ARTIGOS

ESTIMATING PEDESTRIAN SUSCEPTIBILITY TO TRAFFIC ACCIDENTS IN CURITIBA, BRAZIL*

Cassiano Bastos Moroz[1], Jorge Tiago Bastos[2],
Tatiana Maria Cecy Gadda[3]

Abstract: even though pedestrians represented 40% of all urban displacements in Brazil in 2017, they are still highly vulnerable to traffic accidents, with a mortality rate of 2.89 per 100 thousand inhabitants in 2018. The literature suggests a correlation between the occurrence of traffic accidents and demographic, socioeconomic, and urban structure variables. This study aimed to investigate the pedestrian susceptibility to fatal traffic accidents in the City of Curitiba, in Southern Brazil, based on the correlation between these events and available demographic, socioeconomic, and urban structure spatial variables. The methodology involved the integration of a data-driven statistical method (logistic regression) with geospatial techniques in a GIS software. By adopting broadly available spatial information, the proposed methods were robust in estimating the events, presenting an area under the ROC curve of 0.82 in the cross-validation. Additionally, the results highlighted a strong and statistically significant correlation between the pedestrian crashes and the analysed variables of road system hierarchy, presence of BRT routes, land-use, population density and per capita income.

Keywords: Road safety. Logistic regression. Spatial analysis. Geographic Information System.

ESTIMANDO A SUSCETIBILIDADE DE PEDESTRES A ACIDENTES DE TRÁFEGO EM CURITIBA, BRASIL

Resumo: apesar do transporte a pé ter representado 40% dos deslocamentos urbanos no Brasil em 2017, os pedestres ainda são altamente vulneráveis aos aci-

DOI 10.18224/baru.v7i1.8895

* Recebido em: 07/04/2021. Aprovado em: 22/07/2021.
1 Institute of Environmental Science and Geography, University of Potsdam
2 Federal University of Parana
3 Universidade Tecnológica Federal do Paraná

dentes de trânsito, com uma taxa de mortalidade de 2.89 a cada 100 mil habitantes em 2018. A literatura sugere uma correlação entre a ocorrência de acidentes de trânsito e variáveis demográficas, socioeconômicas e de estrutura urbana. Nesse estudo, essa correlação foi investigada por meio de um modelo estatístico (regressão logística) integrado a ferramentas de análise espacial no ambiente SIG, os quais foram aplicados para estimar a susceptibilidade dos pedestres aos atropelamentos no Município de Curitiba, na Região Sul do Brasil. Adotando informações espaciais amplamente disponíveis em portais de dados abertos, os métodos propostos apresentaram resultados robustos ao estimar os atropelamentos, conforme demonstrado pela área abaixo da curva ROC de 0.82 no processo de validação cruzada. Adicionalmente, os resultados destacaram uma forte correlação, estatisticamente significante, entre os atropelamentos e as variáveis adotadas de hierarquia do sistema viário, presença de rotas de BRT, uso do solo, densidade populacional e renda per capita.

Palavras-chave: Segurança viária. Regressão logística. Análise espacial. Sistema de Informações Geográficas.

ESTIMACIÓN DE LA SUSCEPTIBILIDAD DE LOS PEATONES A ACCIDENTES DE TRÁFICO EN CURITIBA, BRASIL

Resumen: aunque los peatones representaron el 40% de todos los desplazamientos urbanos en Brasil en 2017, siguen siendo altamente vulnerables a los accidentes de tránsito, con una tasa de mortalidad de 2.89 por cada 100 mil habitantes en 2018. La literatura sugiere una relación entre la ocurrencia de accidentes de tránsito y variables demográficas, socioeconómicas y de estructura urbana. En este estudio, esta relación fue investigada a través de un modelo estadístico basado en datos (regresión logística) combinado con análisis espacial GIS, aplicado para estimar la susceptibilidad peatonal a accidentes de tránsito en la ciudad de Curitiba, en el sur de Brasil. Al adoptar información espacial ampliamente disponible, los métodos propuestos fueron robustos para estimar los eventos, presentando un área bajo la curva ROC de 0.82 en la validación cruzada. Además, los resultados destacaron una correlación fuerte y estadísticamente significativa entre los choques peatonales y las variables analizadas de jerarquía del sistema vial, presencia de rutas BRT, uso del suelo, densidad de población e ingreso per cápita.

Palabras clave: Seguridad vial. Regresión logística. Análisis espacial. Sistema de Información Geográfica.

The world has been experiencing a fast urbanization process combined with a growing population, which has created a large number of metropolises, especially in the low- and middle-income countries (UNITED NATIONS, 2018). In the Brazilian context, since the 1980s quality of life in urban centres have been challenged by the increased use of motorized private vehicles and the stagnation of the urban public transport system, driven by an economic slowdown and policy orientation (VASCONCELLOS, 1997; MARICATO, 2008). Despite this national trend, the City of Curitiba, after the 1970s, experienced an increase in the supply and efficiency of public transport, driven by its first Master Plan signed in 1966, which culminated in the implementation of exclusive bus lanes (LEVINSON *et al.*, 2002; MERCIER *et al.*, 2016), worldly renowned as Bus Rapid Transit (BRT) system.

The first federal legislation addressing sustainable urban mobility in Brazil were implemented only at the beginning of the 21st century, including Articles 1

and 2 of the Federal Law number 10,257 (BRAZIL, 2001) – entitled *Estatuto das Cidades*, and the Federal Law number 12,587 (BRAZIL, 2012) – entitled *Lei de Mobilidade Urbana*. These instruments supported a new concept on urban mobility in the country, resulting in policy efforts to change the planning and design strategies in Brazilian cities by focusing on non-motorized means of transportation (SILVA; COSTA; MACEDO, 2008).

According to the National Association of Public Transport (2020), in the municipalities with more than 60 thousand inhabitants, the non-motorized modes contributed to 42.5% of total displacements in 2017, of which 40% was constituted by pedestrians and 2.5% by cyclists. Although these means of transportation are more common in smaller urban areas, where distances are shorter, they are still significant in the large metropolises with more than 1 million inhabitants, encompassing 36.8% of total displacements in the same year (BRAZILIAN NATIONAL ASSOCIATION OF PUBLIC TRANSPORT, 2020).

Despite their representativeness, pedestrians remain highly exposed to traffic accidents in Brazil. As specified by the DATASUS Mortality Information System from the Ministry of Health of the Brazilian Government, pedestrians represented 18.43% of total deaths caused by traffic occurrences in 2018 (the absolute number of pedestrian fatalities was 6,018), corresponding to a mortality rate of 2.89 per 100 thousand inhabitants (BRAZILIAN MINISTRY OF HEALTH, 2020). These contradictory figures indicate that there is a lot to be improved in Brazilian cities to meet the goal of reducing road traffic deaths by at least 50% from 2020 to 2030 (UNITED NATIONS, 2020).

The literature suggests the relationship between the spatial distribution of pedestrian crashes accidents and demographic, socioeconomic, and urban structure variables. In the context of demographic variables, population density information tends to be easily available and it is alternatively used as a measure of susceptibility since pedestrian direct exposure data is rarely available. Thus, greater population density is consistently associated with a higher frequency of pedestrian crashes, since it is usually related to increased exposure level due to commuting and commercial activities (LASCALA, GERBER, GRUENEWALD, 2000; HA, THILL, 2011; CHIMBA, MUSINGUZI, KIDANDO, 2018; DING, CHEN, JIAO, 2018).

Concerning socioeconomic variables, previous studies indicate that low-income areas have a higher probability of pedestrian casualties (SIDDIQUI, ABDEL-ATY, CHOI, 2012; NOLAND, KLEIN, TULACH, 2013; DAI, JAWORSKI, 2016; GRISÉ *et al.*, 2018). Noland et al. (2013) argue that the low-income population is associated with a lower proportion of households owning a vehicle, which increases the pedestrian exposure level since people are walking instead of driving. The reduced car ownership straightforwardly increases the demand for public transport; consequently, public transport facilities such as bus stops tend to generate a greater pedestrian activity leading to a relative increased number of pedestrian crashes around the area (UKKUSURI *et al.*, 2012; CHEN, ZHOU, 2016; DAI, JAWORSKI, 2016).

The influence of urban structure variables on pedestrian safety is investigated throughout several approaches, including the impact of land-use patterns and geometric

design parameters. Researches indicate that land-use types capable of generating pedestrian activity (e.g. commercial and retail facilities, high-density housing, schools and parks) have been related to a higher frequency of pedestrian crashes (KIM, BRUNNER, YAMASHITA, 2006; WEDAGAMA, BIRD, METCALFE, 2006; LOUKAITOU-SIDERIS, LIGGETT, SUNG, 2007). On the geometric design parameters, the most widely investigated in the literature is the number of lanes or road width, since it is criticality related to the amount of exposure of pedestrians when crossing streets (SCHNEIDER, RYZNAR, KHATTAK, 2004; UKKUSURI *et al.*, 2012; AZIZ, UKKUSURI, HASAN, 2013; RANKAVAT, TIWARI, 2016; MOHAN, BANGDIWALA; 2017; BASSANI, ROSSETTI, CATANI, 2020).

In this context, the main purpose of this study is to investigate the pedestrian susceptibility to fatal traffic accidents in the City of Curitiba, in Southern Brazil, based on the correlation between these events and available demographic, socioeconomic, and urban structure spatial variables. Moreover the study (1) presents a methodology that can be used in any urban area – with open-source datasets – helping to orientate urban planning strategies aligned with better road safety-oriented practices; (2) introduces a data-driven statistical method to support the identification of traffic accident hotspots a at the municipal scale; (3) investigates the triggering factors associated with pedestrian traffic accidents in the urban environment and their interconnections; and (4) based on these factors, potentially points out the need for changes in urban policy, especially in underdeveloped neighborhoods.

Many authors (SZE, WONG, 2007; TAO *et al.*, 2015; AGARWAL, KACHROO, REGENTOVA, 2016; YILDIZ, ATEŞ, 2020) applied the model to investigate the factors associated with traffic accidents in the built environment. While these studies significantly contributed to identifying critical elements in road safety, they rarely focused on the spatial component through the integration of the regression model with geospatial tools.

METHODOLOGY

The proposed methodology encompassed the application of multiple logistic regression modelling with geospatial tools to associate the fatal pedestrian crashes with spatial factors of the built environment. The logistic regression is a data-driven statistical analysis method broadly adopted to understand the response of a binary dependent variable to multiple predictors. To develop the analysis, the City of Curitiba was selected as a case study area. Curitiba is the capital of the State of Parana and the largest city in the Southern Region of Brazil, with an estimated population of more than 1.9 million inhabitants in 2019 (BRAZILIAN INSTITUTE OF GEOGRAPHY AND STATISTICS, 2020). The city has historically developed a linear urban occupation along structural axes through the integration of high-density and mixed-use developments with structural roads and the BRT system (MOTTA, 2017). This urban configuration provides a promising scenario to investigate the proposed urban challenge .

The method was performed in five steps, including the (1) analysis of the traffic accident database followed by a random selection of training and validation samples for the statistical approach; (2) pre-processing of the required input data, including the dependent and independent variables, through a set of geospatial tools; (3) conversion of the categorical variables to numeric; (4) execution of the logistic regression analysis with all training and validation samples and assessment of the model accuracy; and (5) running of the final logistic regression model to estimate the pedestrian susceptibility to traffic accidents in the study area.

Traffic Accident Database: Analysis and Sample

The traffic accident database was retrieved from the Project *Vida no Trânsito* from the Ministry of Health of the Brazilian Government, made available by the Curitiba Institute of Research and Urban Planning (2018a). The database encompasses the fatal traffic accidents that occurred in the City of Curitiba from 2010 until 2018.

Initially, the database was filtered according to the category *type of accident* and there were selected only the pedestrian crashes, which are the object of this study. To enable the assessment of the model performance, the *k*-fold cross-validation was adopted. The validation approach, in its simplest way, randomly splits the sample into a single training set and a single validation set, the latter being used to analyse the model predictive capacity. In this approach, the variability in the distribution of observations in the training and validation sets can generate highly distinct results. Therefore, the cross-validation acts as a refinement of the validation approach by introducing a larger number of sets (JAMES *et al*., 2013).

The *k*-fold cross-validation was performed following the method described by James et al. (2013) by setting the *k* equals to five. First, the selected traffic accidents (a total of 718 occurrences) were randomly split into five folds. Then, five training and validation samples were created by interactively adopting 80% (*k*-1 folds) of the dataset for training and 20% (remaining fold) for validation. In summary, the split process was repeated five times, considering a different fold each time until all of them served as the validation sample.

The samples of pedestrian crashes were imported to ArcGIS® by adopting the X and Y coordinates provided in the database, thus generating a vector file with accidents represented as points (see Figure 1). For each one of the training and validation samples, the point features were converted to a raster dataset with 10 meters of spatial resolution based on a binary classification: 1 for traffic accidents and 0 for no traffic accidents. This process generated the input layer that represents the dependent variable in the logistic regression model. The spatial resolution was defined considering the scale of the analysis. While a higher resolution would significantly increase the model processing time, jeopardizing the model performance in a large study area, a lower spatial resolution would merge many accidents into a single pixel, thus resulting in data loss.

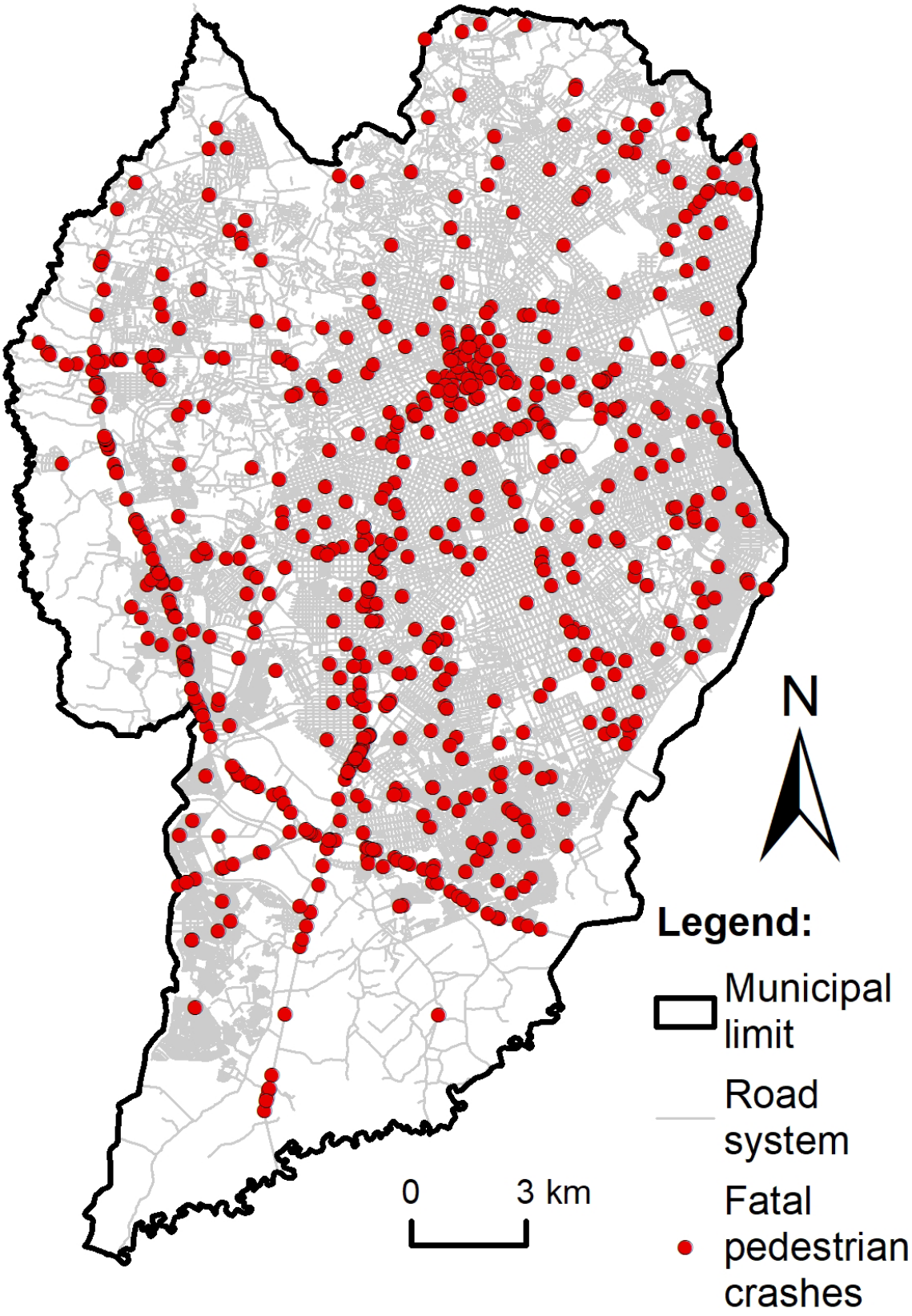

Figure 1. Location of fatal pedestrian crashes registered in the City of Curitiba from 2010 to 2018. Source: Author.

Factor Maps: Acquisition and Pre-processing

Data Acquisition

Based on the literature, five spatial factors were select to investigate their relationship with pedestrian crashes. This includes urban structure (road system hierarchy, presence of BRT routes and land-use), socioeconomic (per capita income) and demographic (population density) variables. The selected spatial factors were pre-processed in ArcGIS®.

To represent the road system, the street axes database from the Curitiba Institute of Research and Urban Planning (2019a) was adopted, which includes all public streets within Curitiba, categorized according to a group of attributes. As information about the number of lanes and the street width was not presented in the acquired data, we sought to identify the attribute that could be related to these parameters, and the classification by *road hierarchy* was selected. This classification divides the street axes into a hierarchy that ranges from 1 (highways) to 4 (minor streets).

The BRT routes were vectorised based on the bus network map provided by Curitiba Urbanisation (2015). To enable this image interpretation process, the original map was converted to an image format, which was then georeferenced and geocoded by adopting an evenly distributed number of street intersections as Ground Control Points (GCPs). The street axes were adopted as a reference when vectorising the BRT routes, thus ensuring that the routes always overlap its corresponding street.

In the case of land-use, the most recent Curitiba Zoning Plan was selected to represent this factor (CURITIBA INSTITUTE OF RESEARCH AND URBAN PLANNING, 2018b). The plan is regulated by the Municipal Law number 15,511 (CITY OF CURITIBA, 2019), which addresses the zoning and land-use in the city. The Zoning Plan is a regulatory framework that orientates urban development by defining zones and sectors that share common characteristics, including densification thresholds and priority land-use classes (CITY OF CURITIBA, 2019). Despite the time divergence between the traffic accident database (2010-2018) and the implementation of the Zoning Plan (2019), the most recent plan was adopted to enable the investigation of the critical areas based on the existing urban configuration.

Finally, the per capita income and population density values were extracted from the last demographic census (2010 Census) carried out by the Brazilian Institute of Geography and Statistics (2011). The database includes several demographic and socioeconomic indexes, which are aggregated according to the census tract code, and a polygonal vector file representing the census tracts. The desired indexes were combined to the attribute table of the vector file through the *Join Field* tool based on a common primary/foreign key: the census tract code.

Data Conversion

To enable the performing of a pixel-by-pixel sampling, each acquired variable was converted to a raster dataset with the same spatial resolution and extent as the pedestrian crashes binary map. This process created the input factor maps for the logistic regression model.

Initially, the street axes and BRT routes were converted into a polygonal feature by adopting the *Buffer* tool with 15 meters. This procedure was required to correctly associate an accident point with its corresponding line (street or BRT route), given that some traffic accidents, after georeferenced, did not precisely intersect the street axes. Then, all five selected factors were converted to a raster dataset. In this process, the major streets and the presence of BRT routes were always prioritized in case of overlapping polygons, as happens in the street intersections. In other words, we indicated to the model that accidents taking place in overlapping areas must always be associated with major streets instead of minor streets, and to the presence of BRT routes instead of its absence.

Data Clipping

An important step in the pre-processing phase was the correct delimitation of the areas that are exposed to traffic accidents. By adopting the entire city in the analysis, without attempting to the road network density, we would wrongly assume that traffic accidents can occur in spaces such as airports, large blocks with military or industrial installations, green areas, among others. In this case, the statistical analysis would result in a low probability of occurrences in these regions because, as expected, the road density is minimal or even null. However, it does not mean that, in the small road sections that cross these areas, there is a low number of observed accidents.

In this context, all factor maps and accident binary maps were clipped by applying the 15-meter buffer polygon as a mask, which was previously generated from the street axes. In summary, this process resulted in multiple maps that maintained the previous pixel values within the buffer extent, while attributed a null value to the pixels outside of it.

Categorical Variables: Conversion to Numeric

When performing a logistic regression analysis, the input data is composed of a matrix that correlates, in each row, the values of the dependent and independent variables (JAMES *et al.*, 2013). Therefore, in the context of this study, the first column represents the pedestrian crashes, which are expressed as a binary. The remaining columns represent the factor maps, or the independent variables for which coefficients will be estimated to predict the probability of accidents occurrence. However, in the case of the categorical variables, which express qualitative instead of quantitative characteristics, the estimation of a single coefficient for the whole column is not possible, as these variables do not possess a numerical meaning (SELTMAN, 2018).

To overcome this limitation, the categorical variables were converted to numeric variables through bivariate statistical analysis by adopting the weight of evidence (WoE) method. The WoE was performed based on the theory described by Bonham-Carter (1994), which applies the prior and conditional probabilities to understand the importance of each factor for the occurrence of an analysed event. The prior probability ($P\{E\}$) represents the probability of an event E occurrence based on similar events

that happened in the past. In a GIS environment, this is represented by the number of cells with events divided by the total number of cells in the study area. The conditional probability ($P\{E \mid C\}$) is calculated when another source of information (e.g. land-use classes, road system hierarchy) is adopted for the probability estimation (REGMI; GIARDINO; VITEK, 2010). Therefore, it represents the probability of having an event E while being in a specific class C, and it can be calculated as the number of cells with events that intersect the analysed class divided by the total number of cells in the class.

Based on these probabilities, Bonham-Carter (1994) defined two equations to estimate the positive and negative weights (W_e^+ and W_e^-) for each class of the analysed spatial factor. The positive weight indicates how important is the presence of the class for the occurrence of the event, while the negative weight indicates how important is its absence. The equations are described as follows:

$$W_e^+ = \log_e \frac{P\{E \mid C\}}{P\{E \mid \overline{C}\}} \tag{1}$$

Where $P\{E \mid C\}$ represents the probability of having an event E while being in class C, and $P\{E \mid \overline{C}\}$ represents the probability of having an event E while not being in class C.

$$W_e^- = \log_e \frac{P\{\overline{E} \mid C\}}{P\{\overline{E} \mid \overline{C}\}} \tag{2}$$

Where $P\{\overline{E} \mid C\}$ represents the probability of not having an event C while being in class C, and $P\{\overline{E} \mid \overline{C}\}$ represents the probability of not having an event E while not being in class C.

Based on the positive and negative weights, the final weight ($W_{e,final}$) for each class of a factor map is calculated as (BONHAM-CARTER, 1994):

$$W_{e,final} = W_e^+ - W_e^- + W_{e,total}^- \tag{3}$$

Where $W_{e,final}$ is the final weight in the class, W_e^+ is the positive weight in the class, W_e^- is the negative weight in the class, and $W_{e,total}^-$ is the sum of the negative weights from all classes represented in a factor map.

According to the equations and concepts described before, a model was developed in ArcGIS® to interactively calculate the final weights of evidence for all classes from the categorical variables. Additionally, for each categorical variable, six WoE maps were generated: five for the training samples of pedestrian crashes and one for the complete sample. Therefore, in the context of the *k*-fold cross-validation, the WoE map estimated from the training sample *k* for the factor F (categorical) was adopted as a factor map F_{WoE} (numeric) in the logistic regression model that is performed with the same training sample *k*.

In this study, three factors are categorical variables (road system hierarchy, BRT routes and land-use) and two are numeric (per capita income and population density).

Because BRT routes represent an indicator variable it should be considered as numeric in the logistic regression model, as the coefficient will already represent the behaviour of the cells indicating the presence of the routes, the value 1 (SELTMAN, 2018). For this reason, only the road system hierarchy and land-use were converted to numeric variables.

Logistic Regression Model

The logistic regression model is a powerful method to estimate the probability that a certain event happens, based on independent variables (JAMES *et al.*, 2013; SELTMAN, 2018). In this sense, the probability values calculated through this method range from 0 to 1, where 0 expresses 0% and 1 expresses 100% of the probability of an event occurrence. In the case of a multiple logistic regression, where more than one independent variable, or predictor, is considered to estimate the response of the dependent variable, the equation is defined as follows (JAMES *et al.*, 2013):

$$p(x) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}} \quad (4)$$

Where $p(x)$ represents the probability that an event happens, $X_1,...,X_p$ represents p predictors, and $\beta_0,\beta_1,...,\beta_p$ represents the regression coefficients related to the predictions.

The regression coefficients are initially unknown. For this reason, they must be estimated based on the available training data (JAMES *et al.*, 2013). After this, the final logistic regression equation, containing the estimated coefficients, must be validated with a new sample, the validation data, to analyse the predictive capacity of the model. Therefore, if the model accurately predicts the behaviour of a new dataset, which was not used for training, it is assumed that it has a good predictive capacity, and can be applied to investigate a certain phenomenon or to simulate new scenarios (FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION, 2013).

In this study, six logistic regression models were calibrated, five for the cross-validation and one to estimate the final pedestrian susceptibility to traffic accidents. To create the input dataset for the statistical analysis, the binary maps representing the pedestrian crashes were sampled, pixel-by-pixel, with all the spatial factors. Initially, this process was performed for each one of the five combinations of training and validation samples. This procedure resulted in five matrixes, where each pixel of the map was translated into a row correlating the dependent and independent variables.

In sequence, the statistical analysis was performed through a code created in RStudio®. First, the independent variables were normalized to ensure that they would equally contribute to the analysis, thus enabling the comparison between the regression coefficients. The normalization scales the values in a way that all variables present a

mean of 0 and a standard deviation of 1, while it does not affect the shape of the distribution (COX, 2007). In this study, the z-score normalization method was adopted, which was calculated by subtracting the mean value of the variable, and then dividing the result by the standard deviation of the same variable.

After the normalization, the logistic regression model was calibrated, with each training sample, by applying the *glm()* function with the binomial family. This process resulted in the regression coefficients for the independent variables. Based on the estimated coefficients, the logistic regression equation was adopted to estimate the pedestrian susceptibility to traffic accidents for each sample. The susceptibility values, ranging from 0 to 1, were calculated to all rows of the matrix. After this, the receiver operating characteristics (ROC) curve was displayed to analyse the overall performance of the model, following the method described by James et al. (2013).

To create the ROC curve, several thresholds *t*, representing a specific probability, were defined by adopting an equal interval of 0.05 (i.e. 0.05, 0.10, …, 0.95, 1.00). For each threshold, the study area was divided into two classes: (1) susceptible when the susceptibility was higher than *t*; (2) not susceptible when the susceptibility was lower than *t*. Then, for each threshold classification, the true positive rate (TPR) and false positive rate (FPR) were calculated based on the equations presented by James et al. (2013), as described below:

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

Where TPR is the true positive rate, TP is the number of true positives, and FN is the number of false negatives.

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

Where FPR is the false positive rate, FP is the number of false positives, and TN is the number of true negatives.

The points obtained from each combination of TPR and FPR for a specific value of *t* were displayed in a scatter plot, with the FPR represented in the x-axis and the TPR in the y-axis.

After evaluating the performance of the logistic regression model through the generation of the ROC for all training samples of the k-fold cross-validation, a final model was calibrated with the complete sample of pedestrian crashes. The statistical analysis was conducted based on the same code in RStudio®, as described before. The regression coefficients presented were then applied in the logistic regression equation to estimate the pedestrian susceptibility to traffic accidents. In this sense, a susceptibility value was calculated for each pixel of the study area, represented as a row in the matrix. To enable the spatial visualization of the results, these values were imported to ArcGIS® by adopting the X and Y coordinates as a reference.

Finally, the susceptibility values were split into three classes by adopting thresholds according to the following rules: (1) the high susceptibility class should contain a large number of the historical pedestrian crashes in a small area, (2) the medium susceptibility class should contain a small number of pedestrian crashes in a small area, and (3) the low susceptibility class should contain a small number of pedestrian crashes in a large area. Notably, these rules can only be fulfilled with a good model performance (typically an AUC over 0.8), when the historical events are correctly associated with the areas presenting the highest susceptibility. In the context of this study, the model enabled the adoption of thresholds so that the high susceptibility class contained 82% of the events in 20% of the study area with the highest values, the medium susceptibility class contained 8% of the events in the following 20% of the study area, and the low susceptibility class contained 10% of the events in 60% of the study area with the lowest values.

Results and Discussion

WoE maps for the adopted independent variables

In an initial stage in the analysis, the data pre-processing and the conversion of the categorical variables to numeric generated a set of input maps representing the five spatial factors adopted as independent variables in the logistic regression. Figure 2 exemplifies these input maps. In the figure, maps A and C were obtained through the WoE method. In map A (road hierarchy), it is possible to observe that most of the streets presented a low WoE value, as represented by the green colour. These streets refer to a road hierarchy 4, which represents minor roads. On the other, the areas with a high WoE value, highlighted in red, represent the highways in the study area, classified as road hierarchy 1. In the case of map C (land-use), the visualization of the orange and red areas indicates that most of the higher WoE values were estimated in regions that are related either to the presence of highways (see map A), BRT routes (see map B), or the central area of the city (the red/orange region in the center of map C). The other spatial factors of BRT routes (map B), per capita income (map D), and population density (map E) were not converted to WoE values because they already represent numerical or binary variables.

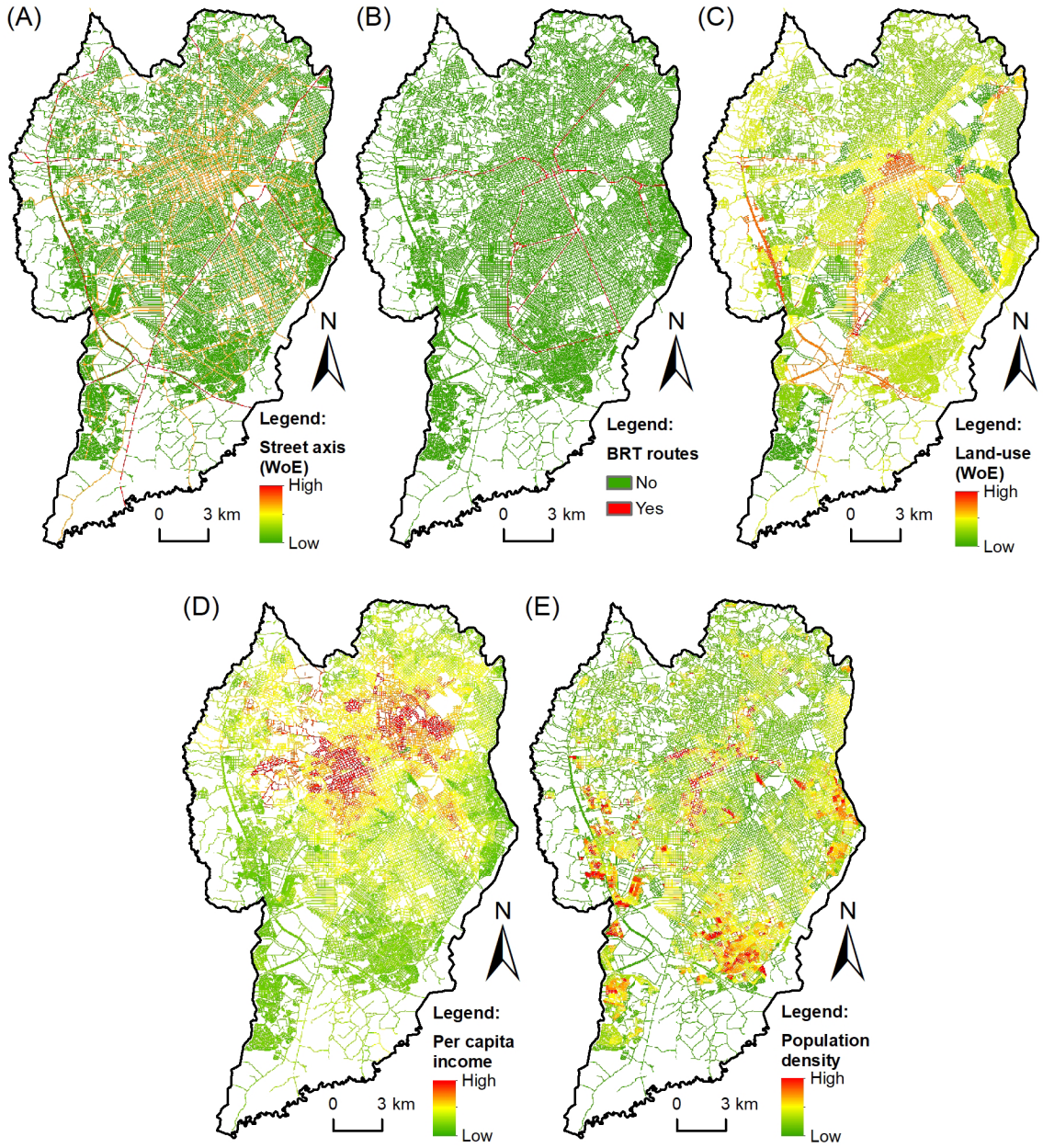

Figure 2. Spatial factors adopted as independent variables in the logistic regression model: (A) road system hierarchy; (B) presence of BRT routes; (C) land-use; (D) per capita income; and (E) population density. Source: Author.

Logistic Regression results

In the following step of the methodology, five logistic regression models were calibrated to perform the *k*-fold cross-validation. Table 1 presents the coefficients estimated for the adopted spatial factors, with their corresponding p-values, for each training sample. The land-use presented the strongest correlation ($\beta = 1.165$) with the occurrence of pedestrian crashes, followed by the road system hierarchy ($\beta = 0.908$). Despite the lower correlation of the remaining predictors, they were all statistically significant, with a p-value lower than 0.05. The only exception was the population density calibrated with the training sample 5, which resulted in a p-value equals to 0.166. However, this

predictor was still adopted in the final logistic regression model, as it was statistically significant when adopting all the other four training samples.

Table 1. Estimated logistic regression coefficients (β) and p-value for each spatial factor per cross-validation sample (1, 2, …, 5). Mean and standard deviation values are presented to indicate the variation among the analyzed samples. Source: Author.

Spatial factor	Training sample					
	1		2		3	
	β	p-value	β	p-value	β	p-value
Per capita income	-0.256	<< 0.001	-0.210	<< 0.001	-0.199	<< 0.001
Population density	0.070	0.013	0.054	0.031	0.059	0.038
Road system hierarchy (WoE)	0.907	<< 0.001	0.940	<< 0.001	0.908	<< 0.001
BRT routes	0.064	0.000	0.054	0.002	0.049	0.006
Land-use (WoE)	1.454	<< 0.001	1.202	<< 0.001	1.165	<< 0.001

Spatial factor	Training sample				Mean	Standard deviation
	4		5			
	β	p-value	β	p-value		
Per capita income	-0.248	<< 0.001	-0.201	<< 0.001	-0.223	0.024
Population density	0.063	0.040	0.045	0.166	0.058	0.008
Road system hierarchy (WoE)	0.914	<< 0.001	0.888	<< 0.001	0.911	0.017
BRT routes	0.069	<< 0.001	0.047	0.010	0.056	0.009
Land-use (WoE)	1.331	<< 0.001	1.327	<< 0.001	1.296	0.103

The coefficients presented in Table 1 were then applied to estimate the pedestrian susceptibility to traffic accidents for each one of the five models. These maps of pedestrian susceptibility were integrated with the validation samples (the remaining 20% of the pedestrian crashes) to generate the ROC curve, as presented in Figure 3. The hatched area represents the range between the minimum and the maximum values extracted from the five models, while the continuous black line illustrates the average value. An ideal ROC curve must be close to the top left corner of the graph, which results in an area under the (ROC) curve (AUC) close to 1. An AUC equals to 0.5 indicates that the model performs no better than chance (JAMES *et al.*, 2013), similar to the line illustrated as a black dashed line in Figure 3. In this study, the average AUC was calculated as 0.82 for the validation samples, thus indicating the good performance of the logistic regression model.

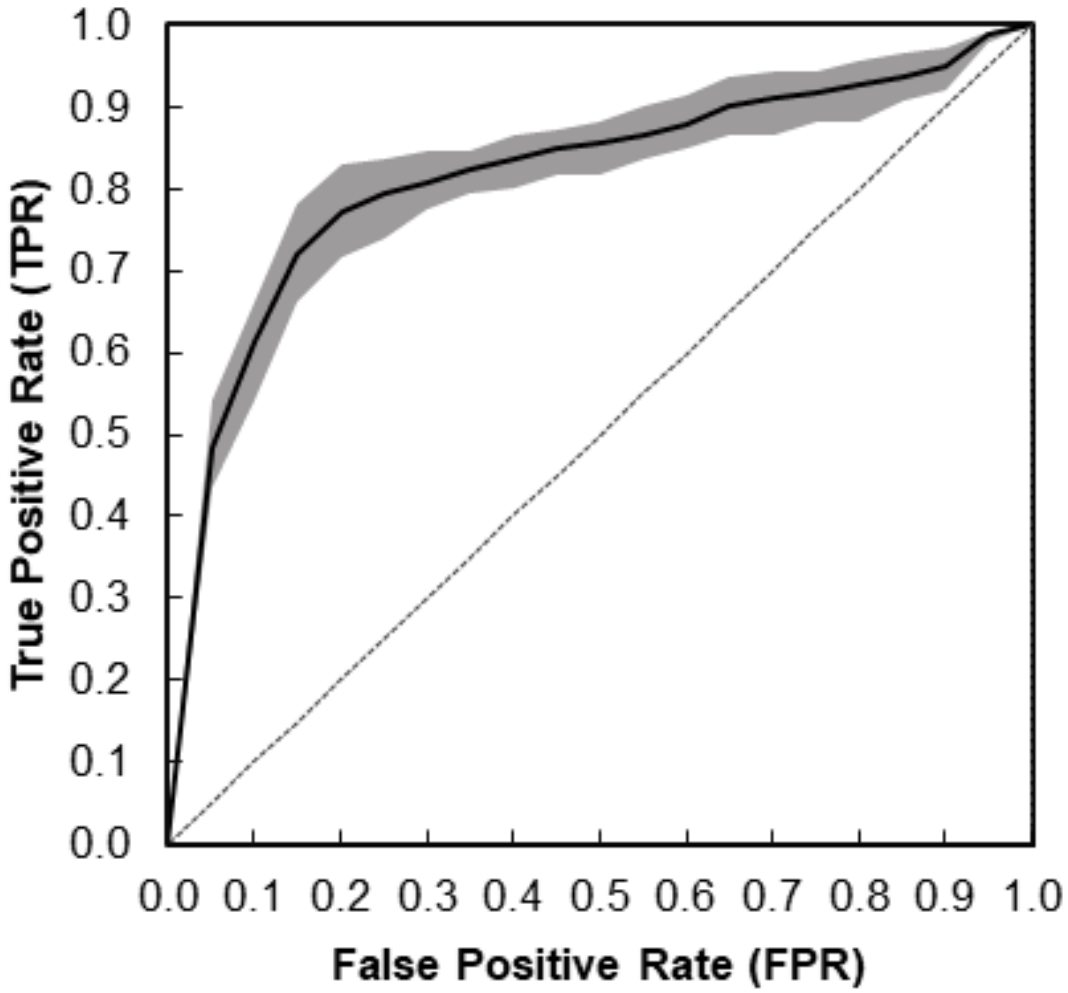

Figure 3. ROC curve (continuous black line) of the logistic regression model for the mean value of the cross-validation samples. The shaded area represents the interval between the minimum and the maximum values from all validation samples. Source: Author.

Based on the good performance of the cross-validation, a new logistic model was calibrated with the entire sample to estimate the pedestrian susceptibility to traffic accidents in the City of Curitiba. Table 2 presents the estimated regression coefficients. Analogously to the training cross-validation, the land-use ($\beta = 1.125$) and the road system hierarchy ($\beta = 0.915$) presented the strongest correlation with the pedestrian crashes, and all predictors were considered statistically significant, with a p-value lower than 0.05.

Table 2. Estimated regression coefficients (β) and p-value for each spatial factor for the final logistic regression model, trained with the complete sample. Source: Author.

Spatial factor	β	p-value
Intercept	-8.775	<< 0.001
Per capita income	-0.222	<< 0.001
Population density	0.059	0.028
Road system hierarchy (WoE)	0.915	<< 0.001
BRT routes	0.057	< 0.001
Land-use (WoE)	1.125	<< 0.001

Estimation of the Pedestrian Susceptibility to Traffic Accidents

The final step of the research was the generation of the susceptibility map based on the calibrated logistic regression model. Figure 4 A illustrates the estimated pedestrian susceptibility to traffic accidents in the City of Curitiba, with a focus on the areas highlighted with a dashed black line. The legend of these areas is highlighted in Figure 4B. The combined analysis of the maps presented in Figures 2 and 4 allows the identification of the impacts on pedestrians' road safety levels due to the combination of the investigated spatial factors throughout different regions of the city.

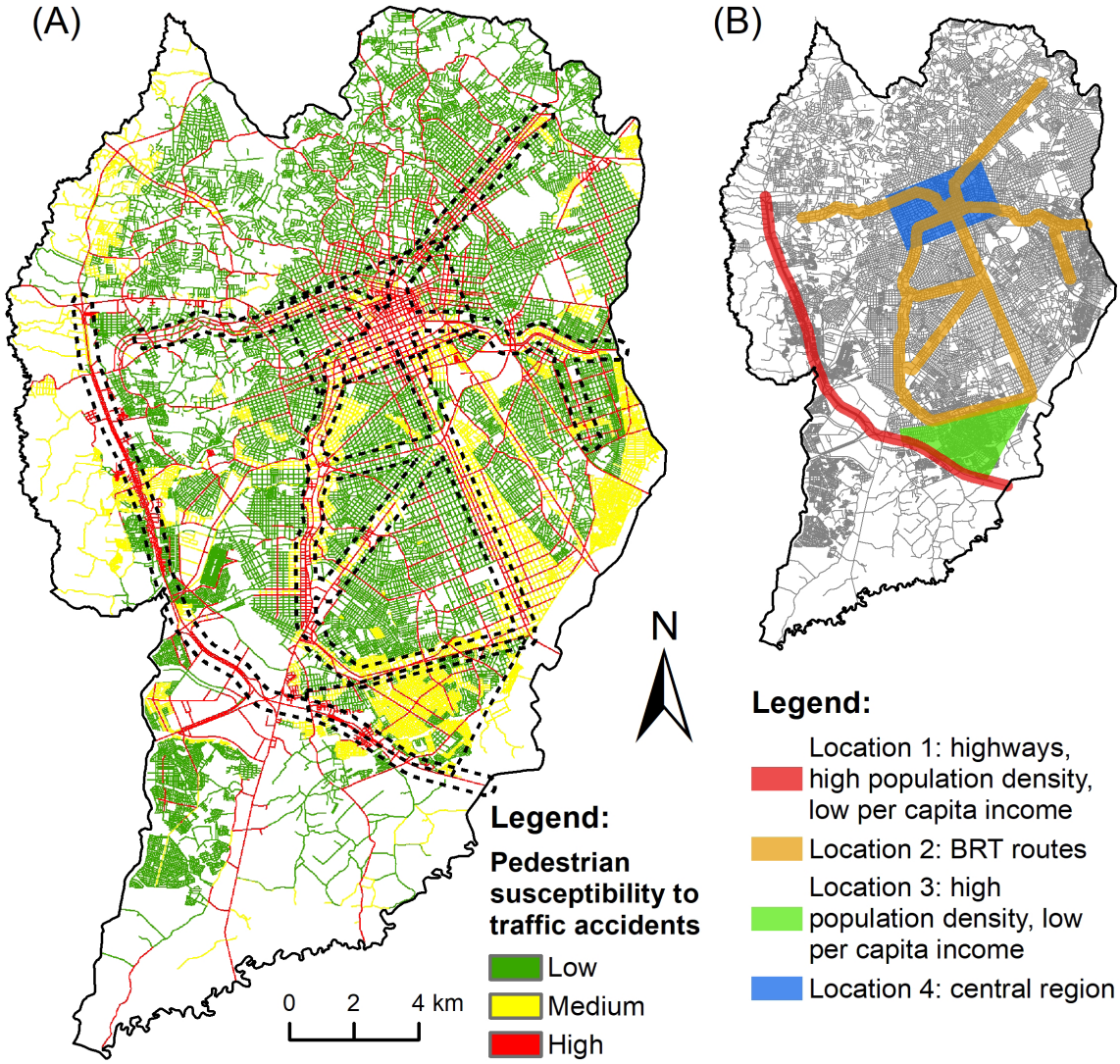

Figure 4. Final map of pedestrian susceptibility to traffic accidents in the City of Curitiba. Source: Author.

For example, the BRT routes are associated with the zoning since the high-capacity transit supply was a decisive factor for establishing land-use guidelines on the BRT routes' surrounding areas in the City of Curitiba. Also, the road hierarchy was defined based on a main road system composed of arterial and bus-exclusive roads. The combination of such factors contributes to a higher susceptibility of pedestrians to traffic accidents in the areas adjacent to the BRT routes, as presented in Figure 4.

On the other hand, the analysis of the population density leads to the identification of regions with high population density where there is no BRT service supply or it is substantially far, despite the principle of associating high population density (demand) and transit supply. The Southern and Eastern peripheral areas of the map are an example of such principle detachment. These areas are characterized by a lower per capita income level (map D in Figure 2), which evidence the occupation of lower-value areas far from the city centre poorly supplied in terms of the transit system. The susceptibility map of Figure 4 suggests that such a combination also created a risky environment for pedestrians, as these areas are mostly classified as medium or high susceptibility. In this context, a possibly relevant aspect that might explain the failure of the urban planning guidelines is the land value, given that the areas close to the transit supply tend to be more expensive to live in, resulting in the movement of the low-income population to peripheral areas. Also, regions with higher land value tend to present a lower per capita land-use rate, despite the higher construction density.

Thus, there is enough evidence to suggest that there is a reasonable correspondence between the factors directly defined by urban planning guidelines (i.e. street axis, BRT routes and land-use). Conversely, the correspondence of these variables with measured parameters (i.e. income level and population density) seemed to be impaired by uncontrolled variables. Both situations produce high susceptibility to pedestrian accidents. The main road hierarchy is associated with a high-speed limit, in this case, equal to 60 km/h or even higher. Therefore, the combination of high-speed limits, high population density and low per capita income created a high susceptibility environment for traffic accidents involving pedestrians.

The presence of BRT routes in combination with high population density, apart from the income level, also seems to contribute to the susceptibility of pedestrian fatal accidents. Two aspects might explain this association. The first of them is related to the physical and operational characteristics of the BRT, a surface-level system in which interaction with pedestrians in crossing areas is frequent. It indeed constitutes a risky environment, since 13% of pedestrian fatal accidents from 2010 to 2018 involved buses (a discretization between normal buses and BRT vehicles is not present). The second aspect concerns the association between the BRT service supply and high-speed limit roads that compose the structural axes.

CONCLUSION

This study generated an updated map of the pedestrian susceptibility to fatal traffic accidents in the City of Curitiba based on broadly available spatial information on demographic (population density), socioeconomic (per capita income) and urban structure factors (presence of BRT routes, road system hierarchy and land-use). The application of a data-driven statistical model combined with GIS spatial analysis presented accurate results when estimating the susceptibility, as demonstrated by an average area under the ROC curve of 0.82 for the validation samples. Therefore, given the great performance of the proposed methodology and the simplicity of its variables, it is possible to affirm that

the model can be used in any urban areas, supporting urban planning practices focused on road safety.

In the context of the urban planning practices, there were identified some externalities that influence the occurrence of traffic accidents in the built environment. In this sense, the urban planning guidelines on public transportation and land-use, even though well-succeeded in many aspects, might produce negative side effects in terms of road safety of pedestrians. This includes, for example, the high pedestrian susceptibility to traffic accidents in the regions adjacent to BRT routes, especially those with high population density. Another relevant aspect is the association between the BRT service supply and high-speed limit roads, which also presented high susceptibility values.

Additionally, this study enabled the spatial visualization of situations in which the urban planning guidelines themselves were not capable of orientating the population growth. This constitutes low-income peripheral areas, with high population density, distant from the transit supply and close to high-speed roads or highways, which, as already mentioned, creates a risky combination for severe pedestrian crashes.

For future studies, it is recommended to adopt the proposed methodology to conduct a spatial-temporal analysis of the pedestrian crashes in the City of Curitiba. By adding the temporal scale, it might be possible to further understand how pedestrian crashes have evolved over the years within the city, or how specific measures that were implemented in the past (e.g. policies and legislations) influenced the pedestrian susceptibility throughout the city. Another suggestion is the adoption of non-fatal traffic accidents in the analysis, which opens a broad range of research questions in the field of road safety. This includes, for example, understanding the sensitivity of the susceptibility map according to the severity of the accidents. Additionally, future studies can also address the adoption of other variables to map the susceptibility of pedestrians to traffic accidents. Several studies – including the references in this paper – can be used as a guideline to select the possible variables. Finally, the traffic accidents hotspots delimited in this study – the high susceptible areas – can be further investigated in future research on a more detailed scale, which can enable the identification of local elements that might be associated with a higher pedestrian susceptibility.

REFERENCES

AGARWAL, S.; KACHROO, P.; REGENTOVA, E. A hybrid model using logistic regression and wavelet transformation to detect traffic incidents', *IATSS Research*. International Association of Traffic and Safety Sciences, v. 40, n. 1, p. 56-63, 2016. DOI: 10.1016/j.iatssr.2016.06.001.

AZIZ, H. M. A.; UKKUSURI, S. V.; HASAN, S. Exploring the determinants of pedestrian-vehicle crash severity in New York City. *Accident Analysis and Prevention*. Elsevier Ltd, n. 50, p. 1298-1309, 2013. DOI: 10.1016/j.aap.2012.09.034.

BASSANI, M.; ROSSETTI, L.; CATANI, L. Spatial analysis of road crashes involving vulnerable road users in support of road safety management strategies. *Transportation Research Procedia,* n. 45, p. 394-401, 2020. DOI: 10.1016/j.trpro.2020.03.031.

BONHAM-CARTER, G. F. *Geographic Information Systems for geoscientists*: modeling with GIS. Oxford: Pergamon, 1994.

BRAZIL. *Estatuto das Cidades [Statute of Cities]*. Brazil. 2001. Available at: http://www.planalto.gov.br/ccivil_03/leis/leis_2001/l10257.htm.

BRAZIL. *Lei da Mobilidade [Mobility Law]*. Brazil. 2012. Available at: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12587.htm.

IBGE: Brazilian Institute of Geography and Statistics. *Características da população e dos domicílios: resultados do universo. Agregados por setores censitários [Characteristics of population and domiciles: universal results. Aggregated by census tract]*, *Censo 2010 resultados*. 2011. Available at: https://censo2010.ibge.gov.br/resultados.html. Accessed: 10 Jan. 2020.

IBGE*:* Brazilian Institute of Geography and Statistics. 2019. *Cidades: Curitiba [IBGE cities: Curitiba]*. Available at: https://cidades.ibge.gov.br/brasil/pr/curitiba/panorama. Accessed: 15 Jul. 2020.

BRAZILIAN MINISTRY OF HEALTH. 2020. *Mortes no trânsito - dados definitivos de 2018 [Deaths on traffic - definitive data for 2018]*, *MS/SVS/CGIAE - Sistema de Informações sobre Mortalidade - SIM*. Available at: http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sim/cnv/ext10uf.def Accessed: 29 Apr. 2020.

BRAZILIAN NATIONAL ASSOCIATION OF PUBLIC TRANSPORT. 2020. *Sistema de informações da mobilidade urbana da Associação Nacional de Transportes Público - relatório geral 2017 [Information system of urban mobility from the National Agency of Public Transport - general report 2017]*. Available at: http://files.antp.org.br/simob/sistema-de-informacao-de-mobilidade-urbana-da-antp--2017.pdf.

CHEN, P.; ZHOU, J. Effects of the built environment on automobile-involved pedestrian crash frequency and risk', *Journal of Transport & Health*. Elsevier, v. 3, n. 4, p. 448-456, 2016. DOI: 10.1016/j.jth.2016.06.008.

CHIMBA, D.; MUSINGUZI, A.; KIDANDO, E. Associating pedestrian crashes with demographic and socioeconomic factors', *Case Studies on Transport Policy*. Elsevier, v. 6, n. 1, p. 11-16, 2018. DOI: 10.1016/j.cstp.2018.01.006.

CITY OF CURITIBA. *Lei de Zoneamento, Uso e Ocupação do Solo [Zoning, Land-use and Occupation Law]*. Brazil. 2019. Available at: https://ippuc.org.br/visualizar.php?doc=http://admsite2013.ippuc.org.br/arquivos/documentos/D311/D311_015_BR.pdf.

COX, N. J. *Transformations*: an introduction. 2007. Available at: http://fmwww.bc.edu/repec/bocode/t/transint.html. Accessed: 2 Mar. 2020.

CURITIBA INSTITUTE OF RESEARCH AND URBAN PLANNING. *Acidentes de trânsito com vítimas fatais*: Município de Curitiba [Traffic accidents with fatal victims. Municipality of Curitiba], Projeto Vida no Trânsito. 2018a. Available at: http://www.ippuc.org.br/mapasinterativos/AcidentesDeTransito/dashboard.html. Accessed: 10 Jan. 2020.

CURITIBA INSTITUTE OF RESEARCH AND URBAN PLANNING. *Zoneamento* - polígonos [Zoning - polygons], Dados Geográficos. 2018b. Available at: https://ippuc.org.br/geodownloads/geo.htm. Accessed: 10 Jan. 2020.

CURITIBA INSTITUTE OF RESEARCH AND URBAN PLANNING. *Eixos de rua [Street axis]*, *Dados Geográficos*. 2019. Available at: https://ippuc.org.br/geodownloads/geo.htm. Accessed: 10 Jan. 2020.

CURITIBA URBANISATION. *Curitiba. Rede integrada de transporte coletivo de Curitiba [Curitiba. Integrated public transport network of Curitiba]*. 2015. Available at: https://www.urbs.curitiba.pr.gov.br/PORTAL/publicador/intranet/BOLETRANS/boletim/upload/1867-20150415135315_5.pdf. Accessed: 15 Feb. 2020.

DAI, D.; JAWORSKI, D. 'Influence of built environment on pedestrian crashes: a network-based GIS analysis'. *Applied Geography*, n. 73, p. 53-61, 2016. DOI: 10.1016/j.apgeog.2016.06.005.

DING, C.; CHEN, P.; JIAO, J. 'Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: a machine learning approach', *Accident Analysis and Prevention*. n. 112, p. 116-126, 2018. DOI: 10.1016/j.aap.2017.12.026.

ITC: Faculty of Geo-Information Science and Earth Observation. 'Models and modelling'. *The core of GIScience: a systems-based approach*. Enschede, p. 41-48, 2013.

GRISÉ, E. *et al.* 'A geography of child and elderly pedestrian injury in the City of Toronto, Canada'. *Journal of Transport Geography*. Elsevier, n. 66, p. 321-329, 2018. DOI: 10.1016/j.jtrangeo.2017.10.003.

HA, H.; THILL, J. 'Analysis of traffic hazard intensity: a spatial epidemiology case study of urban pedestrians', *Computers, Environment and Urban Systems*. Elsevier Ltd, v. 35, n. 3, p. 230-240, 2011. DOI: 10.1016/j.compenvurbsys.2010.12.004.

JAMES, G. *et al. An introduction to statistical learning: with applications in R*. Springer T. New York: Springer Science+Business Media, 2013. DOI: 10.1007/978-1-4614-7138-7.

KIM, K.; BRUNNER, I. M.; YAMASHITA, E. Y. 'Influence of land use, population, employment, and economic activity on accidents'. *Journal of the Transportation Research Board*, 1953, p. 56-64, 2006. DOI: 10.3141/1953-07.

LASCALA, E. A.; GERBER, D.; GRUENEWALD, P. J. 'Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis'. *Accident Analysis and Prevention*, n. 32, p. 651-658, 2000. DOI: 10.1016/S0001-4575(99)00100-1.

LEVINSON, H. *et al.* 'Bus Rapid Transit: an overview', *Journal of Public Transportation*, v. 5, n. 2, p. 1-30, 2002. DOI: 10.5038/2375-0901.5.2.1.

LOUKAITOU-SIDERIS, A.; LIGGETT, R.; SUNG, H. 'Death on the crosswalk in Los Angeles', *Journal of Planning Education and Research*, p. 338-351, 2007. DOI: 10.1177/0739456X06297008.

MARICATO, E. 'O automóvel e a cidade [The automobile and the city]', *Ciência e Ambiente*, n. 37, p. 5-12, 2008.

MERCIER, J. *et al.* 'Policy tools for sustainable transport in three cities of the Americas: Seattle, Montreal and Curitiba', *Transport Policy*. Elsevier, n. 50, p. 95-105, 2016. DOI: 10.1016/j.tranpol.2016.06.005.

MOHAN, D.; BANGDIWALA, S. I. 'Urban street structure and traffic safety', *Journal of Safety Research*, n. 62, p. 63-71, 2017. DOI: 10.1016/j.jsr.2017.06.003.

MOTTA, B. G. *A bikeability index for Curitiba (Brazil)*. University of Twente. 2017.

NOLAND, R. B.; KLEIN, N. J.; TULACH, N. K. 'Do lower income areas have more pedestrian casualties ?'. *Accident Analysis and Prevention,* n. 59, p. 337-345, 2013. DOI: 10.1016/j.aap.2013.06.009.

RANKAVAT, S.; TIWARI, G. 'Pedestrians risk perception of traffic crash and built environment features – Delhi , India'. *Safety Science*, n. 87, p. 1-7, 2016. DOI: 10.1016/j.ssci.2016.03.009.

REGMI, N. R.; GIARDINO, J. R.; VITEK, J. D. 'Modeling susceptibility to landslides using the weight of evidence approach: Western Colorado, USA', *Geomorphology*. Elsevier B.V., v. 115, p. 1-2, p. 172-187, 2010. DOI: 10.1016/j.geomorph.2009.10.002.

SCHNEIDER, R. J.; RYZNAR, R. M.; KHATTAK, A. J. 'An accident waiting to happen: a spatial approach to proactive pedestrian planning', *Accident Analysis and Prevention*, n. 36, p. 193-211, 2004. DOI: 10.1016/S0001-4575(02)00149-5.

SELTMAN, H. J. *Experimental design and analysis*. Pittsburgh: Carnegie Mellon University, 2018. Available at: http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf.

SIDDIQUI, C., ABDEL-ATY, M. AND CHOI, K. (2012) 'Macroscopic spatial analysis of pedestrian and bicycle crashes', *Accident Analysis and Prevention*. Elsevier Ltd, 45, pp. 382–391. doi: 10.1016/j.aap.2011.08.003.

SILVA, A. N. R. DA, COSTA, M. DA S. AND MACEDO, M. H. (2008) 'Multiple views of sustainable urban mobility: the case of Brazil', *Transport Policy*, 15, pp. 350–360. doi: 10.1016/j.tranpol.2008.12.003.

SZE, N. N. AND WONG, S. C. (2007) 'Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes', *Accident Analysis and Prevention*, 39, pp. 1267–1278. doi: 10.1016/j.aap.2007.03.017.

TAO, L. *et al.* 'The traffic accident hotspot prediction: based on the logistic regression method', *ICTIS 2015 - 3rd International Conference on Transportation Information and Safety, Proceedings*, May, p. 107-110, 2015. DOI: 10.1109/ICTIS.2015.7232194.

UKKUSURI, S. *et al.* 'The role of built environment on pedestrian crash frequency', *Safety Science*, v. 50, n. 4, p. 1141-1151, 2012. DOI: 10.1016/j.ssci.2011.09.012.

UNITED NATIONS. *World urbanization prospects*, *Demographic Research*. 2018. DOI: 10.4054/demres.2005.12.9.

UNITED NATIONS. ‘Stockholm Declaration’, in *Third Global Ministerial Conference on Road Safety: Achieving Global Goals 2030*, p. 19-20, 2020.

VASCONCELLOS, E. A. ‘The demand for cars in developing countries’, *Transportation Research Part A: Policy and Practice*, v. 31, n. 3, p. 245-258, 1997. DOI: 10.1016/S0965-8564(96)00021-3.

WEDAGAMA, D. M. P.; BIRD, R. N.; METCALFE, A. V. ‘The influence of urban land-use on non-motorised transport casualties’, *Accident Analysis and Prevention*, n. 38, p. 1049-1057, 2006. DOI: 10.1016/j.aap.2006.01.006.

YILDIZ, K.; ATEŞ, A. D. ‘Evaluation of level crossing accident factors by logistic regression method: a case study’, *Iranian Journal of Science and Technology - Transactions of Civil Engineering*. Springer International Publishing, Apr. 2020. DOI: 10.1007/s40996-020-00367-z.

CASSIANO BASTOS MOROZ
Mestrando em Ciências da Geo-Informação e Observação da Terra. Engenheiro Civil. *E-mail*: c.bastosmoroz@student.utwente.nl.

JORGE TIAGO BASTOS
Doutor em Engenharia de Transportes. Engenheiro Civil. *E-mail*: jtbastos@ufpr.br.

TATIANA MARIA CECY GADDA
Doutora em Ciências Ambientais Humanas e da Terra. Arquiteta e Urbanista. *E-mail*: tatianagadda@utfpr.edu.br.