Assessing the impacts of enriched information on crash prediction performance

Peer-reviewed author version

MARTINS GOMES, Monique; PIRDAVANI, Ali; BRIJS, Tom & Souza Pitombo, Cira (2019) Assessing the impacts of enriched information on crash prediction performance. In: Accident analysis and prevention, 122, p. 162-171.

DOI: 10.1016/j.aap.2018.10.004
Handle: http://hdl.handle.net/1942/27270

Assessing the impacts of enriched information on crash prediction performance

Monique Martins Gomes[1], Ali Pirdavani[2], Tom Brijs[3], Cira Pitombo[4]

[1]UHasselt, Transportation Research Institute (IMOB), Agoralaan, 3590 Diepenbeek, Belgium, and São Carlos School of Engineering, Univ. of São Paulo, Av. Trabalhador São Carlense, 400, São Carlos, Brasil (corresponding author). E-mail: moniqmartins@hotmail.com

[2]UHasselt, Faculty of Engineering Technology, Agoralaan, 3590 Diepenbeek, Belgium. E-mail: ali.pirdavani@uhasselt.be

[3]UHasselt, Transportation Research Institute, Agoralaan, 3590 Diepenbeek, Belgium. E-mail: tom.brijs@uhasselt.be

[4]São Carlos School of Engineering, Univ. of São Paulo, Av. Trabalhador São Carlense, 400, São Carlos, Brasil. E-mail: cirapitombo@gmail.com

Abstract:

While high road safety performing countries base their effective strategies on reliable data, in developing countries the unavailability of essential information makes this task challenging. As a result, this drawback has led researchers and planners to face dilemmas of "doing nothing" or "doing ill", therefore restricting models to data availability, often limited to socio-economic and demographic variables. Taking this into account, this study aims to demonstrate the potential improvements in spatial crash prediction model performance by enhancing the explanatory variables and modelling casualties as a function of a more comprehensive dataset, especially with an appropriate exposure variable. This includes experimental work, where models based on available information from São Paulo, Brazil, and Flanders, the Dutch speaking area of Belgium, are developed and compared with each other. Prediction models are developed within the framework of Geographically Weighted Regression with the Poisson distribution of errors. Moreover, casualties and fatalities as the response variables in the models developed for Flanders and São Paulo, respectively, are divided into two sets based on the transport mode, called active (i.e., pedestrians and cyclists) and motorized transport (i.e., motorized vehicle occupants). In order to assess the impacts of the enriched information on model performance, casualties are firstly associated with all available variables for São Paulo and the corresponding ones for Flanders. In the next step, prediction models are developed only for Flanders considering all the available information in the Flemish dataset. Findings showed that by adding the supplementary data, reductions of 20% and 25% for motorized transport, and 25% and 35% for active transport resulted in AICc and MSPE, respectively. Considering the practical aspects, results could help identify hotspots and relate most influential factors, suggesting sites and data, which should be prioritized in future local investigations. Besides minimizing costs with data collection, it could help policy makers to identify, implement and enforce appropriate countermeasures.

Keywords: Crash prediction models, Geographically Weighted Regression, Road safety, Enriched data

1. Introduction

Despite efforts to improve road safety, an estimated 1.25 million victims of road crashes worldwide still die every year. Developing countries, which have low and middle incomes, account for 90 percent of this number (WHO, 2015), which is likely to rise even more if proper safety countermeasures and investments are not made.

Among the countries attempting to prevent road fatalities, Brazil alone is responsible for up to fifty thousand deaths and five hundred thousand injured people every year. They are casualties of over one million crashes per year in the country (DATASUS, 2014; WHO, 2015). As in other developing countries, this problem has been attributed to an insufficient development of supportive road infrastructure, policy changes and enforcement which have not taken into account urban intensification and the steady increase in vehicle use. In spite of the growing awareness on the urgency to

reverse these trends and efforts toward road safety programs and campaigns, the country's performance remains below expectations leading to an exponential rise in the number of casualties.

Road safety has long been a priority in developed countries. In addition to ongoing efforts regarding the implementation and good practice of successful countermeasures involving infrastructure, vehicle and road user behavior, developed countries have invested a great amount of time and effort developing their road safety strategies at the planning level, for instance by collecting and making comprehensive sources of reliable data available. Specifically in Europe, these strategies at both safety-planning and operational levels have led to a steady reduction in the number of deaths, therefore allowing the European fatality rates to decrease far below the global average (9.3 per 100,000 population, relative to the global rate of 17.4) (WHO, 2015).

In this context, spatial Crash Prediction Models (CPM) are a critical component in terms of safety planning considering both prediction and impact analysis purposes. This argument is valid as CPMs enable the estimation of values while providing an insight of the spatially varying relationship between crashes and related factors. Hence, an appropriate set of potential explanatory variables is crucial. As suggested in the literature, it should include variables that have a major influence on the dependent variables in previous studies, which can be measured in a valid and reliable way, are not endogenous (Elvik, 2007), and above all, consider what people are exposed to that could result in a crash (Carroll, 1971; Chapman, 1973; Hauer, 1982; Hauer, 1995; Hauer et al, 1996; Stewart, 1998; Qin et al, 2004; de Guevara et al., 2004; Elvik, 2007), such as the fact that absence of an exposure variable can lead to biased results (Jovanis and Chang, 1986; Fristrøm et al., 1995; Miaou et al., 2003). In this respect, there are six major groups of influential factors, namely: human, vehicle related, road design, environmental, time and traffic related factors (Kononov, 2002; Valent et al., 2002; Yau, 2004; Shankar et al., 1995; Delen et al., 2006; Elvik, 2007).

Human factors are commonly associated to driver behavior, e.g., alcohol and drug use, negligent and careless vehicle operations, failure to properly use protection devices, using cell phones or texting while driving, fatigue, etc. (Petridou and Moustaki, 2000; Odgen, 1996; Redelmeier and Tibshirani, 1997; Movig et al., 2004). Vehicle related factors refer to the characteristics of the vehicle such as its safety design standards, i.e., active and passive vehicle safety systems (Harvey and Durbin, 1986; Robertson, 1996; Langley et al., 2000; Bédard et al., 2002; Richter et al., 2005). Traffic related volumes are commonly represented by the Average Annual Daily Traffic (AADT) or the Vehicle Miles Travelled (VMT), and are both commonly used as exposure variables in prediction models (Hauer, 1995; Zhou and Sisiopiku, 1997; Martin, 2002; Qin et al., 2004; Pei et al., 2012; Xu et al., 2012; Ahmed et al., 2012; Pirdavani et al., 2013a). Road design refers to the road geometry and roadside conditions, such as well-designed curves and grades, wide lanes, adequate sight distance, clearly visible striping, appropriate design speeds and road categorization, flared guardrails, roadsides free of obstacles,

well-located crash attenuation devices, and well-planned use of traffic signals (Miaou, 1994; Taylor et al., 2000; Amundsen and Ranes, 2000; Kloeden et al., 2001; Karlaftis and Golias, 2002; Nilsson, 2004; Aarts and van Schagen, 2006; Rengarasu et al., 2007). Environmental related factors are weather and light conditions, for example (Shankar et al., 1995; Andrey and Knapper, 2003; Golob and Recker, 2003; Ahmed et al., 2012; Brijs et. al., 2008). Finally, time factors are related to the season or month of the year, weekends or weekdays, or the time of crash occurrence (Doherty et.al, 1998; Qin et al., 2006; Hao et al., 2016).

In spite of the scientific and technological advances toward safety promotion, unfortunately there is a gap between research and practice. Especially in developing countries where data availability has been an issue, this has discouraged researchers and policy makers, as they often find themselves in situations where they have to choose between doing nothing or restricting CPM to data availability. However, even considering efforts made in associating crashes with the available explanatory variables in such circumstances, this drawback leads to modelling errors, and thus unreliable predictions. Besides not being statistically reliable, they fail in terms of impact analysis that could further help to implement appropriate safety countermeasures. One explanation could be the existence of omitted variable bias, which in particular plays an important role in CPM reliability, generating biased and inconsistent estimates and coefficient signs (Washington et al., 2010; Mitra and Washington, 2012).

In order to highlight the importance of a comprehensive set of explanatory variables within CPM, this study aims to assess the potential impacts of enriched information on model performance, scrutinizing their improvements in terms of statistical and practical contributions. This includes empirical work, where models based on crash-related available information from São Paulo, Brazil, and Flanders, the Dutch speaking area of Belgium, are developed and compared. CPMs are developed within the framework of Geographically Weighted Regression (GWR) with the Poisson distribution of errors (GWPR).

Subsequently, GWPR models followed by a sensitivity analysis allow us to identify the statistical contribution of all information that was entered in the prediction models. To the best of our knowledge, this practice has only been explored in terms of microscopic-level analysis, focusing on the influence of the Highway Safety Manual (HSM) data variables on safety predictions. Some approaches found in previous studies include the Fractional Factorial Method (Akgüngör and Yıldız, 2007), Boosted Regression Trees (BRT) (Saha et al., 2015) and the “change one-factor-at-a-time” approach, which is the most commonly used sensitivity method in literature (Alluri and Ogle, 2012; Findley et al., 2012; Jalayer and Zhou, 2013, Williamson et al., 2015). In the present study, this investigation is conducted by analytically adding each variable to the prediction models, altering the other ones and evaluating the statistical contribution of each variable one at a time and their interactions, thus accounting for simultaneous variation of the input variables.

2. Study area and data

Spatial CPMs were developed based on available information of crashes, road networks and socio-economic and demographic variables from the state of São Paulo and Flanders. In the models developed for São Paulo, this information was geographically aggregated to the centroids of 644 municipalities out of 645 that are comprised by the state. São Paulo city itself was not included in the analyses, given its atypical values, which are far higher than the ones for other cities. Flemish models were produced at a zonal level, comprising 2,198 Traffic Analysis Zones (TAZs). The average size of a TAZ is 6.09 square kilometers with a standard deviation of 4.78 kilometers, and an average number of inhabitants equal to 2,416 persons.

For both regions, casualties and fatalities as the response variables in the models developed for Flanders and São Paulo, respectively, were divided into two sets based on the transport mode, called active transport and motorized transport. Casualties/fatalities for active transport included pedestrians and cyclists, while for motorized transport they were associated with motorized vehicle occupants. Moreover, records from a period of three years were used to produce the dependent variable for São Paulo (2009-2011) and Flanders (2010-2012). In the Brazilian models, this information was gathered from the Mortality Information System (*Sistema de informações de Mortalidade* – SIM), which is a public source created by DATASUS (2014). For the Flemish models, this and the remaining information were provided by the Ministry of Mobility and Public Works (MOBIEL VLAANDEREN).

Other socio-economic and demographic information from São Paulo was gathered from the last census index of 2010, made available by the Brazilian Institute for Geography and Statistics (IBGE, 2014). Given the limitations to obtain road feature information in Brazil, we included a "link length" as a proxy variable of the road network. To this end, we used the available road network available in OPENSTREETMAP. For both regions, the link length for motorized transport included information concerning trunk, highway, primary, secondary and tertiary roads, as well as the link length of residential and living streets. For active transport, the same road features were implemented, although highways and respective trunk length information were replaced by cycle paths and link length information of other roads designed only for pedestrians (e.g., footway), according to the OPENSTREETMAP classification.

Tables 1 and 2 show a list of variables of all the variables collected for São Paulo and Flanders, respectively, together with their definition and descriptive statistics. Variables that were included in the final CPM developed for both regions of the study are marked in bold in the tables.

Table 1: Descriptive statistics of variables collected for São Paulo

	Variable	Description	Average	Min	Max	SD*
Fatality	**Active transport**	**Total number of fatalities of active transport mode users over 3 years**	**7.34**	**0**	**295**	**23.428**
	Motorized transport	**Total number of fatalities of motorized transport mode users over 3 years**	**11.88**	**0**	**317**	**28.716**
Network	**Link length of active transport**	**Total link length of active transport in a municipality (km)**	**141.13**	**5.04**	**2626.15**	**210.95**
	Link length of motorized transport	**Total link length of motorized transport in a municipality (km)**	**153.04**	**5.29**	**2831.03**	**227.25**
	Area	**Total surface area in the municipality (km^2)**	**383.03**	**5**	**1977**	**317.07**
Socio-economic and demographic	**Population**	**Total number of inhabitants in the municipality**	**46597.35**	**805**	**1221979**	**108465.83**
	Male population	Total number of male inhabitants in the municipality	22902.55	422	595043	52538.13
	Female population	Total number of female inhabitants in the municipality	23694.81	383	626936	55938.55
	Population density	Total population per square kilometers per municipality	291.13	3.73	12519.10	1166.18
	AAGR	Average Annual Growth Rate 2000-2010 (%) in the municipality	1.03	-2.15	10.92	1.25
	Percentage male population	Percentage of male inhabitants in the municipality	50.52	45.76	81.09	2.52
	Percentage female population	Percentage of female inhabitants in the municipality	49.48	18.91	54.24	2.52
	Percentage proportion population	Rate between the number of men and woman in the municipality	102.97	84.36	428.86	17.88
	Urban population	Total number of inhabitants in the urban zone of the municipality	44150.48	627	1221979	107468.51
	Rural population	Total number of inhabitants in the rural zone of the municipality	2446.88	0	46284	3609.38
	HDI	Human Development Index	0.739	0.639	0.862	0.032
	GNP	Gross National Product	22501.11	7131.54	287646.17	18418.14
	Employed people	Total number of inhabitants with income	12678.37	155	405980	35725.41
	Occupied people	Total number of inhabitants who perform some activity (with income or not)	14931.77	211	471267	41144.02
Vehicle fleet	Motorcycle	Total fleet of motorcycles and tricycles	4744.68	24	100831	10938.16
	Microbus	Total fleet of microbuses	90.76	0	3544	264.87
	Car	Total fleet of cars	13536.09	133	487044	38052.31
	Truck	Total fleet of trucks	705.21	11	18144	1544.29
	Bus	Total fleet of buses	135.84	3	4445	330.34
	Total number of vehicles	Total number of vehicles	19212.58	220	612097	50296.09
Fuel consump.**	Gasoline	Total gasoline consumption	7961187.11	0	256246033	21723939.41
	Diesel oil	Total diesel oil consumption	15343179.63	0	295769873	32673917.02
	Fuel oil	Total fuel oil consumption	822438.64	0	44127640	3078410.70
	GLM	Total liquefied petroleum gas consumption	2304087.98	0	62823861	5948082.76
	Ethanol	Total ethanol consumption	9746540.07	0	342168947	25378940.38

*SD: Standard deviation; **Fuel consumption in liters

138 **Table 2: Descriptive statistics of variables collected for Flanders**

	Variable	Description	Average	Min	Max	SD*
Crash	**Active transport**	**Total number of casualties of active transport mode users observed in a TAZ over 3 years**	**15.04**	**0**	**298**	**25.06**
	Motorized transport	**Total number of casualties of motorized transport mode users observed in a TAZ over 3 years**	**45.36**	**0**	**500**	**53.83**
Network	**Capacity**	**Hourly average capacity of links in a TAZ (Passenger car per direction/h)**	**1790.10**	**1200**	**7348**	**554.60**
	Link length of active transport	**Total link length of active transport in a TAZ (km)**	**14.85**	**0**	**88**	**10.31**
	Link length of motorized transport	**Total link length of motorized transport in a TAZ (km)**	**15.87**	**0.39**	**87.95**	**10.80**
	Intersection density	**Number of intersections per square kilometer**	**1.75**	**0**	**50.63**	**3.37**
	Speed	Average speed limit in a TAZ (km/h)	69.40	31	120	10.91
	Area	Total surface area of a TAZ (km^2)	6.09	0	45	4.78
	Link density	Total link length in a TAZ (km^2)	3.37	0	20.44	2.41
	Intersection	Total number of intersection in a TAZ	5.80	0	40	5.90
	Highway	Presence of a highway in a TAZ, described as: "No" represented by 0 and "Yes" by 1		0	1	
	Urban	Is the TAZ in the urban area? "No" represented by 0 and "Yes" by 1		0	1	
	Suburban	Is the TAZ in the suburban area? "No" represented by 0 and "Yes" by 1		0	1	
Exposure	**Number of trips of active transport**	**Average daily number of trips originating/destined from/to a TAZ involving active mode**	**1103.40**	**0**	**8630**	**1316.12**
	Number of trips of motorized transport	**Average daily number of trips originating/destined from/to a TAZ involving motorized transport**	**2750.09**	**0**	**22650**	**2642.17**
	Vehicle Kilometers travelled (VKT) - Highway	**Total VKT on highways in a TAZ**	**27471.82**	**0**	**946153**	**84669.53**
	VKT - Other roads	**Total VKT on roads other than highways in a TAZ**	**26662.85**	**0**	**303238**	**28133.04**
Socio-economic and demographic	**Car ownership**	**Car ownership per household in a TAZ**	**1.13**	**0**	**14.00**	**0.47**
	School children	**Total number of children living in a TAZ that attend some school**	**364.09**	**0**	**92.45**	**772.59**
	Population	**Total number of inhabitants in a TAZ**	**2614.53**	**0**	**15803**	**2582.60**
	Households	Total number of households in a TAZ	1091.15	0	8062	1177.90
	Employees	Total number of employed people in a TAZ	888.73	0	16286	1575.31
	Income level	Average income of residents in a TAZ described as below:				
		"Monthly salary less than 2249 euro" represented by 0		0	1	
		"Monthly salary more than 2250 euro" represented by 1		0	1	

139 *SD: Standard deviation

3. Methodology

Geographically Weighted Regression

The spatially varying impacts of different risk factors across the study areas were explored within the framework of the local modelling approach, Geographically Weighted Regression (GWR) (see Brunsdon et al., 1996, Fotheringham et al., 1996, Fotheringham et al., 1997; Fotheringham et al., 2002) using the GWR 4.0 software package (Nakaya et al., 2005). Given that the number of casualties and fatalities as the response variables were the count data with discrete and non-negative integer values, GWR models were performed using the Poisson distribution error (GWPR).

Developed by Fotheringham and Brunsdon (Fotheringham et al., 2002), GWR models intend to address the non-stationary relationship between variables found in Generalized Linear Models (GLM). These models capture this spatial variation by fitting a regression model at each sample point. The result of this process is a set of local spatial parameters, described by Equation (1).

$$ln[E(C)(l_i)] = ln\big(\beta_0(l_i)\big) + \beta_1(l_i)ln(Exposure) + \beta_2(l_i)x_1 + \cdots + \beta_n(l_i)x_n \quad (1)$$

Where E(C) is the expected crash frequency, β_0, β_1, β_2 and β_n are model parameters for determined location l_i, *Exposure* is the exposure variable, and x_1 and x_n correspond to other explanatory variables.

One of the assumptions behind GWR is motivated by the First Law of Geography from Tobler (1970), which argues that "everything is related to everything else but closer things are more related to each other". The closer the observed data is from the location of the parameter to be estimated, the greater the influence on the estimation of β at location i compared to those that are far from it. Hence, this influence is determined based on geographic weights, which are assigned in function of all neighboring observations using a kernel function (Fotheringham et al., 2002), e.g., Gaussian (Equation 2) and bi-square (Equation 3), which are the two most common choices of weighting schemes (Hadayeghi et al., 2010).

Gaussian function:

$$W_{ij} = e^{-0.5(\frac{w_{ij}}{b})^2} \quad (2)$$

Bi-square function:

$$W_{ij} = \begin{cases} (1 - (\frac{d_{ij}}{b})^2)^2 & \text{if } d_{ij} < b \\ 0 & otherwise \end{cases} \quad (3)$$

Where W_{ij} is the measure of contribution of location *j* when calibrating the model for location *i*. d_{ij} is the Euclidian distance between locations *i* and *j* and *b* is the bandwidth size defined by a distance metric measure.

In GWR, the bandwidth controls the size of the kernel (number of observations around each data point) and the rate at which weights decay with increasing distances. Thus, similar to the weighting scheme, the choice of the bandwidth size plays an important role in the performance of the GWR models, as it involves a trade-off between bias and variance. The size of the bandwidth is optimized either by distance (fixed kernel), or by the number of neighboring observations (adaptive kernel) (Fotheringham et al., 2002; Guo et al., 2008; Hadayeghi et al., 2010). In this study, GWR was performed using the bi-square form with an adaptive bandwidth, such that the bandwidth varies according to the data density, and the number of areas included in the kernel is kept constant.

Within different approaches to select the optimal bandwidth, minimizing cross validation (CV) or the corrected Akaike Information Criterion (AICc) are the most widely used (Pirdavani et al., 2014). However, while the former is given by the difference between observed and estimated values, the latter, additionally to the statistical Goodness-of-fit, rewards the complexity of the model, by imposing a penalty for increasing the number of estimated parameters (Fotheringham et al., 2002), expressed by the formulation in Equation (4).

$$AICc = D(b) + 2K(b) + 2\,\frac{K(b)(K(b)+1)}{n-K(b)-1} \quad (4)$$

Where D and k denote the deviance and the effective number of parameters in the model with bandwidth *b*, respectively. Moreover, n denotes the number of observations. In this study, the bandwidth is calculated by means of AICc.

Methodological procedure

In order to demonstrate the potential improvements in the performance of spatial prediction models by enhancing the potential explanatory variables, the methodological procedure was divided into two main stages.

In the first stage, GWPR models were developed for São Paulo and Flanders, by only taking into account the same explanatory variables available in both datasets. Given the limitations of the Brazilian dataset, the results of this stage would reveal the best we could do with the available information of São Paulo, while there would be plenty of room yet to improve the Flemish models. In the second stage, and in order to highlight the importance of having data which is as complete as possible, GWPR models were developed for Flanders only, by considering all available variables in the Flemish dataset.

At both stages, a multicollinearity test was conducted prior to the modelling steps, enabling us to select the most significant variables to compose the final models. As a common practice, the Variance Inflation Factor (VIF) was used to quantify how much the variance of the estimated regression coefficients increased if predictors were correlated. As a common rule of thumb, 10 was defined as a cut off value, meaning that if VIF was higher than 10, then multicollinearity was high (Kutner et al., 2004) and, therefore, these variables should not be present in the model simultaneously. Subsequently, at the end of the first stage, models, for which we used the minimum data, were developed which were called basic models.

In the second stage, the affluence of available information in the Flemish dataset enabled us to develop various distinct models, and choose the one with the best overall fit. This exercise was conducted by having the intercept term as our starting point and analytically combining variables with a VIF lower than 10. Hereafter, due to the greater complexity of the GWR estimation procedure that conceivably causes interrelationships among local coefficient estimates when there is no collinearity among the explanatory variables (Hadayeghi et al., 2010; Pirdavani et al., 2013b), at this stage, evidence of multicollinearity among the produced local coefficient estimates was also observed. Hence, among all developed CPMs, the best-fitted one was selected as it met the criteria of non-multicollinearity among variables and produced local coefficient estimates, and subsequently based on the lowest AICc value. As a common rule-of-thumb, the difference between the models was considered significant when the difference of AICc values between two models was higher than 4 (Charlton and Fotheringham, 2009). At the end of this stage, two models were developed for Flanders, which were called improved models.

Finally, the performance of the improved models was compared with the basic models by means of AICc, Mean Squared Prediction Error (MSPE), and the Pearson Correlation Coefficient (PCC).

Sensitivity Analysis

Flemish GWPR models followed by a sensitivity analysis helped us identify the statistical contribution of each variable in the casualty prediction. In previous studies, this practice was commonly associated with micro-level analysis,

focusing on the influence of the HSM data variables on safety predictions. This exercise was held by altering the value of one predictor variable at its maximum, minimum, and/or average, hence estimating the change in output relative to the output generated from using the actual values of the variable. Therefore, the most influential variables were identified as those that produced meaningful changes in the predicted values for the frequency of crashes (Saha et al., 2015).

In our study, this investigation was conducted by analytically adding each variable to the prediction models while altering the other ones, and evaluating their statistical contribution by themselves and by their interactions, thus accounting for simultaneous variations of the input variables. To this end, the intercept term was used as a starting point. Hence, explanatory variables were analytically added to the prediction models and ranked according to their contribution in terms of the model's performance. This contribution was measured by means of the AICc variations (%), where the larger the reduction in AICc by the inclusion of a variable, the greater its contribution on the model performance. Subsequently, this process was repeated with the remaining variables, but taking into account their interactions. Thereupon, variables were tabulated according to their relative percentage of influence on the models in relation to the intercept term, namely Relative I, and in relation to its previously best fitted model composition, namely Relative II.

4. Results

In the next sections, the results of the modelling practices for both case studies will be explained and discussed. However, due to the limit of space in this paper, maps of significance and coefficient estimates are limited to motorized transport in Flanders. Moreover, although it is not the main aim of this paper, we will briefly discuss the effect sizes in relation to prior literature.

Results of modelling – Stage 1, Basic Models

In spite of a great amount of available information in Brazil, most of it was limited to socio-economic and demographic variables. As a result of this, most of the pieces of information collected were found to be correlated with each other, therefore presenting high VIF values. Hence, produced basic models were limited to information concerning the link length and population only. At this stage, the population was used as the exposure variable in its Natural Logarithm (ln) form.

Table 3 shows the local parameter estimates of the basic GWPR models, for both dependent variables in São Paulo and Flanders. This information is described by five number summaries: minimum, maximum (lower quartile, median quartile and upper quartile), tabulated in this format and sequence.

Table 3: Local parameter estimates - Basic Models

	Active Transport		Motorized Transport	
Parameters	**Brazil**	**Flanders**	**Brazil**	**Flanders**
Intercept	-24.094, 15.716 (-13.546, -10.626, -8.081)	-8.890, 11.749 (0.153, 1.813, 2.966)	-16.228, 0.031 (-8.576, -6.837, -4.896)	-5.981, 8.582 (1.894, 3.336, 4.246)
Ln Population	-2.249, 2.463 (0.874, 1.155, 1.438)	-1.646, 1.668 (-0.046, 0.130, 0.365)	0.019, 1.898 (0.672, 0.871, 1.053)	-1.252, 1.307 (-0.063, 0.077, 0.267)
Link length	-0.006, 0.038 (-7.8e-04, 1.5e-04, 1.2e-03)	-0.264, 0.130 (-0.038, -0.012, 0.010)	-0.009, 0.010 (2.2e-04, 4.4e-04, 1.4e-03)	-0.149, 0.139 (-0.031, -0.008, 0.267)

Figure 1 shows maps of the local coefficient estimates as well as their significance at 0.05 level, for motorized transport in Flanders. In order to determine where relationships were significant (in blue) and where they were not (in brown), we computed the t-statistics. Thus, t-statistics between -1.96 and 1.96 were considered insignificant at 0.05 level, and values beyond this scale were considered significant at 0.05 level. Subsequently, Figure 2 shows maps with observed and predicted number of casualties.

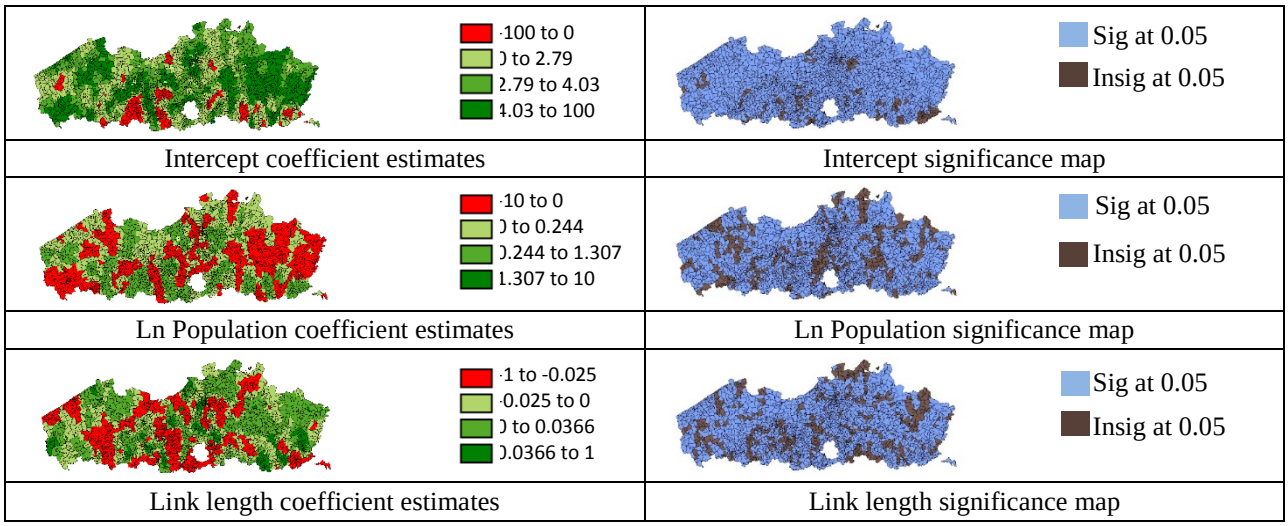

Figure 1: Local coefficient estimates and significance maps (Motorized transport - Basic model - Flanders).

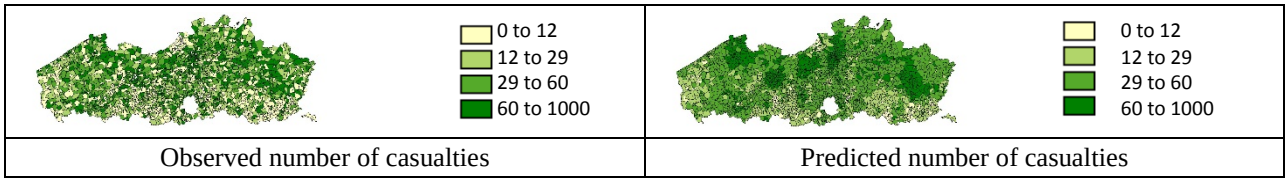

Figure 2: Observed and predicted number of casualties (Motorized transport - Basic model - Flanders).

Results of model performance at this stage revealed a negative correlation between casualties and both explanatory variables in a large number of TAZs, which in turn were significant at 0.05 confidence level. Given their rather direct association with exposure, we would expect that both explanatory variables were likely to have a positive correlation with casualties in most TAZs of this study, as also found in Wang et al. (2009 and Pirdavani et al. (2013b), for example.

One explanation for these counterintuitive signs could be that some variables, at some locations, are locally correlated while no global multicollinearity is observed among them (Guo et al., 2008; Hadayeghi et al., 2010; Pirdavani et al., 2014). Such local correlation has been attributed as the major reason for problems with counterintuitive signs. In order to address this issue, at the second stage, we excluded variables at which high VIF values were found among the produced coefficient estimates. The limitation of variables within the basic models restrained us from conducting this exercise at the first stage. This drawback highlights the importance of a more diverse set of explanatory variables. Evidence in favor of this assumption is shown in the second stage.

Such counterintuitive signs could also be a response to the omission of important variables, which leads to omitted variable bias. Although not thoroughly explored in this study, one could assume the correlation of link length and other road features or exposure variables that were omitted, therefore producing bias. This argument is valid as the exclusion of essential variables, especially an exposure variable, could systematically invalidate further conclusions that could be derived from the results (Washington et al., 2010; Mitra and Washington, 2012). A more in-depth investigation concerning this problem could help to provide a better insight of the direction of these effects, which in this study, remains speculative.

Results of modelling – Stage 2, Improved Models

The Flemish dataset, in addition to significant information related to socio-economic, socio-demographic and road networks (i.e., income level, speed, capacity, number of links, links and intersection density, presence of highways, urbanization degree, to name a few) provided foremost diverse and suitable exposure variables, i.e., number of trips (NOTs), vehicles flow and VKT. This enabled us to produce different models with different combinations of variables, and choose the best fitted one.

After carrying out the VIF tests among variables and produced coefficient estimates, the final improved models for active and motorized transport modes comprised the following information: NOTs, children attending school (school children), road capacity (capacity), intersection density, car ownership, and VKT which was used as our exposure variable. The respective coefficient estimates found for each explanatory variable are presented in Table 4 in the five number summary format, and are followed in Figure 3 by their local coefficient estimates and significance maps. Subsequently, Figure 4 presents the obtained maps for the observed and predicted number of casualties.

Table 4: Local parameter estimates - Improved Models (Flanders)

Parameters	Active Transport	Motorized Transport
Intercept	-14.835, 13.933 (-0.928, 1.580, 4.068)	-7.604, 11.041 (1.585, 3.312, 4.973)
NOTs	-0.002, 0.005 (-2.78-04, 4.8e-05, 4.4-04)	-9.4e-04, 0.001 (-1.04e-04, 2e-05, 1.23e-04)
School children	-0.007, 0.003 (-9.16-04, -2.33e-04, 3.4e-04)	-0.004, 0.003 (-5.74e-04, -1.43e-04, 2.33e-04)
Capacity	-0.008, 0.007 (-3.46e-04, 7.3e-05, 5.47e-04)	-0.007, 0.005 (-3.89-04, 1.5e-05, 4.5-04)
Intersection density	-4.183, 1.686 (-0.139, 0.023, 0.152)	-1.441, 1.171 (-0.086, 0.025, 0.131)
Ln VKT	-0.653, 1.486 (-0.054, 0.128, 0.306)	-0.666, 0.888 (-0.092, 0.076, 0.246)
Car Ownership	-7.378, 6.584 (1.649, -0.375, 0.445)	-6.763, 4.722 (-1.084, -0.128, 0.456)

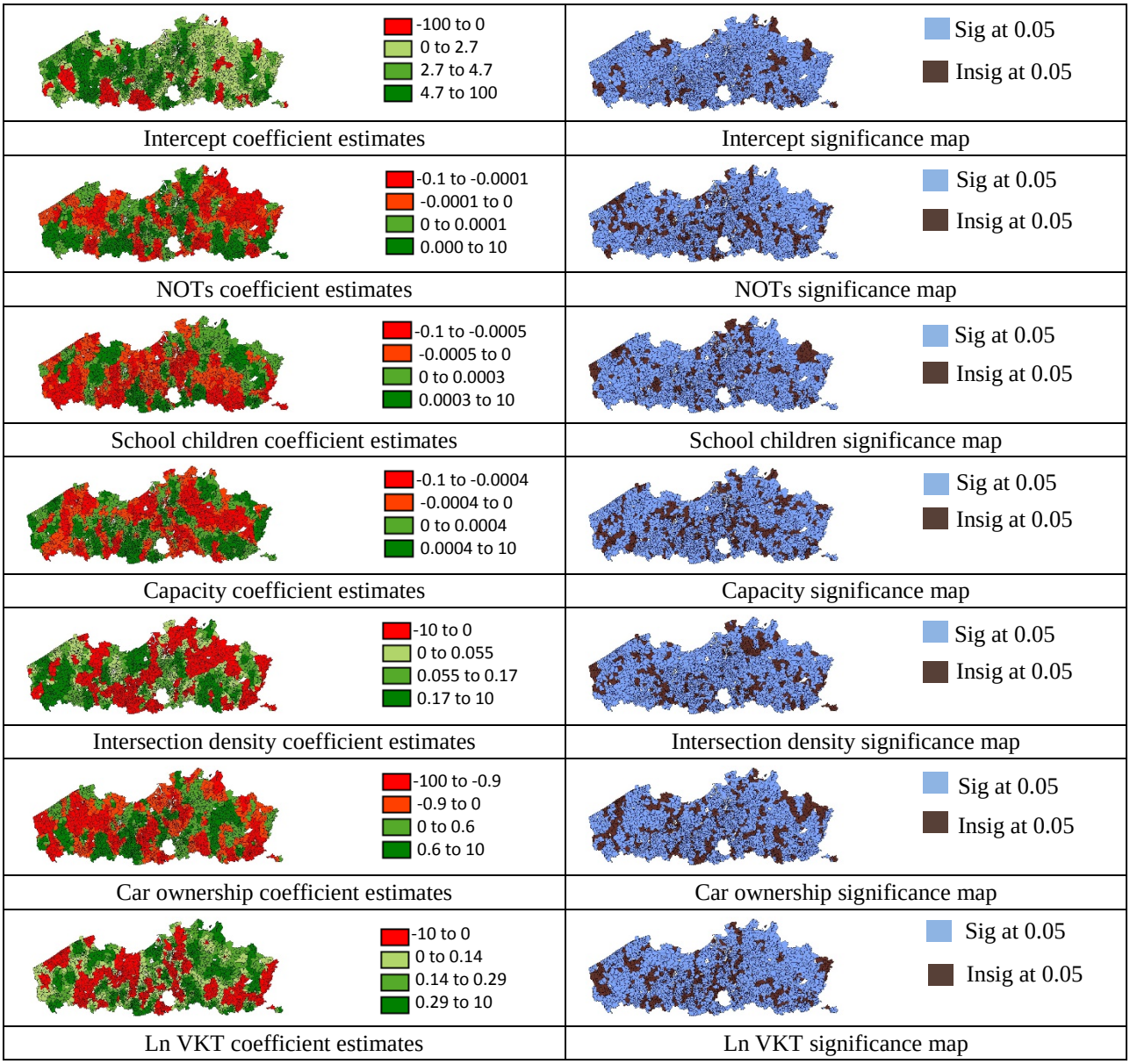

Figure 3: Local coefficient estimates and significance maps (Motorized transport - Improved model - Flanders).

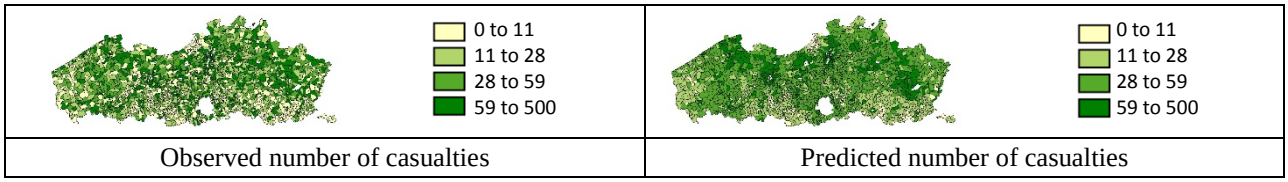

Figure 4: Observed and predicted number of casualties (Motorized Transport - Improved model - Flanders).

In general, results at this stage revealed a better performance in terms of overall model fit (see Table 5). From a statistical point of view, the improved models outperformed the basic models for both dependent variables. In the model developed for motorized transport, reductions of approximately 20% and 25% were observed compared to the basic model, for AICc and MSPE respectively. Likewise for the active mode, 25% and 35% reductions were obtained for AICc and MSPE, respectively.

Table 5: Comparison of parameters between basic and improved models (Flanders)

Parameters	Active Transport		Motorized Transport	
	Basic	**Improved**	**Basic**	**Improved**
GWPR AICc	29345.571	22553.607	60993.481	49525.611
Global AICc	50570.229	45486.734	102673.889	94634.071
MSPE	384.87	261.17	1786.66	1354.12
PCC	0.629	0.771	0.626	0.738

Moreover, results of a more diverse dataset, especially including an exposure variable, enabled the development of coherent coefficient estimates and signs. In this study, positive associations with casualties were found in most subzones, for intersection density, capacity, NOTs and VKT (likewise in Hadayeghi et al., 2003; Aguero -Valverde and Jovanis, 2006; Pirdavani et al., 2013a; Pirdavani et al., 2013b; Shariat-Mohaymany et al., 2015; Xu et al., 2017). On the contrary, casualties were found to have a negative correlation with school children and car ownership, in a large number of TAZs. This negative association with other social standing variables, i.e., income level, has been found in other previous studies (see Li et al., 2013; Pirdavani et al., 2013a; Pirdavani et al., 2013b; Pirdavani et al., 2014; Pirdavani et al., 2016), meaning that less casualties are expected to occur than in more affluent areas (in this study, where car ownership is higher). Given the negative association of casualties and school children, we would assume that speed limit and human factors associated to the driver's behavior, might have an influence.

In order to corroborate these assumptions and enable us to better understand the interactions between variables, such a macro-level analysis could be used as a basis for local investigations. This could help to enforce appropriate countermeasures, especially in areas where higher estimates were found. For instance, at subzones where casualties were found to have a positive association with school children, micro-level analysis could suggest changes in the speed limits

or signalizing intersections. This could be identified as the major contribution of having a more complete and diverse dataset. In spite of more reliable models, they would allow policy makers to prioritize subzones, and depending on the targets, specific TAZs could be used to investigate the interaction between variables, both within and outside of the models.

Subsequently, GWPR improved models followed by a step-wise approach helped us identify the statistical contribution of each variable in the crash prediction performance. The improvements in model performance by means of the percentage reductions found on AICc, for motorized transport, are shown in Figure 5. We considered the percentage difference by means of AICc in relation to the intercept term, namely Relative I, and in relation to its previously best fitted model composition, namely Relative II. In this illustration, we included part of the sensitivity analysis, for which capacity, for instance, showed the highest reduction by means of AICc in relation to the other variables, becoming the chosen variable with the intercept term. Subsequently, the improvements in model performance for active transport are shown in Table 6.

Model No.	Variables	AICc	Relative I (%)	Relative II (%)
1	**Intercept**	69996.71	-	-
2	Model 1 + **capacity**	65149.59	-6.92	-6.92
3	Model 2 + **VKT**	61120.22	-12.68	-6.18
4	Model 3 + **school children**	57616.19	-17.69	-5.73
5	Model 4 + **intersection density**	53909.95	-22.98	-6.43
6	Model 5 + **car ownership**	51521.46	-26.39	-4.43
7	Model 6 + **NOTs**	49525.61	-29.25	-3.87

Variables	AICc	%
Intercept	69996.71	-
Intercept + **capacity**	65149.59	-6.92
Intercept + **VKT**	65345.23	-6.65
Intercept + **intersection density**	57616.19	-5.61
Intercept + **car ownership**	53909.95	-5.57
Intercept + **school children**	51521.46	-5.56
Intercept + **NOTs**	67022.43	-4.25

Figure 5: Ranking of variables according to their contribution in the model for motorized transport.

Table 6: Ranking of variables according to their contribution in the model for active transport

Model No.	Variables	AICc	Relative I (%)	Relative II (%)
1	**Intercept**	33593.51	-	-
2	Model 1 + **capacity**	31360.08	-6.65	-6.65
3	Model 2 + **car ownership**	29132.76	-13.28	-7.10
4	Model 3 + **intersection density**	27310.70	-18.70	-6.25
5	Model 4 + **school children**	25276.7	-24.76	-7.45
6	Model 5 + **NOTs**	23876.5	-28.93	-5.54
7	Model 6 + **VKT**	22553.61	-32.86	-5.54

The results show that road capacity has the highest statistical contribution in the performance of CPM for active and motorized transport modes, suggesting that this information has priority over others. Secondly, VKT statistically contributes more to motorized transport models, while car ownership contributes more to the active transport models, and so on. This practice could be useful as it would help policy makers prioritize data collection, for instance by targeting variables that add higher statistical contributions to one specific travel mode or both, thus reducing costs with data collection.

5. Discussion and conclusions

The difficulty in obtaining crash-related information in Brazil, and its consequences in terms of model performance and development of potential studies that could help understand the crash phenomena and enforcement of appropriate countermeasures were the major reasons for carrying out this study. Although some data can be found in different road departments, police, health and census online sources (i.e., DATASUS, IBGE, DENATRAN), there is no link between their databases, and none of them are able to provide a full and effective data source with regard to accidents and fatalities in the country. Therefore, the absence of a comprehensive and complete database hampered the evaluation and follow-up of national road safety programs, as well as the development of studies that could contribute to national goals toward road safety.

Particularly concerning the explanatory variables collected in this study, most available information was restricted to socio-economic variables. In spite of their merits, socio-economic variables are often highly correlated with each other, and are not appropriate for safety-planning purposes, e.g., implementing safety countermeasures. Other developing countries have faced the same challenges, and this drawback has unfortunately led to the use of poor and unreliable CPMs to promote road safety in these countries.

In view of the above, this study aimed to demonstrate the potential improvements in the performance of spatial crash prediction models by means of a more diverse dataset, including some potential explanatory variables. To this end, benchmarking was carried out based on macro level CPMs developed with available fatality/crash-related information from São Paulo and Flanders. In contrast to developing countries, European countries such as Belgium and other high road safety performing countries have invested a great deal of time and money in obtaining crash-related information and make them available through public channels and to academia. This practice has led to outcomes such as new strategies and studies, and this trade-off has brought improvements in traffic safety by reducing the number of crashes, injuries and fatalities.

Whereas the Brazilian dataset in this study was mostly limited to socio-economic variables, the Flemish dataset, in addition to significant information related to socio-economic, socio-demographic and road networks, provided foremost diverse and suitable exposure variables. As a result of this, improved models revealed lower values of AICc and MSPE, for both dependent variables. Moreover, Flemish models at the second stage, presented a significant set of coefficient estimates together with suitable coefficient signs. One potential outcome of the resulting macro-level CPMs could be the identification of hotspots together with their major influence factors. Such results could be used as a reference to microscopic investigations, and the implementation of suitable safety countermeasures' enforcement in a long term transportation planning process.

Above all, this study addressed the strong dependence of CPMs on suitable and diverse input information, enabling these models to perform as a "powerful tool", as is usually found in literature. Crashes are caused by multiple factors that vary locally, and this complexity implies that ideally casualties are best predicted through a set of appropriate predictive variables, including at least one potential exposure variable. By modelling casualties in Flanders based on the equivalent available data in São Paulo, results were found not to be suitable. Apart from producing unreliable coefficient estimates, they would also not be useful for safety planning and practical aspects. In other words, despite the efforts to improve the statistical fit of the crash prediction models and associate crashes with the available explanatory variables in such circumstances, these models fail to comprehensively explain the road casualties and, therefore, diminish the ability of applying suitable safety countermeasures.

On the contrary, by modelling casualties based on the entire available data in Flanders, results revealed a better model overall fit for both active and motorized transport modes. This suggests that a more diverse set of appropriate explanatory variables, including a relevant exposure variable, is advantageous as it could address problems with counterintuitive signs and omitted variable bias. In the ideal scenario, it could help policy makers determine local appropriate countermeasures

toward safety promotion (e.g., by altering the speed limits and intensifying local speed enforcement, improving intersections, installing traffic management and control systems, implementing crosswalks, etc).

It is also worth mentioning that variables included in the Flemish models, besides having been found to be significant in previous studies as well as in the present one, are less expensive and more accessible than those used in microscopic analysis (e.g., driving data, braking and steering information or variables related to weather conditions). Moreover, they are just examples of other potential information that could be used to develop those predictive/descriptive models. Variables that are used in the models developed for Flanders could at most be interesting suggestions for extra data collection in Brazil, as other local variables could also play a significant role, other than those included in the Flemish models.

Last but not least, a sensitivity analysis was carried out allowing us to assess the statistical contribution of each variable in the prediction model performance. Especially for countries where crash data is limited, either because of the lack of financial resources or other imposed conditions, this practice could empower policy makers and responsible offices to prioritize data collection. For instance, results revealed that data regarding road capacity would signify a major statistical contribution for models, for both dependent variables. This is different from NOTs, for instance, which often have priority in data collection, but as revealed in this study, would not bring such a significant contribution to the Flemish models, neither for active, nor for motorized transport.

This investigation could have more value if similar analyses were carried out in different regions, based on their available information. The consolidation of the produced results would enable, for instance, the development of a solid benchmark, therefore, validating the priorities outlined in this study and helping to determine the importance of different variables to model performance, in different areas.

In addition to this, for further studies, we suggest a more in-depth investigation addressing problems, such as endogeneity and omitted variables to model performance, as they could help to verify the validity of the assumptions of this study. One possible investigation could be for instance, to perform micro-level analysis in the identified hot zones, therefore assessing model performance by adding and removing variables to the models.

ACKNOWLEDGMENTS

This research was supported by the Brazilian National Council for Scientific and Technological Development - CNPq.

References

Aarts, L. van Schagen, I., 2006. Driving speed and the risk of road crashes: a review. Accid. Anal. Prev. 38: p. 215-224.

Aguero-Valverde, J., Jovanis, P., 2006. Spatial Analysis of Fatal Injury Crashes in Pennsylvania. Accid. Anal. Prev. 38: p. 618-25.

Ahmed, M., Abdel-Aty, M., Yu, R., 2012. Assessment of the interaction between crash occurrence, mountainous freeway geometry, real-time weather and AVI traffic data. Transportation Research Record. 2280: p. 51–59.

Akgüngör, A. Yıldız, O, 2007. Sensitivity Analysis of an Accident Prediction Model by the 15 Fractional Factorial Method. . Accid. Anal. Prev. 39 (1), p. 63-68.

Allluri, P., Ogle, J., 2012. Effects of state-specific SPFs, AADT estimations, and overdispersion parameters on crash predictions using SafetyAnalyst. Proceedings of the 91st Annual Meeting of the Transportation Research Board, Washington, DC.

Amundsen, F. H., Ranes, G., 2000. Studies on traffic accidents in Norwegian road tunnels. Tunnelling and Underground Space Technology, 15, 3-11.

Andrey, J., Knapper, C.K., 2003.Weather and Transportation in Canada. Department of Geography Publication Series, no. 55. ISBN: 0-921083-65-3.

Bédard, M., Guyatt, G.H., Stones, M.J., Hirdes, J.P., 2002. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accid. Anal. Prev. 34: p. 717-727.

Brijs, T., Karlis D., Wets G., 2008. Studying the effect of weather conditions on daily crash counts using a discrete time-series model. Accid. Anal. Prev. 40: p. 1180 – 1190.

Brunsdon, C., Fotheringham, A.S., Charlton, M., 1996. Geographically weighted regression: a method for exploring spatial non-stationarity. Geographical Analysis, p. 281-298.

Carroll, P.S., 1971. Techniques for the use of driving exposure information in highway safety research. HSRI, University of Michigan.

Chapman, R.A., 1973. The concept of exposure. Accid. Anal. Prev. 5 (2), p. 95–110.

Charlton, M., Fotheringham, A.S., 2009. Geographically Weighted Regression:White Paper. National Centre for Geocomputation, National University ofIreland Maynooth.

DATASUS, 2014. Ministério da saúde - Departamento de informática do SUS - Sistema de Informações sobre Mortalidade (2010) - Estatísticas Vitais. Available in: <tab-net.datasus.gov.br>. Acessed in: 15 abr. 2014.

de Guevara, F.L., Washington, S.P., Oh, J., 2004. Forecasting crashes at the planninglevel: simultaneous negative binomial crash model applied in Tucson, Arizona.In: Transportation Research Record, Journal of the Transportation ResearchBoard, No. 1897. Transportation Research Board of the National Academies.Washington, D.C, p. 191–199.

Delen, D., Sharada, R Bessonov, M., 2006. Identifying Significant Predictors of Injury Severity in Traffic Accidents Using a Series of Artificial Neural Networks. Accid. Anal. Prev. 38: p. 434-444.

DENATRAN. Departamento Nacional de Transito. Available in: <http://www.denatran.gov.br/>.

Doherty, S.T., Andrey, J.C., MacGregor, C., 1998. The situational risks of young drivers: the influence of passenger, time of day and day of week on accident rates. Accid. Anal. Prev. 30 (1): p. 45–52.

Elvik, R., 2007. State of the art approaches to road accident black spot management and safety analysis of road networks. Institute of Transport Economics Norwegian Centre for Transport Research – rapport 883, Oslo.

Findley, D., Zegeer, C., Sundstrom, C., Hummer, J., Rasdorf, W., 2012. Applying the Highway Safety Manual to two-lane roa curves. J. Transp. Res. Forum 51 (3), 25-38.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 1996. The geography of parameter space: an investigation of spatial non-stationarity. International Journal of Geographical Information Systems. 10: p. 605-627.

Fotheringham, A.S., Charlton, M., Brunsdon, C., 1997. Two techniques for exploring nonstationarity in geographical data, Geographical Systems. 4: p. 59-82.

Fotheringham, A.S., Brunsdon, C., Charlton, M.E., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationship. Wiley: New York.

Fristrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., Thomsen, L.K., 1995. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accid. Anal. Prev. 27: p. 1–20.

Golob, T.F., Recker,W.W., 2003. Relationship among urban freeway accidents, traffic flow, weather, and lighting conditions. J. Transp. Eng. 129: p. 342–353.

Guo, L., Ma, Z., Zhang, L., 2008. Comparison of bandwidth selection in application of geographically weighted regression: a case study. Canadian Journal of Forest Research. 38: p. 2526 – 2534.

Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2003. Macrolevel accident prediction models for evaluating safety of urban transportation systems. Transportation Research Record: Journal of the Transportation Research Board. 1840: p. 87–95.

Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2010. Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. Accid. Anal. Prev. 42: p. 676 – 688.

Hao, W., Kamga, C., Wan, D., 2016. The effect of time of day on driver's injury severity at highway-rail grade crossings in the United States. Journal of Traffic and Transportation Engineering. 3 (I): p. 37-50.

Harvey, A.C., Durbin, J., 1986. The effects of seat belt legislation on british road casualties: a case study in structural time series modelling. Journal of the royal statistical society. Series A (general), 149: p. 187-227.

Hauer, E., 1982. Traffic conflicts and exposure. Accid. Anal. Prev. 14 (5), p. 359–364.

Hauer, E., 1995. On exposure and accident rate. Traffic Eng. Contr. 36 (3), p. 134–138.

Hauer, E., Ng, J.C.N., Lovell, J., 1996. Estimation of safety at signalized intersection. Transport. Res. Board 1285: p. 42–51.

IBGE, 2014. Instituto Brasileiro de Geografia e Estatística. Censo Demográfico, 2010. Available in: <www.censo2010.ibge.gov.br/resultados>. Acessed in: 17 jul. 2014.

Jalayer, M., Zhou, H., 2013. A sensitivity analysis of crash prediction models input in the Highway Safety Manual. Paper presented at the 2013 ITE Midwestern District Meeting, Milwaukee, WI.

Jovanis, P., Chang, H.L., 1986. Modeling the relationship of accidents to miles traveled. Transportation Res. Rec. 1068, p. 42–51.

Karlaftis, M. G., Golias, I., 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. Accid. Anal. Prev. 34: p. 357-365.

Kloeden, C. N., Ponte, G., McLean A. J., 2001. Travelling speed and the risk of crash involvement on rural roads. Department of Transport and Regional Services Australian Transport Safety Bureau, Report no. CR 204, ISSN 1445-4467.

Kononov, J., Janson, B., 2002. Diagnostic Methodology for the detection of safety problems at intersections, Proceedings of the Transportation Research Board, Washington D.C.

Kutner, M.H., Nachtsheim, C.J., Neter, J., 2004. Applied Linear Regression Models , 4th ed. McGraw-42 Hill.

Langley, J., Mullin, B., Jackson, R. Norton, R., 2000. Motorcycle engine size and risk of moderate to fatal injury from a motorcycle crash. Accid. Anal. Prev. 32: p. 659-663.

Li, Z., Wang, W., Liu, P., Bigham, J.M. Ragland, D.R., 2013. Using Geographically Weighted Poisson Regression for county-level crash modeling in California. Safety Science. 58: p. 89-97.

Martin, J.L., 2002. Relationship between crash rate and hourly traffic flow on interurban motorways. Accid. Anal. Prev. 34: p. 619-629.

Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections: poisson versus negative binomial regressions. Accid. Anal. Prev. 26: p. 471-482.

Miaou, S.P, Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping: a space–time modeling approach. J. Transportation Stat. 6 (1), p. 33–57.

Mitra, S., Washington, S., 2012. On the significance of omitted variables in intersection crash modeling. Accid. Anal. Prev. 49: p. 439-448.

MOBIEL VLAANDEREN. Mobiliteit en openbare werken. Available in: < http://www.mobielvlaanderen.be>.

Movig, K. L. L., Mathijssen, M.P.M., Nagel, P.H.A., Van Egmond, T., de Gier, J.J., Leufkens, H.G.M., Egberts, A.C.G., 2004. Psychoactive substance use and the risk of motor vehicle accidents. Accid. Anal. Prev. 36: p. 631-636.

Nakaya, T., Fotheringham, A.S., Brunsdon, C., Charlton, M., 2005. Geographically weighted Poisson regression for disease association mapping. Statistics in Medicine. 24: p. 2695–2717.

Nilsson, G., 2004. Traffic safety dimensions and the power model to describe the effect of speed on safety. Phd, lund institute of technology.

Odgen, K.W., 1996. Safer roads: a guide to road safety engineering, burlington, ashgate publishing company.

OpenStreetMap. Available in < https://www.openstreetmap.org>.

Pei, X., Wong, S.C., Sze, N.N., 2012. The roles of exposure and speed in road safety analysis. Accid. Anal. Prev. 48: p. 464-471.

Petridou, E., Moustaki, M., 2000. Human factors in the causation of road traffic crashes. European Journal of Epidemiology. 16: p. 819-826.

Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., Wets, G., 2013a. Evaluating the Road Safety Effects of a Fuel Cost Increase Measure by means of Zonal Crash Prediction Modeling. In: Accid. Anal. Prev. 50, p. 186-195.

Pirdavani, A., Brijs, T., Bellemans, T., Wets, G., 2013b. Spatial analysis of fatal and injury crashes in Flanders, Belgium: Application of geographically weighted regression technique. In: The 92th Annual Meeting of Transportation ResearchBoard, Washington, DC.

Pirdavani, A., Bellemans, T., Brijs, T., Wets, G., 2014. Application of geographically weighted regression technique in spatial analysis of fatal and injury crashes. Journal of Transportation Engineering.

Pirdavani, A., Daniels, S., van Vlierden, K., Brijs, K., Kochan, B., 2016. Socioeconomic and sociodemographic inequalities and their association with road traffic injuries. In: Journal of Transport & Health, 4, p. 152-161.

Qin, X., Ivan, J. N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accid. Anal. Prev. 36: p. 183-191.

Qin, X., Ivan, J.N., Ravishanker, N., 2006. Bayesian estimation of hourly exposure functions by crash type and time of day. Accid. Anal. Prev. 38 (6): p. 1071 – 1080.

Redelmeier, D.A., Tibshirani, R.J., 1997. Association between cellular-telephone calls and motor vehicle collisions. New england journal of medicine. 336: p. 453-458.

Rengarasu, T. M., Hagiwara, T., Hirasawa, M., 2007. Effects of road geometry and season on head-on and single-vehicle collisions on rural two lane roads in hokkaido, japan. Journal of the Eastern Asia Society for Transportation Studies, Vol. 7, p. 2860-2872.

Richter, M., Pape, H.C., Otte, D., Krettek, C., 2005. Improvements in passive car safety led to decreased injury severity – a comparison between the 1970s and 1990s. Injury. 36: p. 484-488.

Robertson, L.S., 1996. Reducing death on the road: the effects of minimum safety standards, publicized crash tests, seat belts, and alcohol. American journal of public health. 86: p. 31-34.

Saha, D.; Alluri, P.; Gan, A, 2015. Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees. Accid. Anal. Prev. 79: p. 133–144. Shankar, V., Mannering, F. Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accid. Anal. Prev. 27: p. 371-389.

Shankar, V., Mannering, F. Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accid. Anal. Prev. 27: p. 371-389.

Shariat-Mohaymany, A., Shahri, M., Mirbagheri, B., Matkan, A.A., 2015. Exploringspatial non-stationarity and varying relationships between crash data andrelated factors using geographically weighted Poisson regression. Trans. GIS 19 (2), 321–337.

Stewart, D.E., 1998. Methodological approaches for the estimation evaluation, interpretation and accuracy assessment of road travel 'basic risk', 'relative risk', and 'relative risk odds-ratio' performance measure. Indicators: a 'risk analysis and evaluation system model' for measuring, monitoring, comparing and evaluating the level(s) of safety on Canada's roads and highways. Transport Canada Publication No. TP 13238 E.

Taylor, M. C., Lynam, D.A., Baruya, A., 2000. The effects of drivers' speed on the frequency of road accidents. Transport Research Laboratory Report 421, ISSN 0968-4107.

Tobler, W., 1970. A computer movie simulating urban growth in the Detroit region. Economic Geography. 46: p. 234–40.

Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, S. Barbone, F. 2002. Risk factors for fatal road traffic accidents in udine, Italy. Accid. Anal. Prev. 34: p. 71-84.

Xu, C., Liu, P., Wang, W., Li, Z., 2012. Evaluation of the impacts of traffic states on crash risks on freeways. Accid. Anal. Prev. 47: p. 162– 171.

Xu, P., Huang, H., Dong, N., Wong, SC., 2017. Revisiting crash spatial heterogeneity: A bayesian spatially varying coefficients approach. Accid. Anal. Prev. 98: p. 330-337.

Yau, K.K.W., 2004. Risk factors affecting the severity of single vehicle traffic accidents in hong kong. Accid. Anal. Prev. 36: p. 333-340.

Wang, C., Quddus, M.A., Ison, S.G., 2009. Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. Accid. Anal. Prev. 41 (4): p. 798–808.

Washington, S., Karlaftis, M., Mannering, F., 2010. Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed. Chapman & Hall, Boca Raton.

Williamson, M., Jalayer, M., Zhou, H., Pour-Rouholamin, M., 2015. A Sensitivity Analysis of Crash Modification Factors of Access Management Techniques in Highway Safety Manual. Access Management Theories and Practices, 76.

WHO, 2015. World Health Organization. Global Status Report on Road Safety 2015. In: <http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/>, accessed in September 2017.