American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

Spatial Analysis of the Variables Involved in the Frequency and Severity of Traffic Accidents on Rural Highways in Pernambuco

Márcia Macedo[a]*, Maria Maia[b], Emilia Kohlman Rabbani[c], Manoel Marinho[d]

[a,c]*Post-Graduate Program in Civil Engineering, UPE, Pernambuco, Brazil*
[b]*Post-Graduate Program in Civil Engineering, UFPE, Pernambuco, Brazil*
[d]*Post-Graduate Program in System Engineering, UPE, Pernambuco, Brazil*
[a]*Email: marcia.macedo@upe.br*
[b]*Email: nonamaia@gmail.com*
[c]*Email: emilia.rabbani@upe.br*
[d]*Email: marinho75@poli.br*

Abstract

Traffic safety depends on a lot of factors associated with traffic accidents and where it takes place. Analyzing how variables related to traffic accidents influences on its frequency and severity may help on the proposition of significant improvement to the effective reduction of said accidents. The goal of this research is to analyze the impact of contributing factors to traffic accidents of any kind, reducing the number of variables related to the statistic model, adjusting it to the brazilian reallity. The methodology was applied in a case study in a 255km patch of a simple lane countryside highway in the state of Pernambuco. Statistics trials were taken to quantify its possible effects on the frequency and severity of traffic accidents. The analysis showed significant factors that contribute to the frequency and severity of the observed accidents. These factors were the amount of traffic (VDMA), radius of the horizontal curve, greide, age range and day of the week. Even though most of the accidents happened in tangent patches, the most severe accidents take place in turns. It also shows that young people between 18 and 30 years old are 22,7% more likely to get involved in fatal accidents than adults over 50 years old, and that in the weekends the chances of an accident occuring is 67% higher than during a week day. The analysis may be used to provide information on future reviews of parameter selection guidelines, especially regarding turns, based on the main parameters of the highway design to reduce risk of accidents in turns.

Keywords: Crash; Variables; Frequency of the crash; Severity of the crash; Rural Highways.

---

* Corresponding author.

1. Introduction

In a context where resources for infrastructure are increasingly scarce, Brazil continues to incur high social costs due to traffic accidents, spending more than 40 billion per year. Brazil has had one of the highest rates of traffic crashes and fatalities in the world for decades. Although they contain approximately 60% of the world's vehicles, low- and middle-income countries account for 93% of traffic deaths, which are the leading cause of death among youth aged 5 to 29 [1]. Of the crashes that occurred on federal highways (56% of the total), 4% had fatalities and 37% had injured victims, while 59% were crashes having only property damage. Approximately 67% of the fatal crashes occurred in rural areas [2]. Most statistical analyses assume that the variables are independent of one another. However, when considering spatial variables, such as the location where the accident occurred, spatial correlation must be taken into account. Spatial correlation is the measure of correlation of one observation with another through space. Recent studies have shown that ignoring this factor has been the main cause of biases in parameter estimation for Accident Prediction Models [3]. Although more recent studies reveal that most highway accidents occur on straight sections, Radimsky, Matuszkova and Budik [4] noted that curves tend to be riskier, especially when they follow a long tangent section, i.e., a straight road leading to a curve. Existing models for rural highways do not consider the impact of spatial relationships on horizontal curve safety, and the absence of a factor that characterizes the location does not mean that this factor has no effect on accident frequency. It may only indicate that the effect may not be fully known or has not yet been quantified. This study differs from others because it intends to incorporate the impact of spatial relationships into the analysis of the frequency and severity of traffic accidents on rural two-lane highways in Pernambuco. The study will analyze the homogeneous segments using the Spatial (Kernel density-KDE) method. It will address the hierarchical structure of accident data through homogeneous segmentation and capture the indirect impact of the characteristics without conditioning the model to a specific accident type, reducing the number of variables involved in modeling to fit the Brazilian reality. For this study, accident data recorded on rural segments of two-lane highways in Pernambuco from 2006 to 2016 were used. The analysis developed can be used to evaluate the safety effects of geometric road elements and the improvement of geometric designs, thereby enabling the selection of interventions in the highway system that will maximize the benefits achieved.

2. Variables involved in modeling

The occurrence of traffic accidents and their consequences are associated with several factors, of which can be mentioned individual factors and external or environmental factors. At the individual level, the most common are variables associated with drivers (age, gender, driving time, socioeconomic status, etc.), while the external factors include time of day, weather conditions, road characteristics, and vehicle characteristics. Previous studies have shown that the chance of a fatal crash occurring increases with speed and with driver's age, with men more likely to suffer fatal injuries than women. These studies also indicate that there is a higher chance for fatal crashes to occur in the hours between 6 pm and midnight and especially during the early morning hours [3, 4, 5]. For studies carried out on the frequency of accident occurrence, the response variable is quantitative. In some models, especially those on rural roads, when risk factors and traffic accidents have a high degree of complexity, traffic volume is the variable that contributes most to explaining the variability of accidents [6, 7].

Among the geometric characteristics, the horizontal curve is the main area of focus for transportation agencies to improve their road network safety, mainly due to the severity of accidents on curves. A number of studies have focused on the relationship between horizontal curve characteristics and curve safety performance, including design attributes [8]. Straight highway sections appear to have fewer accidents, and roads with more curves have an increased likelihood of accidents. Curves tend to be most dangerous when they follow a long straight section [3, 8, 9]. Several models developed in the past have delineated accidents on horizontal curves based on variables that include traffic volume, curve length, and degree of curvature. The results from these studies showed that increasing the center angle, the transition, and the banking increases accident frequency [9, 10, 11), while increasing the radius reduces accident frequency [9, 12, 13]. Grade is also a variable used in some studies. Steep grades are generally associated with higher accident rates. Grades with slopes of 2.5% and 4% increase accidents by 10% and 20%, respectively, compared to nearly horizontal roads [13]. Considering the same variables and the influence of geometric parameters on accident severity, increased slope in increases accident severity [14], while increased radius [15] and increased banking reduces accident severity [14]. Other variables commonly included in crash prediction models are lane width, shoulder width, and number of lanes. Larger shoulders are assumed to decrease the occurrence of accidents on two-lane highways [15], however, the influence on fatal accidents is lower than that estimated by the models [8]. All of these studies considered different variables, but no spatial variables were included, which could have provided a better context between crash frequency and crash location. However, the main initiatives for this type of modeling indicate problems related to the selection of explanatory variables and challenges regarding how to estimate and incorporate the effect of spatial dependence on traffic accidents into the process [14, 16]. Including spatial relationships in a safety analysis can be an important consideration for a more accurate and comprehensive approach to analyzing actual safety performance [15, 17].

2.1. Spatial analyses

Recent studies have shown that accidents are rarely random events. Spatial analysis of point events, referred to as Point Pattern Analysis (PPA), has been used to analyze the distribution of a set of points (accidents) on a surface (network) [18]. In recent years, the PPA Kernel Density Estimation (KDE) method, developed by Okabe [19], has been widely used in road safety studies to detect dangerous accident locations [20, 21]. Kernel Density Estimation can be considered a spatial statistics technique that demonstrates where concentrations of a given event are allocated within a plane. This feature generates an interpolation, identifying the phenomenon located in the geographical space and highlighting the location where the highest or lowest concentrations of the phenomenon occurs, based on perceptions of intensity of coloration [22]. *Using the absolute number of traffic accidents and their degree of severity along with spatial clustering tools that identify critical points, some studies in Brazil have identified transition regions and spatial trends in traffic accident growth for new areas. The application of spatial methods of analysis, such as that of point patterns, makes it possible to better understand the phenomenon of spatial distribution of traffic accidents, making it possible to identify places with significant concentrations, whether at intersections, sections, corridors, or areas [23].*

3. Methodology

3.1. Study area

The analysis was carried out on a stretch of Highway BR-232, between kilometers 141 and 356, latitudes 8°02'30" S and 8°39'27" S, and longitudes 36°11'56" W and 37°48'57" W (Figure 8). This 255-km stretch of rural highway passes through the municipalities of São Caetano, Pesqueira, Arcoverde, Cruzeiro do Nordeste, and Custódia in northeastern Brazil. The region's economy is primarily commerce and services, and it is passing through a crisis worse than that affecting the rest of Brazil. The Gross Domestic Product (GDP) of the Northeast Region grew last year by only half of Brazil's average. In 2018, the northeastern GDP grew by 0.6%, while Brazil's rose by 1.1%. With an average real monthly worker income of R$1536.00, northeastern Brazil still depends significantly on the federal public sector for investments in infrastructure and road safety [24].

Figure 1: Study area

3.2. Data collection

Accident information was obtained from the Federal Highway Police database for the years between 2007 and 2016, which contains the accident reports, as well as from the DNIT highway base, the OSM cartographic base, and the digital terrain model provided by the Condepe/Fidem Agency. The Average Annual Daily Traffic (AADT) volumes for the years 2014, 2015, and 2016 were obtained from the National Department of Transportation Infrastructure (DNIT), with volumetric and classificatory traffic counts. Traffic counts are permanent and are based on the installation of traffic counting equipment which count and classify the volume of vehicles passing a certain point. For previous years (2007 to 2013), as there was no active collection point within the study area, the AADT values from ANTT's Annual Report (2015) were considered.

3.3. Variable selection

Statistical tests conditioned the selection of variables and the individual contribution of each of the variables

was analyzed. A descriptive analysis was performed and the normality of the data was analyzed using histograms, boxplots, and analysis of the distance between mean and median to measure the degree of deviation or departure from symmetry. In addition to the descriptive methods, the Kolmogorov-Smirnov and Shapiro-Wilks hypothesis tests were be used to assess normality. If the p-value associated with the test statistic is less than the significance level α, the hypothesis that a distribution is normal is rejected. Thus, for significance level α, a decision is made by comparing the two p-values:

If p calculated ≥ p tabulated → H0: sample distribution is normal

If p calculated ≤ p tabulated → H1: sample distribution is non-normal

If the normal distribution cannot be assumed after the preliminary analysis, the non-parametric tests, which don't assume any theoretical distribution of the data, were used. The Mann Whitney non-parametric test was the alternative to the Student t-test, Wilcoxon was the alternative to the paired t-test, and Kruskall-Wallis was the alternative to the one-way ANOVA. Statistical measures were used to assist in the evaluation of the final results, starting with the assessment of the quality of fit between the number of accidents observed and the variables involved in the modeling. A scatter diagram can be used to analyze the correlation between two quantitative variables. If the distribution of the two samples is normal, which assumes a linear behavior for the relationship between the variables, the most commonly used technique is Pearson's r correlation coefficient. Otherwise, for a non-normal distribution, Spearman's ρ coefficient is used. Spearman's ρ coefficient measures the intensity of the relationship between ordinal variables, considering the order of the observations and not the observed value, and is not sensitive to asymmetries in the distribution nor to the presence of outliers. Spearman's ρ converts measurements and calculates the level of correlation between the ranked variables. It is given by Equation 1:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \qquad (1)$$

Where:

n= number of (xi, yi) pairs;

d_i = (rank of xi among the x values) - (rank of yi among the y values);

It is common for accident data samples to not follow a normal distribution (e.g., curve radius, age group, days of the week, grade, number of accidents). In this case, Spearman's rank order and Kendall's rank correlation coefficient (Tau-b) can be used to analyze the correlation between variables, replacing the original data by ordered ranks. These methods are used when at least one of the variables has ordinal characteristics (e.g., age group, land use, time of day). The fit of the variables to the accident data was analyzed by looking at the ρ values. The signs of the coefficients indicate either a direct or inverse relationship between each of the explanatory variables and the response variable and range from -1 to 1. The closer the value is to one of the extremes, the greater the association between the variables. A negative correlation means that the variables vary in the opposite direction, that is, higher values of one variable are associated with lower values of the other. The

significance level of these variables must be below 0.05 for a desired confidence level of 95%. The choice of a 95% confidence level is for comparative purposes, as this is the standard that most other studies have used. The significance level of the intercept, in principle, should be kept, even if it is below the desired value. It should only be eliminated from the modeling if it is almost completely certain that the regression line passes through the origin of the graph [25].

3.4. Identification of homogenous sections

Road segments were grouped according to type of roadway, land use, road layout, and grade. Vertical profiles were prepared from the altimetric database of the Condepe/Fidem agency to obtain the grade values. To differentiate straight sections and curves, the method based on Kernel density was used.

4. Results and discussions

A lack of standardization among the accident reports, as well as a lack of rigor in filling them out, diminish their reliability and their ability to be used for research. An analysis was therefore done to identify any absence or inconsistency in the information recorded in the reports. Tables that did not present all of the necessary information, such as location, type, and date of the accident, were excluded from the sample. A database was developed containing detailed information on road widths, shoulder conditions, road curvature, grade, and AADT on 215 km of rural roads in Pernambuco. This was achieved using geoprocessing and Geographic Information System (GIS) tools to extract relevant attributes from the road alignment, spatial characteristics of the surroundings, and traffic flow, combining it with the accident database created for the study [26]. The accident data included in the database contained all accidents recorded during the 10-year period from January 2007 to December 2016. The sample size considered was 428 observations. Although the database was designed to receive information from many variables, the tests focused on a model that could work with the variables that could be collected within the Brazilian reality. Two groups of variables were considered, one related to road geometry (group 1) and the other to spatial variables (group 2). The following variables were tested in group 1: lane width, shoulder width and type, segment length, grade and banking, curve radius and curve length, including the length of the transition spiral, if any. The spatial variables considered were: cause of crash, age group, crash type, day of the week, time of day, road layout, condition, cause 1, direction, road type, land use, and phase of the day. The explanatory variables of group 2, such as phase of the day, weather conditions, and land use were also used to generate thematic maps in order to better understand the spatial distribution of these variables. For example, in 2016, there were 387 accidents on the section of road under study. The dependent variable (total accidents) was distributed according to severity into the categories Fatal, Victims with serious injuries, Victims with moderate injuries, Victims with minor injuries, and Property damage without victims. The choice of tests to be performed on group 1, whether parametric or non-parametric, depends on two basic conditions: normality and homogeneity. First, descriptive analyses were performed to verify whether the frequency of accidents on rural two-lane highways in the state of Pernambuco has a normal or non-normal distribution, as the theoretical basis assumes a non-normal distribution. The null hypothesis, that is, $p>0.05$, means that the data follow a normal distribution. Both the Shapiro-Wilk and Kolmogorov-Smirnov normality tests showed results of $p<0.05$, so the null hypothesis was rejected, meaning the data distribution is non-normal

(Table 1).

Table 1: Values of normality tests of accident frequency on two-lane highways in the State of Pernambuco, grouped by severity

Normality Tests						
	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	GL	ρ	Statistic	GL	ρ
Total Accidents	.318	243	,000	.671	243	.000
Moderate Injuries	.438	243	,000	.564	243	.000
Serious Injuries	.445	243	,000	.464	243	.000
Minor Injuries	.273	243	,000	.845	243	.000
Property damage only	.540	243	,000	.234	243	.000
Fatal Accidents	.513	243	,000	.381	243	.000

Figure 2 shows the boxplot and expected normal curve for the distribution of the number of occurrences of traffic accidents with severe injuries, moderate injuries, and minor injuries, recorded in the study area during the period from 2007 to 2016. Figure 3 shows the boxplot and expected normal curve for the distribution of the number of occurrences of traffic accidents with fatalities and with property damage only, recorded in the study area during the period from 2007 to 2016.

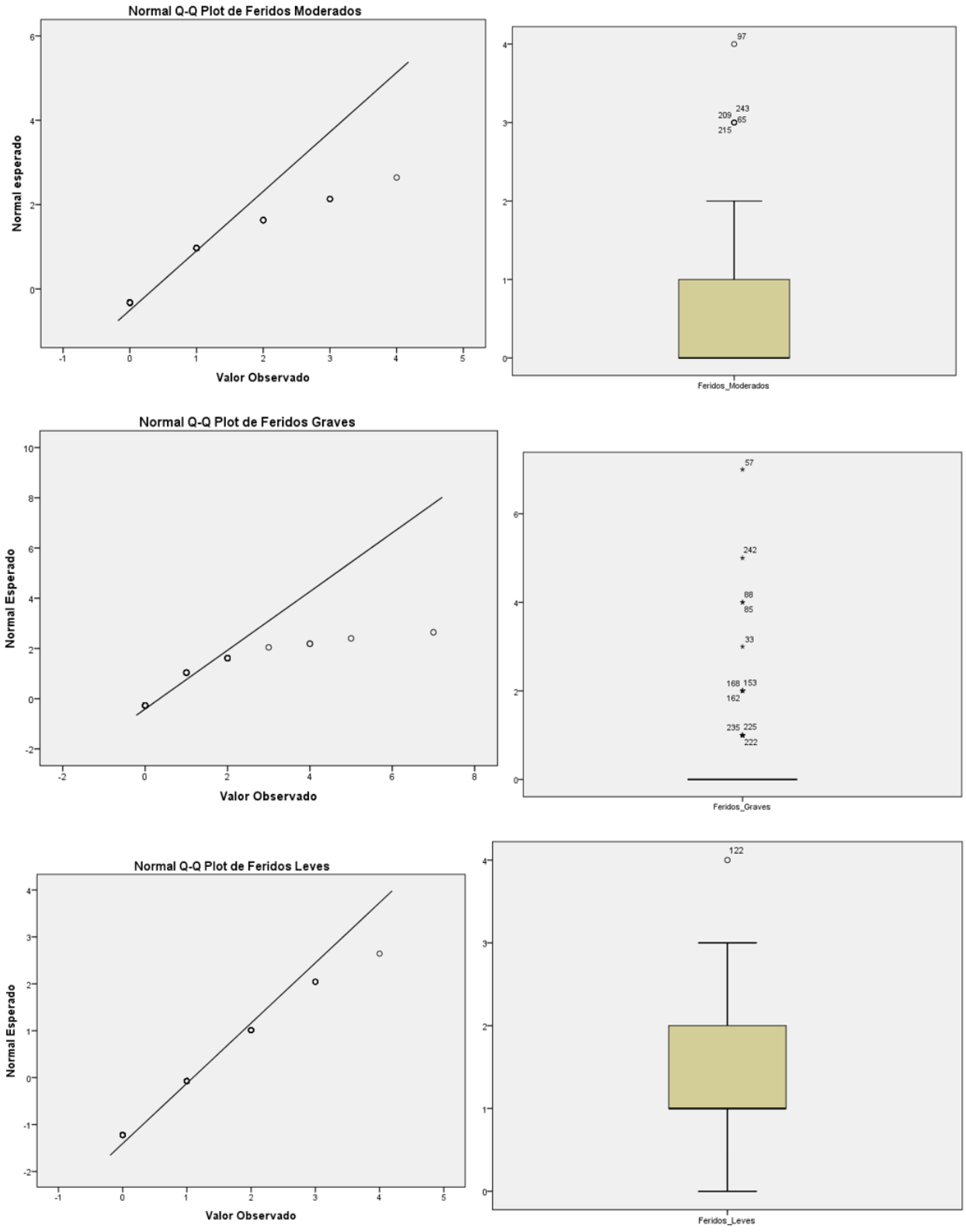


Figure 2: Boxplot for the distribution of the number of traffic accidents with serious injuries, moderate injuries, and minor injuries, recorded in the study area from 2007 to 2016.

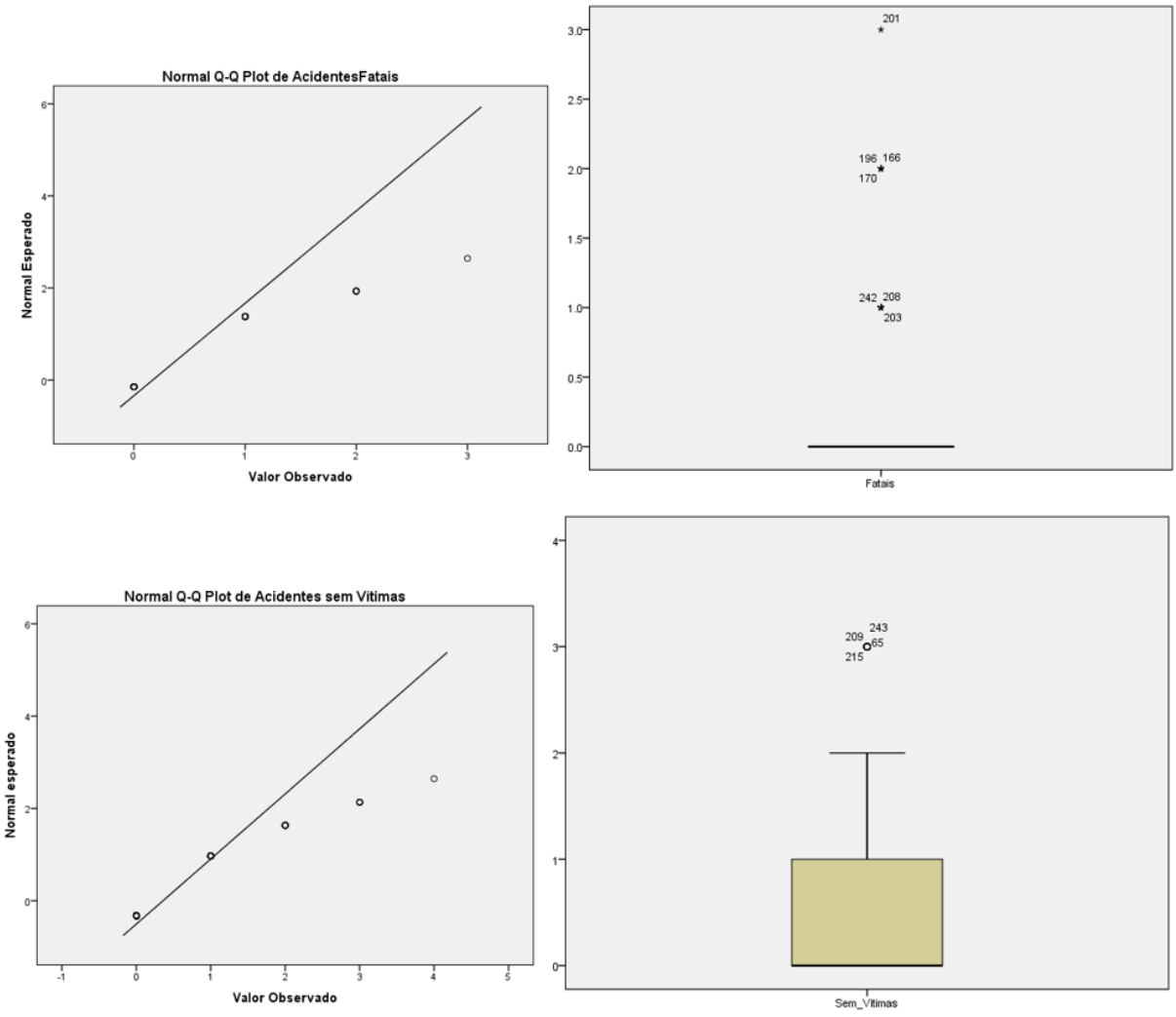


Figure 3: Boxplot for the distribution of the number of fatal accidents and accidents having property damage only, recorded in the study area from 2007 to 2016.

From these figures, it can be seen that the five types of events studied contained outliers, that is, points outside of 1.5 times the interquartile deviation plus the third quartile. There were seven outliers in the distribution of traffic accidents with fatalities, 11 outliers in the distribution with severe injuries, five outliers in the distribution with moderate injuries, one outlier in the distribution with minor injuries, and four outliers in the distribution of traffic accidents having property damage only. The confirmation of a non-normal distribution meant that non-parametric tests should be used. Nevertheless, parametric alternatives with greater statistical power were initially sought because non-parametric tests tend to be less robust and reliable than the parametric ones, and are also more difficult to apply when the analysis is more complex, such as comparisons of means that involve more than one factor. This alternative is valid when the deviations from normality are small and it does not compromise the results of the analysis. The influence of the route (whether straight or curved) on accident severity was also analyzed. The descriptive statistics presented in Table 2 show the number of accidents on straight segments is higher than on curves, however, the average values for accidents that occurred on straight sections were very close to the averages on curves.

Table 2: Descriptive statistics of accident frequency on two-lane highways in the state of Pernambuco, grouped by severity

	Route	N	*Mean*	*Std. Deviation*	*Std. Error Mean*
Moderate Injuries	Straight	202	.36	.728	.051
	Curve	41	.34	.617	.096
Severe Injuries	Straight	202	.31	.776	.055
	Curve	41	.56	1.163	.182
Minor Injuries	Straight	202	1.13	.764	.054
	Curve	41	.90	.831	.130
Property Damage Only	Straight	202	.05	.227	.016
	Curve	41	.05	.218	.034
Fatal Accidents	Straight	202	.17	.500	.035
	Curve	41	.17	.495	.077
Total Accidents	Straight	202	2.0248	1.28302	.09027
	Curve	41	2.0244	1.54091	.24065

Because the sample size is larger than 30, the t-test can be used without regard for the non-normality of the data. The t-independent test showed that, on average, straight sections have a greater influence on traffic accidents with moderate injuries ($t(241)=0.164$; $p<0.05$), minor injuries ($t(241)=1.740$; $p<0.05$), and property damage only accidents ($t(241)=0.147$; $p<0.05$). Fatalities ($t(57)=-0.028$; $p<0.05$) and serious injuries ($t(241)=0.164$; $p<0.05$) are more frequent on curves (Table 3).

Table 3: Results from the t-independent test for analyzing the influence of road design on the accident severity on two-lane highways in Pernambuco

	t	Degrees of Freedom	Sig. (2-tailed)	Difference from the mean	Standard deviation the mean	95% Confidence Interval from *Lower*	*Upper*
Moderate Injuries	.164	241	.870	.020	.122	-.220	.260
Severe Injuries	-1.340	47.475	.187	-.254	.190	-.635	.127
Minor Injuries	1.740	241	.083	.231	.133	-.031	.493
Property Damage Only	.147	241	.884	.006	.039	-.071	.082
Fatal Accidents	-.028	57.792	.977	-.002	.085	-.172	.168

The second condition to be met to continue the parametric tests is homogeneity of variances. However, as shown in Table 4, Levene's test presented values of $p<0.05$, not satisfying the null hypothesis that the variances are equal.

Table 4: Results of the Levene Test to analyze the homogeneity of variance of accidents on two-lane highways in Pernambuco, grouped by severity

	Levene Test	df1	df2	Sig.
Moderate Injuries	2.572	6	236	.020
Severe Injuries	4.460	6	236	.000
Minor Injuries	2.368	6	236	.031
Property Damage Only	5.065	6	236	.000
Fatal Accidents	4.623	6	236	.000

Concluding that the samples do not obey a normal distribution and are not homogeneous, new non-parametric tests were performed to consider the possible interactions between the variables in group 1. The Mann-Whitney test showed that the road design interferes with the frequency of crashes where severe injures ($U = 3811$; $p<0.05$), fatalities ($U = 4127$; $p<0.05$) and minor injuries ($U = 3384$; $p<0.05$) occur, while crashes where moderate injuries ($U = 4090$; $p>0.05$) or property damage only ($U = 4117$; $p>0.05$) occurs, did not suffer this interference. The results are presented in Table 5.

Table 5: Results of the Mann-Whitney non-parametric test to analyze the influence of the road design on the severity of accidents on two-lane highways in Pernambuco

	Moderate Injuries	Serious Injuries	Minor Injuries	Property Damage Only	Fatal Accidents
Mann-Whitney U	4090.500	3811.000	3384.000	4117.500	4127.500
Wilcoxon W	24593.500	24314.000	4245.000	4978.500	24630.500
Z	-.161	-1.124	-2.006	-.147	-.058
ρ	.872	.021	.045	.883	.003

These variables could be analyzed separately by performing, for example, a chi-square analysis of independence, to know the percentage of accidents on the road section studied, according to various combinations (severity, day of the week, age group, etc.), but this would increase the chance for error. The option selected was to explain to the model that these variables belong to a single set with multiple responses by adopting the non-parametric related samples test. The Friedman test showed that the frequencies differ according to accident severity [$X2(4) = 334.914$; $p<0.001$]. Figure 4 shows an example of the analysis of accident frequency according to severity using the Friedman test.

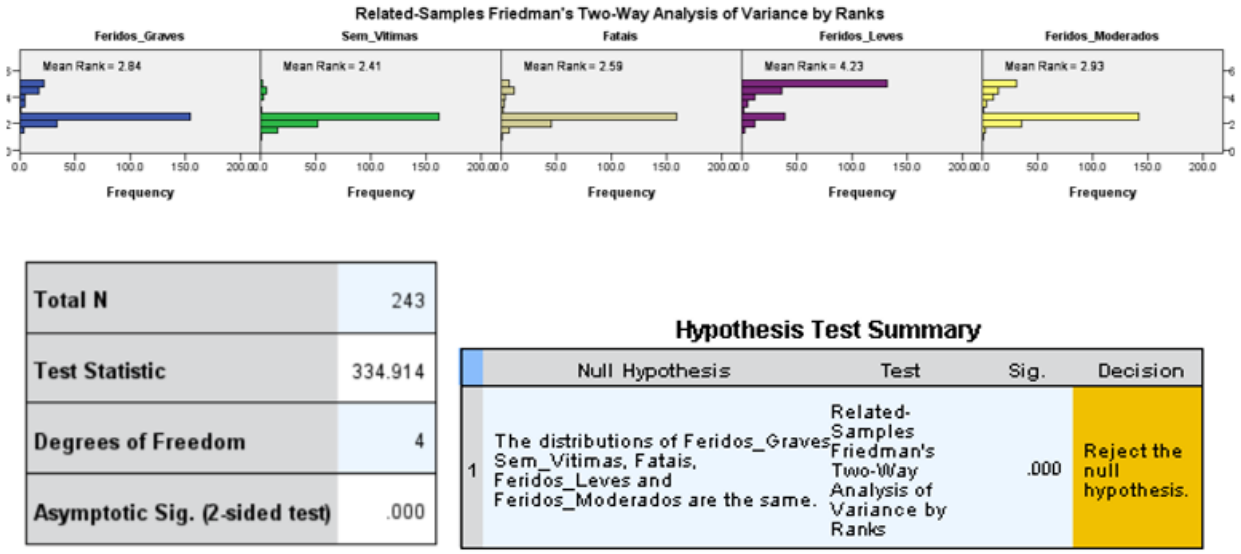


Figure 4: Results of Friedman's non-parametric test analyzing the frequency of accidents on two-lane highways in Pernambuco, grouped by severity

The results of the analyses of the combinations made between the variables of group 1 were compiled and, from these, the percentage distribution of the accident frequency by severity on two-lane roads in Pernambuco (Table 6) and the percentage distribution by accident type and severity on two-lane roads in Pernambuco (Table 7) were obtained. Analyses were also performed to evaluate whether or not the variables in group 1 were dependent among themselves. The following questions were answered: a) if accident frequency is directly linked to age group; b) if day of the week influences accident frequency; and c) if weather conditions influence accident frequency.

Table 6: Percentage distribution of accident frequency on two-lane highways in Pernambuco, grouped by severity

Accident Severity	Percentual of total accidents
Fatal	3.6 %
Victims with serious injuries	10.90 %
Victims with moderate injuries	20.15 %
Victims with minor injuries	26.88 %
Property damage only	38.40 %
Total	100.00%

Table 7: Percentage distribution of accidents on two-lane highways in Pernambuco, grouped by type and severity

Accident Type	Degree of Severity		
	Victims	Property damage	Total
Accidents involving only one vehicle			
Collision with animal	3.3%	5.0%	3.8%
Collision with cyclist	0.4%	0.2%	0.3%
Collision with pedestrian	2.7%	7.5%	4.3%
Rollover	9.7%	5.7%	8.3%
Run-off-road	8.7%	8.4%	8.6%
Others	9.2%	4.5%	7.6%
Total (only one vehicle)	34.0%	31.5%	33.0%
Accidents involving more than one vehicle			
T-bone collision	15.0%	7.2%	12.3%
Head-on collision	9.2%	2.4%	6.9%
Read-end collision	23.1%	35.3%	27.2%
Sideswipe collision	12.6%	12.8%	12.6%
Others	5.8%	10.6%	7.5%
Total (more than one vehicle)	65.9%	68.4%	66.7%
Total (all accidents)	100%	100%	100%

For age group, three categories were considered: (1) between 18 and 30 years; (2) between 30 and 50 years; and (3) over 50 years. The residual values were expected to be in the range from -1.96 to 1.96. For the 30 to 50 age group, the residuals are very small, i.e. the expected accident frequency for this age group is very close to the actual value. For the age group over 50, the residual is a slightly higher (0.5), but still close to the expected value, while for the age group 18 to 30, the residual was 2.7. For the weekday group, two categories were considered: (1) Weekday (Monday, Tuesday, and Wednesday) and (2) Weekend (Thursday, Friday, Saturday, and Sunday). During the week, the residuals are very small, that is, fewer accidents occur than expected, while the weekend residual values are high, that is, more accidents occur on weekends than during the week. It can be concluded that, using the over 50 age group as a reference, young people between 18 and 30 years old have a 22.7% greater chance of being involved in a fatal accident, while adults between 30 and 50 years old have a 34% lower chance of getting involved in an accident. On weekends, the chance of an accident occurring is 67% greater than during the week. The results of the statistics are presented in Tables 8 and 9. Regarding weather conditions, this information began to be recorded in 2016 and the classes are very differentiated (sunny, clear sky, rain, cloudy, fog, etc.). This stratification of classes made the sample sizes very small, with fewer than 15 samples per class and, for this reason, this group was not considered.

Table 8: Analysis of accident frequency distribution by age group on two-lane highways in Pernambuco

			Age group		
			18 - 30 years	30 - 50 years	> 50 years
Accident Frequency	No	Count	81	90	33
		Expected Count	82.9	90.7	44.8
		% Accident Frequency	39.7%	44.1%	16.2%
		% Age Group	73.0%	75.0%	55.0%
		% of Total	29.7%	33.3%	12.1%
		Residual Adj.	-.6	-.2	-4.0
	Yes	Count	30	30	22
		Expected Count	21.2	29.3	20.5
		% Accident Frequency	36.6%	45.5%	26.8%
		% Age Group	47.6%	25.0%	36.1%
		% of Total	12.3%	11.1%	9.0%
		Residual Adj.	2.7	.2	.5
Total		Count	111	120	60

Table 9: Analysis of accident frequency distribution by day of the week on two-lane highways in Pernambuco

			Day of the week	
			Weekday	Weekend
Accident Frequency	No	Count	81	90
		Expected Count	73.3	97.7
		% Accident Frequency	47.4%	52.6%
		% Age Group	90.0%	75.0%
		% of Total	38.6%	42.9%
		Residual Adj.	2.8	-1.7
	Yes	Count	9	30
		Expected Count	16.7	22.3
		% Accident Frequency	23.1%	76.9%
		% Age Group	10.0%	25.0%
		% of Total	4.3%	14.3%
		Residual Adj.	-4.8	3.8
Total		Count	111	90

With regard to group 2, consistent relationships were expected between the potential variables in the model, i.e., curve radius, curve length, and grade, parameters that were measured in this study. Where the correlation between variables is high, there is nothing to be gained from including both of them, so some parameters were excluded from the model early on. The design speed, for example, was not considered due to its correlation with curve radius (used in the calculation of this parameter). An analysis of the correlations between the selected variables was performed to check for redundancy. Spearman's Rho was used in preference to Pearson's r due to the non-normal distribution of the variables and the non-linear relationships between them. Spearman's Rho converts measurements into rankings and calculates the level of correlation between the ranked variables. The closer the Rho value is to ±1.00, the greater the degree of correlation between the two variables. When two variables are strongly correlated, one can be excluded from the model as a redundant variable. Generally, there were weak correlations between most variables, as shown in Table 10. The exceptions were a strong negative

correlation between lane width and shoulder width (-0.317) and between curve radius and curve length (-0.228). The former correlation reflects the application of the road design cross section standards. It was not possible to precisely identify the type of shoulder and its width, so standard values used by the High Safety Manual (HSM) were considered, which could justify a strong correlation between these variables. The second correlation was related to the fact that curve radius is used as a parameter to calculate the curve length. Based on these strong correlations, shoulder width and curve length were excluded from the model. For most of the other variables, the correlation with traffic volume was weak, and AADT was included in the model because it was considered a strong predictor variable.

Table 10: Correlation matrix for the independent variables (Spearman's Rho)

Variables	Lane Width	Shoulder Width	Curve Length	Grade	Curve Radius	AADT
Lane Width	1.000					
Shoulder Width	-0.317	1.000				
Curve Length	-0.040	-0.077	1.000			
Grade	-0.002	-0.002	-0.006	1.000		
Curve Radius	0.085	0.088	-0.228	0.003	1.000	
AADT	0.328	-0.168	-0.170	0.000	0.164	1.000

To determine the best shape for the model, variables were added using the stepwise method, forward direction. This is an automatic variable selection procedure that starts from a null model and adds one variable at a time. The null model was a base model, with only one parameter representing the same mean value μ for all observations yi, in this case, the accidents that occurred on the respective segment. Wald's Chi-square statistic was used to test the statistical significance of the overall model and each of the independent variables contained within it. This test indicates whether the inclusion of an independent variable makes a statistically significant difference in the model's accident predictions. Some independent variables were included in the model because they provided a better fit and contributed to a parsimonious model, even though they did not have a statistically significant difference. It is likely that the non-significant variables had a modifying relationship with other predictors in the model. To be able to spatialize accident severity as a function of segment type (straight or curve) and slope (downhill or uphill), groupings were performed by subsections as a function of roadway type, land use, terrain type, roadway layout, and grade. To ensure that the segmentation considered all of the spatial characteristics, without regard to the accident frequency, a risk Index was created. According to the characteristics presented most often in the literature and their respective ranges, values varying from 1 to 3were established, with 1 being low risk, 2 being medium risk, and 3 being high risk (Table 11). The index itself ranges from 3 to 8, with 3 being the lowest risk and 8 being the highest. For example, a stretch 1880 m in length with an AADT of 4800 vpd on a slope has a risk index of 5, while a stretch with a VDMA of 4800 vpd on a slope with a curve of radius 500 m has a risk index of 7, as shown in Table 12.

Table 11: Estimated values for calculating the risk index

Variables	Categories	Estimated values
AADT [10]	≤ 5500 vpd	1
	>5500 vpd	2
Curve radius (m) [4]	≤ 600	3
	600 – 1500	2
	> 1500	1
Grade (%) [15]	Negative	3
	Positive or zero	1
Segment length (m) [7]	≤ 200	1
	200-1000	2
	≥ 1000	3

Table 12: Risk index composition example

Variables	AADT ≤ 5500 vpd	Curve radius (m) ≤ 600	Grade (%) Negative	Segment length (m) ≥ 1000	Day Weekend
Values	1	-	3	1	5
	1	3	3	-	7

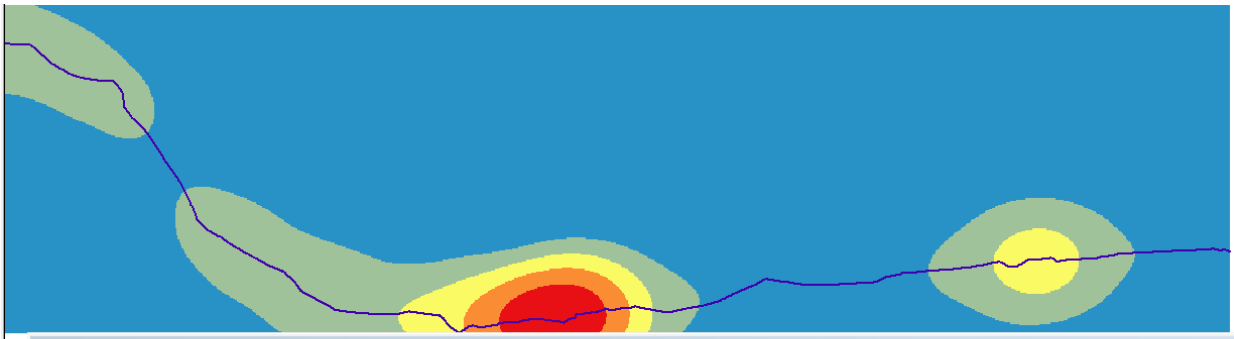
A risk index was also created for the categorical variables, according to the statistical results. Values were established ranging from 1 to 3, with 1 being low risk, 2 being medium risk, and 3 being high risk. Table 13 shows the risk index for the day of the week and age group variables.

Table 13: Estimated values for calculating the risk index for the categorical variables Day of the Week and Age Group

Variables	Categories	Estimated values
Day of the Week	Weekday	1
	Weekend	2
Age group (years)	18 - 30	3
	30 - 50	1
	> 50	2

The kernel estimation technique was applied, based on the index, in order to identify areas with the same spatial characteristics, according to Figure 5. In crossing the spatial variable with the geometric variables, for example, variables "road layout" and "grade", the kernel estimation technique was also applied in order to identify the concentrations between the road layouts and the presence of steepness or slope, verifying differences in concentration. The procedure was repeated, taking various combinations of clusters into account. The accidents

that fit within the selected segment of each highway were associated with it.



Table

Join_Output_2

FID	Shape	BR232_KM_148_356_ACIDENTE	FID_	TIPO	NUME	KM_INICI	KM_FIN	NOTE	Nube	Type	Length	Radiu	L	Cota_TS_	Cota_
0	Polyline	0	0	CURV	7	150,1	150,2		7	Curv	52,2947	3000	0	580,405	0
1	Polyline	1	0	CURV	2	148,7	149		2	Curv	271,478	600	5	591,302	591,48
2	Polyline	2	0	CURV	5	149,4	149,7		5	Curv	298,179	700	5	600,32	599,14
3	Polyline	3	0	CURV	9	150,4	150,6		9	Curv	211,187	600	4	590,045	591,71
4	Polyline	4	0	CURV	11	154,6	154,6		11	Curv	15,2856	5000	0	603,738	0
5	Polyline	5	0	CURV	13	155,7	155,7	MENOR Q	13	Curv	29,5435	2200	0	598,209	0
6	Polyline	6	0	CURV	15	155,8	155,8	MENOR Q	15	Curv	21,2374	2200	0	595,895	0

Figure 5: Table of homogeneous segments according to spatial criteria

With the structuring of the database, it was possible to compare the distribution of the severity and frequency of accidents during the period studied on the curved sections, considering the slope of the terrain, as shown in Figure 6. The results show that approximately 68% of the accidents occurred on straight sections and 32% on curved sections, however, attention should be drawn to the severity of the accidents. Of the accidents that occurred on straight sections, 29% were serious and 9% fatal, compared to 35% and 18%, respectively, for curved sections (Figure 6). The analyses also showed that, of the 41% of accidents on curved sections that had a descending slope, 40% were serious accidents and 19% fatal, while on straight sections, the percentage was less than 1% for all cases (Figure 7).

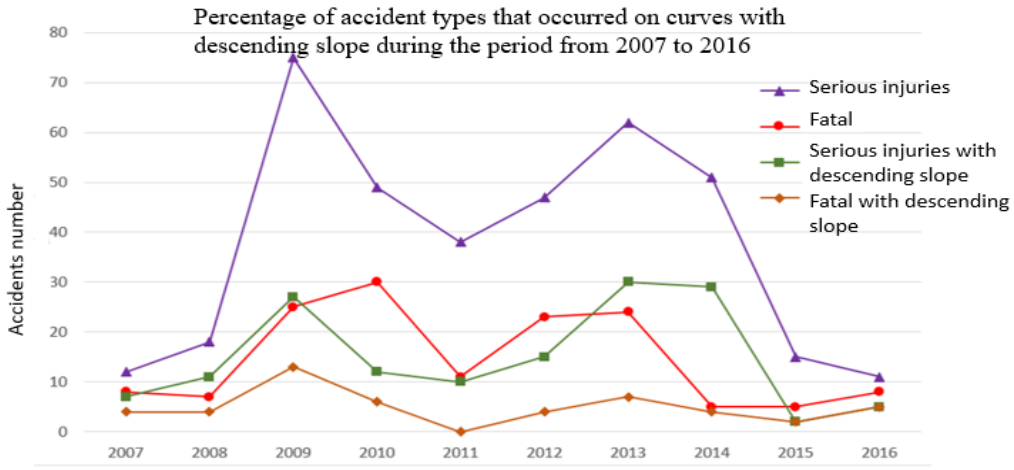


Figure 6: Graph of the percentage of accident types that occurred on curves with descending slope during the period from 2007 to 2016

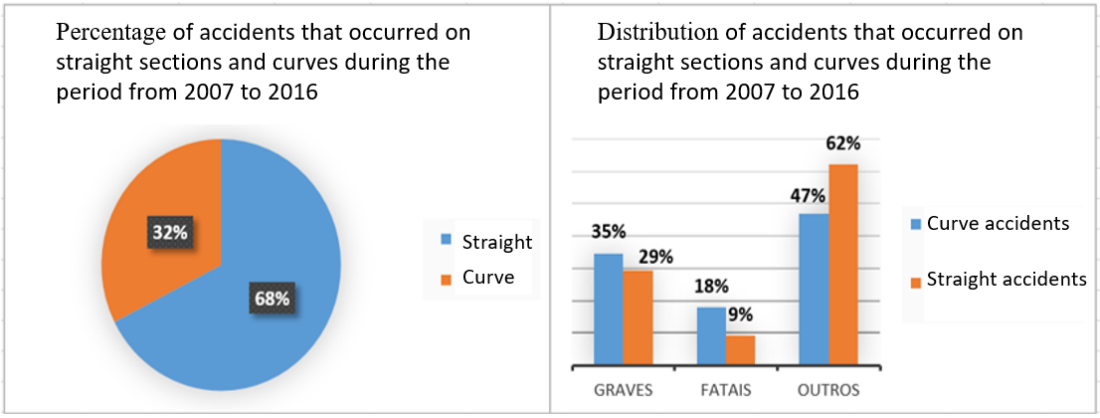


Figure 7: Graph of the distribution of accidents that occurred on straight sections and curves during the period from 2007 to 2016

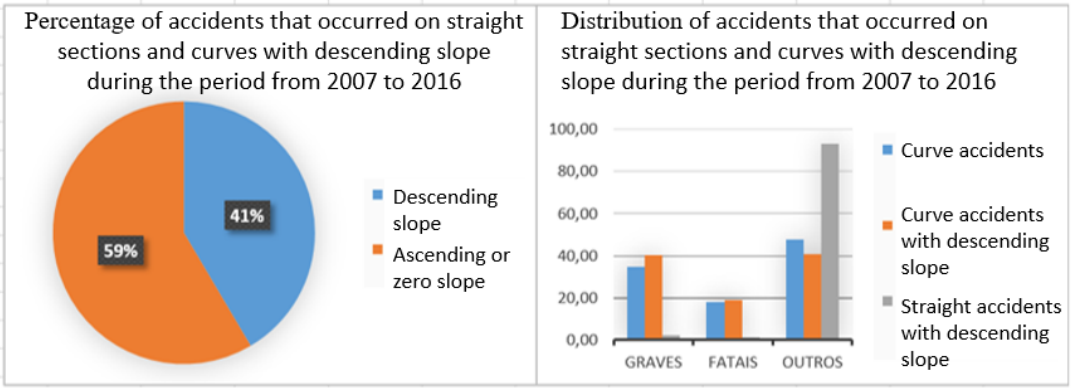


Figure 8: Graph of the distribution of accidents that occurred on straight sections and curves with descending slope during the period from 2007 to 2016

5. Conclusions

Brazil ranks fifth in deaths from traffic accidents, according to the United Nations. To reduce this number, effective road safety management actions are needed. This management depends on many factors associated with accidents and the places where they occur, of which the geometric design of the road and spatial analysis of the contributing factors are indispensable. For rural two-lane highways, which were the focus of this study, it was observed that, not only did these areas concentrate the highest number of fatal accidents, especially on curves, but the lack of accurate data often prevents the use of consolidated international methodologies for safety analysis that would be able to suggest significant improvements for reducing traffic accidents and lowering mapping costs. The structuring of the database using a GIS focused on accident data, compared through accident type, accident rates, accident indexes, the situation of those involved, weather conditions, vehicles, and in relation to the referenced period. Its structure sought to visualize the geometric parameters, principally of curves, not only through designs that did not always reflect reality, but through semi-automated processes proposed in this study that combined several available and current databases. By addressing these concepts, this study sought to better understand the influence of contributing factors on the frequency and

severity of accidents on two-lane highways, considering two groups of variables, one related to road layout (straight or curve) and the other to spatial variables. The influence of geometric parameters (horizontal curvature, traffic volume), geometric variables (radius, grade, lane width, shoulder width, and length) and categorical variables (day of the week and age group) were investigated. The chosen site was a section of rural two-lane highway on BR 232 in Pernambuco, which runs from km 141.00 in São Caetano to km 356.00 in Custódia. Regarding the question of how to perform segmentation of sections with similar geometric characteristics (homogeneous sections) while considering the impact of spatial relationships on horizontal curve safety, kernel estimation was less complex and had fewer steps, however, it was more sensitive to input data, with accurate spatial location information being essential to obtain good results. The structuring of the database, the automation of the process, and its visual output facilitated the interpretation of the results. The finding that homogeneous segmentation based on kernel estimation offers good results shows that it is possible to perform a hierarchization and establish which geometric characteristics have the greatest influence on the occurrence and severity of traffic accidents on Brazilian rural two-lane highways. The selection of the explanatory variables to be included in the model was made using a stepwise methodology, initially entering all available variables and testing each of the three segmentations in order to keep only variables that were significant. This method was applied using different sets of variables to avoid problems due to correlation of variables. The results showed that, for the sample size considered (428 observations), the models were able to capture the significant factors that contributed to the observed accident frequencies. These factors were traffic volume (AAFT), horizontal curve radius, grade, and the length of the section. Since these factors result in positive coefficients in the models, when they increase, it is reasonable to expect that the accident frequency will also increase. This model can be used to provide information for future revisions of the road design selection guidelines, based on the principal road design parameters available in the Brazilian database. The results from the modeling can be used especially for curve selection, to reduce the risk of accidents on curves. One advantages of this statistical analysis was to consider the trend over time. In general, the analysis period depends on the availability of traffic data and accident data, but in the literature, numerous studies have shown that periods longer than five years can introduce bias into the mathematical model for the spatial variables, linked to the location in the network and/or the natural trend over time. This study found that rural highways in the state of Pernambuco are still 3.3 times more likely to have fatal accidents than urban highways. About 58% of fatal road accidents occur on horizontal curves, according to visual inspection when filling out the reports, which means that this number may be higher. Based on the age range of those over 50, young people between 18 and 30 are 22.7% more likely to be involved in fatal accidents, while adults between 30 and 50 are 34% less likely to be involved in accidents. On weekends, the accidents are 67% more likely than during the week. The analysis presented represents an important step towards revising road design guidelines. For example, given a set of AAFT, curvature, and land use classification constraints, a designer can manipulate a range of radii to arrive at an optimal solution with lower cost and a lower risk of accidents. It should be emphasized that any improvement initiative based on results from the model needs to be subject to a full economic evaluation of alternative options, crash performance monitoring, and post-implementation evaluation. It is concluded that, according to international experience, there are many more variables that can be used to calibrate sophisticated models, which have their own advantages and disadvantages. A statistical model with more variables can help to better understand how these factors influence and describe road safety. However, including too many variables in a model can make it

unstable and prone to perform poorly when applied to a new sample. This study has achieved its goal by evaluating a simpler model that helps improve road safety management. This study can be considered a starting point for similar actions and more detailed studies using other categories of roads and other variables. As future studies, it is intended to expand the area of analysis and apply the methodology to other regions with similar characteristics to northeastern Brazil and other developing countries, not for the model's transferability, but for the model's adequacy and variables of interest at the regional level and subsequent adequacy at the national level.

References

[1]. Organização Pan Americana da Saúde (OPAS). (2019). Segurança no Trânsito nas Américas.

[2]. Instituto de Pesquisa Econômica Aplicada (IPEA). (2019). Acidentes de Trânsito nas Rodovias Federais Brasileiras Caracterização, Tendências e Custos para a Sociedade. Brasília: IPEA e PRF.

[3]. Abdulhafedh, A. "A Novel Hybrid Method for Measuring the Spatial Autocorrelation of Vehicular Crashes: Combining Moran's Index and Getis-Ord G*i Statistic." Open Journal of Civil Engineering, vol. 7, pp. 208-221, 2017.

[4]. Radimsky, M., Matuszkova, R. and Budik, O. "Relationship between horizontal curves design and accident rate." Jurnal Teknologi, vol. 78, pp. 5–2, 2016.

[5]. Kong, C. and Yang, J. "Logistic regression analysis of pedestrian casualty risk in passenger vehicle collisions in China." Accident Analysis & Prevention, vol. 42, pp. 987–993, 2010.

[6]. Greibe, P. "Accident prediction models for urban roads." Accident Analysis & Prevention, vol. 35 (2), pp. 273-285, 2003.

[7]. Lord, D., Manar A. and Vizioli A. "Modelling crash-flow-density and crash-flow-V/C ratio relatinships for rural and urban freeways segments." Accident Analysis & Prevention, 37 (1), 185-199, 2005.

[8]. Sameen, M. I. and Pradhan, B. "Forecasting severity of traf5ic accidents using road geometry extracted from mobile laser scanning data," in Conference on Remote Sensing (ACRS), 2016.

[9]. Aram, A. "Effective safety factors on horizontal curves of two-lane highways." Journal of Applied Sciences, vol. 10 (22), pp. 2814–2822, 2010.

[10]. Dong, C; S. S. Nambisan; S. H. Richards and Ma, Z. "Assessment of the effects of highway geometric design features on the frequency of truck involved crashes using bivariate regression." Transportation Research Part A: Policy and Practice, vol. 75, pp. 30–41, 2015.

[11]. Schneider, W. H., Savolainen, P. T. and Moore, D. N. "Effects of Horizontal Curvature on Single-Vehicle Motorcycle Crashes along Rural Two-Lane Highways." Transportation Research Record: Journal of the Transportation Research Board, vol. 2194 (1), pp. 91–98, 2010.

[12]. Ma, J. and Li, Z. "Bayesian Modeling of Frequency-Severity Indeterminacy with an Application to Traffic Crashes on Two Lane Highways." American Society of Civil Engineers, 2010.

[13]. Hosseinpour, M., Yahaya, A. S., Sadullah, A. F., Ismail, N. and Ghadiri, S. M. R. "Evaluating the effects of road geometry, environment, and traf5ic volume on rollover crashes." Transport, vol. 31 (2), pp. 221–232, 2016.

[14]. Chikkakrishna, N. K., Parida, M. and| Jain, S. S. "Identifying safety factors associated with crash frequency and severity on nonurban four-lane highway stretch in India." Journal of Transportation

Safety & Security, vol. 9 (6), pp. 32-30, 2017.

[15]. Anastasopoulos, P. C., Shankar, V. N, Haddockc, J. E. and Mannering, F. L. "A multivariate tobit analysis of highway accident injury-severity rates." Accident Analysis & Prevention, vol. 45, pp. 110–119, 2012.

[16]. Hadayeghi, A., Shalaby, A.S. and Persaud, B. "Development of planning level transportation safety tools using Geographically Weighted Poisson Regression." Accident Analysis and Prevention, vol. 42, pp. 676-688, 2007.

[17]. Findley, D. J., Hummer, J. E., Rasdorf, W., Zegeer, C. V. and Fowler, T. J. "Modeling the impact of spatial relationships on horizontal curve safety." Accident Analysis & Prevention, vol. 45, pp. 296-304, 2012.

[18]. Ervin, D. (2015). Advanced Spatial Analysis.

[19]. Okabe, A., Satoh, T. and Sugihara, K. "A kernel density estimation method for networks, its computational method and a GIS-based tool." Int. J. Geogr. Inf. Sci., vol. 23 (1), pp. 7–32, 2009.

[20]. [20] Nie, K., Wang, Z., Du, Q., Ren, F. and Tian, Q. "A network-constrained integrated method for detecting spatial cluster and risk location of traffic crash: a case study from Wuhan, China." Sustainability, vol. 7 (3), pp. 2662–2677, 2015.

[21]. Vemulapalli, S.S. "GIS-based spatial and temporal analysis of aging-involved crashes in Florida." Doctoral Thesis, The Florida State University, 2015.

[22]. Agbelie, B. R. D. K. "A comparative empirical analysis of statistical models for evaluating highway segment crash frequency." Journal of Traffic and Transportation Engineering, vol. 3 (4),pp. 374–379, 2016.

[23]. Queiroz, M. P., Loureiro, C. F. G. and Yamashita, Y. "Caracterização de padrões pontuais de acidentes de trânsito aplicando as ferramentas de análise especial," in XVIII ANPET, 2018.

[24]. UOL Economia (2020). Available:https://economia.uol.com.br/noticias/redacao/2019/04/06/pib-economia-nordeste.htm. [Jan 19, 2020].

[25]. Gujarati, D. N. and Porter, D. C. (2011). Econometria básica. 5. ed. Porto Alegre: AMGH, 924p.

[26]. Macedo, M.R.O.B.C., Maia, M. L. A., Kohlman Rabbani, E. R., and Lima Neto, O. C. C. "Remote Sensing Applied to the Extraction of road geometric features based on OPF classifiers, Northeastern Brazil." Journal of Geographic Information System, vol. 12, pp. 15-44, 2020.