ANDREA ARÉVALO-TÁMARA, M.Sc.[1]
E-mail: andrea.arevalo@usantotomas.edu.co
MAURICIO OROZCO-FONTALVO, M.Sc.[2]
(Corresponding autor)
E-mail: mauricio.orozco@unimilitar.edu.co
VICTOR CANTILLO, Ph.D.[3]
E-mail: victor.cantillo@uninorte.edu.co
[1] Santo Tomás University
Road Infrastructure Master Programme
Cra. 9 #51-11, Bogotá, Colombia
[2] Nueva Granada Military University
Department of Civil Engineering
Carrera 11#101-80, Bogotá, Colombia
[3] Universidad del Norte
Department of Civil and Environmental Engineering
Km 5 via Puerto Colombia, Barranquilla, Colombia

Safety and Security in Traffic
Original Scientific Paper
Submitted: 23 Oct. 2019
Accepted: 13 May 2020

FACTORS INFLUENCING CRASH FREQUENCY ON COLOMBIAN RURAL ROADS

ABSTRACT

Traffic crashes in Colombia have become a public health problem causing about 7,000 deaths and 45,000 severe injuries per year. Around 40% of these events occur on rural roads, taking note that the vulnerable users (pedestrians, motorcyclists, cyclists) account for the largest percentage of the victims. The objective of this research is to identify the factors that influence the frequency of crashes, including the singular orography of the country. For this purpose, we estimated Negative Binomial (Poisson-gamma) regression, Zero-inflated model, and generalized the linear mixed model, thus developing a comparative analysis of results in the Colombian context. The data used in the study came from the official sources regarding records about crashes with consequences; that is, with the occurrence of fatalities or injuries on the Colombian roads. For collecting the highway characteristics, an in-field inventory was conducted, gathering information about both infrastructure and operational parameters in more than three thousand kilometres of the national network. The events were geo-referenced, with registries of vehicles, involved victims, and their condition. The results suggest that highways in flat terrain have higher crash frequency than highways in rolling or mountainous terrain. Besides, the presence of pedestrians, the existence of a median and the density of intersections per kilometre also increase the probability of crashes. Meanwhile, roads with shoulders and wide lanes have lower crash frequency. Specific interventions in the infrastructure and control for reducing crashes risk attending the modelling results have been suggested.

KEY WORDS

traffic crashes; crash frequency; Colombian rural roads; Negative Binomial model; Zero-inflated model; generalized linear mixed model; traffic safety;

1. INTRODUCTION

Traffic crashes may occur due to three main factors: humans, vehicles, and the environment. In Colombia, the government tried to control the vehicle factor with the mandatory "technical-mechanical" revisions for any motor vehicle. The human factor is difficult to control as it refers to the user behaviour while driving, walking, or cycling, but it has been a subject of study in the recent years, giving it the importances it deserves [1]. On the other hand, the influence of the environment, particularly road infrastructure, on the frequency and severity of traffic crashes, deserves great attention, given that future events will occur under the same conditions as in the past situations [2]. Efforts made in that direction are focused on reducing errors in highway design and construction, with the purpose of diminishing crash rates and lessen its severity. Among the characteristics of the infrastructure and the environment, influencing the crash frequency are the width of the lane, pavement surface condition, grade, road marking, and traffic volume, among others, such as the surrounding land use.

The particular Colombian orographic conditions may have a strong influence on the potential risk of roads, to which is added the fact that most of them are two-lane roads. Regarding roads quality, Colombia is ranked 102 and 92 in road connectivity index among 140 evaluated countries [3]. One of the reasons of this position is due to orographic conditions, which present a challenge for road designers. The present study aims to evaluate the factors influencing the crash frequency, including the type of terrain (flat, rolling, or mountainous). This analysis is useful to define the policies and actions destined to decrease the current rates on the rural roads.

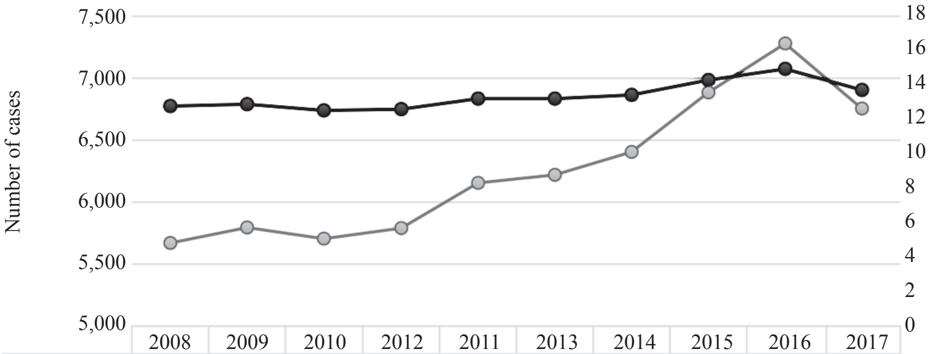
Analysing the crash rate per population, in 2016 the country recorded the highest rate of the last ten years with an ascendant tendency with 14.93 deaths per every 100,000 inhabitants (*Figure 1*). It shows a setback in road safety in the country, given that the previous top rate was 14.9 in 1996 [4].

Aware of this situation, several authors have studied different factors influencing crash occurrence in Colombia, mainly in urban contexts. The research conducted by Guerrero et al. [5] determined the influence of factors such as road geometry, environment, traffic volume and speed on the crash frequency on urban roads, while Cantillo, Garcés & Márquez [6] studied the factors influencing the occurrence of traffic crashes in Cartagena, one of the largest cities of the country. Their results show that variables like roadway width, land use, the number of intersections, pavement type, and average speed have a significant influence on the crash frequency. However, no studies about factors influencing traffic crash frequency on the Colombian rural roads could be found.

Other variables related to the environment, such as weather conditions, also influence the crash frequency. Xi et al. [7] concluded that there is a strong correlation between the environmental conditions and crashes. Including the terrain type or orography as a variable in the prediction models was necessary because despite the terrain type being one of the major factors determining the design speed of rural roads in Colombia and worldwide, their specific safety effects on actual crash occurrences have not been fully investigated, apart from the works developed by Ackaah & Salifu [8] and Choi et al. [9].

The use of statistical methods for estimating the crash frequency has been a research subject for many years. Different specifications of econometric models have been proposed to evaluate the relevant factors involving characteristics of traffic factors, control, highways, environment, and involved users. Several studies used the Poisson distribution, concluding that it is adequate to model the crash frequency, although the phenomenon of over-dispersion may occur. Other discrete distributions like the Negative Binomial are also widely used [10]. Poisson and Negative Binomial models are suitable econometric approaches for crash assessment and forecasting under certain conditions and are still being used in the literature [11]. In general, the Poisson models are more appropriate for homogeneous conditions, while the negative binomial models behave better when the conditions are heterogeneous [10].

Lord and Mannering [12] emphasize that the Poisson distribution is appropriate if mean and variance are equal. When this assumption is substantially violated, the NB distribution can provide an improvement over the Poisson distribution. On the other hand, the Zero-inflated models are



	2008	2009	2010	2012	2011	2013	2014	2015	2016	2017
● Cases	5,670	5,796	5,704	6,152	5,792	6,219	6,402	6,884	7,280	6,754
○ Rate	12.76	12.89	12.53	13.21	12.58	13.2	13.43	14.28	14.93	13.7

Figure 1 – Evolution of deaths and mortality rate (deaths per 100,000 inhabitants) in Colombia

econometric approximations to consider because on a large number of road sections (RS) the frequency of crash is null [13, 14]. Moreover, several authors claim that the Zero-inflated Poisson model is an adequate candidate for the study of road crashes when an excess of zeros exists in the data but can be challenging to interpret [15]. However, some authors such as Lord et al. [16] stated that Zero-inflated models should be avoided for modelling motor vehicle crashes on highway entities, except when forecasting is the sole research objective. There is vast literature discussing the modelling approach for crash frequency. In this sense, the review provided by Lord and Mannering [12] is really useful to select the most appropriate model to apply, based on its pros, cons, and more important the characteristics of the data gathered (sample size, over/under-dispersion, sample mean, amount of zeros) and estimation process complexity.

The results reported in the literature are diverse when comparing alternative models to evaluate the crash frequency. Miranda & Moreno [17] concluded that the Negative Binomial Zero-inflated is particularly useful when there is a frequency of zeros in the data, and the results are just a little better than the traditional Negative Binomial model, inferring that the traditional model can still be used for predictions with a high frequency of zeros in the data. On the other hand, Aguero-Valverde [18] compared the Bayes Poisson-Gamma and Poisson lognormal distributions, and the Zero-inflated models for measuring precision in the prediction of crash frequency, concluding that the model with better adjustment was the one that used the Zero-inflated – the Poisson distribution. Meanwhile, considering that over-dispersion is almost always present in crash data and that mixed-effects (fixed and random) could exist, the use of generalized linear mixed models should be considered as a modelling alternative. [19-22]

In Colombia, where the orographic conditions are the main criteria for road design in most cases, there are no studies that evaluate its influence on the crashes. For this reason, we will evaluate several variables and their influence through econometric modelling, including the terrain type as an explanatory variable. Data used for modelling purposes include crashes with consequences (injuries and fatalities) during the years 2015 - 2016. A generalized linear mixed model was estimated and compared with the traditional and validated NB and ZINB models to evaluate its applicability to this kind of study. The results are particularly useful for the authorities in charge of road infrastructure when defining strategies that lead to the reduction of crash effects by proposing appropriate infrastructure interventions. To our best knowledge, this is the first study on this topic in such a context.

2. METHODS

Data regarding crashes used in the study were obtained from official sources. They consist of crashes with consequences (fatalities or injuries, excluding property-damage-only crashes), that happened on the Colombian rural roads during a period of two (2) years (2015-2016) provided by the Instituto de Medicina Legal y Ciencias Forenses of Colombia. The events were geo-referenced. Additionally, the database contains information about environmental conditions. Crashes with only property damages (without injury) were not considered because such registries are not complete and are, therefore, sub-represented. These databases are often used for forensic or legal cases; hence, they usually only include high severity events like deaths or serious injuries [23].

In addition to the crash database, a complete and also geo-referenced registry of road conditions was made, obtained by a vehicle equipped with cameras and GPS covering about 40% of the rural arterial roads in the country. It collected information on road geometry (width and number of lanes, median types existence and shoulders width, gradient, curvature, pavement condition), and about the environment (presence of lateral obstacles, use of surrounding ground, pedestrian flows, lighting, signals, accesses).

To build the database used for modelling, homogeneous road sections were considered, on average 5 km in length. Road sections crossing urbanized areas (*Figure 2*) were considered differently, regardless

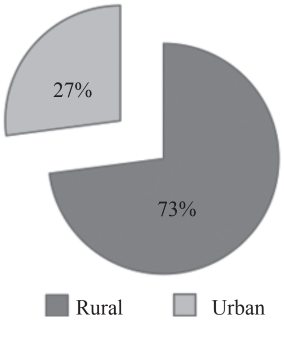


Figure 2 – Land use distribution

of their length. Segmentation was made based on the methodology proposed by Cafiso et al. [24]. The average daily traffic (ADT), section width (W) and an indicator of the number of curves per length of the section were used [25]. To determine the section homogeneity, the sections with these parameters constant along 5 km were grouped. The RSH index proposed in this methodology could not be calculated because the equipment used determined only the distance to roadside obstacles, not identifying the kind of obstacle or the danger present.

In total, the selected sections of this study add up to 3,162 kilometres that make up to 414 sections, 73% being rural roads and 27% being rural roads crossing urban zones, covering the entire country (*Figure 3*). The sample is representative of the Colombian rural roads and allows us to correctly estimate the parameters with the selected techniques [12]. These characteristics vary from region to region (because of the orography, ADT, vehicle typology, among other factors).

On the selected sections, data report 1,409 injury or fatal crashes during two years. In this study, the crashes of each year were considered separately. For this reason, the number of observations used for modelling add up to 828 sections, comparable to the study performed by Aguero-Valverde [18], where 865 rural segments of two lane roads were analysed. *Table 1* shows the different locations studied.

Figure 4 shows the number of crashes reported in each section studied, from the 828 registers, 395 (48%) showed zero events. This result is consistent with the study performed by Miaou and Lum [26], who found that more than 80% of road sections did not report crashes during a year.

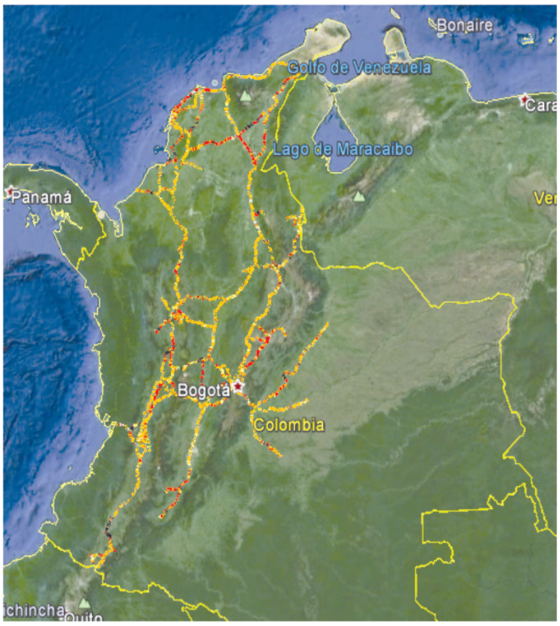


Figure 3 – Roads analyzed

2.1 Influential variables

Next, the description and characteristics of the used variables in the study are shown, divided between continuous and mute variables, including factors already known to have a significant effect on the crash [27]. *Table 2* shows all the variables taken into account in the study. These variables were selected based on the previous studies [2, 8, 11, 28]. Some of them were not significant in the modelling results. Speed corresponds to the percentile 85, but it is not the specific information about the vehicles involved in the crash.

Information about traffic counts (ADT) were obtained from INVIAS (Colombian official roads agency), while pedestrian flows and speed were measured in the field. Terrain variable was assigned based in the INVIAS road classification.

Table 1 – Road segment distribution

Location (Colombia)	Length [km]	Road segments	Orography (terrain)
Valle del Cauca (West)	168.9	53	Plain
Atlántico & Bolivar (Northern coast)	150.5	39	Flat
Nariño & Cauca (Southwest)	219	58	Rolling
Santander & Cesar (Northeast)	404.6	1-35 36- 109	Mountainous Flat
Atlántico & Bolívar (Northern coast)	109.6	27	Flat
Santander (Northeast)	89.1	20	Mountainous
Santander & Boyacá (Center)	127.8	28	Mountainous
Antioquia	185.1	45	Mountainous
Santander & Norte de Santander (Northeast)	126.6	35	Mountainous
TOTAL	1,581.20	414	

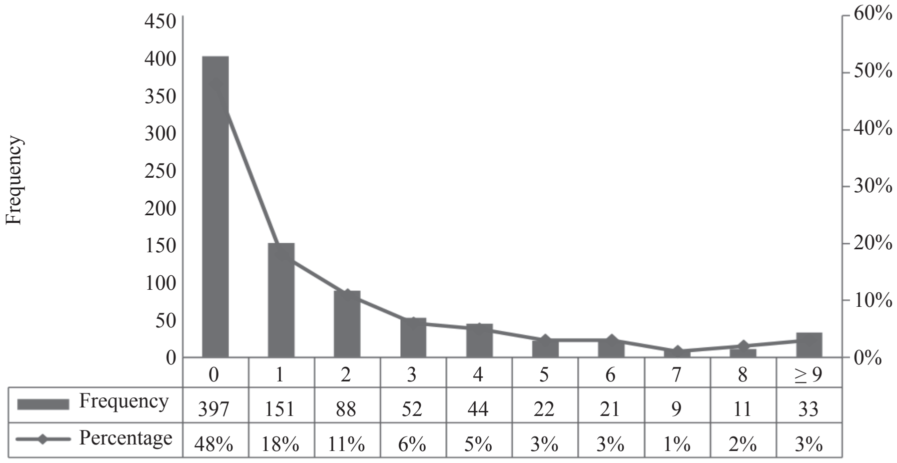


Figure 4 – Distribution of crash frequency

Table 2 – Road-related variables

CONTINUOUS VARIABLES					
Variable	Description	Max.	Min.	Mean	Standard deviation
Crashes (dependent variable)	Number of crashes per year on the section	21	0	1.72	2.82
Curves/length	Number of horizontal curves per kilometre	39	0	4.05	6.56
Lane width	Lane width in metres	3.6	2.7	3.51	0.23
Intersections/length	Number of intersections per kilometre	51	0	2.61	4.96
Pedestrians	Number of observed pedestrians on the section [ped/day]	1,668	0	73.39	35.57
ADT	Average daily traffic	36,767	2,523	8,036	5,242
Motorcycles	Number of motorcycles per length of the section [veh/day]	50	8	27	13
Speed	km/h	95	30	61.53	15.74
Slope	%	8.5	0	2.88	1.42
Cyclists	Number of cyclists observed	1,050	0	125	164
BINARY VARIABLES					
Variable	Description	Frequency		Percentage	
Road shoulder	0: Not present	488		60%	
	1: Present	332		40%	
Median	1: Present	138		17%	
	0: Not present	682		83%	
Road marking conditions	0: Bad condition	236		29%	
	1: Good condition	584		71%	
Pavement	0: Bad condition	38		5%	
	1: Good condition	782		95%	
Land use	0: Urban	216		26%	
	1: Rural	604		74%	
Lighting	0: Not present	680		83%	
	1: Present	140		17%	
Terrain	1: Flat	386		47%	
	0: Rolling or mountainous	442		53%	

2.2 Models for data analysis

Zeng & Huang [29] who applied the Bayesian techniques for traffic crash modelling on urban roads in Florida and compared the results with the Poisson and NB model to conclude that the traditional models have adequate performance for the prediction and assessment of road safety, proving its efficiency. In this study, a generalized linear mixed model is presented with a negative binomial distribution (GLMMB) and compared with a Negative Binomial (NB) and a Zero-inflated Negative (ZINB).

Negative Binomial Model (NB)

The Negative Binomial model is an alternative approach for modelling crash frequency. It is useful when over-dispersed data are present, a case in which the Poisson model can result in biased and inconsistent parameter estimates. The NB model is derived from *Equation 2*, in such a way that for each section *i* and time *j*, it has the following equation [30]:

$$\lambda_{ij} = EXP(\beta X_{ij} + \varepsilon_{ij}) \tag{1}$$

In *Equation 1*, $EXP(\varepsilon_{ij})$ is a term of distribution error gamma with mean 1 and variance α^2. The addition of this period allows the differentiation of the variance from the mean in the following way [30]:

$$VAR[n_{ij}] = E[n_{ij}][1 + \alpha E[n_{ij}]] = E[n_{ij}] + \alpha E[n_{ij}]^2 \tag{2}$$

The NB distribution has the following equation [30]:

$$P(n_{ij}) = \frac{\Gamma((1/\alpha) + n_{ij})}{\Gamma(1/\alpha) n_{ij}!} \left(\frac{1/\alpha}{(1/\alpha) + \lambda_{ij}} \right)^{1/\alpha} \left(\frac{\lambda_{ij}}{(1/\alpha) + \lambda_{ij}} \right)^{n_{ij}} \tag{3}$$

where Γ(.) is a gamma function. The mean and variance of the random variable of NB are given by [31]:

$$E(n_{ij}) = \lambda_{ij} \tag{4}$$

$$Var(n_{ij}) = \lambda_{ij} + \alpha \lambda_{ij}^2 \tag{5}$$

Particularly, when $\alpha \rightarrow 0$, the variance of crashes is equal to the mean, and this model converges to the standard Poisson regression model.

Zero-inflated (ZI)

The ZI regression model is particularly useful in the case of data with an excess of zeros [32]. The standard Poisson model does not differentiate the causes of an excess of zeros, but the ZI model does. In this research, the sample records 397 sections (48%) without crashes. Thanks to this observation, it was deemed better to apply the ZI model combined with the Poisson and NB.

The models of combined distributions are convex linear combinations of distributions of probabilities. The zeros in the data may be a mix of structural zeros and the sample ones. From this kind of model, the so-called Zero-inflated Poisson regression model (ZIP) is widely used for counting data with an excess of zeros. This model was proposed by Lambert (1992).

For the description of the ZIP model, P_{ij} is the probability of excess of zeros in segment *i* in period *j*, and (1-P_{ij}) is the probability of crash frequency derived from the Poisson probability function. The model is described as shown in *Equation 6* [33].

$$P(N = n_{ij}) = \begin{Bmatrix} P_{ij} + (1 - P_{ij}) e^{\lambda_{ij}} \\ (1 - P_{ij}) \dfrac{e^{-\lambda_{ij}} \lambda_{ij}^{n_{ij}}}{n_{ij}!} \end{Bmatrix} \begin{matrix} n_{ij} = 0 \\ n_{ij} > 0 \end{matrix} \tag{6}$$

Variable n_{ij} is the number of crashes for segment *i* in year *j*, and λ_{ij} are the expected crashes in segment *i* in the function of the covariance λ_{ij}=exp(βX_{ij}), and P_{ij} is the probability of having zero events and is given on the logistic regression model, in the following manner [33]:

$$P_{ij} = \frac{\exp(\theta K_{ij})}{1 + \exp(\theta K_{ij})} \tag{7}$$

where $K_{it} \equiv (K_{it1}, \ldots, K_{itm})$ is a set of explanatory variables, and $\theta \equiv (\theta_1, \ldots, \theta_m)$ are the parameters to be estimated.

The other proposed distribution under this approach is the Zero-inflated Negative Binomial distribution. This distribution is especially adequate when besides the excess of zeros, there is also over-dispersion. However, the interpretation of these models is not as simple as it occurs with the two-part models.

Similar to the ZIP model, the function of the density of probability of the ZINB models is as shown in *Equation 8* [33]:

$$P(N = n_{ij}) = \\ = \begin{Bmatrix} P_{ij} + (1 - P_{ij}) \dfrac{1}{(1 + \alpha \lambda_{ij})^{\frac{1}{\alpha}}} \\ (1 - P_{ij}) \dfrac{\Gamma(n_{ij} + (1/\alpha))}{\Gamma(n_{ij} + 1)\Gamma(1/\alpha)} \dfrac{(\alpha \lambda_{ij})^{n_{ij}}}{(1 + \alpha \lambda_{ij})^{n_{ij} + (1/\alpha)}} \end{Bmatrix} \tag{8}$$

$$n_{ij} = 0$$
$$n_{ij} > 0$$

In this case, α is the dispersion parameter, and Γ(.) is the gamma function for the ZINB model. The proposed models explain the number of crashes with consequences (deaths or injuries) in each section for one year, which is an entire non-negative variable. Since there are two years of data, there are two registries for each section.

Several authors have used the zero-inflated distribution, arguing that it provides a better goodness-of-fit than the traditional models such as Poisson and Negative Binomial [14]. Zero-inflated models assume that a dual-process is responsible for generating crash data [34-37].

Besides, Raihan et al. [38] state that there exists the possibility that datasets with a high amount of zeros are not eligible for analysis with traditional models, while the ZINB model is ideal for datasets with excessive dispersion or zeros. Meanwhile, Lord et al. [16] concluded that although this model promises a better goodness-of-fit than traditional models, it should only be used if the only objective of the research is crash frequency prediction and should not be used for crash modelling on highways. As part of this research, we decided to estimate a ZINB model given that we wanted to consider the convenience of using this model versus the negative binomial.

In this framework, the dataset was rigorously revised, considering time/space scales and the exposition problem [38] in order to apply the ZINB model correctly.

Generalized linear mixed models

Generalized Linear Models (GLMs) have the limitation of including only fixed effects, an extension of which are the Generalized Linear Mixed Models (GLMMs) that include both fixed and random parameters, allowing response variables from different distributions [39]. In this study, a GLMM with an NB distribution was estimated. In this case, the mean response for the number of crashes is assumed to have a log-linear relationship with the covariates and is structured as follows [20]:

$$\ln(\mu) = \beta_0 + \sum_{s=1}^{q} \beta_s X_s \tag{9}$$

where X_s is a traffic and geometric variable of a particular site, and β_s is a regression coefficient to be estimated.

Defining the variance as [40]:

$$var(Y) = \emptyset \frac{p}{(1-p)^2} = \frac{1}{\emptyset}\lambda^2 + \mu \tag{10}$$

The probability mass function of the NB distribution can be given as:

$$P(Y=y;\emptyset,p) = \frac{\Gamma(\phi+y)}{\Gamma(\phi)\cdot y!}(p)^{\emptyset}(1-p)^{y};\emptyset > 0, 0 < p < 1 \tag{11}$$

Parameter p is defined as the probability of failure:

$$p = \frac{\emptyset}{\mu+\emptyset} \tag{12}$$

where μ is mean response of the observation and ϕ is the inverse of the dispersion parameter α; that is, $\phi=1/\alpha$.

The probability mass function of the NB distribution and its GLM is defined by (as a Poisson-gamma model) (*Equations 3* and *4*):

$$P(Y=y,\mu,\emptyset) = NB(y;\emptyset;\mu) = \\ = \frac{\Gamma(\phi+y)}{\Gamma(\phi)\Gamma(y+1)}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\phi}{\mu+\phi}\right)^{y} \tag{13}$$

Elasticities

The analysis of elasticities allows evaluating the relative impact of each variable in the model and for continuous variables, it can be estimated for the Poisson model as follows:

$$E_{x_{ijk}}^{\lambda_{ij}} = \frac{\Delta\lambda_{ij}}{\lambda_{ij}} \cdot \frac{x_{ijk}}{\Delta x_{ijk}} = \frac{\partial\lambda_{ij}}{\partial x_{ijk}} \cdot \frac{x_{ijk}}{\lambda_{ij}} = \beta_k x_{ijk} \tag{14}$$

where $E_{x_{ijk}}^{\lambda_{ij}}$ represents the elasticity of the response variable λ_{ij} regarding its explicative variable x_{ijk}; β_k is the estimated parameter for the k-th independent variable.

This elasticity is calculated for each road section. In terms of analysis, it is convenient to estimate the elasticity as the average of the elasticities calculated for each observation.

Equation 14 is only feasible for continuous variables (i.e. lane width, number of intersections, ADT, etc.), for binary variables such as shoulder presence, median presence, pavement condition, etc., a pseudo-elasticity needs to be calculated, which estimates the change of crash frequency due to changes of these variables, and is calculated when using the Poisson regression model as follows [30]:

$$E_{x_{ijk}}^{\lambda_{ij}} = \frac{EXP(\beta_k) - 1}{EXP(\beta_k)} \qquad 15)$$

3. RESULTS

Results for negative binomial, zero-inflated and generalized mixed models are presented in the following subsections. The elasticities are presented as well.

3.1 Negative Binomial model (NB)

The first model presented is a Negative Binomial regression (NB), which is shown in *Table 3*. The variables considered were the number of curves per RS length, terrain type, number of pedestrians, presence of road shoulder, ADT*RS length, lane width and number of intersections. The speed was not

Table 3 – NB model

Variable	Model 1		
	Estimate	Z value	Pr(>IzI)
(Intercept)	1.8900	2.51	0.0120
Curves / RS length	-0.0727	-1.54	0.1230
Terrain	0.687	5.09	0.0000*
Number of pedestrians	0.0011	2.86	0.0042*
Shoulder	-0.2310	-2.01	0.0450*
ADT*RS length	0.00001	6.16	0.0000*
Lane width	-0.6190	-2.96	0.0031*
Intersections/RS length	0.0334	3.10	0.0019*
Log-likelihood	-2,757.4160		

**p-value<0.05, **p-value<0.1*

significant. Its effect is probably assumed by the conditions of the orography since on flat terrain the speeds are greater than on mountainous terrain roads.

All variables included in the model, excluding the density of curves, have a p-value < 0.05, which means all of them influence the crash frequency at a 95% confidence level. The model shows that the number of crashes is higher on flat terrain and highways with a lower number of curves per kilometre. This result matches the one obtained by Ackaah & Salifu [8] in Ghana where they applied the generalized linear models and concluded that crash probability (with injuries or deaths) is higher on flat terrain than on the rolling and mountainous terrains. Regarding vulnerable users such as pedestrians, the model shows that on rural roads, the crash probability increases with the increase in their number. This result is similar to the research conducted by Sasidharan & Menéndez [41], who developed a study of traffic crash severity with pedestrians and concluded that lighting, road signs, and age, are the main factors why pedestrians are involved in such events.

According to the results, the absence of road shoulder significantly increases the crash probability. The most recent Colombian normativity regarding the geometric design of highways states that flat terrain should have broad shoulders with a minimum of 2.0m width and 1.5 m for rolling and mountainous terrain. Nevertheless, in the data gathered from fieldwork, 60% of RS did not have broad shoulders, as shown in *Table 2*. The model also shows, as expected, that a narrow lane width negatively influences the crash frequency, as a wider lane provides comfort and better manoeuvrability and the number of intersections per kilometre positively influences crash, as it increases the number of conflicts plus the fact that the operational speed is usually higher on rural roads.

3.2 Zero-inflated Negative Binomial (ZINB)

The variables considered for this model were the number of lanes, number of pedestrians, lane width, presence of median, ADT*RS length, number of curves, and terrain type (*Table 4*).

The signs, significance, and estimates obtained in the ZINB model were similar to NB. When comparing both presented models using the Akaike Information Criterion (AIC), the results are also similar (NB AIC: 2,775.416, ZINB AIC: 2,775.469) with a slightly lower value for NB. According to the model, sections on flat terrain roads are less prone to have zero crashes than those on the rolling or mountainous terrain.

Table 4 – Zero-inflated negative binomial model

Variable	Estimate	Z value	Pr(>IzI)
Count model coefficients			
(Intercept)	2.190	3.06	0.002*
Curves / RS length	-0.140	-3.30	0.001*
Number of pedestrians	0.001	1.78	0.074**
Number of intersections	0.028	2.14	0.032*
Number of lanes	0.348	3 .33	0.001*
Lane width	-0.556	-2.73	0.006*
Log (theta)	-0.115	-0.96	0.335
Zero-inflated model coefficients (binomial)			
(Intercept)	1.089	3.06	0.002*
Terrain	-2.4300	-2.24	0.025*
Median	0.6270	1.19	0.235
ADT*RS length	-0.00009	-3.64	0.000*
Log-likelihood	-1,377		

**p-value<0.05, **p-value<0.1*

3.3 Generalized linear mixed model with negative binomial

The results using the generalized linear mixed model with negative binomial follow the same tendencies as the previous NB and ZINB. Consequently, similar conclusions are obtained (*Table 5*).

Table 5 – Generalized linear mixed model with negative binomial

Variable	Estimate	Z	P>[z]
Terrain	0.6409	4.73	0.000*
Curves / RS length	-0.0711	-1.52	0.129
ADT*RS length	0.00001	4.57	0.000*
Number of pedestrians	0.0017	3.67	0.000*
Shoulder	-0.2587	-2.15	0.031*
Median	0.2461	1.64	0.102**
Lane width	-0.6100	-2.82	0.005*
(Intercept)	1.9073	2.49	0.013*
Ln(α)	0.4063	4.60	0.000
Log-likelihood	-1,380.7		

**p-value<0.05, **p-value<0.1*

All model variables proved to be statistically significant and all coefficients have reasonable and intuitive signs. The terrain coefficient is the highest among the covariates and has a positive sign meaning that more crashes occur on the flat terrain, straight roads with no slopes or curves making speeding a common event on these roads in Colombia, followed by lane width with a negative coefficient as expected, meaning that narrow lanes are more dangerous. The presence of a median is an interesting result, as it increases the crash frequency (probably related with speeding on multilane highways). However, since crash severity is not considered, it is inconvenient to suggest against installing the medians. On the contrary, medians are essential on rural roads with more than one lane per direction to avoid frontal collisions. Considering that in Colombia all multilane highways are divided, the models suggest that they have a higher risk of crashes. Therefore, the authorities should increase the control over them.

3.4 Elasticity analysis

Table 6 shows the results of the elasticities computation for the NB model. The reported values are averages of the analysed RS. Elasticities are a reliable indicator of the influence of the variables included in the models presented. The results show that the lane width, ADT*RS length, and flat terrain are the most sensitive variables for crash frequency. Regarding ADT*RS length, when increasing 1%, the crash frequency would increment by 0.32%. The flat terrain increases the crash probability by 0.49% when compared to the rolling or mountainous terrain. Elasticity analysis evidences high sensitivity of crash frequency to lane width.

Table 6 – NB elasticities

Negative Binomial - Elasticities	
Curves / RS length	-0.074
Terrain	0.497
Number of pedestrians	0.082
Shoulder	-0.260
ADT*RS length	0.318
Lane width	-2.175
Number of intersections	0.029

As presented in *Table 7*, the elasticities obtained using the ZINB model are similar to those obtained before. In this regard, the conclusions are rather the same.

Table 8 shows the elasticities computed for GLMM. It shows how investment in upgrading lane and shoulder conditions should be prioritized and suggest giving more attention to roads located on the flat terrain.

Table 7 – ZINB elasticities

Variable	Elasticities
Count model coefficients	
Curves / RS length	-0.142
Number of pedestrians	0.070
Number of intersections per kilometre	0.070
Number of lanes	0.420
Lane width	-1.950
Zero inflation model coefficients (binomial)	
Terrain	0.248
Presence of median	-0.077
ADT*RS length	-2.388

Table 8 – GLMM elasticities

Generalized linear mixed model - elasticities	
Curves / RS length	-0.071
Terrain	0.473
Number of pedestrians	0.124
Shoulder	-0.295
ADT*RS length	0.312
Lane width	-2.143
Median	0.218

4. CONCLUSION

The results of the models used to explain the fatal and injury crash frequency on the Colombian rural roads show that all models presented are suitable for estimating the crash frequency. GLMM has been estimated and compared with the traditional NB and ZINB models. The results show that all variables included in the GLMM are significant with a p-value < 0.05, while in the other models there is one variable below this value (Curves*section length in the NB and presence of the median in ZINB). The GLMM proved to be slightly better than the ZINB but the best of the three models presented is the NB, which has a higher log-likelihood. This means that, in this case, the most parsimonious model considered is the one that better fits the data. Regarding the variables, terrain type, ADT*section length, pedestrian number and lane width proved to be highly significant in all the models presented. The ADT coefficient is below 1 in all the models which means that as the number of vehicles increases, the number of crashes increases, but at a decreasing rate. This result is consistent with the findings of Shirazi et al. [19]. In this case, the use of a more complex model (GLMM) did not provide a better fit.

Other variables that proved to be significant were lane width and the number of lanes. The exposition factors, given by the number of vehicle-kilometre (ADT*RS length) and the number of pedestrians, as expected, were relevant in explaining the number of traffic crashes.

Results demonstrated that the variable related to the number of curves per kilometre has a negative marginal effect on crashes with fatalities and injuries. This can be explained because on the roads with continuous curves the drivers must reduce speed and drive with more caution, which is very common on the Colombian roads, particularly on roads built on the mountainous terrain in the Andean zone of the country, where operational speeds are low. The above is in keeping with the findings of Ackaah & Salifu [8] in Ghana. Interestingly, orography was also a significant variable. In Colombia, the roads on flat terrain are more prone to the occurrence of crashes with fatalities and injuries than the ones in rolling or mountainous environment. This result may be related to speed, indicating that mountainous and rolling terrains tend to generate fewer crashes with fatalities and injuries despite the perception of being more dangerous. According to these results, the Colombian authorities should increase controls on flat terrain roads and multilane highways, such as speed cameras.

On the Colombian rural highways, the infrastructure for pedestrians is not safe enough. Moreover, rural roads frequently cross urban centres and many schools are located in their surroundings without proper markings and facilities. This issue should be addressed in the short term by implementing traffic calming policies at these locations and investing in pedestrian infrastructure and drivers' sensitization. Medium and long-term land use should be revised to avoid these situations while building alternate roads to divert cars and heavy traffic out of the urban crossings. In conclusion, road infrastructure in Colombia needs major investment and, given the modelling results, the country position in the ranking mentioned at the beginning of this paper is not surprising. Most of the Colombian rural roads have no shoulders, narrow lanes, inadequate pedestrian infrastructure, and land use incompatibility in proximities. Such situations represent risk for their users.

Future research may consider the factors influencing the severity of traffic crashes. It is also interesting to consider the role of motorcycles, which is present in one of each two fatal crashes in Colombia.

ANDREA ARÉVALO-TÁMARA, M.Sc.[1]
E-mail: andrea.arevalo@usantotomas.edu.co
MAURICIO OROZCO-FONTALVO, M.Sc.[2]
E-mail: mauricio.orozco@unimilitar.edu.co
VICTOR CANTILLO, Ph.D.[3]
E-mail: victor.cantillo@uninorte.edu.co
[1] Universidad Santo Tomás
Maestría en Infraestructura Vial
Cra. 9 #51-11, Bogotá, Colombia
[2] Universidad Militar Nueva Granada
Programa de Ingeniería Civil
Carrera 11#101-80, Bogotá, Colombia
[3] Universidad del Norte
Departamento de Ingeniería Civil y Ambiental
Km 5 via Puerto Colombia, Barranquilla, Colombia

FACTORES INFLUYENTES EN LA FRECUENCIA DE CHOQUES EN LAS VÍAS RURALES COLOMBIANAS

RESUMEN

Los accidentes de tránsito en Colombia se han convertido en un problema de salud pública que causa alrededor de 7,000 muertes y 45,000 lesiones graves por año. Alrededor del 40% de los eventos ocurren en carreteras rurales, teniendo en cuenta que los usuarios vulnerables (peatones, motociclistas, ciclistas) son el mayor porcentaje de las víctimas. El objetivo de esta investigación es identificar los factores que influyen en la frecuencia de los choques, incluida la orografía singular del país. Para tal fin, se estimó una Regresión Binomial Negativa (Poisson-gamma), un modelo Cero Inflado y un modelo Mixto Lineal Generalizado, desarrollando así un análisis comparativo de resultados en el contexto colombiano. Los datos utilizados en el estudio, provienen de fuentes oficiales de los registros de choques con consecuencias; es decir, con que involucran muertes o lesiones en las carreteras colombianas. Para recopilar las características de las carreteras, se realizó un inventario vial, recolectando información sobre infraestructura y parámetros operativos en más de tres mil kilómetros de la red nacional colombiana. Los choques fueron geo-referenciados, con registros de vehículos, víctimas involucradas y su condición. Los resultados sugieren que las carreteras en terreno plano tienen una mayor frecuencia de choque que las carreteras en terreno ondulado o montañoso. Además, la presencia de peatones, la existencia de separador central y la densidad de intersecciones por kilómetro también aumentan la probabilidad de accidentes de tránsito. Mientras tanto, las carreteras con berma y calles anchas tienen menos frecuencia de choques. Como recomendación, se sugieren intervenciones *específicas en la infraestructura y el control para reducir el riesgo de accidentes de tránsito atendiendo a los resultados del modelado.*

PALABRAS CLAVE

Choques de tráfico; frecuencia de choques; vías rurales colombianas; Binomial negativo; Zero-Inflado; Modelo mixto lineal generalizado; seguridad vial;

REFERENCES

[1] Batrakova A, Gredasova O. Influence of Road Conditions on Traffic Safety. *Procedia Engineering*. 2016;134: 196-204. Available from: doi:10.1016/j.proeng.2016.01.060

[2] Yakar F. Identification of Accident-Prone Road Sections by Using Relative Frequency Method. *Promet - Traffic &Transportation*. 2015;27: 539-47. Available from: doi:10.7307/ptt.v27i6.1609

[3] World Economic Forum. *Global Competitiveness Report*; 2018.

[4] Medicina Legal. *Forensis*; 2017.

[5] Guerrero Barbosa TE, Espinel-Bayona Y, Palacio-Sánchez D. Effects of the Attributes Associated with Roadway Geometry, Traffic Volumes and Speeds on the Incidence of Accidents in a Mid-Size City. *Ingenieria y Universidad*. 2015;19: 105. Available from: doi:10.11144/Javeriana.iyu19-2.eaar

[6] Cantillo V, Garcés P, Márquez L. Factors influencing the occurrence of traffic accidents in urban roads: A combined GIS-Empirical Bayesian approach. *DYNA*. 2016;83: 21-8. Available from: doi:10.15446/dyna.v83n195.47229

[7] Xi J, Zhao Z, Li W, Wang Q. A Traffic Accident Causation Analysis Method Based on AHP-Apriori. *Procedia Engineering*. 2016;137: 680-7. Available from: doi:10.1016/j.proeng.2016.01.305

[8] Ackaah W, Salifu M. Crash prediction model for two-lane rural highways in the Ashanti region of Ghana. *IATSS Research.* 2011;35: 34-40. Available from: doi:10.1016/j.iatssr.2011.02.001

[9] Choi J, Kim S, Heo T-Y, Lee J. Safety effects of highway terrain types in vehicle crash model of major rural roads. *KSCE Journal of Civil Engineering.* 2011;15: 405-12. Available from: doi:10.1007/s12205-011-1124-x

[10] 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.* 2006;38: 751-66. Available from: doi:10.1016/j.aap.2006.02.001

[11] Zou Y, Ash JE, Park B-J, Lord D, Wu L. Empirical Bayes estimates of finite mixture of negative binomial regression models and its application to highway safety. *Journal of Applied Statistics.* 2018;45: 1652-69. Available from: doi:10.1080/02664763.2017.1389863

[12] Lord D, Mannering F. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice.* 2010;44: 291-305. Available from: doi:10.1016/j.tra.2010.02.001

[13] Lee J, Mannering F. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. *Accident Analysis & Prevention.* 2002;34: 149-61. Available from: doi:10.1016/S0001-4575(01)00009-4

[14] Lord D, Washington SP, Ivan JN. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis & Prevention.* 2005;37: 35-46. Available from: doi:10.1016/j.aap.2004.02.004

[15] Miaou S-P. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention.* 1994;26: 471-82. Available from: doi:10.1016/0001-4575(94)90038-8

[16] Lord D, Washington S, Ivan JN. Further notes on the application of zero-inflated models in highway safety. *Accident Analysis & Prevention.* 2007;39: 53-7. Available from: doi:10.1016/j.aap.2006.06.004

[17] Miranda-Moreno LF, Fu L. A Comparative Study of Alternative Model Structures and Criteria for Ranking Locations for Safety Improvements. *Networks and Spatial Economics*. 2006;6: 97-110. Available from: doi:10.1007/s11067-006-7695-2

[18] Aguero-Valverde J. Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates. *Accident Analysis & Prevention.* 2013;50: 289-97. Available from: doi:10.1016/j.aap.2012.04.019

[19] Shirazi M, Lord D, Dhavala SS, Geedipally SR. A semi-parametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. *Accident Analysis & Prevention.* 2016;91: 10-8. Available from: doi:10.1016/j.aap.2016.02.020

[20] Geedipally SR, Lord D, Dhavala SS. The negative binomial-Lindley generalized linear model: Characteristics and application using crash data. *Accident Analysis & Prevention.* 2012;45: 258-65. Available from: doi:10.1016/j.aap.2011.07.012

[21] Mussone L, Bassani M, Masci P. Analysis of factors affecting the severity of crashes in urban road intersections. *Accident Analysis & Prevention.* 2017;103: 112-22. Available from: doi:10.1016/j.aap.2017.04.007

[22] Haule HJ, Sando T, Kitali AE, Richardson R. Investigating proximity of crash locations to aging pedestrian residences. *Accident Analysis & Prevention.* 2019;122: 215-25. Available from: doi:10.1016/j.aap.2018.10.008

[23] Barffour M, Gupta S, Gururaj G, Hyder AA. Evidence-Based Road Safety Practice in India: Assessment of the Adequacy of Publicly Available Data in Meeting Requirements for Comprehensive Road Safety Data Systems. *Traffic Injury Prevention.* 2012;13(sup.1): 17-23. Available from: doi:10.1080/15389588.2011.636780

[24] Cafiso S, Di Graziano A, Di Silvestro G, La Cava G, Persaud B. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. *Accident Analysis & Prevention.* 2010;42: 1072-9. Available from: doi:10.1016/j.aap.2009.12.015

[25] Cafiso S, D'Agostino C, Persaud B. Investigating the

influence of segmentation in estimating safety performance functions for roadway sections. *Journal of Traffic and Transportation Engineering* (English Edition). 2018;5: 129-36. Available from: doi:10.1016/j.jtte.2017.10.001

[26] Miaou S-P, Lum H. Modeling vehicle accidents and highway geometric design relationships. *Accident Analysis & Prevention.* 1993;25: 689-709. Available from: doi:10.1016/0001-4575(93)90034-T

[27] Wu L, Lord D, Zou Y. Validation of Crash Modification Factors Derived from Cross-Sectional Studies with Regression Models. *Transportation Research Record: Journal of the Transportation Research Board.* 2015;2514: 88-96. Available from: doi:10.3141/2514-10

[28] ChikkaKrishna NK, Parida M, Jain SS. Identifying safety factors associated with crash frequency and severity on nonurban four-lane highway stretch in India. *Journal of Transportation Safety & Security.* 2017;9: 6-32. Available from: doi:10.1080/19439962.2016.1150927

[29] Zeng Q, Huang H. Bayesian spatial joint modeling of traffic crashes on an urban road network. *Accident Analysis & Prevention*. 2014;67: 105-12. Available from: doi:10.1016/j.aap.2014.02.018

[30] Washington S, Karlaftis MG, Mannering FL. *Statistical and econometric methods for transportation data analysis*. 2nd ed. Boca Raton, FL: CRC Press; 2011.

[31] Geedipally SR. *Examining the application of conway-maxwellpoisson models for analyzing traffic crash data.* Submitted to the Office of Graduate Studies of Texas A&M University in partial fulfillment of the requirements for the degree of Doctor of Philosophy; 2008. 24 p.

[32] Lambert. Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. *Technometrics.* 1992;34(1): 1-14. Available from: doi:10.1080/00401706.1992.10485228

[33] Hosseinpour M, Yahaya AS, Sadullah AF. Exploring the effects of roadway characteristics on the frequency and severity of head-on crashes: Case studies from Malaysian Federal Roads. *Accident Analysis & Prevention.* 2014;62: 209-22. Available from: doi:10.1016/j.aap.2013.10.001

[34] Kumara SSP, Chin HC. Modeling Accident Occurrence at Signalized Tee Intersections with Special Emphasis on Excess Zeros. *Traffic Injury Prevention*. 2003;4: 53-7. Available from: doi:10.1080/15389580309852

[35] Lee J, Mannering F. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. *Accident Analysis & Prevention.* 2002;34: 149-61. Available from: doi:10.1016/S0001-4575(01)00009-4

[36] Qin X, Ivan JN, Ravishanker N. Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accident Analysis & Prevention.* 2004;36: 183-91. Available from: doi:10.1016/S0001-4575(02)00148-3

[37] Shankar V, Milton J, Mannering F. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. *Accident Analysis & Prevention.* 1997;29: 829-37. Available from: doi:10.1016/S0001-4575(97)00052-3

[38] Raihan MA, Alluri P, Wu W, Gan A. Estimation of bicycle crash modification factors (CMFs) on urban facilities using zero inflated negative binomial models. *Accident Analysis & Prevention.* 2019;123: 303-13. Available from: doi:10.1016/j.aap.2018.12.009

[39] McCullagh P, Nelder JA. *Generalized linear models.* 2nd ed. Boca Raton: Chapman & Hall/CRC; 1998.

[40] Casella G, Berger RL. *Statistical inference.* 2nd ed. Australia, Pacific Grove, CA: Thomson Learning; 2002.

[41] Sasidharan L, Menéndez M. Partial proportional odds model—An alternate choice for analyzing pedestrian crash injury severities. *Accident Analysis & Prevention.* 2014;72: 330-40. Available from: doi:10.1016/j.aap.2014.07.025