher impacts in increasing accident frequency compared to residential land use. The junction density also increases accident frequency, with the highest contribution made by crossroad/km and T-junction signalised/km, followed by T-junction signalised/km and access/km in Model (1). Signalised junction/km made highest contribution to accident frequency, followed by unsignalised junction/km in Model (2). Such results might derive from higher traffic flows at the signalised junctions than at the unsignalised ones.

4.3. Model performance assessment

CURE plot residuals were developed for each variable in both models based on the method by Snirivasan et al. [46] to assess the residuals of accident frequency in the studied case. The plots present the cumulative residual (observed accident frequency minus estimated accident frequency) over the values of a given explanatory variable sorted in ascending order. Fig. 8(a)-(g) presents the CURE plots of variables in Model (1) and Fig. 9(a)-(e) presents the CURE plots of variables in Model (2).

CURE plots indicate that the cumulative residuals do not exhibit an increasing or decreasing trend as expected, and they fall within the confidence intervals for most of the explanatory variables. However, exceptions are observed for lower values of length (below 7.0 km) and access/km (below 2.0) in Model (1), as well as length (below 7.0 km) and unsignalised junctions (below 2.5) in Model (2). Following the early work by Hauer [47] and Lyon [50], Snirivasan et al. [46] suggested that if 5% or more of cumulative residuals fall outside the confidence intervals of CURE plots (i.e., below or above the confidence bounds), further investigation into the functional form of the predictive models is warranted for specific variables, although this evidence should be explored further. Practitioners should carefully analyse estimations within these intervals, as different models may be needed for specific intervals depending on the dataset and application. Table 8 presents the Variance Inflation Factor (VIF) per variable and model to assess the multi-collinearity between the independent variables of a regression.

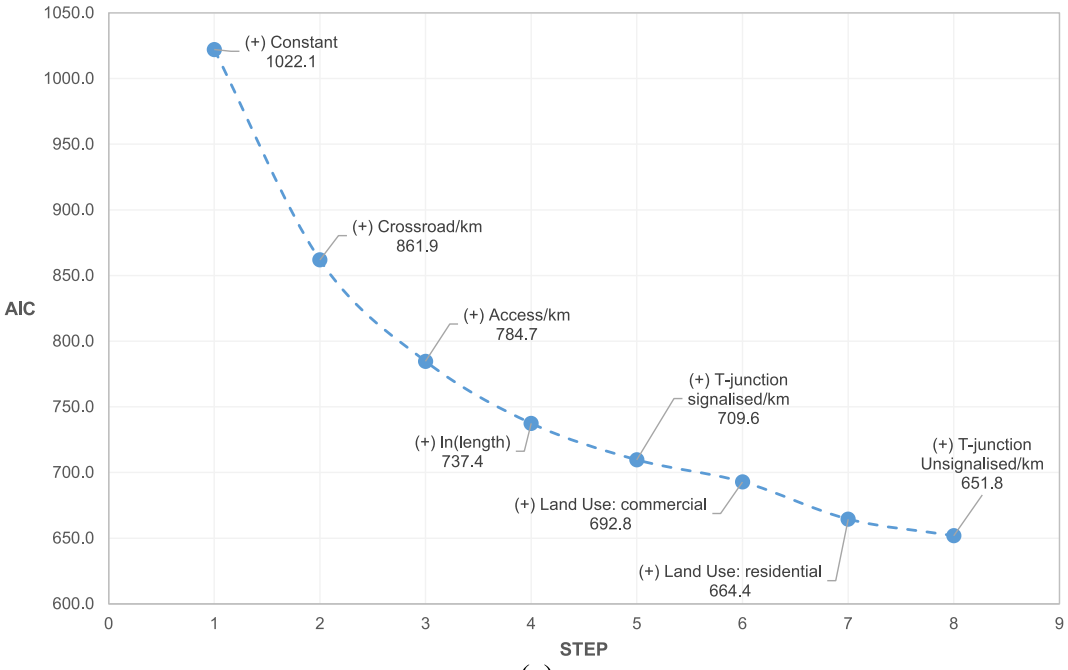
There is a lack of a consensus on acceptable values in literature for the Variance Inflation Factor. Montgomery [51], Kutner et al. [52], O'Brien [53], Chatterjee and Simonoffart [54], Marcoulides and Raykov [55] discussed the commonly used rule of thumb and suggested that VIF values below 5.0 are generally considered acceptable, indicating a lower proportion of variance that the i^{th} independent variable shares with other independent variables in the model. Therefore, we consider the results of our model acceptable, as the VIF for all variables in the models are below 3.00.

Finally, Table 9 presents the RMSE of the cross-validation procedure when each year is considered as validation dataset and the remaining years are used as calibration dataset. Results show that both models have similar RMSE close to 1.2 accident on average in the validation dataset.

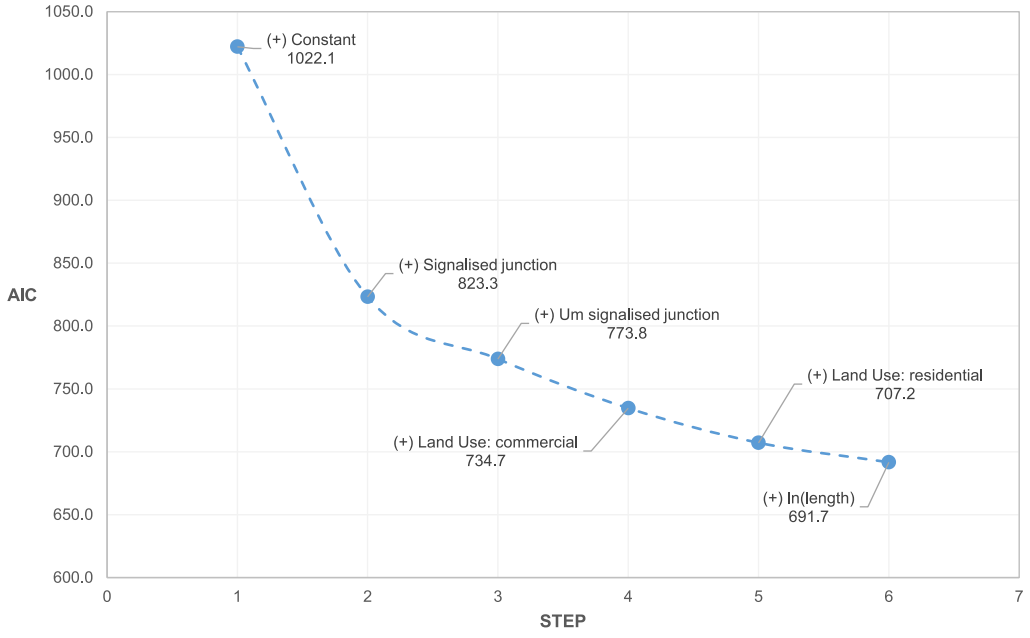
In conclusion, results show that the estimations of accident frequency are more accurate when the variables related to geometry (length as proxy), number of intersections per type and land use are included in the models. It is worth noting that the models perform equally well whether or not the junctions are merged into the unsignalised and signalised categories. Roadside environment such as availability for parking and the traffic make-up such as percentage of HGV are found in literature [26] to have a high influence on crashes. These can be tested for their significance in the model should the data be available.

4.4. Field survey

A field survey was carried out over the Avenida Radial Leste on Friday 7th April 2023. 340 images were obtained from the survey and later linked to a GIS, which enabled visual assessment of the site for roadside traffic environment, such as the number of lanes. A sample of those images are presented in Fig. 10(a) and the variables of land use are illustrated in Fig. 10(b) for residential use and Fig. 10(c) for commercial use. Appendix C of the supplementary material presents a set of images per segment and direction (Eastbound and Westbound) illustrating the road geometry and land use.



(a)



(b)

Fig. 7. AIC in each step of the stepwise variable selection procedure.

Table 6
Measures of performance of models.

Model	Final AIC	Number of variables (k)	Final log-likelihood	LR test	$\chi^2_{2;95\%}$ (degrees of freedom = 7–5 = 2)
1	651.8	7	−317.9	0.6	5.991
2	691.7	5	−318.2		

The number of lanes is not a significant variable in any model as to explain the accident frequency, mainly because the road geometry is not homogeneous in the sections, where several segments contain barriers dividing a set of lanes in the same direction as seen in Fig. 10(c).

5. Conclusion and recommendations

The geometric features of a road, junction and roadside land use are important factors to influence the frequency of accident occurrence on urban roads. A set of prediction models is developed in this study to analyse the most influential factors on accident frequency. The purpose

Table 7
Coefficients estimated for the models.

Variable	Model (1)		Model (2)	
	Estimation	z-value	Estimation	z-value
Constant	−10.474	−9.691	−10.516	−10.011
ln(length)	1.087	9.056	1.089	9.272
Land Use: Residential	1.980	5.385	2.042	5.656
Land Use: Commercial	2.598	6.872	2.532	7.515
Access/km	0.210	5.720	–	–
T-junction unsignalised/km	0.235	3.927	–	–
T-junction signalised/km	0.717	5.960		
Crossroad/km	0.798	10.673	–	–
Unsignalised junction/km	–	–	0.225	10.737
Signalised junction/km	–	–	0.786	11.31

is to identify any adverse traffic and roadside environment, which cause abrupt changes in drivers' workload and subsequently increase crash occurrences. The literature review by Lord and Mannering [28] provided a wide options of modelling techniques for crash prediction. The multiple linear regression models are used to testify whether one or more independent variables, e.g. speed, radius, can be used to significantly predict a dependent variable, e.g. crash frequency or crash rate. Poisson distribution is widely found an appropriate alternative to linear regression because accident frequencies are (a) integers, (b) relatively small numbers and (c) nonnegative. However, when the distribution curve has a long tail, the negative binomial distribution can provide an improvement over the Poisson distribution. It allows greater variance in the data (thereby deals with over-dispersion), compared with linear or Poisson regression. How different modelling methods could reveal similar or divergent results from those presented in this paper is an area for further study.

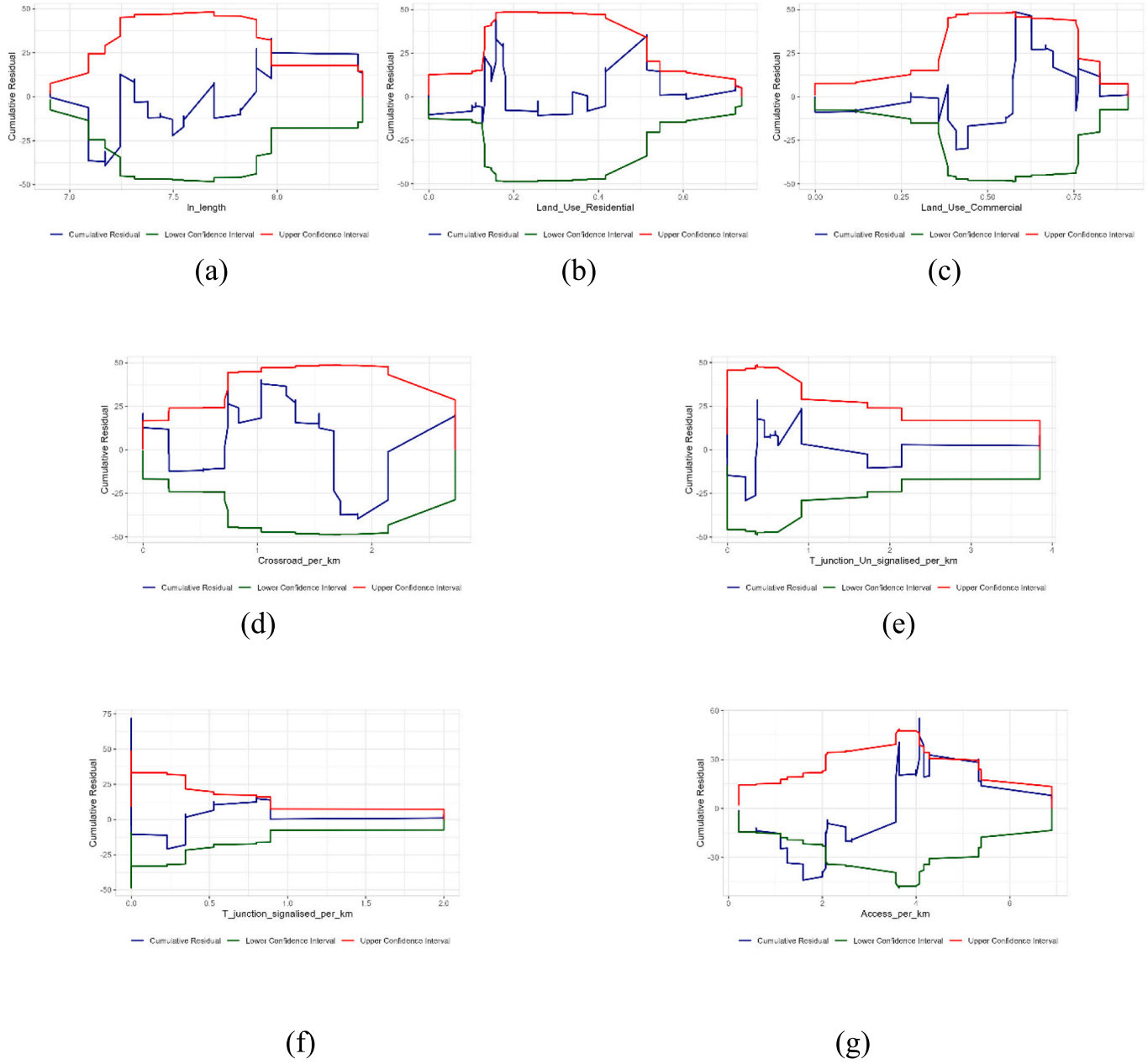


Fig. 8. CURE plots of variables of Model (1): (a) ln(length), (b) Land Use: Residential, (c) Land Use: Commercial, (d) Crossroads/km, (e) T-junction unsignalised/km, (f) T-junction signalised/km and (g) Access/km.

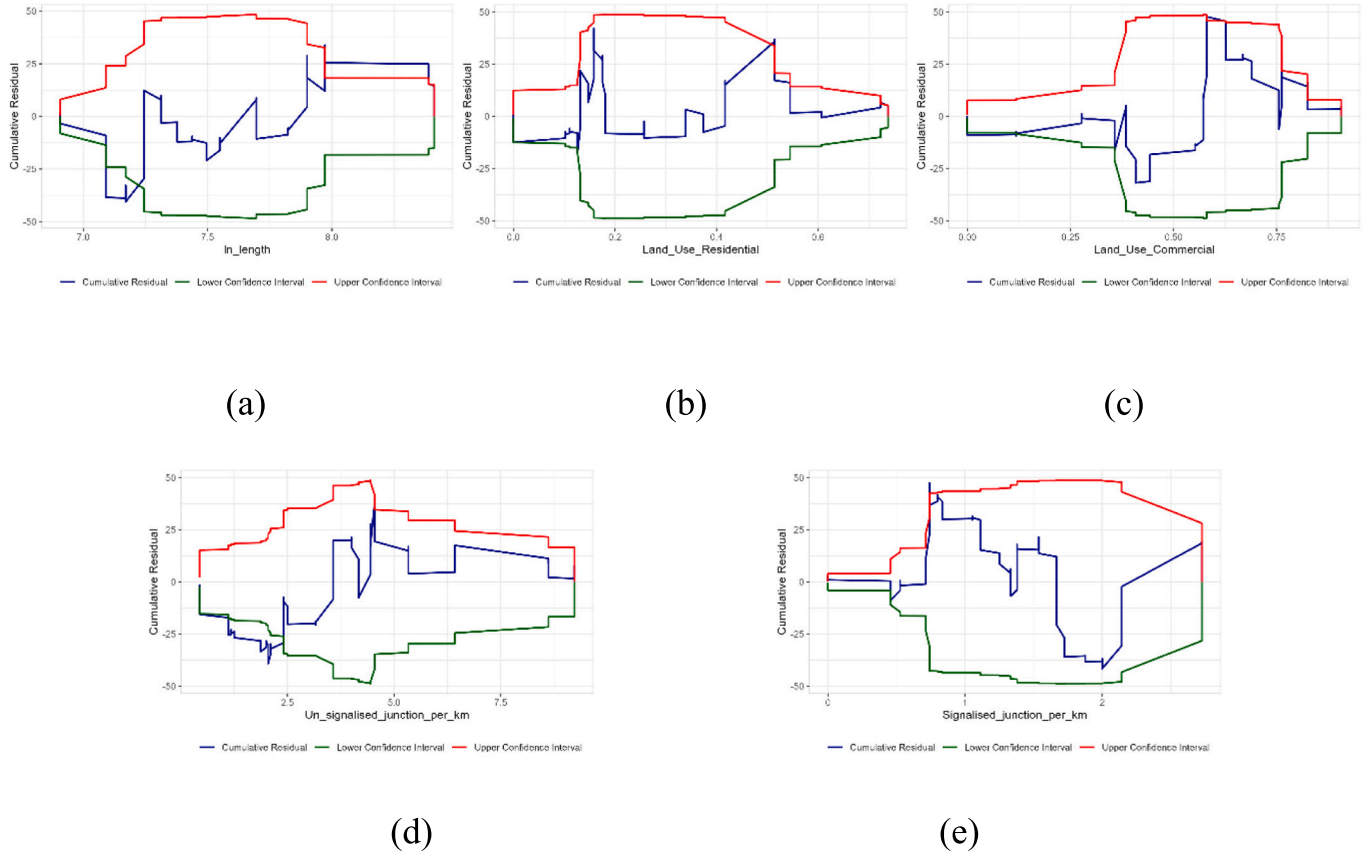


Fig. 9. CURE plots of variables of Model (2): (a) ln(length), (b) Land Use: Residential, (c) Land Use: Commercial, (d) Unsignalised junction/km and (e) Signalised junction/km.

Table 8
VIF per explanatory variable and model.

Variable	Model (1)	Model (2)
ln(length)	1.29	1.26
Land Use: Residential	2.76	2.70
Land Use: Commercial	3.17	2.57
Crossroads/km	2.28	–
T-junction unsignalised/km	2.27	–
Access/km	2.48	–
Unsignalised junction/km	–	1.59
Signalised junction/km	–	1.24

Table 9
RMSE of the estimated models using a cross-validation procedure.

Model	RMSE with dataset of 5 years					Average RMSE
	2017	2018	2019	2020	2021	
1	1.58	1.05	1.13	0.91	1.37	1.21
2	1.59	1.05	1.13	0.89	1.36	1.20

Regression analysis of the most influential factors is carried out in this paper, and the models are validated using data collected from a main urban road in São Paulo. Our prediction model initially considered variables such as average radius, the difference in 85th percentile speed between successive sections, AADT, percentage of land use (residential or commercial) and number of intersections per type (crossroad, T-junction signalised unsignalised, access). Results show that length, land use and junctions per type significantly affect the crash frequency.

Field observation was carried out in this study which collected primary data on the site for model development, but more can be done in particular in the analysis of vulnerable road users (VRUs), e.g. motorcyclists. There is a need to investigate fully the accidents involving motorcycles, specific to trip purposes, e.g. commuting, goods delivery, motorcycle taxi, and thereafter to propose intervention measures. Knowledge of the accident type (e.g. side/rear-end collision) and vehicle type involved is also important for understanding the causes and for implementing safety design. Developing a framework to assess road network-wide risks is suggested in literature [56]. Improvement should be made where a location is known for high crash incidents. Use of GIS and digital twins of road infrastructure, as well as partnership with other public sectors such as health and police, will enable timely and cost-effective measures be taken to address new challenges in road safety.

Additionally, cycling has increased in popularity in Brazil over the past decade. This adds additional challenges to road safety in dealing with more diversified road users. Traffic management and law enforcement, such as license management and prohibiting riding between lanes, are very important alongside the provision of infrastructures. Despite of the safety benefits brought by emerging vehicle technologies, crash statistics in Brazil suggest that the protection of VRUs should be prioritised in road safety programmes, given their growing share of road injuries and deaths.

CRediT authorship contribution statement

Cassiano Augusto Isler: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yue Huang:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Lucas Eduardo Araújo de Melo:** Writing – review & editing, Visualization, Validation, Data curation.

(a)

(b)

(c)

Fig. 10. Site observation of Avenida Radial Leste in April 2023.

Declaration of competing interest

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Acknowledgements

This research was supported by São Paulo Research Foundation (FAPESP - grant #2019/05515-5). The first author acknowledges the Brazilian National Council for Scientific and Technological Development for the scholarship (CNPq - grant #306552/2022-1). The authors acknowledge Ciclocidade for the assistance with traffic data processing and validity analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.iatssr.2024.07.002.

References

[1] Departamento Nacional de Transportes (DENATRAN), Frota de veículos, 2019 http://www.denatran.gov.br/estatistica/237-frota-veiculos (accessed August 18, 2023).
[2] European Comission, Annual Accident Report. https://road-safety.transport.ec.europa.eu/document/download/287aa31e-48c2-4e04-a9cc-e2ca24d29cc2_en?filename=ERSO_annual_report_20220509.pdf (accessed 18 August 2023).
[3] J.T. Bastos, Y. Shein, E. Hermans, T. Brijs, G. Wets, A.C.P. Ferraz, Traffic fatality indicators in Brazil: state diagnosis based on data envelopment analysis research, Accid. Anal. Prev. 81 (2015) 61–73, https://doi.org/10.1016/j.aap.2015.01.024.
[4] World Health Organization, Brazil Road Safety Country Profile https://www.who.int/publications/i/item/9789241565066. (accessed 25 August 2023).
[5] Companhia de Engenharia de Tráfego (CET), Acidentes de Trânsito – Relatório anual 2016. https://www.cetsp.com.br/media/562061/relatorioanualacidentestransito-2016.pdf. (acessed 25 August 2023).
[6] H. Naci, D. Chisholm, T.D. Baker, Distribution of road traffic deaths by road user group: a global comparison, Inj. Prev. 15 (2009) 55–59, https://doi.org/10.1136/ip.2008.018721.
[7] L. Fridstrøm, J. Ifver, S. Ingebrigtsen, R. Kulmala, L.K. Thomsen, Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts, Accid. Anal. Prev. 27 (1995) 1–20, https://doi.org/10.1016/0001-4575(94)E0023-E.
[8] M. Moeinaddini, Z. Asadi-Shekari, M. Zaly Shah, The relationship between urban street networks and the number of transport fatalities at the city level, Saf. Sci. 62 (2014) 114–120, https://doi.org/10.1016/j.ssci.2013.08.015.
[9] H. Barbosa, F. Cunto, B. Bezerra, C. Nodari, M.A. Jacques, Safety performance models for urban intersections in Brazil, Accid. Anal. Prev. 70 (2014) 258–266, https://doi.org/10.1016/j.aap.2014.04.008.
[10] NCHRP, The Highway Safety Manual (HSM), 1st Edition ed, National Cooperative on Highway Research Program, 2010.
[11] R. Lamm, E.M. Choueiri, J.C. Hayward, A. Paluri, Possible design procedure to promote design consistency in highway geometric design on two-lane rural roads, Transp. Res. Rec. 1195 (1988) 111–122. http://onlinepubs.trb.org/Onlinepubs/trr/1988/1195/1195-011.pdf.
[12] R. Lamm, B. Psarianos, S. Cafiso, Safety evaluation process for two-lane rural roads: a 10-year review, Transp. Res. Rec. 1796 (2002) 51–59, https://doi.org/10.3141/1796-06.
[13] M. Castro, L. Iglesias, R. Rodriguez-Solano, J.A. Sánchez, Highway safety analysis using geographic information systems, Proc. Inst. Civil Eng.– Transport. 161 (2008) 91–97, https://doi.org/10.1680/tran.2008.161.2.91.
[14] J. Morrall, R. Talarico, Side friction demanded and margins of safety on horizontal curves, Transp. Res. Rec. 1435 (1994) 145. http://onlinepubs.trb.org/Onlinepubs/trr/1994/1435/1435-019.pdf.
[15] R. Lamm, B. Psarianos, T. Mailaender, Highway Design and Traffic Engineering Handbook, McGraw-Hill, New York, 1999.
[16] J. De Oña, L. Garach, Accidents prediction model based on Speed reduction on Spanish two-lane rural highways, Procedia Soc. Behav. Sci. 53 (2012) 1010–1018, https://doi.org/10.1016/j.sbspro.2012.09.950.
[17] I.B. Anderson, K.M. Bauer, D.W. Harwood, K. Fitzpatrick, Relationship to safety of geometric design consistency measures for rural two-lane highways, Transp. Res. Rec. 1658 (1999) 43–51, https://doi.org/10.3141/1658-06.
[18] M.A. Hadi, J. Aruldhas, L.-F. Chow, J.A. Wattleworth, Estimating safety effects of cross-section design for various highway types using negative binomial regression 1500, 1995, pp. 169–177. http://onlinepubs.trb.org/Onlinepubs/trr/1995/1500/1500-021.pdf.
[19] J.L. Martin, Relationship between crash rate and hourly traffic flow on interurban motorways, Accid. Anal. Prev. 34 (2002) 619–629, https://doi.org/10.1016/S0001-4575(01)00061-6.
[20] 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 (2021) 493–502, https://doi.org/10.1016/j.iatssr.2021.05.001.
[21] K. Fitzpatrick, L. Elefteriadou, D.W. Harwood, J.M. Collins, J. McFadden, I. B. Anderson, R.A. Krammes, N. Irizarry, K.D. Parma, K.M. Bauer, K. Passetti, Speed prediction for two-lane rural highways. No. FHWA-RD-99-171, Federal Highway Administration, United States, 2000.
[22] M.C. Taylor, A. Baruya, J.V. Kennedy, Relationship between speed and accidents on rural single-carriageway roads - TRL 511, in: Transport Research Laboratory, 2002 https://cis.ihs.com/cis/document/257222 (accessed 15 September 2023).
[23] J.C.W. Ng, T. Sayed, Effect of geometric design consistency on road safety, Can. J. Civ. Eng. 31 (2004) 218–227, https://doi.org/10.1139/l03-090.
[24] F.J. Camacho-Torregrosa, A.M. Pérez-Zuriaga, J.M. Campoy-Ungría, A. García-García, New geometric design consistency model based on operating speed profiles for road safety evaluation, Accid. Anal. Prev. 61 (2013) 33–42, https://doi.org/10.1016/j.aap.2012.10.001.
[25] T. Ben-Bassat, D. Shinar, Effect of shoulder width, guardrail and roadway geometry on driver perception and behavior, Accid. Anal. Prev. 43 (2011) 2142–2152, https://doi.org/10.1016/j.aap.2011.06.004.
[26] 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. 2010 (42) (2010) 1072–1079, https://doi.org/10.1016/j.aap.2009.12.015.
[27] L. Garach, J. De Oña, G. López, L. Baena, Development of safety performance functions for Spanish two-lane rural highways on flat terrain, Accid. Anal. Prev. 95 (2016) 250–265, https://doi.org/10.1016/j.aap.2016.07.021.
[28] D. Lord, F. Mannering, The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives, Transp. Res. A Policy Pract. 44 (5) (2010) 291–305, https://doi.org/10.1016/j.tra.2010.02.001.
[29] E. Hauer, F.M. Council, Y. Mohammedshah, Safety models for urban four-lane undivided road segments, Transp. Res. Rec. 1897 (2004) 96–105, https://doi.org/10.3141/1897-13.
[30] M. Guo, X. Zhao, Y. Yao, P. Yan, Y. Su, C. Bi, D. Wu, D, a study of freeway crash risk prediction and interpretation based on risky driving behavior and traffic flow data, Accid. Anal. Prev. 160 (2021) 106328, https://doi.org/10.1016/j.aap.2021.106328.
[31] M. Hossain, M. Abdel-Aty, M.A. Quddus, Y. Muromachi, S.N. Sadeek, Real-time crash prediction models: state-of-the-art, design pathways and ubiquitous requirements, Accid. Anal. Prev. 124 (2019) 66–84, https://doi.org/10.1016/j.aap.2018.12.022.
[32] C. Caliendo, M. Guida, A. Parisi, A crash-prediction model for multilane roads, Accid. Anal. Prev. 39 (2007) 657–670, https://doi.org/10.1016/j.aap.2006.10.012.
[33] R. Fu, Y. Guo, W. Yuan, H. Feng, Y. Ma, The correlation between gradients of descending roads and accident rates, Saf. Sci. 49 (2011) 416–423, https://doi.org/10.1016/j.ssci.2010.10.006.
[34] P. Greibe, Accident prediction models for urban roads, Accid. Anal. Prev. 35 (2003) 273–285, https://doi.org/10.1016/S0001-4575(02)00005-2.
[35] P. Savolainen, F. Mannering, Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes, Accid. Anal. Prev. 39 (2007) 955–963, https://doi.org/10.1016/j.aap.2006.12.016.
[36] P. Morency, L. Gauvin, C. Plante, M. Fournier, C. Morency, Neighborhood social inequalities in road traffic injuries: the influence of traffic volume and road design, Am. J. Public Health 102 (2012) 1112–1119, https://doi.org/10.2105/AJPH.2011.300528.
[37] D.N. Moore, W.H.T. Schneider, P.T. Savolainen, M. Farzaneh, Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations, Accid. Anal. Prev. 43 (2011) 621–630, https://doi.org/10.1016/j.aap.2010.09.015.
[38] K. Haleem, M. Abdel-Aty, Examining traffic crash injury severity at unsignalized intersections, J. Saf. Res. 41 (2010) 347–357, https://doi.org/10.1016/j.jsr.2010.04.006.
[39] G. Nilsson, Traffic Safety Dimensions and the Power Model to Describe the Effect of Speed on Safety, Lund Institute of Technology and Society, Traffic Engineering, 2004.
[40] M.H. Cameron, R. Elvik, Nilsson's power model connecting speed and road trauma: applicability by road type and alternative models for urban roads, Accid. Anal. Prev. 42 (2010) 1908–1915, https://doi.org/10.1016/j.aap.2010.05.012.
[41] F. Famoye, D.E. Rothe, Variable selection for Poisson regression model, J. Mod. Appl. Stat. Methods 2 (2003) 380–388, doi: 10.56801/10.56801/v2.i.95.
[42] E.L. Frome, M.H. Kutner, J.J. Beauchamp, Regression analysis of Poisson-distributed data, J. Am. Stat. Assoc. 68 (1973) 935–940, https://doi.org/10.1080/01621459.1973.10481449.
[43] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov, F. Csaki (Eds.), International Symposium on Information Theory, 1973, pp. 267–281.
[44] T. Yamashita, K. Yamashita, R. Kamimura, A stepwise AIC method for variable selection in linear regression, Commun. Stat.—Theory Methods. 36 (13) (2007) 2395–2403, https://doi.org/10.1080/03610920701215639.
[45] M.H. Kutner, C.J. Nachtsheim, J. Neter, Applied Linear Regression Models, McGraw-Hill Irwin, New York, 1983.
[46] R. Srinivasan, T. Saleem, J. Bonneson, V. Gayah, K. Kersavage, NCHRP 17–93: Updating Safety Performance Functions for Data-Driven Safety Analysis, in: Working White Paper: Describing How to Calibrate or Update a Crash Prediction Model, 2023 https://onlinepubs.trb.org/onlinepubs/nchrp/17-93/

GuidanceforStandardizedDatabase and PrioritizingInputDataElements.docx. (accessed May 31, 2024).
[47] E. Hauer, The Art of Regression Modeling in Road Safety, Springer International, New York, 2015.
[48] D.Y. Lin, L.J. Wei, Z. Ying, Model-checking techniques based on cumulative residuals, Biometrics 58 (1) (2002) 1–12, https://doi.org/10.1111/j.0006-341x.2002.00001.x.
[49] W.H. Greene, Econometric Analysis, Prentice Hall, 2012.
[50] C. Lyon, B. Persaud, F. Gross, The Calibrator—An SPF Calibration and Assessment Tool, Updated User Guide. Report No. FHWA-SA-17-016, Federal Highway Administration, Washington, D.C, 2018.
[51] D.C. Montgomery, E.A. Peck, Introduction to Linear Regression Analysis, 2nd ed., John Wiley & Sons, 1992.
[52] M.H. Kutner, C.J. Nachtsheim, J. Neter, Applied Linear Regression Models, 4th ed, McGraw-Hill/Irwin, 2004.
[53] R.M. O'Brien, A caution regarding rules of thumb for variance inflation factors, Qual. Quant. 41 (2007) 673–690, https://doi.org/10.1007/s11135-006-9018-6.
[54] S. Chatterjee, J.S. Simonoff, Handbook of regression analysis, New York, Wiley, NY, 2013.
[55] K.M. Marcoulides, T. Raykov, Evaluation of variance inflation factors in regression models using latent variable modeling methods, Educ. Psychol. Meas. 79 (5) (2019) 874–882, https://doi.org/10.1177/0013164418817803.
[56] M. Bonera, B. Barabino, G. Yannis, G. Maternini, Network-wide road crash risk screening: a new framework, Accid. Anal. Prev. 199 (2024) 107502, https://doi.org/10.2139/ssrn.4573447.