PERSPECTIVES ON THE DEVELOPMENT OF SAFETY PERFORMANCE MODELS FOR BRAZILIAN ROADS

Heloisa Barbosa[1]*, Flávio Cunto*[2]*, Bárbara Bezerra*[3]*, Christine Nodari*[4]*, Maria Alice Jacques*[5]

1 Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, Brazil
2 Universidade Federal do Ceará, Fortaleza, Brazil
3 Universidade Estadual Paulista, Bauru, Brazil
4 Universidade Federal do Rio Grande do Sul, 90035-190, Porto Alegre, Brazil
5 Universidade de Brasília, Brasília, Brazil
Email for correspondence: heloisa@etg.ufmg.br

ABSTRACT

This paper presents the modelling effort for developing SPM for urban intersections for five major Brazilian cities. The proposed methodology for calibrating SPM for Brazilian roads has been structured into four steps: 1. Identifying geometric and operational attributes for homogeneous groups (entities) such as signalized and unsignalized urban intersections; 2. Estimating the annual average daily traffic (AADT) for available intersection groups; 3. Obtaining crash data from jurisdictional accident data system; 4. SPM calibration for the jurisdictions involved in the task force. This initial SPM calibration effort has indicated a number of database structural challenges as well as research initiatives in order to allow a more practical use of this somewhat consolidated framework for safety assessment for the Brazilian urban environment, as the need for an integrated accident and traffic flow database and the need for a reliable and precise spatial crash location system. This research has showed the most significant independent variables to be used for calibrating SPM for Brazilian cities were the exposure variables AADT, segment length and, depending on the jurisdiction, other variables such as the number of lanes, central median and other were found to be significant. The SPM were obtained for three cities Fortaleza, Belo Horizonte and Brasília. The SPM developed for signalized intersections in Fortaleza and Belo Horizonte had the same structure and the most significant independent variables were AADT entering the intersection and numbers of lanes, also the coefficient of the best models were in the same range of values. In Brasilia due the sample size the signalized and unsignalized intersections were grouped and the AADT was split in minor and major approaches which were the most significant variables. Moreover, this study of SPM acts as an important step towards the implementation of this tool in decision-making process allowing more efficient allocation of resources for road safety engineering interventions in Brazil.

Keywords: safety performance models, road safety, generalized linear models, observational road safety studies.

INTRODUCTION

One of the main goals for transportation researchers and practitioners is to ensure adequate safety performance of the various transportation components to all road users given the resources available. Historically, safety has been defined and measured in terms of observed number of crashes in part by the intuitive and logical link between these two. This type of approach relies heavily on the reliability and overall quality of accident data systems as well as on statistical models aiming to estimate the expected number of crashes as a function of geometric and operational attributes of the traffic system components, also known as safety performance models (SPMs), safety performance functions (SPFs) or accident prediction models (APMs).

It is known that the occurrence of accidents in a given location has a strong random component. In this context, SPMs can contribute to measure its real safety performance by attenuating the effect of randomness in the observed crash frequency (Hauer, 2002).

SPMs are developed based on police reported crashes and geometric and operational road attributes as covariates. These models have the potential to improve road safety by the possibility of comparing alternative road projects regarding its expected relative safety performance (AASHTO, 2010). Thus, it is possible to explicitly include road safety criterion in the decision-making process for selecting those projects that are expected to have fewer accidents during a given operation period.

A great deal of the recent worldwide interest in developing SPMs can be accredited to the release of the first edition of the Highway Safety Manual - HSM (AASHTO, 2010). The HSM devotes much of its content to justification, premises, development and application of SPFs in the transportation systems planning process at the strategic, tactical and operational level.

If on one hand SPMs have been explored for more than two decades in countries like Canada, USA, England and Sweden, on the other hand, in Brazil, this methodology stills in its infancy. Early modeling efforts for the Brazilian environment can be accredited to the doctoral research of Cardoso (2006) and summarized in the later work of Cardoso and Goldner (2007), in which urban arterial segments in the city of Porto Alegre were analyzed.

It is believed that this discrepancy can be attributed in part to problems with the availability and quality of information on traffic accidents associated with the relative scarcity of procedures for calibration and validation of such models nationwide. This situation, along with the increase in crash frequencies in Brazilian urban areas, have resulted in a joint research effort started in 2009 and sponsored by the National Council of Technological and Scientific Development (CNPq). The research group consists of six Brazilian Universities and four international institutions namely: University of Waterloo and Ryerson University from Canada, Lund University from Sweden and *Universidade do Minho* from Portugal.

This paper presents the modelling effort and initial results for the development of safety performance models for urban streets in five Brazilian cities. Given the limited data available in most cities, this project also aimed to generate alternative solutions to obtain data and at the same time, to guide future developers on the best approach to defining the functional form of the models, techniques to be adopted in the calibration process and investigating model spatial and temporal transferability.

ROAD SAFETY MODELING WITH OBSERVATIONAL DATA

From the engineering perspective, methodologies for safety assessments are heavily influenced by the way safety is defined and measured. Traditionally, the level of safety of a given entity has been defined as “the number of crashes by kind and severity, expected to occur on the entity during a specific period” (Hauer, 2002). Representing safety throughout crash events is therefore the natural domain of observational studies.

Observational studies can be viewed as a passive learning process where the knowledge comes from meticulous analysis of the outcome of events that have not been formally designed to address the problem. According to Davis (2004), an underlying assumption of these studies is that crashes are individually unpredictable, although groups of crashes observed on a given location can produce predictable statistical pattern. Basically, these groups of crashes may be related to a single time period (number of crash related to one or more years are considered globally) or may include longitudinal data. In the latter situation, data belonging to different years for a given location can be treated and analyzed as time series events. Additionally, the safety condition of different locations may be related to each other. The presence of temporal or spatial correlations in the database imposes specific statistical considerations for model development (Lord and Persaud, 2000; Wang and Abdel-Aty, 2006).

Under this paradigm, several methods for linking accidents and their consequences to human, vehicle, roadway and environmental attributes have been proposed over the last two decades. These include the use of contingency tables, linear multivariate regression models, logistic models, hierarchical loglinear models, induced exposure models, generalized linear models, among others. A broad review of these methodological alternatives for global crash-frequency data can be found in Lord and Mannering (2010). In addition, the work of Savolainen *et al.* (2011) presents a general review on methodological alternatives for specific analysis of crash-injury severities.

Due to the relative ability to deal with some aspects of the inherent stochastic rare random nature of crashes, such as the regression to the mean phenomenon and the crash frequency over dispersion, the generalized linear modelling approach (GLM) has recently become widely applied (Hauer, 2004; Sawalha and Sayed, 2006; Hadayeghi *et al.*, 2007; AASHTO, 2010). Furthermore, a procedure derived from the GLM, called generalized estimating equations (GEE), has been successfully applied in the presence of longitudinal and/or spatial correlations in the database (Lord and Persaud, 2000; Wang and Abdel-Aty, 2006). The most commonly found general expression for SPFs can be written as (Hakkert *et al*., 1996; Sawalha e Sayed, 2006; AASHTO, 2010):

$$Y = \alpha \left[\prod_i (\boldsymbol{A}_i)^{\beta_i} \right] \cdot e^{\sum_j (\gamma_j \boldsymbol{B}_j)} \tag{1}$$

where
Y = expected number of crashes over a specific time interval (year);
A e B = predictive variables;
α, β_i, γ_j = model's coefficients.

Initially SPMs were developed assuming the error structure to be compatible with the Poisson distribution, however, other studies have shown better results by assuming the negative binomial distribution (also known as poisson-gama) error structure in cases where crash frequency presents a considerable dispersion (variance greater than average) among similar entities (Bonneson and McCoy, 1993; Persaud and Mucsi, 1995). For all distributions, one important aspect that imposes difficulties for the modeling procedure is the presence of many records of zero crashes in the database. Different approaches have been adopted to deal with this situation (Shankar et al., 1997; Lee and Mannering, 2002; Kumara and Chin, 2003). However, some of them present limitations when considered under the traffic engineering point of view (Lord *et al*. 2005, Lord *et al*., 2007).

SPM expressions have been used as one of the most important component of a methodology to improve crash estimations known as the Empirical Bayes method (EB method), which applies concepts of conditional probability to both the reference population (represented by SPMs) and specific sites to produce a weighted value of the expected number of crashes. The EB estimate of crashes is given by (Hauer, 2002):

$$E(m|x) = w \cdot E(m) + (1 - w) \cdot x \tag{2}$$

where
E(m|x) = the expected number of crashes for entity *m* given that *x* crashes have been observed for the same entity;
E(m) = crash estimate obtained from regression model developed using crash data of similar sites (SPMs);
w = weight assigned to *E(m)* ($0 \leq w \leq 1$);
x = observational crash data for the site.

The expression that yields the best estimate of *w* is given by

$$w = \frac{1}{1 + {VAR(m)}/{E(m)}} \tag{3}$$

Where *VAR(m)* is the variance associated with the regression model developed. Basically, the weight *w* in Equations 2 and 3 is a function of the variability found in the data used to develop the crash prediction model. The lower the variation in these data the higher the weight placed on the model estimates of crashes, i.e., higher level of confidence in the model results. The EB method has been largely adopted in hot-spot identification (network screening) and in before-after analysis (AASHTO, 2010).

Despite the considerably large body of research on SPM development, a general methodology for SPM calibration is currently not available mostly due to a general lack of consensus regarding strategic questions such as: 1) What would be the minimum sample size required to develop “acceptable” models?; 2) How to select the most important variables to be considered in the model formulation?; 3) What is the most adequate model structure?; 4) How to confirm the model usefulness and acceptability (model validation) and; 5) When is appropriate to draw causal inferences as opposed to predictive inferences? The following section describes the general methodology adopted for SPM calibration for Brazilian urban intersections.

METHODOLOGY

The proposed methodology for calibrating SPM for Brazilian roads has been structured into the following steps: 1. Identifying geometric and operational attributes for homogeneous groups (entities) such as signalized and unsignalized urban intersections; 2. Estimating the annual average daily traffic (AADT) for available intersection groups; 3. Obtaining crash data from jurisdictional accident data system; 4. SPM calibration for the jurisdictions involved in the task force.

Identifying geometric and operational attributes for homogeneous groups

This initial research effort has focused primarily on estimating SPM for urban intersections manly due to data availability regarding traffic flow and other attributes considered important for the initial definition of homogeneous groups such as: number of approaches, number of traffic lanes, intersection angle, central median configuration, type of traffic control, land use, bus stops and parking configuration. In fact, the literature reviewed has indicated a lack of consensus regarding an objective criterion to consider a given attribute as a group variable rather than a predictive variable.

Some of the tools used in this step were aerial photographs using applications such as Google Earth ©, maps and reports provided by the local traffic center, as well as specific legislation relating to land use and functional classification of each municipality. Table I presents a summary of group and predictive variables as well as the sample size for each jurisdiction.

Table I – Sample size and predictive variables for homogeneous entities

Jurisdiction	Entity	Group Variables	Predictive Variables	Sample Size
Belo Horizonte	Signalized	Commercial land use, central median	AADT, #lanes, #approaches, central median	220
Brasília	Intersections	nd	Major/Minor AADT, #lanes/approach, type of traffic control; red light enforcement; land use	32
Fortaleza	Signalized	Mixed land use, four legged intersections	Major/Minor AADT, total AADT, #lanes, #approaches, central median	101
Fortaleza	Unsignalized	Mixed land use, four legged intersections	Major/Minor AADT, total AADT, #lanes, central median	132
Porto Alegre	Unsignalized	central median, mixed land use, parking	Total AADT, # lanes # approaches	238
São Carlos	Signalized	four legged intersections	Major/Minor AADT # lanes	100

Average Annual Daily Traffic (AADT) Estimation

It has been verified a significant discrepancy between the databases of vehicular flow on the cities involved in this research. Due to the multiple collection methods (technology, frequency and scope) and data manipulation leading to the estimate of AADT, local research group efforts varied from the establishment of partnerships with traffic management agencies to analyse the data to the establishment of plans for manual counting on samples of the entities to be modelled. The main challenge presented for the AADT estimation was the scarcity of data records of traffic volume in time and space. To solve this limitation, each city has adopted specific procedures reported below.

Brasília

The estimative of AADT was made from short counts of three hours in each intersection approach (V_{3h}). Expansion factors were determined from historical data collected by the nearest red light surveillance camera for three hours of observation at each site, taking into account the day of the week, month and year of collection (2005- 2010) Thus, each spot had its own expansion factors for the V_{3h} collected: Daily factor for the day of the week; Monthly factor for the month of the year; and an Annual factor related to temporal variation for each year. These factors are, therefore, only applied to the three-hour counts performed aiming at estimating the total annual traffic for each year of the period 2005-2010. The AADT for each intersection approach per year is then obtained by:

$$\text{AADT} = \frac{V_{3h} \times \text{Daily Factor} \times \text{Monthly Factor} \times \text{Annual Factor}}{365} \qquad (4)$$

When the data of the red light surveillance equipment do not allow for an estimate of the AADT for one or more year, this value is obtained by a regression analysis.

Belo Horizonte

The AADT estimation for Belo Horizonte used the methodology proposed by McShane *et al.* (1998). This method allowed work with a sample of 24 hours (V_{24}) for each intersection combined with data obtained from permanent counting stations. The expansion factor (F_w) was determined weekly by dividing Average Daily Traffic ADT by the estimated Annual Average Daily Traffic ($AADT_{est}$) for each day of the week in the spot. The factor of monthly (F_M) expansion is obtained by dividing the Average Daily Traffic per month (ADT_M) by the Estimated Annual Average Daily Volume ($AADT_{est}$). Equation 5 is employed:

$$\mathrm{AADT} = \mathrm{V}_{24} \times \mathrm{F}_{\mathrm{w}} \times \mathrm{F}_{\mathrm{M}} \tag{5}$$

Fortaleza

In Fortaleza, daily flows (V_{24} in equation 5) were estimated from short duration counts using daily expansion factors calculated for 5 homogeneous groups of intersections depending on geographic location and type of collection equipment (loop detectors for traffic signals or red light enforcement). The expansion factors of short daily counts, presented were estimated by surveying conducted in two intersections of Fortaleza during the 24 days to obtain the daily traffic flow pattern. The V_{24} estimates were subsequently expanded to AADT using weekly (F_w) and montlhy (F_M) factors obtained in the studies by Oliveira (2004). Tables II, III and IV present the expansion factors applied to the Fortaleza data set.

Porto Alegre

The estimation of AADT was based on expansion factors calculated specifically for the city of Porto Alegre (presented in Tables II, III and IV). These factors were calculated from data obtained by automatic traffic counters installed along the speed surveillance equipment in two intersections. The same factors for expansion were used for all intersections considered in the study. The procedures for obtaining the expansion factors for estimating the AADT are detailed in Holz *et al.* (2012). Besides the application of the expansion factors, it was also necessary to correct the variation related to increase in traffic through the estimation of annual growth factors.

São Carlos

The estimation of AADT was based on short counts in typical weekday (Tuesday through Thursday) expanded by a time correction factor to obtain the ADT, this factor was obtained from counts made by surveillance cameras for 16 weeks during the year 2009 in three spots. The ADT is multiplied by 341 and divided by a factor 365 in order to obtaining the AADT.

The daily and monthly expansion factors obtained are presented in Tables II and III respectively. In Table IV are shown the hour expansion factors.

Table II – Daily expansion factors for Belo Horizonte, Fortaleza and Porto Alegre

Day of week	Belo Horizonte	Fortaleza	Porto Alegre

Sun	1.85	1.65	n.a.
Mon	0.91	0.93	0.98
Tue	0.92	0.93	0.98
Wed	0.89	0.92	1.00
Thu	0.90	0.91	1.00
Fri	0.86	0.91	1.04
Sat	1.20	1.18	n.a.

Table III – Month expansion factors for Belo Horizonte, Fortaleza and Porto Alegre

Month	Belo Horizonte	Fortaleza	Porto Alegre
Jan	1.09	1.03	0.93
Feb	1.07	1.01	0.91
Mar	0.98	1.01	1.06
Apr	1.64	1.01	1.00
May	1.00	1.01	0.98
Jun	1.03	1.00	0.99
Jul	0.99	1.03	0.95
Ago	1.00	0.99	1.01
Set	0.93	1.00	1.04
Out	0.95	0.99	1.05
Nov	0.92	0.96	1.05
Dec	0.98	0.95	1.02

Table IV – Hourly expansion factors for Fortaleza, Porto Alegre and São Carlos

Hour	Fortaleza	Porto Alegre	São Carlos
0 -1 h	1.3	1.2	0.4
1- 2 h	0.9	0.6	0.2
2 - 3 h	0.7	0.3	0.1
3 - 4 h	0.6	0.2	0.1
4 - 5 h	0.7	0.3	0.1
5 - 6 h	1.2	0.8	0.2
6 - 7 h	3.2	4.0	1.9
7 - 8 h	6.0	7.7	6.3
8 - 9 h	6.0	7.1	5.2
9 - 10 h	5.9	5.6	5.2
10 - 11 h	6.0	4.8	5.3
11 - 12 h	6.4	4.8	6.2
12 - 13 h	6.1	4.9	7.5
13 - 14 h	5.7	5.5	8.0
14 - 15 h	5.6	5.2	7.5
15 - 16 h	5.6	5.1	6.5
16 - 17 h	5.8	5.6	6.9
17 - 18 h	6.6	7.2	8.1
18 - 19 h	6.7	7.7	8.3
19 - 20 h	5.6	6.4	5.5
20 - 21 h	4.6	4.7	3.9
21 - 22 h	4.1	3.8	3.4
22 - 23 h	3.0	3.8	2.4
23 - 24 h	1.7	2.3	1.2

Accident data systems

The greatest challenges to the consolidation of observational studies applied to road safety are associated to problems related to the use of databases of traffic accidents as a primary source of data. Among the most common problems found in the literature are: (i) low reportability index, (ii) incomplete and inaccurate information, and (iii) errors during data entry, among others (Davis, 2004; Hauer, 2002 and Hirst *et al*., 2004).

The development of SPM directly depends on the existence of reliable accident database showing traffic accidents recorded in recent years. These databases were obtained from local authorities at each municipality related to this study and the database information acquired are: Location, Type of accident, Date, Day of week, Time, Type of vehicle, Severity of casualty, Weather, Light conditions; the last two were not available for Belo Horizonte.

Different degrees of difficulty were encountered at each municipality concerning the format and the interpretation of the contents of the database. In Belo Horizonte and Brasilia the accident database system (ADS) does not record property damage only accidents. This procedure tends to reduce the total number of accident in the intersections considered in the study.

In medium sized cities such as Bauru, São Carlos and Jaú, data quality can vary considerably according to the year, since changes in the political environment (Mayors and respective Chiefs) can affect data collection procedures, sometimes carried out by the municipal office or by private companies. This indicates the need of a strong regulation about the procedures and local jurisdiction responsibilities for the accident database management.

The determination of the number of accidents from the database information is not a direct activity, due to data storage process, especially for Belo Horizonte jurisdiction. In this case, the attempt to determine the number of accidents by creating a routine in MS Excel was not successful due to data format. Thus, the coordinates of the accident were used. MS Excel database was used in combination to Global Mapper software to plot on the city map the occurrence of accident using geographic coordinates of the accidents. The identification of the number of accidents was made by manual counting.

In general, crash reports and traffic flow information are available from different databases which make it difficult to manipulate databases individually, connected to the time spent to associate volume and number of accidents information in the same file.

Table V present a summary of predictive and response variables obtained from the available sample for Belo Horizonte, Brasília, Fortaleza, and Porto Alegre jurisdictions.

Table V – Variables description and basic statistics for the intersections studied

Variable	Description	Average(*)	Sd.	Mín.	Máx.
BELO HORIZONTE: Signalized Intersections*					
AADT	AADT – all entering vehicles	39,211	21,975	2079	117,236
#lanes	Total number of lanes	6		2	14
#app	Number of approaches	3	0.7	2	5
Y_{IF}	Injury+ fatal crashes	4	2.6	0	20
Y_{IF}	Injury+ fatal crashes (2007-2010)	4	3.0	0	24
BRASÍLIA: **					
$AADT_{major}$	Major street AADT	21,659	12,688	6,387	61,479
$AADT_{minor}$	Minor street AADT	4,960	5,501	133	18,434
AADT	AADT – all entering vehicles	26,620	15,227	6,521	73,801
#lanes	Total number of lanes major street	2,43	0,66	2	5
Traffic Control	1=signalized intersection 0=unsignalized intersection	0,49	0,5	0	1
Red light surveillance camera	1 = with camera 0 = without camera	0,43	0,5	0	1
Y_{it}	Number of crashes for intersection i on time period t	2.21	2.61	0	12
FORTALEZA: Signalized Intersections***					
$AADT_{major}$	Major street AADT	23,583	7,850	8,047	44,312
$AADT_{minor}$	Minor street AADT	11, 735	5,267	641	28,563
AADT	AADT – all entering vehicles	35,319	10,438	15,887	65,618
#lanes	Total number of lanes	5.8	1.4	4	12
#app	Number of approaches	2.7	0.6	2	4
mconf	0=no median, 1=major., 2=major and minor	0.7	0.6	0	2
Y_T	Total crashes	7.5	7.7	0	48
Y_{IF}	Injury+ fatal crashes	1.5	1.6	0	7
*Unsignalized Intersections*****					
AADT	AADT – all entering vehicles	17,424	8,375	1,040	41,984
#lanes	Total number of lanes	4.2	0.85	4	8
mconf		0.02	0.25	0	1
Y_T	Total crashes	4.1	3.9	0	19
PORTO ALEGRE:					
Signalized and *Unsignalized* Intersections					
AADT	AADT – all entering vehicles (2011)	46,312	31,627	865	163,176
Y_T	Total crashes (2005-2010)	6.2	8.3	0	47

(*) sample = 220 intersections. Year = 2007 (**) sample = 32 intersections Years = from 2005 to 2010

(***) sample = 101 intersections. Year = 2009 (****) sample = 132 intersections. Year = 2007 XX

SPM calibration for the jurisdictions involved in the task force

Best model parameters estimates were obtained by applying the maximum likelihood method performed by the Newton-Rapson optimization algorithm. The task force has used a number of computational applications such as SAS©, SPSS©, and R© that have pre-established routines for the estimation of parameters depending on the structure assumed for the error (Poisson or negative binomial).

Due to the reduced sample size and consequent need for a longitudinal study, SPMs for Brasilia jurisdiction have been developed using a Generalized Estimation Equations (GEE) procedure which represents an extension of the Generalized Linear Models approach (Liang and Zeger, 1986; Halekoh et al., 2006) The GEE procedure allows for investigating different types of correlation structure in the longitudinal crash data, including the independence condition (Liang and Zeger, 1986; Wang and Abdel-Aty, 2006;).

The calibration methodology followed a sequential process of inserting variables suggested by Hauer (2004). In the base model, the only predictive variable considered was the AADT representing the exposure. Other predictive variables were introduced to the base model forming a set of models with two predictive variables. These models were compared to the base model in terms of the logarithm of the maximum likelihood (*2l*). The 2-variable model with the highest *2l* is then recalibrated with a third variable. This process is repeated until all significant variables are included in the model.

An initial test using the dispersion parameter (σ_d) was carried out on base models in order to verify the most adequate assumption regarding the error structure (Poisson or negative binomial). The dispersion parameter is obtained by the ratio of the generalized Pearson statistic χ_p^2 and the difference between the number of observations and number of model parameters. In this case obtaining values close to 1 indicates that the variance assumed in the model structure is similar to that observed in the data. In all jurisdictions this test yielded σ_d values close to 1 assuming NB error structure.

The goodness of fit analysis for candidate models was performed using the following statistics: the generalized Pearson χ_p^2 statistic, the scaled deviance S_p and the Akaike Information Criteria (AIC). The Pearson χ_p^2 statistic can be used for null hypothesis significance testing regarding the equivalence of the variance assumed in the modeling effort and the sample variance. The scale deviance is useful to compare the proposed model and the saturated model and the AIC compares different models based on the balance between the bias and variance explained by them. For models developed using the GEE approach (Brasilia jurisdiction) a Quasi-likelihood Information Criterion as presented by Pan (2001) was also applied as a goodness of fit criteria.

The model performance over the entire range of predictive variables was assessed throughout the Cumulative Residual Plot (CURE plot). According to Hauer (2004), for reasonably good models, the CURE plot should present a moderate random oscillation around zero ("random walk"). Constant up or downward trend in the graph suggests areas (range of values) where the model can over or underestimates the predicted number of crashes. Additionally, Hauer (2004) recommends developing $+2\sigma_d$ and $-2\sigma_d$ confidence intervals for the CURE plot to verify the model validity over its entire domain. The following section discusses the major results from the model calibration effort focusing on the estimated parameters and correspondent statistics as well as on overall performance of the proposed SPM.

MAJOR RESULTS

SPM for Fortaleza

The model calibration effort for signalized intersections in Fortaleza has yielded a total of 8 different SPM expressions for the total number of crashes (2009) and the set of predictive variables described earlier (Table V). Calibrated model coefficients and other relevant statistics for all estimated expressions can be found elsewhere (Cunto *et al*. 2012).

A strict relative comparison using the generalized Pearson χ_p^2 statistic, the scaled deviance S_p and the Akaike Information Criteria (AIC) would suggest model with all variables (AADT, number of lanes, number of approaches and central median configuration as the best model. However, differences among statistics are subtle revealing no conclusive indicators to choose this expression as the best model. Since a significant correlation among variables is indicated by the changes on parameter coefficients during the model calibration process, the selected model was the one with least variables including exposure in order to avoid cause/effects inferences. Equation 6 represents the SPM for signalized intersections in Fortaleza:

$$Y_T = 0.00591(AADT^{0.52}) \cdot e^{0.28 \cdot \#lanes} \tag{6}$$

Where

Y_T = total number of crashes in 2009;

AADT = annual daily traffic entering the intersection;

#lanes = total number of lanes (all intersection legs).

For the models using the frequency of injury and fatal crashes for signalized intersection in Fortaleza, the calibration did not produce statistically significant coefficients for the set of independent variables. A visual analysis using a plot with AADT and the frequency of injury and fatal crashes revealed a considerably scattered pattern. It is possible that this over dispersed behavior, the inherently low sample mean and the relative small sample size (101 intersections) have contributed significantly to the problems observed for the injury and fatal crashes model.

For the unsignalized intersections in Fortaleza the results (Table VI) indicated that number of lanes and central median configuration were not statistically significant in models 3 and 4. Model 1 (total number of crashes and AADT) chosen basically due to its simplicity and the smaller AIC can be expressed as:

$$Y_T = 0.049(AADT^{0.46}) \tag{7}$$

Where

Y_T = total number of crashes in 2007.

The cure plot for both signalized (Cunto *et al.* 2012) and unsignalized models (Figure 1) for Fortaleza have indicated a reasonable performance over the entire values of the exposure value.

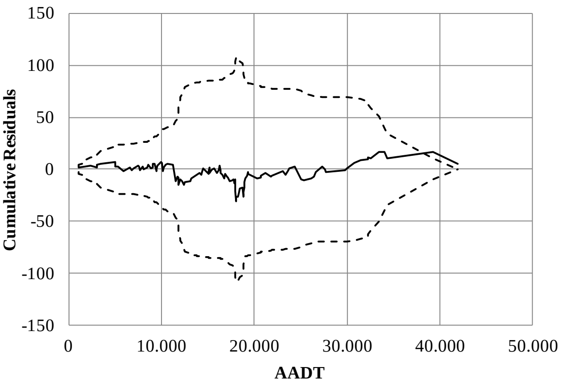

Figure 1 – Cure plot – unsignalized intersections - Fortaleza

Table VI – Calibrated models for total number of crashes - unsignalized intersections – Fortaleza

Variable		Models			
		1	2	3	4
α	$ln(\alpha)$	-3.02	-2.66	-3.00	-2.91
	Coef.	0.049	0.070	0.050	0.055
	$\hat{\sigma}_\beta^{(**)}$	1.28	1.34	1.29	1.59
$AADT_{total}$	Coef.	0.46	0.46	0.46	0.46
	$\hat{\sigma}_\beta$	0.13	0.13	0.13	0.13
#lanes	Coef.	(-)	-0.09	(-)	$-0.02^{(*)}$
	$\hat{\sigma}_\beta$	(-)	0.10	(-)	0.22
mconf	Coef.	(-)	(-)	$-0.33^{(*)}$	$-0.25^{(*)}$
	$\hat{\sigma}_\beta$	(-)	(-)	0.34	0.73
ϕ		1.55	1.57	1.57	1.58
2L		-656.85	-656.09	-650.10	-650.09
$\chi^2_{p.crítico;0.05}$		158	157	157	155
χ^2_p		59	62	61	58
S_p		149	149	150	149
AIC		657	658	658	660
σ_d		1.15	1.16	1.16	1.16

(*)coefficients statistically NOT significant (α=0.05) (**) standard-error
(-) variable not used in the model

SPM for Belo Horizonte

A similar set of independent variables used in Fortaleza was tested for the Belo Horizonte signalized intersection sample (220 observations). A noteworthy difference between these jurisdictions however relies on the fact the, for Belo Horizonte jurisdiction a longitudinal analysis was also performed using four years of injury and fatal crashes for each intersection (2007-2010). Individual models were calibrated for each year using the combination of predictive variables as shown in Table VII and the general model expression similar to Equation 1.

Table VII – Tested models for total number of injury and fatal crashes - signalized intersections – Belo Horizonte

	Models							
Variable	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
$AADT_{YY(*)}$	x	x	x	x	x	x	x	x
#lanes				x	x	x		x
mconf		x				x	x	x
#app			x		x		x	x
Y_{IF} (dependent variable)	x	x	x	x	x	x	x	X

*YY = year

The statistics described earlier (Pearson χ_p^2,S_p and AIC), the maximum likelihood and the cure plot were applied to all developed models in order to check their overall performance. The results suggest that for models 5 through 8 all variables are statistically significant despite the year tested. Among these, model 8 (AADT and number of lanes) presented the largest increase in the logarithm of the maximum likelihood (-2ℓ). Equations 8 to 11 represent the calibrated SPF for Belo Horizonte signalized intersections:

$$Y_{IF07} = 0.0955(AADT^{0.246}) \cdot e^{0.158 \cdot \#lanes} \tag{8}$$

$$Y_{IF08} = 0.0544(AADT^{0.160}) \cdot e^{0.309 \cdot \#lanes} \tag{9}$$

$$Y_{IF09} = 0.0067(AADT^{0.502}) \cdot e^{0.530 \cdot \#lanes} \tag{10}$$

$$Y_{IF10} = 0.0221(AADT^{0.389}) \cdot e^{0.160 \cdot \#lanes} \tag{11}$$

Where

Y_{IFYY} = total number of injury and fatal crashes (YY=year);
AADT = annual daily traffic entering the intersection;
#lanes = total number of lanes (all intersection legs).

Additional models considering the aggregate data from all four years have also been investigated. In this case, a new variable representing the year (*YY*) was included in the general expression and indicated to be statistically significant. The model type 5 and 6 (Table VII) and for all tested models at least one variable was not statistically significant. These results therefore suggest a trend on the data that could deteriorate the overall model prediction strength if one aggregates yearly data for this exercise.

Figure 2 presents the CURE plots (solid line) with $\pm 2\sigma_d$ confidence interval (dashed lines) for 2007 to 2010 models. The results suggest that in general, there is a reasonably good fit until the total AADT around 80k vehicles per day, although the cumulative residuals tend to be negative. Beyond this AADT level, models reduce their estimation strength given the relative small number of observations on this range.

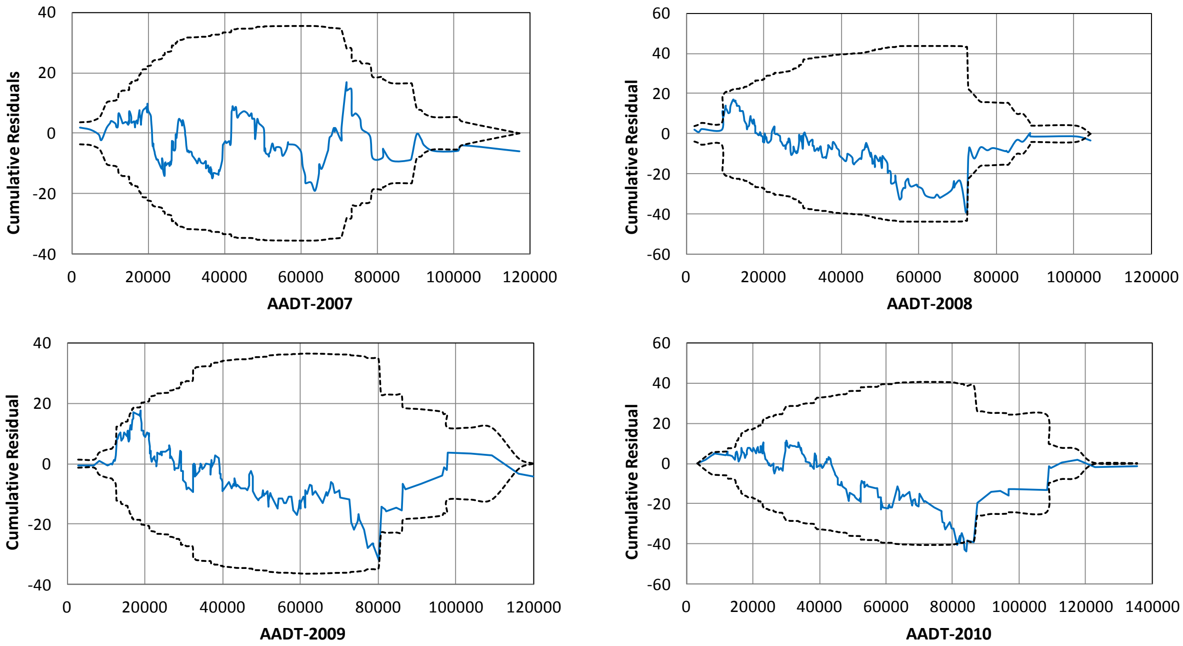

Figure 2 – Cure plot for signalized intersections – Belo Horizonte (2007 -2010)

SPM for Brasilia

Based on the literature review, especially on the work of Lord and Park (2008), 5 (five) alternative model functions were investigated, which are initially only based on exposure measures. They are presented in Equations 12 to 16:

$$Model\ 1: \ln \mu_{it} = \beta_0 + \beta_1 \ln(F1_{it} + F2_{it}) \qquad (12)$$

$$Model\ 2: \ln \mu_{it} = \beta_0 + \beta_1 \ln F1_{it} + \beta_2 \ln F2_{it} \qquad (13)$$

$$Model\ 3: \ln \mu_{it} = \beta_0 + \beta_1 \ln(F1_{it} \times F2_{it}) \qquad (14)$$

$$Model\ 4: \ln \mu_{it} = \beta_0 + \beta_1 \ln(F1_{it} + F2_{it}) + \beta_2 \ln\left(\frac{F2_{it}}{F1_{it}}\right) \qquad (15)$$

$$Model\ 5: \ln \mu_{it} = \beta_0 + \beta_1 \ln F1_{it} + \beta_2 \ln F2_{it} + \beta_3 \ln(F1_{it} \times F2_{it}) \qquad (16)$$

Where:

β_0, β_1, β_2, β_3 – parameters to be calibrated;

μ_{it} - expected number of crashes for intersection i on time period t;

$F1_{it}$, $F2_{it}$ – AADT at intersection *i* on time period *t* for the approaches of road 1 (arterial) and 2 (minor road), respectively.

The investigated models were calibrated with the GEE procedure, as the available data are longitudinal. Three possible working correlation structures among the annual data for each location were considered: independent, exchangeable and autoregressive. It was assumed Negative Binomial Distributions for the estimative errors because the high dispersion on the available crash data (many zero-observations). Other relevant aspects considered for all estimated expressions can be found elsewhere (Claude, 2012).

Models were evaluated by the following statistical tests: the cumulative residual test (CURE); the Akaike's information criterion (AIC) in the GEE, which is called the quasi-likelihood information criterion (QIC); and the R2 statistic calculated based on standardized residuals. This evaluation for the models with only exposure variables showed that Models 1, 2 and 3 are the most acceptable for further developments. Their parameters are statistically significant for α = 5%, they revealed exchangeable work correlation among the data and their CURE plots oscillates around zero and are inside the limits defined by the procedure, as required. The R2 for Models 2 and 3 (0.17 and 0.22, respectively) are significantly higher than for Model 1 (0.06).

These models were them further developed to test the inclusion of three other explanatory variables which are: traffic control (TC=1, in case of signalized intersections; TC = 0, otherwise); red light surveillance camera (EF = 1, with camera; EF = 0, without camera); number of traffic lanes of the arterial road (NL), presenting minimum value of 2 and maximum value of 5. Only variable EF presented associated regression coefficients statistically significant when attached to the original Models 2 and 3, which will further referred to Models 6 and 7, respectively. Both models are statistically acceptable based on the statistical tests performed and their R^2 is similar (0.18 for Model 6 and 0.19 for Model 7). Therefore, both of them can be used and their final expressions are given in Equations 17 and 18. As compared to the models without the variable EF, it is clear that Model 3 presents the highest R^2 value. Its final expression is shown in Equation 19.

$$\text{Model 6:} \qquad \mu_{it} = e^{-10.0421} \times F1_{it}^{\;0.6160} \times F2_{it}^{\;0.5519} \times e^{-0.6001 \times EF} \tag{17}$$

$$\text{Model 7:} \qquad \mu_{it} = e^{-9.7609} \times \left(F1_{it} \times F2_{it}\right)^{0.5725} \times e^{-0.6096 \times EF} \tag{18}$$

$$\text{Model 3:} \qquad \mu_{it} = e^{-7.8724} \times \left(F1_{it} \times F2_{it}\right)^{0.4583} \tag{19}$$

Because of the small size of the considered intersection sample, the above models must be used with caution for predicting the frequency of crashes for intersections located in Taguatinga. However, the fact that in both Models 6 and 7 the presence of red light camera reduces this frequency gives an important indication of the potential effect of this equipment on promoting traffic safety (Claude, 2012).

SPM for São Carlos and Porto Alegre

In São Carlos was not possible to establish SPM functions due to the small number of intersections which has both AADT and accidents data. In this way any attempt to modeling the SPM will not have statistical significance. The solution to address this question is to aggregate data from cities in the state of Sao Paulo with similar characteristics such as population, car fleet and income per capita.

The estimation of the SPM for Porto Alegre is not already completed due to failures in the accident data information, which are being corrected with visits to accident sites.

CONCLUDING REMARKS

This paper presented the initial joint effort for developing SPM for urban intersections of five Brazilian cities. The SPM were developed for signalized and unsignalized intersections in Fortaleza, for signalized and unsignalized intersections (grouped) in Brasilia and for signalized intersections in Belo Horizonte.

This research has showed the most significant independent variables to be used for calibrating SPM for Brazilian cities. First the exposure variables as AADT for total flow entering the intersection (Fortaleza e Belo Horizonte) and the AADT for minor and major flow (Brasília) are used as based model. The other variables considered as number of lane, number of approaches and central median where included in the SPM in Fortaleza and Belo Horizonte, using the generalized Pearson χ_p^2 statistic, the scaled deviance Sp and the Akaike Information Criteria (AIC) all variables were significant; however, differences among statistics are no conclusive to choose the best model. Since a significant correlation among variables is suggested by changing on parameter coefficients during the model calibration process, the selected model was the one with least variables including exposure in order to avoid cause/effects inferences, in this way the models with the entering AADT and number of lanes were the best fitted, especially because requiring less variables, where the number of lanes is easy obtained. In Brasilia due the number of intersections used (32) the signalized and unsignalized are grouped, thus one variable the type of control was added, also more two variables are used as red light surveillance camera and number of lanes; in the case of Brasilia only the variable red light surveillance camera was significant.

The difference among the models produced we can infer the following: Brasilia due the small number of samples (32) the signalized and unsignalized intersections are grouped, but the AADT were entered for minor and major approaches, this can explain why the number of lanes and the type of control were not significant variables, because the AADT split for minor and major approach could have correlation among the number of lanes and type of traffic control, i.e., is incorporating their effects; Fortaleza and Belo Horizonte the number of samples were not too small and the model structure and coefficients are in the same range of values, and the main variables are the entering AADT and the number of lanes; Brasília in its SPM also indicated that the enforcement with red light camera could make the difference in the number of accidents.

Finally, among the major initiatives to enhance the scope of this research for the development of SPM for Brazilian cities (but that can server for other countries in the same stage) are the need for an integrated accident and traffic flow database and the need for a reliable and precise spatial crash location system. The study also has implications for establishing standards to existing and proposed organizational accident database systems that can easily be used to feed models used for road safety assessment.

Among the subsequent steps expected for the research group are a comparison between SPMs developed for different jurisdictions (worldwide), the calibration of SPMs for urban road segments and vulnerable users. It is also expected to be explored the application of developed SPMs for safety assessments and its usefulness as a tool to improve the transportation planning process at the strategic, tactic and operational levels..

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Brazilian Council for Scientific and Technological Development (CNPq) for funding this research.

REFERENCES

AASHTO (2010). Highway Safety Manual. American Association of State Highway and Transportation Officials. 1st. Edition, Washington, D.C.

Bonneson, J. A., P. T. McCoy (1993). Estimation of safety at two-way stop-controlled intersections on rural highways. Transportation Research Record, n. 1401, TRB, Washington, D.C., p. 83–99.

Cardoso, G. (2006). *Modelos para previsão de acidentes em vias arteriais urbanas*. Tese (Doutorado). Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal do Rio Grande do Sul–UFRGS.

Cardoso, G. e L. Goldner (2007). *Desenvolvimento e aplicação de modelos para previsão de acidentes de trânsito. Transportes,* v. 15, n. 2, p. 43–51.

Claude, G. F. de M. (2012). *Previsão da ocorrência de acidentes de trânsito em interseções de vias arteriais urbanas – O caso de Taguatinga/DF*. Dissertação Mestrado. Programa de Pós-Graduação em Transportes, Universidade de Brasília.

Cunto, F.J.C., M.M. Castro Neto, D.S. Barreira (2012). *Modelos de previsão de acidentes de trânsito em interseções semaforizadas de Fortaleza*. *Transportes*, v. 20, n. 2, p. 57–64. DOI:10.4237/transportes.v20i2.558.

Davis, G. A. (2004) Possible Aggregation Biases in Road Safety Research and a Mechanism Approach to Accident Modeling. Accident Analysis and Prevention, v. 36, p. 1119–1127. http://dx.doi.org/10.1016/j.aap.2004.04.002.

Hadayeghi, A., A.S. Shalaby. B.N. Persaud (2007). Safety Prediction Models: Proactive Tool for Safety Evaluation in Urban Transportation Planning Applications. Transportation Research Record, No. 2019, TRB, Washington, D.C., p. 225-236. http://dx.doi.org/10.3141/2019-27

Hakkert, A. S., L. Hocherman e A. Mensah (1996). Levels of safety on interurban roads. Transportation Research Record, TRB, National Research Council, Washington, D.C., n. 1553, p. 95–102.

Halekoh, U., S. Hϕjsgaard and J. Yan (2006). The R Package geepack for Generalized Estimating Equations. Journal of Statistical Software, Volume 15, Issue 2, pp. 1-11.

Hauer, E. (2002) Observational Before-after Studies in Road Safety. (1ª ed.). Pergamon.

Hauer, E. (2004) Statistical Road Safety Modeling. Transportation Research Record, No. 1897, TRB, Washington, D.C., p. 81-87. http://dx.doi.org/10.3141/1897-11.

Holz, R.F., Medeiros, F.S., Lange, R.L., Nodari, C.T., Lindau, L.A. (2012) Volume Diário Médio Anual (VDMA): Questões Conceituais e Cálculo de Fatores de Expansão. Anais do XXVI Congresso da ANPET, Joinville, SC.

Kumara, S.S.P. and H.C. Chin (2003). Modeling Accident Occurrence at Signalized Tee Intersections with Special Emphasis on Excess Zeros. Traffic Injury Prevention 4, pp.53-57.

Lee, J. and F. Mannering (2002). Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accident Analysis and Prevention 34, 149-161.

Liang, K. and S.L. Zeger (1986). Longitudinal Data Analysis Using Generalized Linear Models. Biometrika 73, pp. 13-22.

Lord, D. and B.N. Persaud (2000). Accident Prediction Models With and Without Trend – Application of the Generalized Estimating Equations Procedure. Transportation Research Record 1717, pp. 102-108.

Lord, D. and F. Mannering (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A, 44, pp. 291-305.

Lord, D., S.P. Washington and J.N. Ivan (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 37, pp. 35-46.

Lord, D., S.P. Washington and J.N. Ivan (2007). Further notes on the application of zero-inflated models in highway safety. Accident Analysis and Prevention 39, pp. 53-57.

Oliveira, M. V. (2004). A Natureza dos Padrões de Variação Espaço-temporal do Volume Veicular em Ambiente Urbano: Estudo de Caso em Fortaleza. Dissertação de Mestrado. Programa de Pós-graduação em Engenharia de Transportes – PETRAN. Universidade Federal do Ceará, Fortaleza.

Pan, W. (2001). Akaike's Information Criterion in Generalized Estimating Equations. Biometrics 57, pp. 120-125.

Persaud, B. and K. Mucsi (1995) Microscopic accident potential models for two-lane rural roads. Transportation Research Record: Journal of the Transportation Research Board, n. 1485, p. 134–139.

Savolainen, P. T., F.L. Mannering, D. Lord, M.A. Quddus (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis and Prevention 43, pp. 1666-1676.

Sawalha, Z. and T. Sayed (2006). Traffic accident modeling: some statistical issues. Canadian Journal of Civil Engineering. v. 33, p. 1115–1124. http://dx.doi.org/10.1139/L06-056.

Shankar, V. N., G.F. Ulfarsson, R.M. Pendyala and M.B. Nebergall (2003). Modeling crashes involving pedestrians and motorized traffic. Safety Science 41, pp. 627-640.

Shankar, V., J. Milton and F.L. Mannering (1997). Modeling Accident Frequencies as Zero-Altered Probability Processes: an Empirical Inquiry. Accident Analysis and Prevention 29, pp.829-837.

Wang, X. and M. Abdel-Aty (2006). Temporal and spatial analysis of rear-end crashes at signalized intersections. Accident Analysis and Prevention 38, pp. 1137-1150.