



PORTUGUESE TWO-LANE HIGHWAYS: MODELLING CRASH FREQUENCIES FOR DIFFERENT TEMPORAL AND SPATIAL AGGREGATION OF CRASH DATA

Jocilene Otilia da Costa¹, Maria Alice Prudêncio Jacques², Paulo António Alves Pereira³,
Elisabete Fraga Freitas⁴, Francisco Emanuel Cunha Soares⁵

^{1,3,4,5}*Territory, Environment and Research Centre (C-TAC), Dept of Civil Engineering,
University of Minho, Portugal*

²*Dept of Civil and Environmental Engineering, University of Brasilia, Brazil*

Submitted 18 July 2014; resubmitted 10 November 2014; accepted 3 March 2015;
published online 18 August 2015

Abstract. The identification of contributory factors to crash frequencies observed in different highway facilities can aid transportation and traffic management agencies to improve road traffic safety. In spite of the strategic importance of the national Portuguese road network, there are no recent studies concerned with either the identification of contributory factors to road crashes or Crash Prediction Models (CPMs) for this type of roadway. This study presents an initial contribution to this problem by focusing on the national roads NR-14, NR-101 and NR-206, which are located in Northern region of Portugal. They are two-lane single carriageway rural roads. This study analysed the crash frequencies, Average Annual Daily Traffic (AADT) and geometric characteristics of 88 two-lane road segments. The selected segments were 200-m-long and did not cross through urbanized areas. The fixed length of 200 meters corresponds to the road length used in Portugal to define a critical point. Data regarding the annual crash frequency and the AADT were available from 1999 to 2010. Due to the high number of zero-crash records in the initial database, the data were explored to identify the best statistical modelling approach to be adopted. The Generalized Estimating Equations (GEE) procedure was applied to 10 distinctive databases formed by grouping the original data in time and space. The results show that the different observations within each road segment present an exchangeable correlation structure type. This paper also analyses the impact of the sample size on the model's capability of identifying the contributing factors to crash frequencies. The major contributing factors identified for the two-lane highways studied were the traffic volume (expressed in AADT), lane width, vertical sinuosity, and Density of Access Points (DAP). Acceptable CPM was identified for the highways considered, which estimated the total number of crashes for 400-m-long segments for a cumulative period of two years.

Keywords: crash contributory factors; generalized estimating equations; crash prediction models; two-lane highways; longitudinal data.

Introduction

The increasing number of traffic crashes in rural and urban roadways has created the need to develop strategies to help highway agencies reducing these events. The World Health Organization (WHO 2013) revealed that more than 1.2 million people die and 50 million people are injured on the world's roads every year and in Portugal, these figures for 2011 were 689 and 42162, respectively, for a country with a population of 10562178.

Properly registered and analyzed road crash data allows for the identification of the areas or sites where safety measures have a greater potential for success and

effectiveness (Thomas *et al.* 2003). The definition of effective safety measures, which are chosen to cope with the necessary reduction of road traffic crashes in a given roadway facility (segment or intersection), benefits from knowledge regarding estimates of the facility's expected crash frequency along with the contribution of facility's physical and operational characteristics to the expected safety performance. The common approaches that are used to provide current and future safety performance estimates of roadway segments or intersections are historical crash data, statistical models based on regression analysis, before-after studies, and expert judgments (Harwood *et al.* 2000).



Crash Prediction Models (CPMs) are important tools in promoting traffic safety in different roadway facilities. They can provide accurate estimates for the total crash frequency for a location per unit of time, which is usually a function of the roadway's traffic and geometric characteristics. However, the interpretation of a CPM coefficient as the true effect of an incremental change in an associated roadway feature is not usually satisfactory (Harwood *et al.* 2000; Hauer 2004). According to the referenced authors, this situation can arise due to problems such as (i) the cause-effect assumed between some roadway characteristics and crashes may not always be true; (ii) the presence of a strong correlation among the model's independent variables; and (iii) the lack of important explanatory variables in the model, which causes the coefficient of one or more variables in the model to represent the unavailable variable rather than their own effect. Even with these limitations, the analysis of the signs and the identification of the relative significance of the independent variables presented in the model can lead to important insights regarding the major contributory factors to road safety.

The development of CPMs is based on random, discrete, nonnegative, and highly over dispersed data. Additionally, in some cases, the available data may present temporal or spatial correlations, which impose specific statistical considerations for the model development (Wang, Abdel-Aty 2006). A comprehensive analysis of the data and the methodological issues regarding the development of analytic approaches to study the factors related to road crashes can be found in the work of Lord and Mannering (2010). Additionally, some modelling difficulties imposed by databases with many records of zero crashes have caused the usage of different statistical modelling approaches for CPM development, which are not easy to justify from a traffic engineering perspective (Lord *et al.* 2005a, 2007). When the use of these approaches cannot be justified and the database is formed by temporal and/or space-related data, it is possible to develop the CPM based on aggregated number of crash observation through some time period or given space. However, the impact of this aggregation on the identification of the significant explanatory variables to the observed crash frequency need to be further investigated.

Due to the intrinsic characteristics of crash data, CPMs are commonly developed using the Poisson and negative binomial regression models (Joshua, Garber 1990; Poch, Mannering 1996; Milton, Mannering 1998; Lord, Bonneson 2007; Gomes *et al.* 2012). The form of the CPM usually consists of the product of the exposure measures' powers multiplied by an exponential term related to the other explanatory variables. Models containing additive terms have also been referred to in the literature, where the additive component aims to account for the influence of hazard points (Hauer 2004; Caliendo *et al.* 2007). The model without the additive term, which is the focus of this work, allows for coefficient estimation using the Generalized Linear Model (GLM). Other modelling approaches can be found in

the literature, as is the case of the generalized additive models (Zhang *et al.* 2012).

An important issue related to the model's parameters is whether one can assume that the parameters vary across observations. For the purpose of crash frequency modelling, it is usually assumed that they are constant from year to year (fixed-parameters models), as referred to by Lord and Persaud (2000) in a study regarding a comparative analysis between fixed-parameters and random intercepts models (only the intercept varies across years and the other parameters remain fixed). They indicate some advantages of the latter models especially for applications as before-and-after studies and traffic crashes trend analysis. Other papers on investigating random-parameters models can be found in the literature (Anastasopoulos, Mannering 2009, 2011; El-Basyouny, Sayed 2009; Anastasopoulos *et al.* 2012; Venkataraman *et al.* 2013). Although this modelling approach is promising, for the purpose of the present paper only fixed-parameters model will be considered.

The crash records can be grouped by taking into account the period of the observation (usually, but not necessarily, the year) or clustered according to some spatial or other characteristics observed at a specific time period (see the work of Wang, Abdel-Aty 2006). The observation within these groups may or may not be statistically independent from each other, but the groups are independent among themselves. The existence and type of correlation within the groups of data that form the entire database is extremely relevant when determining the complexity of the parameters' estimation using different model types.

Observations in a group present a temporal correlation when the data are collected over successive time periods for the same sample element and could, therefore, share unobserved effects. For example, the crash frequency data present temporal correlation when the same roadway facilities (segments or intersections) are observed over time. In this case, each subject of analysis constitutes a group, and the repeated observations made at different years are the elements of that group and form longitudinal crash data (panel data).

When the data within groups are not correlated, the GLM can be applied. In that case, the research of the effect of the data grouping factor (e.g. space or time) on the response variable may be done by including dummy variables in the model (Wooldridge 2010). For the case of longitudinal data, the procedures for estimating the parameters are derived from the GLM procedure. For non-Gaussian outcomes, the procedures may be separated into a marginal model family, which are the Generalized Estimating Equations (GEE) (Lord, Persaud 2000; Lord *et al.* 2005a; Wang, Abdel-Aty 2006; Lord, Mahlawat 2009), as one example, and a random-effects family, a Generalized Linear Mixed Model (GLMM) that is a more complex family of models (Vangeneugden *et al.* 2011). The latter procedure will not be considered in the present work.

The GEE procedure was developed by Liang and Zeger (1986) as an extension of the GLM for the analy-

sis of longitudinal data when the primary focus of the analysis is the dependence on the model's response to the explanatory variables. The GEE can be applied for both Gaussian and non-Gaussian response variables and is a general method for analyzing clustered data where the following are true (Halekoh *et al.* 2006): (i) observations within a cluster may be correlated; (ii) observations in separate clusters are independent; (iii) a monotone transformation of the expectation is linearly related to the explanatory variables; and (iv) the variance is a function of the expectation. Regarding the correlation among the observations in a given cluster, the GEE allows for different choices including the non-correlation condition.

The main objective of this study is to identify the factors that contribute to fatal and injury crash frequency for road segments on Portuguese national roads NR-14, NR-101 and NR-206, which are located in the Northern region of Portugal.

Additionally, this study aims at analysing the impact of different database structures in time and space on the factors identified. The type of correlation within the data related to each road segment will also be evaluated using the GEE equations procedure for modelling the longitudinal data. It is expected that the identification of the correlation type (other than the 'independent' correlation) may call other modellers' attention for the fact that this data characteristic must be evaluated in advance to the selection of the method to estimate the coefficients of CPMs.

Several studies designed to predict accident frequency in two-lane single carriageway rural roads can be found in the international literature (Harwood *et al.* 2000; Cafiso *et al.* 2010; Dinu, Veeraragavan 2011). However, applying them to realities different to that in which their results were obtained, in this case, the roads of Portugal, would not be recommendable.

The importance of studying crashes to improve road safety in the Portuguese northern national road system is justified by the fact that these roads serve a high concentration of cities and industrial zones. Despite its importance, recent studies concerning the promotion of road safety in those roads are scarce. Among

these studies, the work of Gomes and Cardoso (2012) must be highlighted. These authors studied the impact of low-cost engineering measures on the decrease of accidentalness in some stretches of a multilane national road, the NR-6. As for the identification of contributory factors for road crashes in Portuguese road system, the main focus of recent studies made in the country has been on the elements of urban roads (Couto, Ferreira 2011; Gomes *et al.* 2012; Gomes 2012).

Taking into account the theoretical aspects previously addressed and to reach its objectives, the study described in this paper started with the complete characterization of the two-way highway road segments to be analysed. Following, the CPM model formulation and assessment is described, as well as the major results of the study are presented and discussed. Finally, the summary and conclusions of the research are presented.

1. Data Description

The initial traffic crash database includes the number of fatal and injury crashes, volume and geometric characteristics data on two-lane 200-m-long highway road segments belonging to sections of highways located in Northern region of Portugal. This fixed length of 200 meters corresponds to the road length used in Portugal to define a critical point. None of the segments selected for the study has intersections or has stretches that lies within the area of influence of intersections. The area of influence of an intersection was defined as a circle of radius 76 meters with its centre at the intersection in accordance with the criteria adopted by the Crash Prediction Module – CPM (FHWA-TFHRC 2003). Thus, all the accidents unnecessary in this paper refer to occurrences in road segments only. The data are available for the years 1999 to 2010, and do not include pedestrian and cyclist crashes. The highways and respective links considered are presented in Table 1.

In Portugal, according to the National Authority for Road Safety (*Autoridade Nacional de Segurança Rodoviária*), the analysis of critical safety points for two-lane highways must be based on 200-m-long road segments (ANSR 2009). Therefore, the links considered were divided into 200-m-long fixed segments for which

Table 1. Road 200-m-long segments considered for the CPMs development

NR	Link	Length studied ^(a) [km]	Number of segments			
			Influenced by road junctions ^(b)	Studied [m]		
				200	400	600
14	Braga–Famalicão	17.60	75	13	4	0
	Braga–Guimarães	17.60	66	22	8	4
101	Braga–Vila Verde	8.80	39	5	1	1
	Guimarães–Felgueiras	17.80	61	28	11	5
206	Famalicão– Guimarães	19.20	81	15	6	2
	Guimarães–Fafe	6.40	27	5	2	1
	Total	89.40	349	88	32	13

Notes: ^(a) considers only two-lane extensions not located within urban areas; ^(b) 200-m-length segments presenting road junctions or containing part of a given road junction approach.

the geometric characteristics, traffic flow (expressed in Average Annual Daily Traffic – AADT) and number of crashes from the years 1999 to 2010 were registered. Some segments of these links were not included in the sample studied because they present one or more characteristics that do not fit the purpose of this study. These characteristics are the following: (i) more than two-lanes; (ii) cross through urbanized areas; and (iii) contain road junctions or portions of junction approaches (with roadways for accessing cities or with interchanges for the national expressway system). Because of these criteria, only eighty-eight 200-m-long segments are present in the initial (more disaggregated) database, as shown in Table 1. To study alternative spatial variations on crash data collection reference (by means of defining different lengths for the segments considered), the original 200-m-long segments were grouped into consecutive units of 2 or 3 segments, which formed 400-m-long or 600-m-long segments, respectively. As shown in Table 1, the sample size with this grouping strategy turned out to be very small, especially for the case of the 600-m-long segments, which will not be considered in this paper.

1.1. Geometric Characteristics

For the purpose of this study, the following geometric characteristics of each segment were considered:

- Lane Width (LW);
- Shoulder Width (SW);
- Lateral Offset (LO);
- Rate of the length in horizontal tangent per total segment length (RSL) calculated by:

$$RSL = \frac{\sum(r_i)}{SL}, \quad (1)$$

where: r_i is the length of straight line i (tangent in horizontal alignment); SL is the length of the segment considered;

- Horizontal Sinuosity (HS), calculated by:

$$HS = \frac{\sum(|\Omega_i|)}{SL}, \quad (2)$$

where: Ω_i is the road alignment curvature at the horizontal curve i (in degrees); SL is the length of the segment considered;

- Rate of the length in vertical tangent per total segment length (RTL) calculated by:

$$RTL = \frac{\sum t_i}{SL}, \quad (3)$$

where: t_i is the length of straight line i (tangent in vertical alignment); SL is the length of the segment considered;

- Vertical Sinuosity (VS), calculated by:

$$VS = \frac{\sum(|\Delta_i|)}{SL}, \quad (4)$$

where: Δ_i is the algebraic difference in grades (in percentage) observed at the sag or crest vertical curve i ;

- Density of Pedestrian Crossings (DPC), which is defined as the number of pedestrian crossing facilities per segment;
- Density of Access Point (DAP), which is calculated as the number of accesses to private properties (and/or to secondary roadways without exits) per segment.

The geometric data were collected in the field, and some statistics related to the observed values for the 200-m-long segments are presented in Table 2. It is important to highlight that these characteristics were treated as initial explanatory variables for the observed crash frequency for each road segment.

Table 2. Descriptive statistics of the segments' geometric characteristics

Variables	Units	Mean	Standard Deviation (SD)	Min	Max
LW	m	3.66	0.18	3.05	4.23
SW	m	0.74	0.56	0.10	3.45
LO	m	2.32	1.14	0.60	5.68
RSL		0.62	0.26	0.04	1.00
HS	degree/m	300.01	280.46	0.00	1491.50
RTL		0.85	0.31	0.00	1.00
VS	%/m	4.63	8.20	0.00	25.35
DPC	n°/segment	0.15	0.35	0.00	1.00
DAP	n°/segment	4.01	2.93	0.00	15.00

1.2. Traffic Data

The traffic data considered for the model development were the AADT per road segment considering each year of the twelve years in the crash database (from 1999 through 2010).

The AADT was estimated for each year based on the traffic counts available in the Portuguese Roadway, EP (*Estradas de Portugal*), database for the highways studied for the years 2003 to 2008. Besides these data, a direct estimate was made for the volume of traffic in those highways for the year 2011 using an adaptation of EP's own methodology, made available by the company for the conduct of the research. The historical series for each highway for the period between 1999 and 2010 was obtained through a trend analysis made with the available annual data.

To estimate the AADT of each highway in 2011 and also to attain the historical series for all portions in the time period between 1999 and 2010, specific 14-hour counts made in 2011 at 31 locations, covering the roadway extensions in this study were also considered (10 of these locations belong to the set of locations considered by the EP database).

To determine AADT's historical series for each segment, 19 out of 31 collection points were put into place in sites that precede the 200 meters segments (whether they're isolated and successive) belonging to different stretches of the studied NR's (9 points in NR-101,

5 points in NR-206 and 5 points in NR-14). For each point, the AADT was estimated for 2011 and the proportion of that volume was calculated with respect to the AADT of the NR in which it was included (it was estimated for that same year). This proportion was considered to be constant in all years of the studied period and it was then used to estimate the AADT in the point for each year from the correspondent highway's AADT. With the historical series of those 19 points, the historical series of AADT for all the studied road segments were obtained. The AADT of each segment was estimated based on the AADT data of the closest (located in the same roadway) point. Therefore, many segments present the same AADT, which may negatively impact the calibration of the developed CPMs. The AADT values varied from 2165 to 32857 vehicles, with mean and standard deviation of 12936 and 6323, respectively.

1.3. Crash Data

The crash data for this study were provided by the National Authority for Road Safety, ANSR (2009, *Autoridade Nacional de Segurança Rodoviária*), and cover the period from 1999 to 2010. The ANSR maintains a database with information gathered from the Traffic Crash Registration Form, BEAV (*Boletim Estatístico de Acidentes de Viação*), which is filled out at the time of the crash.

The initial database, which has 12 records for each of the eighty-eight 200-m-long segments, is formed using 1056 records, of which 740 recorded zero crashes. That is, the database is zero inflated. Therefore, a preliminary analysis aimed at verifying whether there are plausible reasons for these zeros was performed. This analysis could indicate the convenience of using a zero-inflated regression model (Zero-Inflated Poisson – ZIP, or Zero-Inflated Negative Binomial – ZINB) (Shankar *et al.* 1997; Carson, Mannering 2001; Kumara, Chin 2003; Lord *et al.* 2005b). Taking into account the total number of traffic crashes registered per 200-m-long segment during the overall analysis period (12 years), the frequency distribution of the total number of crashes per segment was determined and is presented in Table 3.

The main characteristics of the four segments with zero crashes were analyzed against correspondent characteristics observed at the nine segments presenting 10 or more crashes. This analysis reveals that, based on the characteristics listed in Table 2 and on the traffic volume levels, there is no technical justification for assuming that the zero-crash segments can be treated as potentially safe segments. Table 4 presents a synthesis of those characteristics for two kinds of segment and it can be seen that there are no discrepancies between them. Furthermore, all segments show similar characteristics in regard to the environment bordering the highway.

The zero-inflated regression models assume that the elements being studied are in two different states – safe (not subject to accidents) and liable to accidents – and that the two need to be modelled separately (Carson, Mannering 2001; Kumara, Chin 2003).

Based on the analysis it was decided that the use of a zero-inflated regression model would be unsuitable for

the data in hand. In this case, as recommended by Lord *et al.* (2005b, 2007), alternative time periods for aggregating the number of crashes were considered for modelling, as a means to reducing the number of records (observations) with zero crashes, as shown in Table 5.

Table 3. Frequency distribution of the number of crashes per segment in 12 years – 200-m-long segments

NC	0	1	2	3	4	5	6	7	8	9	10	11	12	16	22
NS	4	12	12	10	14	11	4	2	3	7	2	4	1	1	1

Notes: NC – Number of Crashes; NS – Number of Segments.

Table 4. Characteristics of segments with zero and with more than nine accidents

Variables	Number of crashes = 0		Number of crashes ≥10	
	Max	Min	Max	Min
AADT	27903.58	7466.17	27903.58	5067.33
LW	3.75	3.45	4.23	3.53
SW	0.70	0.31	1.73	0.50
LO	1.90	0.70	3.90	0.60
RSL	0.94	0.45	1.00	0.22
HS	336.30	65.60	853.20	0.00
RTL	1.00	1.00	1.00	0.00
VS	0.00	0.00	25.35	0.00
DPC	0.00	0.00	1.00	0.00
DAP	5.00	0.00	15.00	0.00

Notes: AADT – Annual Average Daily Traffic; LW – Lane Width; SW – Shoulder Width; LO – Lateral Offset; RSL – rate of the length in horizontal tangent per total segment length; HS – Horizontal Sinuosity; RTL – rate of the length in vertical tangent per total segment length; VS – Vertical Sinuosity; DPC – Density of Pedestrian Crossings; DAP – Density of Access Point.

Table 5. Number of Zero (ZR) and Total Crash (TR) records for different time observation periods

Item	Type of records	Time observation period [in years]					
		1	2	3	4	6	12
200-m-long	ZR	740	273	142	90	37	4
	TR	1056	528	352	264	176	88
	[%] of ZR	70%	52%	40%	34%	21%	5%
400-m-long	ZR	195	63	30	20	4	0
	TR	384	192	128	96	64	32
	[%] of ZR	51%	33%	23%	21%	6%	0%

2. Methodology

The CPM for the Portuguese two-lane highway segments was developed using the GEE with the negative binomial link function.

The verification of the presence and type of correlation structure in the longitudinal crash data was performed by testing the following correlation structures

provided by the GEE procedure: independent, exchangeable, and autoregressive. These structures allow for the specification of the correlation between the number of crashes measured at a given road segment for different periods of time. The correlation structures have been proposed by Liang and Zeger (1986) and are fairly well detailed in the work of Wang and Abdel-Aty (2006).

2.1. Model Formulation

Given the objective of this study, the analysis considers only models with the general expression presented in Eq. (5); the GLM version (Eq. (6)) is derived from this equation:

$$E(y_{mt}) = \alpha \left(\left(\text{Volume}_{mt} \right)^\gamma e^{\sum_j \beta_j x_{j,mt}} \right); \quad (5)$$

$$\ln(E(y_{mt})) = \ln(\alpha) + \gamma \ln(\text{Volume}_{mt}) + \sum_j \beta_j x_{j,mt}, \quad (6)$$

where: $E(y_{mt})$ – the expected number of crashes at segment m over time period t ; Volume_{mt} – AADT observed at segment m over time t ; $x_{j,mt}$ – value of explanatory variable i observed at segment m over time t ; α , γ , β_j – model parameters to be estimated.

The modelling procedure followed a backward elimination starting with the AADT and all candidate variables (presented in Table 2). The final model for each combination of segment length and time observation period, which considered the three correlation structures provided by the GEE, present only the explanatory variables that are statistically significant at 5% significance level.

The identification of the factors affecting the frequency of crashes defined by the combinations of time and space was based on the model that best fit the field data. The best model for a given combination was selected based on the conditions presented in Section 2.2.

The overall best model was selected with the Cumulative REsiduals (CURE) method and with the marginal R^2 (R_{marg}^2). The Quasi-likelihood Information Criterion (QIC) statistic was considered for the correlation structure evaluation. Another important consideration for model selection was the analysis of the model parameters' sign. The parameters' sign must be compatible with the expectation from a traffic engineering point of view.

2.2. Model Assessment

Three elements were considered for examining the goodness-of-fit for each CPM generated: the CURE method, the R_{marg}^2 and the Akaike's Information Criterion (AIC) in the GEE, which is called the QIC.

The CURE test considers the difference between the number of observed and predicted crashes (the residual) as the basic element for judging the CPM fit (Hauer 2004).

The CURE plot allows for the examination of the CUREs against the variable of interest, which is the Volume_{it} (AADT observed at segment i over time t) for

the present study. For this examination, the N residuals are initially sorted in increasing order of the variable of interest. They are then numbered consecutively 1, 2, ..., z , ..., N . The CURE for each z is calculated as the sum of the residuals from 1 to z . A good fit means that the CURE plot oscillates around the zero value of the CUREs.

Additionally, the CURE plot presents two additional curves formed by acceptable limits for the CUREs, which are built as follows:

- compute the squared residual for each of the N sorted residuals;
- compute $\hat{\sigma}^2(z)$ as the sum of these squared residuals from 1 to z ;
- compute:

$$\sigma^*(z) = \hat{\sigma}(z) \sqrt{1 - \frac{\hat{\sigma}^2(z)}{\hat{\sigma}^2(N)}}, \quad (7)$$

where: $\sigma^*(z)$ – CUREs from 1 to z , used to the CURE plot; $\hat{\sigma}(z)$ – square root of the sum of squared residuals from 1 to z ($\hat{\sigma}^2(z)$); $\hat{\sigma}^2(N)$ – sum of squared residuals from 1 to N ;

- add $\pm 2\sigma^*(z)$ limits to the CURE plot.

The marginal R^2 was introduced by Zheng (2000) for models developed by GEE method. It takes the form:

$$R_{marg}^2 = 1 - \frac{\sum_{t=1}^T \sum_{i=1}^n (Y_{it} - \hat{Y}_{it})^2}{\sum_{t=1}^T \sum_{i=1}^n (Y_{it} - \bar{Y})^2}, \quad (8)$$

where: Y_{it} – the observed number of crashes at element i over time period t ; \hat{Y}_{it} – the expected number of crashes at element i over time period t ; \bar{Y} – the marginal mean,

given by: $\bar{Y} = \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (Y_{it})$.

To assess the best correlation structure between data of each grouping strategy (spatial and temporal) the QIC statistic was used. This statistic was proposed by Pan (2001) as a modification for the AIC in the GEE. The modification was developed to address a model selection problem in the GEE concerning the selection of the type of correlation among observations in a given cluster (working correlation structure). The modification involves using the quasi-likelihood constructed from the estimating equations (QIC) using the working independence model and any general working correlation structure in the GEE. To select a working correlation structure in the GEE, it is necessary to calculate the QIC for various candidate working correlation structures. According to Ballinger (2004) the correlation structure to be adopted is the structure that produces the smallest (closest to zero) QIC.

3. Results and Discussion

The main results for the models generated in the cases of the 200-m-long and 400-m-long segments are presented in Tables 6 and 7, respectively. From Table 6, it

Table 6. Model estimates for 200-m-long road segments

Time	Parameters	Correlation structure					
		Ind		Ex		Ar	
		Coeff.	<i>p</i>	Coeff.	<i>p</i>	Coeff.	<i>p</i>
1-year	Intercept	-7.13	0.01	-7.59	<0.01	-7.13	0.01
	ln(AADT)	0.35	0.06	0.38	0.04	0.35	0.06
	LW	0.68	0.07	0.72	0.06	0.68	0.07
	VS	0.03	0.02	0.03	0.02	0.03	0.02
	DAP	0.07	0.01	0.07	0.01	0.07	0.01
	QIC	1558.45		1554.80		1561.56	
	R^2_{marg}	0.04		0.04		0.04	
Number of observations in the database = 1056							
2-years	Intercept	-6.57	0.01	-7.08	0.01	-6.85	0.01
	ln(AADT)	0.37	0.04	0.40	0.03	0.38	0.04
	LW	0.67	0.07	0.71	0.06	0.72	0.06
	VS	0.03	0.03	0.03	0.03	0.03	0.02
	DAP	0.07	0.01	0.07	<0.01	0.07	0.01
	QIC	989.90		987.35		999.18	
	R^2_{marg}	0.07		0.07		0.07	
Number of observations in the database = 528							
3-years	Intercept	-5.99	0.02	-6.30	0.01	-6.24	0.01
	ln(AADT)	0.36	0.05	0.38	0.03	0.37	0.04
	LW	0.64	0.08	0.67	0.08	0.69	0.07
	VS	0.03	0.03	0.03	0.03	0.03	0.03
	DAP	0.07	<0.01	0.07	<0.01	0.07	<0.01
	QIC	631.95		630.19		640.35	
	R^2_{marg}	0.09		0.09		0.09	
Number of observations in the database = 352							
4-years	Intercept	-5.99	0.02	-6.30	0.01	-6.24	0.01
	ln(AADT)	0.36	0.05	0.38	0.03	0.37	0.04
	LW	0.64	0.08	0.67	0.08	0.69	0.07
	VS	0.03	0.03	0.03	0.03	0.03	0.03
	DAP	0.07	<0.01	0.07	<0.01	0.07	<0.01
	QIC	394.44		393.74		396.54	
	R^2_{marg}	0.10		0.10		0.10	
Number of observations in the database = 264							
6-years	Intercept	-5.60	0.02	-5.86	0.02	-5.76	0.02
	ln(AADT)	0.36	0.05	0.38	0.04	0.36	0.04
	LW	0.62	0.09	0.64	0.09	0.65	0.08
	VS	0.03	0.03	0.03	0.03	0.03	0.03
	DAP	0.07	<0.01	0.07	<0.01	0.07	<0.01
	QIC	41.25		41.24		41.24	
	R^2_{marg}	0.13		0.13		0.13	
Number of observations in the database = 176							

Notes: AADT – Annual Average Daily Traffic; LW – Lane Width; VS – Vertical Sinuosity; DAP – Density of Access Point; QIC – Quasi-likelihood Information Criterion statistic.

can be observed that according to the QIC parameter, the correlation structure that best fits the longitudinal data considered is the exchangeable correlation, according to which the correlations between any two observations within a group is constant. Table 7 shows that for the 400-m-long segments the exchangeable correlation is also found for the 1-year, 2-years and 4-years models, while the independent correlation best fits for the

3-years and 6-years models, respectively. That is, the independence correlation, which allows the longitudinal data to be treated as independent records with the aid of the basic GLM procedure, is only suitable for the current database for the 400-m-long segments over 3-years and 6-years time period.

Another important finding is that, as expected, a small number of observations (records) considered for

model calibration can prevent the generation of an acceptable CPM. For the cases involving the 400-m-long segments and 3, 4 and 6-years of cumulative data, the explanatory variable HS was determined to have a negative coefficient value. This result is not expected from a

traffic engineering point of view. When this variable is excluded from the models (which is not justified from a statistical point of view) and new models are generated, only LW and traffic volume variables remain statistically significant. For the case of the 200-m-long segments the

Table 7. Model estimates for 400-m-long road segments

Time	Parameter	Correlation structure					
		Ind		Ex		Ar	
		Coeff.	p	Coeff.	p	Coeff.	p
1-year	Intercept	-12.69	0.01	-12.90	0.01	-12.52	0.01
	ln(AADT)	0.76	0.01	0.77	<0.01	0.75	0.01
	LW	1.31	0.01	1.34	0.01	1.30	0.01
	VS	0.03	0.01	0.03	0.01	0.03	0.02
	DAP	0.08	0.02	0.08	0.02	0.07	0.03
	QIC	653.54		653.10		658.52	
	R^2_{marg}	0.11		0.11		0.10	
Number of observations in the database = 384							
2-years	Intercept	-11.93	0.01	-12.14	0.01	-11.89	0.01
	ln(AADT)	0.76	0.01	0.77	0.01	0.74	0.01
	LW	1.30	0.01	1.32	0.01	1.34	0.01
	VS	0.03	0.01	0.03	0.01	0.03	0.01
	DAP	0.08	0.03	0.08	0.03	0.07	0.04
	QIC	247.50		247.18		256.00	
	R^2_{marg}	0.18		0.18		0.18	
Number of observations in the database = 196							
3-years	Intercept	-11.93	0.01	-11.97	<0.01	-12.41	0.01
	ln(AADT)	0.73	0.01	0.74	<0.01	0.74	0.01
	LW	1.57	0.04	1.57	0.04	1.69	0.03
	HS	-0.01	0.03	-0.01	0.03	-0.01	0.03
	VS	0.02	0.02	0.02	0.02	0.02	0.02
	DPC	0.72	0.03	0.72	0.03	0.73	0.03
	QIC	13.96		14.02		14.51	
R^2_{marg}	0.19		0.19		0.19		
Number of observations in the database = 128							
4-years	Intercept	-11.70	0.01	-11.70	0.01	-11.89	0.01
	ln(AADT)	0.74	0.01	0.74	0.01	0.75	0.01
	LW	1.57	0.04	1.57	0.04	1.61	0.03
	HS	-0.01	0.02	-0.01	0.02	-0.01	0.02
	VS	0.02	0.02	0.02	0.02	0.02	0.02
	DPC	0.73	0.03	0.73	0.03	0.73	0.03
	QIC	-178.53		-178.51		-179.20	
R^2_{marg}	0.21		0.21		0.21		
Number of observations in the database = 96							
6-years	Intercept	-11.67	<0.01	-11.55	<0.01	-11.55	0.01
	ln(AADT)	0.77	0.01	0.76	<0.01	0.76	0.01
	LW	1.61	0.03	1.60	0.03	1.60	0.03
	HS	-0.01	0.02	-0.01	0.02	-0.01	0.02
	VS	0.02	0.02	0.02	0.02	0.02	0.02
	DPC	0.73	0.02	0.73	0.03	0.73	0.03
	QIC	-354.85		-355.50		-355.50	
R^2_{marg}	0.27		0.27		0.27		
Number of observations in the database = 64							

Notes: AADT – Annual Average Daily Traffic; LW – Lane Width; HS – Horizontal Sinuosity; VS – Vertical Sinuosity; DPC – Density of Pedestrian Crossings; DAP – Density of Access Point; QIC – Quasi-likelihood Information Criterion statistic.

results show that the models generated for sample sizes equal to or less than 128 observations present problems for properly identifying the major contributory factors to the crash frequencies observed in the field.

The models acceptable from both a statistical and traffic engineering point of view are those generated from the databases with 200-m-long road segments and 3-years and 4-years time periods, and with 400-m-long road segments and 1-year and 2-years time periods. The latter two models best fit the corresponding data as their R^2_{marg} is greater than those found for the 200-m-long databases' models. Additionally, all of their explanatory variables are statistically significant for $\alpha = 5\%$, while for the 200-m-long segments' models the variable LW is only statistically significant for $\alpha = 7.6\%$ (3-years model) and $\alpha = 8.5\%$ (4-years model).

Considering only the models acceptable it is possible to verify that among the contributing variables studied, the major contributing factors to the crash frequency are the traffic volume, expressed in terms of AADT, the LW, the VS and the DAP. All of these variables have a positive impact on the dependent variable (coefficients with positive sign). One important aspect to highlight is that the LW in the database varies from 3.05 to 4.23 m. What the results show, therefore, is that in this range, larger traffic lanes can have a negative effect on traffic safety, which is probably due the reduction in the lane's capability of helping the vehicles maintain similar trajectories. As a final evaluation of the previously considered acceptable models, the CURE plot for each case was developed. Figs 1 and 2 show these plots.

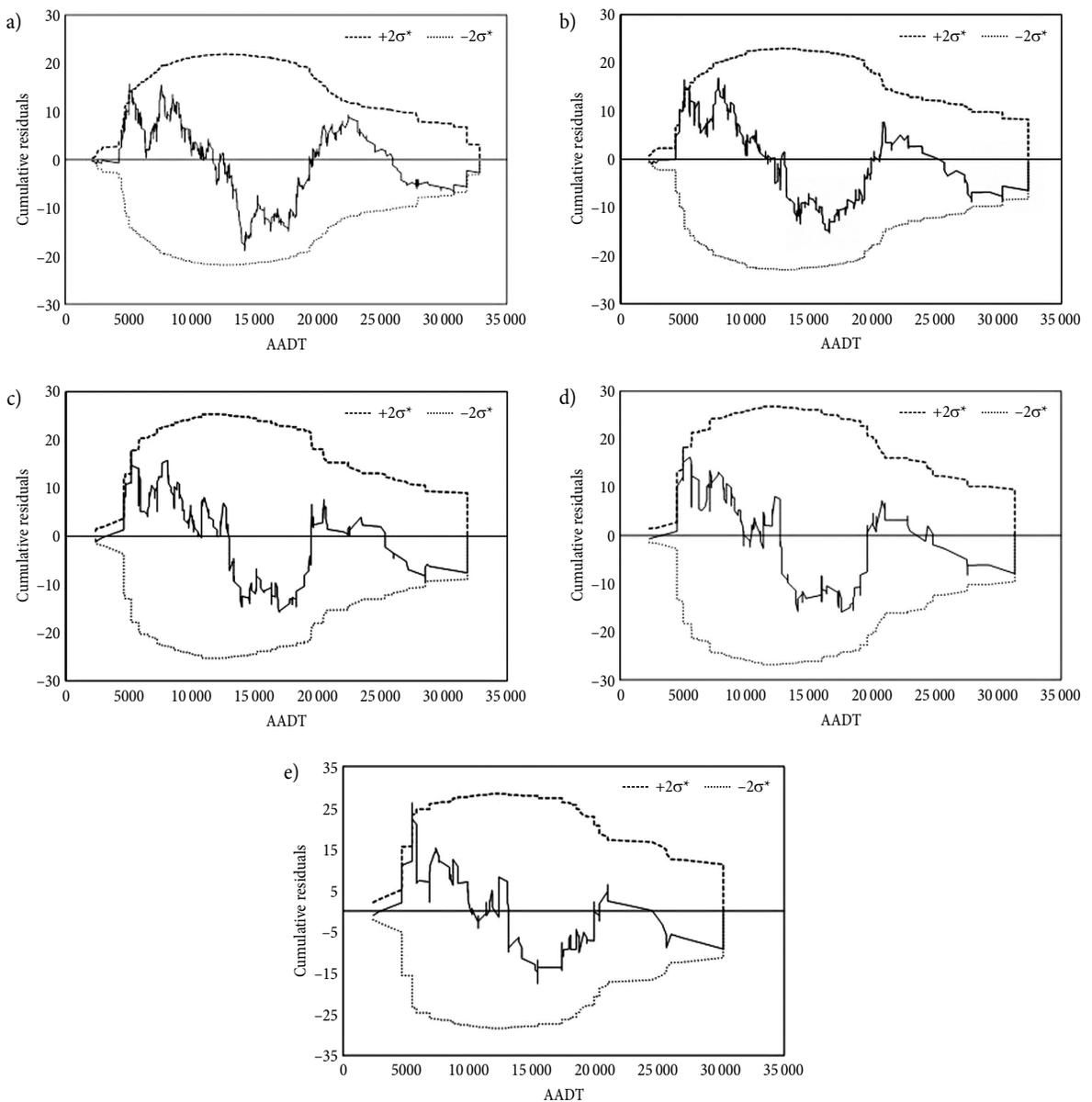


Fig. 1. CURE plots for the models developed for the 200-m-long segments: a – 1-year crash data (ZR = 70%; TR = 1056); b – 2-years crash data (ZR = 52%; TR = 528); c – 3-years crash data (ZR = 40%; TR = 352); d – 4-years crash data (ZR = 34%; TR = 264); e – 6-years crash data (ZR = 21%; TR = 176)

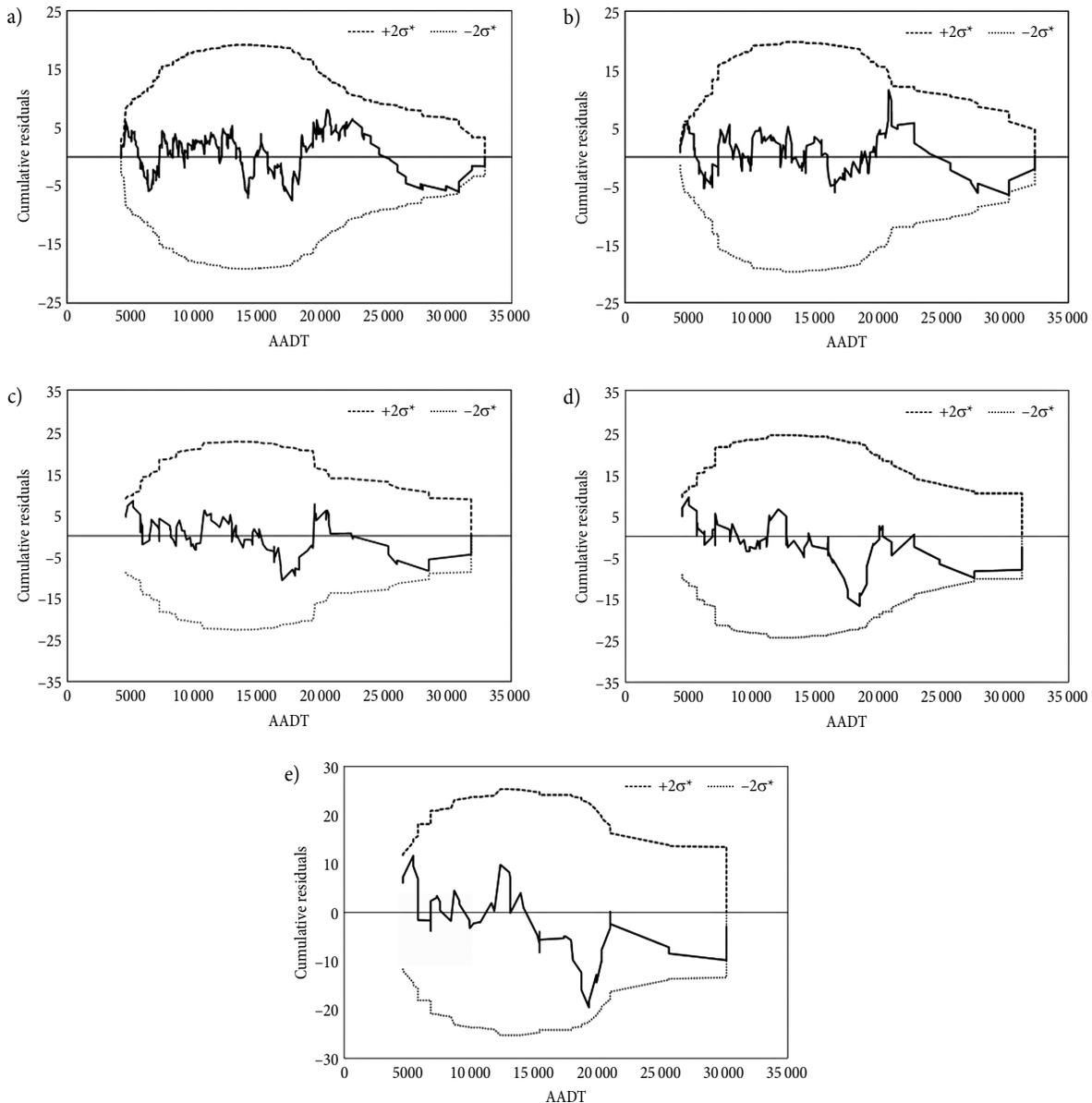


Fig. 2. CURE plots for the models developed for the 400-m-long segments: a – 1-year crash data (ZR = 51%; TR = 384); b – 2-years crash data (ZR = 33%; TR = 192); c – 3-years crash data (ZR = 23%; TR = 128); d – 4-years crash data (ZR = 21%; TR = 96); e – 6-years crash data (ZR = 6%; TR = 64)

From Fig. 1, it is possible to observe that for the 3-years and 4-years crash data models, the CURE plots oscillate around 0 and do not cross the upper and lower limits. Therefore, the models can be considered acceptable. However, their CUREs present points very close to the upper limit. Although the 3-years model presents R^2_{margin} slightly smaller than the corresponding for the 4-years model, the level of significance of its variables is greater. Additionally, for practical applications, a model that can estimate the number of crashes for shorter time period seems to be more convenient. Because of this, the 3-years model can be considered to be the best model for the 200-m-long segments.

According to Fig. 2, for the 400-m-long segments, the CURE plots for the 1-year and 2-years models reveal that these models are acceptable. The other models in

this figure are also acceptable from the CURE plots but they are not from a traffic engineering point of view, as previously discussed.

Comparing the R^2_{margin} of 1-year and 2-years models it is clear that the 2-years model fits the field observation better than the 1-year model (Fig. 2). Also, it fits the field observations better than the 200-m-long segments counterparts and can be selected as the best model for the highway segments studied.

Summary and Conclusions

The objective of the present study was the identification of the major contributory factors to road crash frequency for road segments of Portuguese two-lane highways located in the northern region of the country. The study

also sought to investigate the effects of aggregating data for the identified factors in different ways, that is, in temporal groupings and spatial groupings. The importance of this work is to contribute to the promotion of road safety in the Portuguese northern national road system, which serves many cities and industrial zones.

The initial database considered for this study was formed by the fatal and injury crash frequency (without pedestrian and cyclist crashes), the Average Annual Daily Traffic (AADT) and the geometric characteristics of eighty-eight 200-m-long segments during the years 1999 to 2010. This database contains 1056 data records, of which 740 have zero annual crashes. However, a preliminary analysis of the characteristics of the segments for which there were no crashes registered over the 12-years period revealed that there was no clear technical justification for classifying those segments as inherently safe and, accordingly, the use of zero-inflated regression models was ruled out.

To reduce the number of zero crash records, different databases were developed from the initial database by taking into account variations on the space and time scale of the data, as suggested by Lord *et al.* (2005a, 2005b). In terms of time, four options for aggregating the data were considered, all of which were aimed at including all of the 12-years data available. For the space scale, in addition to the 200-m-long original segments, 400-m-long segments were also considered. Therefore, including the 1-year crash data, 10 different databases were analyzed. As expected, an enlargement in the length of the road segment analyzed or in the time periods from which the crash data are considered allows for a significant reduction in the zero crash observations in the database and contributes to generating acceptable Crash Prediction Models (CPMs). However, the concern with an excess of zeros must be balanced with considerations regarding the impact of grouping data (in space, in time or in both) on the sample size, which can, in turn, make it difficult to develop acceptable CPMs. The GEE procedure was selected to work with the longitudinal data present in all databases. This procedure made it possible to verify that the observations are effectively correlated (exchangeable correlation), except for the 3-years and 6-years data for the 400-m-long segments, where the observations proved to be independent.

From an analysis of the models obtained using the CUMulative REsidual (CURE) plots and marginal R^2 it was found that the models generated on the basis of the 3-years database for the 200-m-long segments and of the 2-years database for the 400-m-long were those with the best fits for the data. Both models were able to capture the same significant contributory factors to the observed crash frequencies. These factors were the traffic volume (expressed in AADT), Lane Width (LW), Vertical Sinuosity (VS), and Density of Access Points (DAP). As these factors result in positive coefficients in the models, which are acceptable, when they increase, it is reasonable to expect that the crash frequency increases as well. Thus, the study proved to be capable of pinpointing those elements that need to be the object of measures taken to promote road safety on national

roads located in Northern region of Portugal. Among such measures the control is increased over the implementation of accesses to lots bordering the road and the use of road signs, not only to channel traffic flows, but also to warn drivers of the presence of stretches of road with vertical sinuosities in close succession.

The main limitation in this study was the number of 200 meters homogenous segments included in the sample, 88, partly due to the high costs associated with the data collection and the lack of logistical resources and equipment. This sample size may have impeded the identification of the significance of some of the researched variables. Other limitation was the lack of detailed information related to all the interventions that took place in the studied NR, which did not allow the inclusion of other variables related to the road conditions in the models (road surface, friction coefficient, etc.).

References

- Anastasopoulos, P. C.; Mannering, F. L. 2011. An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data, *Accident Analysis & Prevention* 43(3): 1140–1147. <http://dx.doi.org/10.1016/j.aap.2010.12.024>
- Anastasopoulos, P. C.; Mannering, F. L. 2009. A note on modeling vehicle accident frequencies with random-parameters count models, *Accident Analysis & Prevention* 41(1): 153–159. <http://dx.doi.org/10.1016/j.aap.2008.10.005>
- Anastasopoulos, P. C.; Mannering, F. L.; Shankar, V. N.; Haddock, J. E. 2012. A study of factors affecting highway accident rates using the random-parameters tobit model, *Accident Analysis & Prevention* 45: 628–633. <http://dx.doi.org/10.1016/j.aap.2011.09.015>
- ANSR. 2009. *Acidentes Rodoviários. Observatório de Segurança Rodoviária* [Road Crashes. Road Safety Observatory]. Autoridade Nacional de Segurança Rodoviária (ANSR) [National Authority for Road Safety], Lisboa, Portugal (Portuguese).
- Ballinger, G. A. 2004. Using generalized estimating equations for longitudinal data analysis, *Organizational Research Methods* 7(2): 127–150. <http://dx.doi.org/10.1177/1094428104263672>
- Cañiso, S.; Di Graziano, A.; Di Silvestro, G.; La Cava, G.; Per-saud, B. 2010. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables, *Accident Analysis & Prevention* 42(4): 1072–1079. <http://dx.doi.org/10.1016/j.aap.2009.12.015>
- Caliendo, C.; Guida, M.; Parisi, A. 2007. A crash-prediction model for multilane roads, *Accident Analysis & Prevention* 39(4): 657–670. <http://dx.doi.org/10.1016/j.aap.2006.10.012>
- Carson, J.; Mannering, F. 2001. The effect of ice warning signs on ice-accident frequencies and severities, *Accident Analysis & Prevention* 33(1): 99–109. [http://dx.doi.org/10.1016/S0001-4575\(00\)00020-8](http://dx.doi.org/10.1016/S0001-4575(00)00020-8)
- Couto, A.; Ferreira, S. 2011. A note on modeling road accident frequency: a flexible elasticity model, *Accident Analysis & Prevention* 43(6): 2104–2111. <http://dx.doi.org/10.1016/j.aap.2011.05.033>
- Dinu, R. R.; Veeragavan, A. 2011. Random parameter models for accident prediction on two-lane undivided highways in India, *Journal of Safety Research* 42(1): 39–42. <http://dx.doi.org/10.1016/j.jsr.2010.11.007>

- El-Basyouny, K.; Sayed, T. 2009. Accident prediction models with random corridor parameters, *Accident Analysis & Prevention* 41(5): 1118–1123. <http://dx.doi.org/10.1016/j.aap.2009.06.025>
- FHWA-TFHRC. 2003. *Interactive Highway Safety Design Model (IHSDM): Crash Prediction Module (CPM) Engineer's Manual*. EUA: IHSDM. Federal Highway Administration (FHWA), Turner-Fairbank Highway Research Center (TFHRC). 66 p. Available from Internet: http://www.wsdot.wa.gov/publications/fulltext/design/IHSDM/CPM_EM.pdf
- Gomes, S. V. 2012. The influence of the infrastructure characteristics in urban road accidents occurrence, *Procedia - Social and Behavioral Sciences* 48: 1611–1621. <http://dx.doi.org/10.1016/j.sbspro.2012.06.1136>
- Gomes, S. V.; Cardoso, J. L. 2012. Safety effects of low-cost engineering measures. An observational study in a Portuguese multilane road, *Accident Analysis & Prevention* 48: 346–352. <http://dx.doi.org/10.1016/j.aap.2012.02.004>
- Gomes, S. V.; Geedipally, S. R.; Lord, D. 2012. Estimating the safety performance of urban intersections in Lisbon, Portugal, *Safety Science* 50(9): 1732–1739. <http://dx.doi.org/10.1016/j.ssci.2012.03.022>
- Halekoh, U.; Højsgaard, S.; Yan, J. 2006. The R package geepack for generalized estimating equations, *Journal of Statistical Software* 15(2): 1–11.
- Harwood, D. W.; Council, F. M.; Hauer, E.; Hughes, W. E.; Vogt, A. 2000. *Prediction of the Expected Safety Performance of Rural Two-Lane Highways*. Publication No FHWA-RD-99-207. US Department of Transportation, Federal Highway Administration (FHWA). 200 p. Available from Internet: <http://www.fhwa.dot.gov/publications/research/safety/99207/99207.pdf>
- Hauer, E. 2004. Statistical road safety modeling, *Transportation Research Record* 1897: 81–87. <http://dx.doi.org/10.3141/1897-11>
- Joshua, S. C.; Garber, N. J. 1990. Estimating truck accident rate and involvements using linear and Poisson regression models, *Transportation Planning and Technology* 15(1): 41–58. <http://dx.doi.org/10.1080/03081069008717439>
- Kumara, S. S. P.; Chin, H. C. 2003. Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros, *Traffic Injury Prevention* 4(1): 53–57. <http://dx.doi.org/10.1080/15389580309852>
- Liang, K.-Y.; Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models, *Biometrika* 73(1): 13–22. <http://dx.doi.org/10.1093/biomet/73.1.13>
- Lord, D.; Bonneson, J. 2007. Development of accident modification factors for rural frontage road segments in Texas, *Transportation Research Record* 2023: 20–27. <http://dx.doi.org/10.3141/2023-03>
- Lord, D.; Mahlawat, M. 2009. Examining application of aggregated and disaggregated poisson-gamma models subjected to low sample mean bias, *Transportation Research Record* 2136: 1–10. <http://dx.doi.org/10.3141/2136-01>
- Lord, D.; Manar, A.; Vizioli, A. 2005a. Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments, *Accident Analysis & Prevention* 37(1): 185–199. <http://dx.doi.org/10.1016/j.aap.2004.07.003>
- Lord, D.; Mannering, F. 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives, *Transportation Research Part A: Policy and Practice* 44(5): 291–305. <http://dx.doi.org/10.1016/j.tra.2010.02.001>
- Lord, D.; Persaud, B. 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure, *Transportation Research Record* 1717: 102–108. <http://dx.doi.org/10.3141/1717-13>
- Lord, D.; Washington, S.; Ivan, J. N. 2007. Further notes on the application of zero-inflated models in highway safety, *Accident Analysis & Prevention* 39(1): 53–57. <http://dx.doi.org/10.1016/j.aap.2006.06.004>
- Lord, D.; Washington, S. P.; Ivan, J. N. 2005b. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory, *Accident Analysis & Prevention* 37(1): 35–46. <http://dx.doi.org/10.1016/j.aap.2004.02.004>
- Milton, J.; Mannering, F. 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies, *Transportation* 25(4): 395–413. <http://dx.doi.org/10.1023/A:1005095725001>
- Pan, W. 2001. Akaike's information criterion in generalized estimating equations, *Biometrics* 57(1): 120–125. <http://dx.doi.org/10.1111/j.0006-341X.2001.00120.x>
- Poch, M.; Mannering, F. 1996. Negative binomial analysis of intersection-accident frequencies, *Journal of Transportation Engineering* 122(2): 105–113. [http://dx.doi.org/10.1061/\(ASCE\)0733-947X\(1996\)122:2\(105\)](http://dx.doi.org/10.1061/(ASCE)0733-947X(1996)122:2(105))
- Shankar, V.; Milton, J.; Mannering, F. 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry, *Accident Analysis & Prevention* 29(6): 829–837. [http://dx.doi.org/10.1016/S0001-4575\(97\)00052-3](http://dx.doi.org/10.1016/S0001-4575(97)00052-3)
- Thomas, P.; Morris, A.; Otte, D.; Breen, J. 2003. Real-world accident data-coordinated methodologies for data collection to improve vehicle and road safety, in *Proceedings: 18th International Technical Conference on the Enhanced Safety of Vehicles*, 19–22 May 2003, Nagoya, Japan, 1–10.
- Vangeneugden, T.; Molenberghs, G.; Verbeke, G.; Demétrio, C. G. B. 2011. Marginal correlation from an extended random-effects model for repeated and overdispersed counts, *Journal of Applied Statistics* 38(2): 215–232. <http://dx.doi.org/10.1080/02664760903406405>
- Venkataraman, N.; Ulfarsson, G. F.; Shankar, V. N. 2013. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type, *Accident Analysis & Prevention* 59, 309–318. <http://dx.doi.org/10.1016/j.aap.2013.06.021>
- Wang, X.; Abdel-Aty, M. 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections, *Accident Analysis & Prevention* 38(6): 1137–1150. <http://dx.doi.org/10.1016/j.aap.2006.04.022>
- WHO. 2013. *Global Status Report on Road Safety 2013: Supporting a Decade of Action*. World Health Organization (WHO), Geneva, Switzerland. 318 p. Available from Internet: http://www.iru.org/cms-filesystem-action/policies/sustainable_development/road_safety/gsrss_en.pdf
- Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press. 1096 p.
- Zhang, Y.; Xie, Y.; Li, L. 2012. Crash frequency analysis of different types of urban roadway segments using generalized additive model, *Journal of Safety Research* 43(2): 107–114. <http://dx.doi.org/10.1016/j.jsr.2012.01.003>
- Zheng, B. 2000. Summarizing the goodness of fit of generalized linear models for longitudinal data, *Statistics in Medicine* 19(10): 1265–1275. [http://dx.doi.org/10.1002/\(SICI\)1097-0258\(20000530\)19:10<1265::AID-SIM486>3.0.CO;2-U](http://dx.doi.org/10.1002/(SICI)1097-0258(20000530)19:10<1265::AID-SIM486>3.0.CO;2-U)