Estimating the Safety Performance of Urban Intersections in Lisbon, Portugal

Sandra Vieira Gomes[1]

Research Assistant
National Laboratory of Civil Engineering
Transportation Department
Av. Brasil, 101
1700-066 – Lisbon, Portugal
Tel. +351 218443528
Fax. +351 8443029
Email : sandravieira@lnec.pt

Srinivas Reddy Geedipally

Engineering Research Associate
Texas Transportation Institute
Texas A&M University
3136 TAMU
College Station, TX 77843-3136
Tel. (979) 862-9892
Fax. (979) 845-6481
Email : srinivas8@tamu.edu

Dominique Lord

Assistant Professor
Zachry Department of Civil Engineering
Texas A&M University
3136 TAMU
College Station, TX 77843-3136
Tel. (979) 458-3949
Fax. (979) 845-6481
Email : d-lord@tamu.edu

Word Count: 5,410 + 2,000 (5 tables + 3 figures) = 7,410 words

July 31, 2009

[1]Corresponding author

ABSTRACT

According to official statistics, a large percentage of crashes are reported in Portuguese urban areas. For instance, from 2004 to 2007, about 70% of all injury accidents and 43% of the fatalities occurred inside urban agglomerations. This important safety problem has also been observed on the urban network located in and around Lisbon. Understanding this significant problem, the Government of the Portuguese Republic via its research grant agency – The Foundation for Science and Technology – funded a project whose primary objective consists of developing tools that would help estimating the safety performance of various components of the urban highway system in Lisbon. This paper documents one component of the safety tools that are currently under development. More specifically, this paper describes the steps that were taken to develop predictive models for estimating the safety performance of signalized and unsignalized intersections located in Lisbon. Several crash predictive models (CPMs) were developed using the Poisson-gamma and Conway-Maxwell-Poisson modeling framework. Two types of models were estimated: flow-only and models with covariates. They were estimated using crash and other related data collected at 29 three-legged and 30 four-legged intersections for the years 2004-2007, inclusively. It was found that some highway geometric design characteristics were associated with the crashes occurring at urban three- and four-legged intersections in Lisbon.

INTRODUCTION

According to official statistics, a large percentage of crashes are reported in Portuguese urban areas. For instance, from 2004 to 2007, about 70% of all injury accidents and 43% of the fatalities occurred inside urban agglomerations (*1 - 4*). This important safety problem has also been observed on the urban network located in and around Lisbon. On the urban network in Lisbon, more than 2,400 crashes occurred during the same time period, with about 30% of all crashes involved a pedestrian.

Understanding this significant problem, the Government of the Portuguese Republic via its research grant agency – The Foundation for Science and Technology – funded a project whose primary objective consists of developing tools that would help estimating the safety performance of various components of the urban highway system in Lisbon. No such tools exist in Portugal for estimating the safety performance of urban networks, hence the funding of this project. The research project titled "IRUMS – Safer Roads in Urban Areas" is carried out by the National Laboratory of Civil Engineering (LNEC) jointly with the Department of Engineering at the University of Coimbra, Coimbra. This project intends to develop methods for managing the safety of urban road networks, particularly those applied to Lisbon. The methods focused on estimating the expected crash frequencies, the identification of hazardous sites (or sites with promise) and subsequently select effective countermeasures to reduce the number and severity of crashes.

This paper documents one component of the safety tools that are currently under development. More specifically, this paper describes the steps that were taken to develop predictive models for estimating the safety performance of signalized and unsignalized intersections located in Lisbon. Several crash predictive models (CPMs) were developed using the Poisson-gamma and Conway-Maxwell-Poisson modeling framework. Two types of models were estimated: flow-only and models with covariates. They were estimated using crash and other related data collected at 29 three-legged and 30 four-legged intersections for the years 2004-2007, inclusively.

The paper is organized as follows. The first section provides a brief background about existing statistical models developed in Portugal and elsewhere in Europe. The second section describes the methodology used for estimating the CPMs. The third section presents the characteristics of the data used in this study. The fourth section summarizes the modeling results. The last section provides a summary of the work accomplished so far in this project.

BACKGROUND

As discussed in previous work (see *5* and *6*), there has been a significant amount of research done on the development and application of crash prediction models for various types of highway safety analyses. Since design standards and operational characteristics (e.g., vehicle size, etc.) are obviously very different between Europe and other places around the world, especially in North America, this background section focuses on predictive models that have estimated and applied in European countries.

Hall (*7*) developed several CPMs for four-arm single carriageway urban signalized intersections in the U.K. The models were estimated using 177 intersections with a speed limit equal to or above 30 miles/hour in urban areas. The author concluded that wider approaching lanes were associated with a larger number of right-angle crashes; a greater number of lanes were associated with higher pedestrian accident rates; and the sight distance was found to be significantly associated with left turn accidents. The author also concluded that the increased displacement of the opposite arm to the left or right was associated with lower accident rates.

Mountain and Fawaz (*8*) developed CPMs using data from 662 intersections with different types of traffic control systems in the UK. The authors developed models relating crashes with traffic flows and other variables, namely: method of control, road class, carriageway type (single or dual), number of arms and speed limit. The authors concluded that only the method of control had a significant effect on crash occurrence, although the best fitted models were ones with traffic flow as the only explanatory variable.

Greibe (*9*) developed CPMs for road segments and urban intersections with three or four legs and with or without traffic signals in Denmark. For the intersections' models, the author found that the geometric variables are not (or less) significantly linked to the occurrence of crashes. This was attributed to the complicated internal correlations in the intersection design data, or a lack of good descriptive variables. The estimated accident prediction models for road links were capable of describing more than 60% of the systematic variation ('percentage-explained' value) while the models for junctions had lower values. The significant variables found in the study were: speed limit, road width, number of exits per km, number of minor side roads per km, parking and land use.

Brüde and Larson (*10*) developed models for pedestrian and cyclist crashes occurring at intersections in Sweden. Data from 285 intersections were used for modeling crashes involving pedestrians, and data from 432 intersections were used for modeling crashes with cyclists. The purpose of this study was to develop CPMs and illustrate their predictive capabilities. They used volumes of motorized vehicles, pedestrians and cyclists as explanatory variables. Interestingly, the authors concluded that the models with a low R^2 value may have high predictive capabilities.

Reurings and Janssen (*11*) developed CPMs for road segments in The Netherlands. Crash data from 524 km of roads on urban and rural areas were used for developing the models. The conclusions of this study were: carriageways inside urban areas with AADT< 25000 generally have a lower crash rate than carriageways outside urban areas; carriageways with a speed limit of 50 km/h or 80 km/h and one driving direction have a lower crash rate than carriageways with the same speed limit but with two driving directions; the average crash rate of urban carriageways with a speed limit of 70 km/h is lower than the crash rate of carriageways with a speed limit of 50 km/h; and the average crash rate of rural carriageways with a speed limit of 60 km/h has almost the same crash rate as that of rural carriageways with a speed limit of 80 km/h and with two driving directions.

Although crash modeling is widely spread all over the Europe, Portugal have no such tools for estimating the safety performance of urban networks. The few models that were developed by Lopes and Cardoso (*12 - 14*) in Portugal were related to road segments on motorways and rural areas. In their study, they used crash data collected from 1999 to 2004. Several exploratory variables were considered, such as number of lanes, type and condition of shoulders and medians, and the presence of an additional lane. The authors concluded that on Portuguese motorways all the variables except the number of lanes were significant. However, for single carriageway rural roads with a median, the number of lanes had a significant effect on crash occurrences.

DATA DESCRIPTION

The data collected at signalized and unsignalized intersections included the following: geometric design characteristics, crash data (severity, manner of collision, etc.) and traffic volumes. With the exception of the crash data, all the data were obtained from on-site visits. All the intersections which had missing traffic flow volumes were excluded from the analysis. Thus, given the costs associated with the data collection process, only 59 intersections were finally used (29 three-legged and 30 four-legged). Note that since four years of crash data were considered as distinct observations, there were 116 observations for three-legged intersections and 120 observations for four-legged intersections (those were considered repeated measurements).

Figure 1 shows the graphical representation of the Lisbon road network and the location of the intersections used in this study. The three-legged intersections are shown with a ‘triangle’ whereas four-legged intersections with a ‘square’. The dark colored intersections indicate that traffic flows were either counted manually or using an automated system (see below). The light colored intersections have the traffic flows that were estimated by the models developed by Martinez (*15*).

Figure 1 - Location of the intersections used in the study

Injury accident data were collected from the official Portuguese accident statistics database and police reports. The crashes were then geocoded in a GIS database. The main reason for using both sources of crash data is that the official accident statistics database includes only the street names, but not the exact location where the accident occurred within or near the intersection. To overcome this problem, the sketch of the accident location from the police report was projected onto the road network (intersections and segments). This allowed for identifying the location of each accident for most of the cases.

Figure 2 presents the spatial distribution of all injury accidents that occurred in Lisbon between 2004 and 2007. Each dot represents an accident. From the figure, it can be seen that crashes were concentrated at intersections and segments located on major arterial roads (Av. Gen. Norton Matos and Eixo Norte-Sul are the major expressways of Lisbon, with a speed limit of 80 km/h and Av República and Av. Almirante Reis are two of the major arterial roads, located right in the heart of the city).

Figure 2 - Spatial distribution of injury accidents in Lisbon between 2004 and 2007.

Given the scope for this part of the study, only crashes that were classified as intersection-related were considered in the development of the statistical models. Unfortunately, there is no exact definition about the radius from the center of the intersection to classify a crash as intersection-related. Previous studies have used different criterion. For instance, Mountain, et al. (*16*) considered that the accidents are intersection related if they occur at a distance of 20m from the curb line. Sayed and Rodriguez (*17*) considered a distance of 30m from the intersection (without specifying the point from which it is measured); Lord (*18*) found that this influence is significant up to 15m, measured from the center of the intersection, and Turner et al. (*19*) used a distance of 50m without specifying the point from which it is measured. For this study, it was decided to use a radius of 40m from the center of the intersection to classify crashes as intersection-related.

Traffic flow counts at intersections were collected using three different sources:

- Automated: with a system that records traffic images and then processes it with a movement detection device;
- Manual: with several operators that counted the inflow on each leg;
- By estimates: Martinez (*15*) developed a traffic assignment model for Lisbon, which covered about 55% of the road network length (see Figure 3), all local roads were excluded from the model. Ten intersections of each type were included in the sample.

All the three sources for counting the traffic flow defined above had some limitations. The automated and manual counts were only done for a single day for each intersection. Since they were all collected in week days, it was assumed that the values obtained represented the Average Annual Daily Traffic (AADT). The changes in AADT over time were considered to be practically null, and so, the same value of AADT was adopted for all four years. The output data obtained from the traffic assignment model estimated by

Martinez (*15*) was not fully adjusted to real traffic movements in some of the intersections included in the modeling sample. However, since the variables used relate to total inflow in each intersection, these values were still considered in the model development. Also, the estimated traffic data was supplied in Peak Hour Volume (PHV) from 8:00 to 9:00 am. From the manual and automated counts a conversion factor was calculated in order to transform the PHV into AADT (ratio between the number of vehicles from 8:00 to 9:00 over the AADT).

Figure 3 – Road network used by Martinez in comparison with the total road network of Lisbon.

Table 1 summarizes important data characteristics for crash data occurring at three-legged and four-legged intersections. During the study period, a total of 140 crashes occurred at three-legged intersections, and 211 occurred at four-legged intersections.

Table 1 - Summary Statistics of the crash dataset

Type of intersection	Year	Accidents (min – max – total)	Major Road Flow (min – max)	Minor Road Flow (min – max)
3 Legged	2004	0 – 8 – 42	7140 - 77082	474 - 20956
	2005	0 – 4 – 27		
	2006	0 – 10 – 37		
	2007	0 – 15 – 25		
4 Legged	2004	0 – 12 – 67	5038 - 56066	1691 - 31627
	2005	0 – 15 - 74		
	2006	0 – 13 – 91		
	2007	0 – 7 - 62		

Table 2 summarizes the key variable statistics of the intersections used in this study. The minimum and maximum AADT values ranged from about 400 to 77,000 vehicles per day. Among the variables collected, they included number of legs, number of lanes per leg, average lane width, median width, number of left turn lanes, number of right turn lanes and type of traffic control device (signalized or unsignalized).

Table 2 - Summary Statistics of the Dataset

Variable	Description		3 Legged Intersections				4 Legged Intersections			
			Min.	Max.	Average (std. dev)	Frequency	Min.	Max.	Average (std. dev)	Frequency
F1	AADT on major		7140	77082	20632 (13925)	29	5038	56066	21939 (13811)	30
F2	AADT on minor		474	20956	5409 (4816)	29	1691	31627	11722 (7542)	30
FT	F1+F2		8793	78977	26041 (15319)	29	9340	80211	33661 (19716)	30
FR	F2/FT		0,02	0,48	0,21 (0,13)	29	0,08	0,50	0,35 (0,10)	30
FQ	F2/F1		0,02	0,92	0,31 (0,26)	29	0,09	0,98	0,59 (0,24)	30
LB	Lane balance	1 – yes	-	-	-	9 (31,03%)	-	-	-	5 (17,24%)
		0 - no	-	-	-	20 (68,97%)	-	-	-	24 (82,76%)
LMAJT2	Total number of entering lanes on major = 4 or 5	1 – yes	-	-	-	11 (37,93%)	-	-	-	8 (27,59%)
		0 - no	-	-	-	18 (62,07%)	-	-	-	21 (72,41%)
LMAJT5	Total number of entering lanes on major = 6 or more	1 – yes	-	-	-	8 (27,59%)	-	-	-	11 (37,93%)
		0 - no	-	-	-	21 (72,41%)	-	-	-	18 (62,07%)
LMINT2	Total number of entering lanes on minor = 4 or 5	1 – yes	-	-	-	1 (3,45%)	-	-	-	5 (17,24%)
		0 - no	-	-	-	28 (96,55%)	-	-	-	24 (82,76%)
LMINT5	Total number of entering lanes on minor = 6 or more	1 – yes	-	-	-	3 (10,34%)	-	-	-	9 (31,03%)
		0 - no	-	-	-	26 (89,66%)	-	-	-	20 (68,97%)
LWMAJ	Average lane width on major (m)		2,85	4,60	3,58 (0,48)	29	2,40	5,20	3,67 (0,58)	29
LWMIN	Average lane width on minor (m)		2,50	7,03	3,86 (1,08)	29	2,55	5,84	3,83 (0,68)	29
MMAJ	Median presence on major	1 – yes	-	-	-	16 (55,17%)	-	-	-	22 (75,86%)
		0 – no	-	-	-	13 (44,83%)	-	-	-	7 (24,14%)
MMIN	Median presence on minor	1 – yes	-	-	-	12 (41,38%)	-	-	-	19 (65,52%)
		0 – no	-	-	-	17 (58,62%)	-	-	-	10 (34,48%)
LTPMAJ	Left turn presence on major	1 – yes	-	-	-	13 (44,83%)	-	-	-	8 (27,59%)
		0 – no	-	-	-	16 (55,17%)	-	-	-	21 (72,41%)
LTPMIN	Left turn presence on minor	1 – yes	-	-	-	7 (24,14%)	-	-	-	6 (20,69%)
		0 – no	-	-	-	22 (75,86%)	-	-	-	23 (79,31%)
RTPMAJ	Right turn presence on major	1 – yes	-	-	-	10 (34,48%)	-	-	-	10 (34,48%)
		0 – no	-	-	-	19 (65,52%)	-	-	-	19 (65,52%)
RTPMIN	Right turn presence on minor	1 – yes	-	-	-	11 (37,93%)	-	-	-	13 (44,83%)
		0 – no	-	-	-	18 (62,07%)	-	-	-	16 (55,17%)
TCD	Traffic control device	1 - signals	-	-	-	15 (51,72%)	-	-	-	22 (75,86%)
		0 – all others	-	-	-	14 (48,28%)	-	-	-	7 (24,14%)

METHODOLOGY

This section provides a brief description of the characteristics of the two models used for estimating the crash prediction models: the Poisson-gamma and the COM-Poisson models, respectively. A small note on Generalized Estimating Equations (GEE) method and the goodness-of-fit (GOF) statistics are also given in later part of this section.

Poisson-gamma Model

The Poisson-gamma (or negative binomial or NB) model has the following modeling structure (*20*): the number of crashes 'Y_{it}' for a particular i^{th} site and time period t when conditional on its mean μ_{it} is Poisson distributed and independent over all sites and time periods

$$Y_{it} \mid \mu_{it} \sim Po(\mu_{it}) \qquad i = 1, 2, \ldots, I \text{ and } t = 1, 2, \ldots, T \tag{1}$$

The mean of the Poisson is structured as:

$$\mu_{it} = f(X;\beta)\exp(e_{it}) \tag{2}$$

where,

$f(.)$ is a function of the covariates (*X*);
β is a vector of unknown coefficients; and,
e_{it} is the model error independent of all the covariates.

With this characteristic, it can be shown that Y_{it}, conditional on μ_{it} and α, is distributed as a Poisson-gamma random variable with a mean μ_{it} and a variance $\mu_{it} + \alpha\mu_{it}{}^{2}$ respectively. (Note: other variance functions exist for the Poisson-gamma model, but they are not covered here since they are seldom used in highway safety studies. The reader is referred to *21* and *22* for a description of alternative variance functions. The probability density function (PDF) of the Poisson-gamma structure described above is given by the following equation:

$$f\left(y_{it};\alpha,\mu_{it}\right) = \frac{\Gamma\left(y_{it}+\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)y_{it}!}\left(\frac{\alpha^{-1}}{\mu_{it}+\alpha^{-1}}\right)^{\alpha^{-1}}\left(\frac{\mu_{it}}{\mu_{it}+\alpha^{-1}}\right)^{y_{it}} \tag{3}$$

Where,

y_{it} = response variable for observation i and time period t;
μ_{it} = mean response for observation i and time period t; and,
α = dispersion parameter of the Poisson-gamma distribution.

Note that if $\alpha \to 0$, the variance equals the mean and this model converges to the standard Poisson regression model

The term α is usually defined as the "dispersion parameter" of the Poisson-gamma distribution or model (Note: that in some published documents, the variable α has also been defined as the "over-dispersion parameter"). This term has traditionally been assumed to be fixed and a unique value applied to the entire dataset in the study. As described above, the dispersion parameter plays an important role in safety analyses, including the computation of the weight factor for the Empirical Bayes method and the estimation of confidence intervals around the gamma mean and the predicted values of models applied to a different dataset than the ones employed in the estimation process.

COM-Poisson model

The COM-Poisson distribution is a generalization of the Poisson distribution and was first introduced by Conway and Maxwell (*23*) for modeling queues and service rates. Shmueli et al. (*24*) further elucidated the statistical properties of the COM-Poisson distribution using the formulation given by Conway and Maxwell (*23*), and Kadane et al. (*25*) developed the conjugate distributions for the parameters of the COM-Poisson distribution. Its probability mass function (PMF) can be given by Equations (4) and (5).

$$P(Y=y)=\frac{1}{Z(\lambda,\nu)}\frac{\lambda^{y}}{(y!)^{\nu}} \tag{4}$$

$$Z(\lambda,\nu)=\sum_{n=0}^{\infty}\frac{\lambda^{n}}{(n!)^{\nu}} \tag{5}$$

where, Y is a discrete count; λ is a centering parameter that is approximately the mean of the observations in many cases; and, ν is defined as the shape parameter of the COM-Poisson distribution .

The COM-Poisson can model both under-dispersed ($\nu > 1$) and over-dispersed ($\nu < 1$) data, and several common PMFs are special cases of the COM with the original formulation. Specifically, setting $\nu = 0$ yields the geometric distribution; $\lambda < 1$ and $\nu \to \infty$ yields the Bernoulli distribution in the limit; and $\nu = 1$ yields the Poisson distribution. This flexibility greatly expands the types of problems for which the COM-Poisson distribution can be used to model count data.

With the original formulation, the first two central moments of the COM-Poisson distribution are given by Equations (6) and (7) below.

$$E[Y]=\frac{\partial \log Z}{\partial \log \lambda} \tag{6}$$

$$Var[Y] = \frac{\partial^2 \log Z}{\partial \log^2 \lambda} \quad (7)$$

The COM-Poisson distribution does not have closed-form expressions for its moments in terms of the parameters λ and ν. However, the mean can be approximated through a few different approaches, including (i) using the mode, (ii) including only the first few terms of Z when ν is large, (iii) bounding E[Y] when ν is small, and (iv) using an asymptotic expression for Z in Equation (1). Shmueli et al. (*24*) used the last approach to derive the approximation in Equation (8).

$$E[Y] \approx \lambda^{1/\nu} + \frac{1}{2\nu} - \frac{1}{2} \quad (8)$$

Using the same approximation for Z as in Shmueli et al. (*24*), the variance can be defined as

$$Var[Y] \approx \frac{1}{\nu} \lambda^{1/\nu} \quad (9)$$

Care should be taken in using these approximations. In particular, they may not be accurate for ν>1 or $\lambda^{1/\nu} < 10$ (*24*).

Sellers and Shmueli (*26*) derived the likelihood function for the COM-Poisson GLM where the centering parameter 'λ' is made dependent on the covariates. This derivation greatly simplifies the estimation of GLMs, as opposed to the Bayesian estimating method (*27*).

The COM-Poisson was used for three-legged intersections, since the regression analyses indicated that the modeling output of the Poisson-gamma showed signs of under-dispersion.

Generalized Estimating Equations

The GEE method was introduced as a method for handling correlated discrete data among observations that generated the data (*28*). The GEE method is an extension of the Generalized Linear Model (GLM) to enable correlated data be analyzed appropriately. The GLM method is based on the maximum likelihood theory for independent observations (*29*), whereas the GEE method is based on the quasi-likelihood theory (*30*) where the response observations may not be independent. The difference between GLM and GEE is related to the standard errors of the coefficients when the dataset does not contain missing values. The standard errors are usually underestimated when temporal effects are not included in the modeling framework (see *31* and *32* for additional information). The first-order autoregressive covariance structure was used in this study. For further information on the application of GEE method, the reader is referred to *31*.

Goodness-of-fit statistics

Different methods were presented for evaluating the GOF of the models. The two methods used in this study include the following:

Akaike Information Criterion (AIC)

The AIC is a measure of the goodness of fit of an estimated statistical model and is defined as (*33*)

$$AIC = -2\log L + 2p \tag{10}$$

Where L is the maximized value of the likelihood function for the estimated model, and p is the number of parameters in the statistical model. The AIC methodology attempts to find the model that best explains the data with a minimum of free parameters and thus it penalizes models with a large number of parameters. The model with the lowest AIC is considered to be the best model among all available models.

Quasilikelihood under the Independence model Criterion (QIC)

The QIC statistic proposed by Pan (*34*) and further discussed by Hardin and Hilbe (*32*) is analogous to the AIC statistic used for comparing models fit with likelihood-based methods. Since the GEE method is not a likelihood-based method, the AIC statistic cannot be used. The QIC is defined as (*34*)

$$QIC = Q + 2p \tag{11}$$

Where Q is the quasilikelihood value and p is the number of parameters in the statistical model. When the QIC is used to compare two models, the model with the smaller statistic is preferred.

ANALYSIS RESULTS

This section presents the estimation results of the models that were developed for three- and four-legged intersections. Flow-only models and models with covariates were developed for each of the intersection type.

Flow-Only models

Flow-only models were developed for the three- and four-legged intersections. Since disaggregated data were used in the study, the GEE were also estimated in addition to the GLM framework. These models reflect the average conditions found in the data. They can be used for cases where the user has limited information about the geometric design features for the particular project under study. For this type of model, accident modification factors (AMFs) can be used to adjust for changes in geometric design features, if they are subsequently known. However, the AMFs need to be re-calibrated or adjusted to reflect the average conditions found in the data (For detailed information about the models and AMFs, the interested reader is referred to *35*).

The functional form used for estimating flow-only models for intersections was the following:

$$\mu_{it} = \beta_0 (F_{1it} + F_{2it})^{\beta_1} (F_{2it} / F_{1it})^{\beta_2} \tag{12}$$

Where,

μ_{it} = estimated number of crashes for intersection i and year t;

F_{1it} = entering traffic flows in vehicles per day (AADT) on the major approaches for intersection i and year t;

F_{2it} = entering traffic flows in vehicles per day (AADT) on the minor approaches for intersection i and year t;

β_0, β_1, β_2 = estimated coefficients.

Table 3 summarizes the modeling results for the flow-only models for three- and four-legged intersections. As seen in this table, there is no significant difference for mean estimate of the coefficients, but the difference exists with the standard errors estimated using the GLM and GEE. Usually, GLMs underestimates the standard errors (as shown by t-values) when correlation exists in the data (see *31*). For three-legged intersection, the dispersion parameter (α) tends towards zero, which means that the data are Poisson distributed (when conditional upon the mean).

Table 3 - Estimates for Four Legged Intersections

			$Ln(\beta_0)$	F_1+F_2 (β_1)	F_2/F_1 (β_2)	α	AIC/QIC
Three-legged Intersections	GLM	Estimates	-13.8213 (1.732)†	1.3711 (0.166)	--	0.0221 (0.081)	304.1
		t-value	-7.98	8.26	--	--	
	GEE	Estimates	-13.7682 (1.995)	1.3664 (0.191)	--	0.0221 (0.081)	155.04
		t-value	-6.9	7.16	--	--	
Four-legged Intersections	GLM	Estimates	-8.0771 (1.448)	0.8872 (0.139)	0.3430 (0.163)	0.2980 (0.099)	472.4
		t-value	-5.58	6.38	2.1	--	
	GEE	Estimates	-9.0533 (2.746)	0.9746 (0.259)	0.3663 (0.191)	0.2980 (0.099)	394.43
		t-value	-3.3	3.77	1.92	--	

† Standard error

Models with covariates

For the model with covariates, the relationship between crashes and geometric design features is captured via the covariates inside the statistical model. The selection of the covariates to be included into the model can be governed by various statistical criteria, such as the statistical significance of each variable and Akaike Information Criterion (AIC), as well as the statistical significance of the coefficients.

The functional form used for estimating the models with covariates for intersections is the following:

$$\mu_{it} = \beta_0 (F_{1it} + F_{2it})^{\beta_1} (F_{2it} / F_{1it})^{\beta_2} e^{\sum_{k=3}^{n} \beta_k x_k} \tag{13}$$

Where,

μ_{it}, F_{1it} and F_{2it} are as defined above.

x_k = model covariates (e.g., right-turning lanes, median, etc.); and,

β_0, β_1, β_2, β_k = estimated coefficients.

In the initial step, a Poisson-gamma model was estimated with all the variables documented in Table 2 for three-legged intersections. Most of the variables were found to be insignificant or counterintuitive. This was attributed to small sample size problem (29*4= 116 observations). In the end, only six variables were found to be significant at the 15% confidence level. Since a slight under-dispersion was observed with the Poisson-gamma model (negative dispersion parameter in Table 4), a COM-Poisson model was used to see if it would improve the model fit (COM-Poisson outperforms NB model when the crash data are under-dispersed, for more details on this issue, the reader is referred to *36*). Due to the recent introduction of COM-Poisson model to the statistical community, the GEE modeling framework for this model type is not yet available. Thus, the standard errors are probably slightly underestimated.

Table 4 presents the estimation results for three-legged intersections. The variables that were found to be significant were the total traffic flow entering the intersection, the total number of entering lanes equal to four or five on the major direction, the total number of entering lanes equal to four or five on the minor direction, the total number of entering lanes equal to six or more on the minor direction, the average lane width on minor direction and the presence of a median in the major direction. There was no major difference in the significance of variables between NB and COM-Poisson models. When correlation was taken into account, the model output with GEE showed that there was a major difference in the significance of variables. The COM-Poisson output showed that the data (conditional upon the mean) is under-dispersed.

Table 4 - Estimates for Three Legged Intersections

Variable	GLM				GEE (NB)	
	NB		COM-Poisson			
	Estimates	t-value	Estimates	t-value	Estimates	t-value
Ln(β_0)	-13.2376 (1.8182)†	-7.28	-14.6939 (2.8917)	-5.08	-13.2061 (2.2806)	-5.79
F_1+F_2 (β_1)	1.4062 (0.1662)	8.46	1.5796 (0.3046)	5.18	1.4042 (0.1951)	7.20
LMAJT2 (β_2)	-0.2514 (0.1930)	-1.30	-0.2946 (0.2150)	-1.37	-0.2547 (0.1539)	-1.65
LMINT2 (β_3)	-0.2468 (0.3951)	-0.62	-0.2915 (0.4423)	-0.66	-0.2432 (0.1719)	-1.41
LMINT5 (β_4)	-0.5938 (0.3476)	-1.71	-0.6746 (0.3855)	-1.75	-0.5953 (0.1614)	-3.69
LWMIN (β_5)	-0.1535 (0.0902)	-1.70	-0.1809 (0.1014)	-1.78	-0.1557 (0.0985)	-1.58
MMAJ (β_6)	-0.3474 (0.2096)	-1.66	-0.3925 (0.2350)	-1.67	-0.3496 (0.1919)	-1.82
Shape parameter‡	-0.0309 (0.0561)	--	1.2169 (0.2778)	--	-0.0309 (0.0561)	--
AIC/QIC	303.1	--	302.3	--	138.03	--

† Standard error
‡ ‘α’ for NB model and ‘ν’ for COM-Poisson model

Similar to the three-legged intersections analysis, a Poisson-gamma model was developed with all the variables documented in Table 2 for the four-legged intersections. Only four variables were found to be significant at 15% confidence level. Table 5 presents the estimation results for four-legged intersections. The significant variables were the total traffic flow in the intersection, the ratio between minor direction and major direction flow, the total number of entering lanes equal to four or five on the minor direction, and the presence of a median in the minor direction. As seen in Table 5, the results by GEE showed a major difference in the significance of variables when correlation is taken into account while modeling.

Table 5 - Estimates for Four Legged Intersections

Variable	GLM (NB)		GEE (NB)	
	Estimates	t-value	Estimates	t-value
Ln(β_0)	-8.8252 (1.4565) [†]	-6.06	-9.4219 (2.5150)	-3.75
F_1+F_2 (β_1)	0.9967 (0.1442)	6.91	1.0408 (0.2346)	4.44
F_2/F_1 (β_2)	0.4322 (0.1697)	2.55	0.4178 (0.2199)	1.90
LMINT2 (β_3)	-0.6737 (0.2353)	-2.86	-0.6043 (0.2172)	-2.78
MMIN (β_4)	-0.3831 (0.1801)	-2.13	-0.2972 (0.2515)	-1.18
α	0.2390 (0.0882)	--	0.2390 (0.0882)	--
AIC/QIC	464.4	--	424.61	--

[†] Standard error

Although it is usually believed that the use of signals as traffic control devices is safer, this variable was not significant for this dataset. The same happened with the variables left/right turn lane presence (on major or minor direction). Again, this was attributed to the small sample size problem.

SUMMARY AND CONCLUSIONS

The purpose of this study was to develop crash prediction models for urban intersections located in Lisbon, Portugal which would describe the expected number of accidents as a function of a range of explanatory variables, namely traffic flow counts and highway geometric design features. Flow-only models and models with covariates were estimated using data collected at 29 three-legged intersections and 30 four-legged intersections. The coefficients were estimated using two modeling approaches: the GLM (Poisson-gamma and COM-Poisson) and the GEE. Traffic flow was found to be highly significant both for three- and four-legged intersections, as expected. Some highway design geometric variables influenced safety. For three-legged intersections, the total number of entering lanes equal to four or five on the major direction, the total number of entering lanes equal to four or five on the minor direction, the presence of a median on the major direction, the total number of entering lanes on the minor direction equal to six or more and the average lane width on minor direction were found to have a positive effect on the number of crashes occurring at three-legged intersections. For four-legged intersections, only the total number of entering lanes equal to four or five on the minor direction and the presence of a median on the minor direction were found to reduce accident occurrence.

It is recognized that a statistically significant association between crash frequency and explanatory variables does not necessarily explain a cause-effect relationship. However, this study can be considered a good starting point about gaining knowledge for better understanding the relationships between crashes and roadway characteristics in Lisbon. The overall project lead by the LNEC is still on-going and other predictive models for estimating the safety performance of roadway links and roundabouts as well as for estimating pedestrian collisions are currently under development.

REFERENCES

1. Direcção Geral de Viação. Sinistralidade rodoviária 2004 – Elementos Estatísticos. Ministério da Administração Interna. Observatório de Segurança Rodoviária, 2004.
2. Direcção Geral de Viação. Sinistralidade rodoviária 2005 – Elementos Estatísticos. Ministério da Administração Interna. Observatório de Segurança Rodoviária, 2005.
3. Direcção Geral de Viação. Sinistralidade rodoviária 2006 – Elementos Estatísticos. Ministério da Administração Interna. Observatório de Segurança Rodoviária, 2006.
4. Autoridade Nacional de Segurança Rodoviária. Ano de 2007 - Sinistralidade rodoviária. Observatório de Segurança Rodoviária, 2007.
5. Lord, D., S.P. Washington, and J.N. Ivan. Poisson, Poisson-Gamma and Zero Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. *Accident Analysis & Prevention*, Vol. 37, No. 1, 2005, pp. 35-46.
6. Lord, D., and P.Y-J. Park. Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. *Accident Analysis & Prevention*, Vol. 40, No. 4, 2008, pp. 1441-1457.
7. Hall, R. D. *Accidents at four-arm single carriageway urban traffic signals*. Contract Report 65, Transportation and Road Research Laboratory, United Kingdom, 1986.
8. Mountain, L., and B. Fawaz. Estimating accidents at junctions using routinely-available input data. *Traffic Engineering & Control*, Vol. 37, No. 11, 1996, pp. 624–628.
9. Greibe, P. Accident prediction models for urban roads. In: *Accident analysis and prevention*, Vol 35, 2003, p. 273-285,.
10. Brüde, U. and J. Larson. Models for predicting accidents at junctions where pedestrians and cyclist are involved. How well do they fit? – *Accident Analysis and Prevention*, Vol. 25, No. 5, 1993, p. 499-509.
11. Reurings, M. and T. Janssen. *Accident prediction models for urban and rural carriageways*. R-2006-14, SWOV, 2006.
12. Azeredo Lopes, S. and J.L Cardoso. *Accident Prediction Models for Portuguese Motorways*. ICT Informação Científica, LNEC, Lisbon, 2007.

13 Azeredo Lopes, S. and J.L Cardoso. *Accident Prediction Models for Portuguese Single Carriageways Roads*. ICT Informação Científica, LNEC, Lisbon, 2007.

14 Azeredo Lopes, S. and J.L Cardoso. *Accident Prediction Models for Bidirectional Data on Portuguese Motorways*. ICT Informação Científica, LNEC, Lisbon, 2007.

15 Martinez, L. - *TAZ Delineation and Information Loss in Transportation Planning Studies*. MSc Thesis, IST, Portugal, 2006.

16 Mountain, L.; M. Maher and B. Fawaz. The influence of trend on estimates of accidents at junctions. In *Accident Analysis & Prevention*, Volume 30, nº 5, 1998, p. 641-649.

17 Sayed, T. and F. Rodriguez. Accident prediction models for urban unsignalized intersections in British Columbia – In: *Transportation Research Record: Journal of the Transportation Research Board*, No 1665, 1999, p. 93-99.

18 Lord, D. The prediction of accidents on digital networks: Characteristics and issues related to the application of accident prediction models. PhD Thesis, Toronto University, 2000

19 Turner, S. A., A. P. Roozenburg and T. Francis. *Predicting Accident Rates for Cyclists and Pedestrians*. Land Transport New Zealand Research Report 289, Christchurch, New Zealand, 2006.

20 Lord, D. Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis & Prevention*, Vol. 38, No. 4, 2006, pp. 751-766.

21 Cameron, A.C., and P.K. Trivedi. *Regression analysis of count data*, Econometric Society Monograph No.30, Cambridge University Press, New York, N.Y., 1998.

22 Maher M.J., and I. Summersgill. A Comprehensive Methodology for the Fitting Predictive Accident Models. Accident Analysis & Prevention, Vol. 28, No. 3, 1996, pp. 281-296.

23 Conway, R.W, and W.L. Maxwell. A Queuing Model with State Dependent Service Rates. *Journal of Industrial Engineering*, Vol. 12, 1962, pp. 132-136.

24 Shmueli, G., T.P. Minka, J.B. Kadane, S. Borle, and P. Boatwright. A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution, *Journal of the Royal Statistical Society: Part C,* Vol. 54, 2005, pp. 127-142.

25 Kadane, J.B., G. Shmueli, T.P. Minka, S. Borle, and P. Boatwright. Conjugate analysis of the Conway-Maxwell-Poisson distribution. *Bayesian Analysis*, Vol. 1, 2006, pp. 363-374.

26 Sellers, K.F., and G. Shmueli. A Flexible Regression Model for Count Data. Robert H. Smith School Research Paper No. RHS 06-060, Georgetown University, Washington, D.C., 2008, (Available at SSRN: http://ssrn.com/abstract=1127359)

27 Guikema, S.D. and J.P. Coffelt. A Flexible Count Data Regression Model for Risk Analysis. *Risk Analysis*, Vol. 28, No. 1, 2008, pp. 213-223.

28 Liang, K., and S. Zeger. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, Vol. 73, 1986, pp. 13-22.

29 McCullagh, P., and J. A. Nelder. *Generalized Linear Models*, 2nd ed. Chapman and Hall, London, 1989.

30 Wedderburn, R. W. M. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*, 1974, 439-447.

31 Lord, D., and B. N. Persaud. Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations Procedure. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1717*, TRB, National Research Council, Washington, D.C., 2000, pp. 102-108.

32 Hardin, J.W.,&Hilbe, J. M. *Generalized estimating equations*. Boca Raton, FL: Chapman and Hall/CRC Press, 2003.

33 Akaike, 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control. vAC-19. 716-723.

34 Pan, W. (2001), "Akaike's information criterion in generalized estimating equations," *Biometrics*, 57, 120-125.

35 Lord, D., Geedipally, S.R., Persaud, B.N., Washington, S.P., van Schalkwyk, I., Ivan, J.N., Lyon, C., and T. Jonsson, Methodology for Estimating the Safety Performance of Multilane Rural Highways. NCHRP Web-Only Document 126, National Cooperation Highway Research Program, Washington, DC, 2008.

36 Lord, D., S.R. Geedipally, and S. Guikema. Extension of the Application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crash Data Exhibiting Under-Dispersion. Working paper, Texas A&M University, College Station, TX, 2009.