Revista Brasileira de Cartografia
RBC | Rev. Bras. Cartografia

DOI: 10.14393/rbcv70n4-46140

Geostatistical approach to estimate car occupant fatalities in traffic accidents

Abordagem geoestatística para estimativa de mortes no trânsito com usuários de carro

Monique Martins Gomes[1]
Cira Souza Pitombo [2]
Ali Pirdavani[3]
Tom Brijs[4]

Recebido em novembro de 2017.
Aprovado em setembro de 2018.

ABSTRACT
Despite the scientific and technological advances toward road crash prediction models, modelling road crashes in Brazil is a challenging task due to unreliable data and unavailability of essential information. Geostatistical approaches fit into this context as they not only incorporate the spatial factor, but also estimate variables in locations where they are not sampled. In order to contribute to this investigation and present geostatistics as a suitable tool in estimating road deaths, this study aimed to explore two different univariate interpolation methods to predict car occupant fatalities. In this study, we considered ordinary kriging and indicator kriging as such approaches. Analyses were held based on data from the state of São Paulo in Brazil. The results revealed a statistical outperformance in favor of indicator kriging, although spatial patterns found on kriging maps for both techniques indicated similarities in terms of hotspots. Furthermore, they were coherent with local aspects observed in the state, for instance those related to highway and vehicle characteristics.

KEYWORDS: Geostatistics. Crash Prediction Models. Kriging. Road crashes.

RESUMO
Apesar dos avanços científicos e tecnológicos em prol dos modelos de previsão de acidentes de trânsito, tal prática é considerada um desafio no Brasil, por motivos como dados não confiáveis ou indisponibilidade de

[1]University of São Paulo, Av. Trabalhador São Carlense, 400, São Carlos, Brazil / UHasselt
E-mail: monique.martinsgomes@uhasselt.be
[2]University of São Paulo, Av. Trabalhador São Carlense, 400, São Carlos, Brazil
E-mail: cirapitombo@usp.br
[3]UHasselt, Faculty of Engineering Technology, Agoralaan, 3590 Diepenbeek, Belgium.
E-mail: ali.pirdavani@uhasselt.be
[4]UHasselt, Transportation Research Institute (IMOB), Agoralaan, 3590 Diepenbeek, Belgium
E-mail: tom.brijs@uhasselt.be

informações. Neste contexto, técnicas de geoestatística se adequam por incorporar o fator espacial e ainda possibilitar a estimação de variáveis em locais onde não há o registro de tais informações. A fim de contribuir com essa investigação e apresentar a geoestatística como uma ferramenta eficaz na previsão de acidentes de trânsito, o objetivo desse estudo foi explorar duas diferentes ferramentas de interpolação univariada na previsão de vítimas fatais de acidentes de trânsito envolvendo usuários de automóveis. Nesse estudo foram consideradas as ferramentas de krigagem ordinária e krigagem indicativa. Análises foram conduzidas com base em informações disponíveis para o estado de São Paulo, Brasil. Os resultados revelaram um melhor desempenho estatístico da krigagem indicativa, embora os padrões espaciais obtidos com ambas as técnicas tenham indicado similaridade nos *hotspots*. Além disso, esses resultados se mostraram condizentes a aspectos locais observados no estado, como aqueles relativos às características das rodovias e dos veículos, por exemplo.

PALAVRAS-CHAVE: Geoestatística. Modelos de previsão de acidentes. Krigagem. Acidentes de trânsito.

* * *

Introdução

Road traffic accidents are a huge problem worldwide accounting for 1.25 million annual deaths. Low and middle-income countries bear a large share of the burden, accounting for 90 percent of this number. Specifically in Brazil, increasingly high numbers of road crashes and related deaths have placed a heavy burden on households, as well as the national economy. Statistics point to around one million road crashes every year, which in turn results in at least five hundred thousand injured and up to fifty thousand annual deaths (WHO, 2015).

Fast economic development, as well as a steady increase in motorization rates and urban intensification are pointed out as the major reasons for the rising numbers of road fatalities in Brazil. If, on the one hand, economic growth has brought fast urbanization and motorization, on the other hand it has increased the number of inexperienced road drivers and road traffic. Moreover, the development of supportive road infrastructure, policy changes and enforcement have not kept pace with the advances in vehicle use. In spite of growing awareness in Brazil as to the urgency of reversing this trend, this drawback will not be overcome if proper safety

countermeasures and investments in road safety are not proportional to the scale of the problem (WHO, 2015).

In this context, road accident statistics have been a topic of interest and have drawn the attention of researchers and policy makers to seek efficient ways to reduce the numbers of fatal victims. Specifically at the strategic planning level, important scientific and technological advances toward crash modelling techniques have been made, such as those involving spatial statistics. Despite the efforts, modeling road crashes in Brazil is a challenge given the unavailability of essential information. This drawback often prompts researchers to adopt alternative strategies, which might not be the best ones, but at least lead to an insight into the problem.

In view of the foregoing, geostatistical approaches are justified as they enable the study of features in which the variables are spatially correlated, thus enabling the estimation of values of a specific variable in areas where they are not sampled (MATHERON, 1963; MATHERON, 1971; YAMAMOTO and LANDIM, 2013; SOARES, 2006; WACKERNAGEL, 2003). Specifically for road safety studies, the use of geostatistics is advantageous as the values from unsampled locations can be inferred. Moreover, the produced kriging maps of the predicted values, could contribute to the identification of hotspots, thus helping policy makers and responsible agencies to prioritize safety countermeasures on areas with higher road fatalities occurrence.

Despite the potential of these approaches, they are usually applied to data with apparent spatial continuity, e.g. temperature, rainfall and land composition, into the fields of geology, hydrology and mining, for example. In spite of this limitation, since the last decades, the use of geostatistics on spatially discrete data has proven to be a potential alternative when adapted to such spatial continuity problems, thus being more and more explored on different fields, e.g. health studies, where kriging techniques have been used for instance to identify areas of contamination or risk of mortality (GOOVAERTS, 2004, 2005, 2006, 2008, 2009).

Likewise, on transportation studies, the application of geostatistical tools and their produced kriging maps, has produced successful results. However, most available literature in this field refers to traffic engineering studies (CIUFFO, PUNZO and QUAGLIETA, 2011; MAZZELLA, PIRAS and PINNA, 2011; ZOU et al., 2012; ZHANG and WANG, 2013), vehicle emission gases (PEARCE et al., 2009; KASSTEELE and VELDERS, 2006; KASSTEELE and STEIN, 2006), and since recently, to travel demand forecasting problems (PITOMBO et al., 2015; LINDNER et al., 2016; GOMES et al., 2016).

Specifically on traffic data, geostatistical tools have been implemented to analyze the spatial structure of the data under explanatory purposes (MAJUMDAR, NOLAND and OCHIENG, 2004; MCMILLAN, HANSON and LAPHAM, 2007; LASCALA, JOHNSON and GRUENEWALD, 2001) or toward confirmatory analysis (MANEPALLI and BHAM, 2011; MATSUMONO and FLORES; 2013; GUNDOGDU, 2014; MOLLA, STONE and LEE, 2014). In most of these studies, ordinary kriging (OK) was used as the main tool to estimate traffic accidents.

In Manepalli and Bham (2011) kriging was performed in order to estimate the frequency of accidents on Highway I-630 in Arkansas, United States. Statistics from three years (2000-2002) were used. The estimation was carried out considering the frequency and severity of accidents, called the Crash Severity Index (CSI). This index was determined based on the product of the frequency of accidents under pre-determined characteristics and weights assigned to the severity of the accident. Lastly, the authors compared the performance of kriging with results from an empirical Bayesian method, and concluded that kriging performed better.

In Matsumoto and Flores (2013), geostatistical techniques were used in order to identify areas of the highest concentration of road accidents. Such investigation was carried out based on data concerning traffic accidents in Presidente Prudente, Brazil. By preparing semivariograms, the existence of isotropy or anisotropy was firstly verified. Then, assuming isotropic process

cartographic visualizations were generated, which in turn enabled the interpretation of the areas most affected by the accidents, therefore more prone to risks of occurrence of the phenomenon.

Gundogdu (2014) applied OK to explore the spatial distribution of traffic accidents in Konya, Turkey. The database used consisted of 10 years of information and it was based on the Accident Criteria (AC) determined as follows: fatal victims, injuries, accidents with damage and total number of accidents. Results obtained from the kriging maps enabled the author to identify the areas of greatest risk of crash occurrence. In addition, the results contributed to determining new routes and safety zones.

In this paper, we address the application of kriging estimators on road fatality data in the state of São Paulo, Brazil. The aim of this study is to present the geostatistics technique as a suitable tool to estimate road deaths, exploring the statistical contribution of two different univariate interpolation methods to predict car occupant fatalities: OK and indicator kriging (IK). Since OK have been used in previous studies, the originality of this study lies in the consideration of IK as an effective estimation tool in relation to OK. Thus, IK could be more appropriate taking into account outliers and the data asymmetry.

Besides this introduction, three other sections comprise this paper. In section 2, information regarding the data and software programs used is presented, as well as the study area followed by the adopted methodological procedures. Section 3 presents the results obtained by using the approaches, and the discussion. Section 4 draws the main conclusions of this study and points out the methodological limitations found.

2 Materials and Method

2.1 Software packages

Different software packages were used to develop this study. Non-spatial statistics (characterization of the database through descriptive measures, histograms, hypothesis test) were carried out using the IBM SPSS 24 software. GeoStatistical Modeling Software - geoMS - version 1.0, was used in the geostatistical modeling stages: visualization of point map, development, modeling and adjustment of variograms, spatial estimation of data and cross validation. Finally, Geographic Information Systems (SIG), such as TransCAD 5.0, Qgis 2.8.1 and ArcGIS 10.1 helped us to develop the thematic maps.

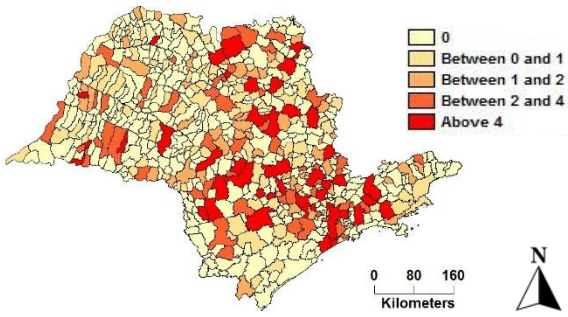
2.2 Study area and data

São Paulo is one of the 26 states that compose the Republic Federative of Brazil. Located in the southeast region of the country, São Paulo has a road network of more than 35000 kilometers, used to connect the 645 municipalities of the state. Thereupon, São Paulo has the greatest population of Brazil. At the last census, in 2010, statistics pointed to around 41 million inhabitants in a land area close to 248000 square kilometers. In 2017, the population had already exceeded 45 million (IBGE, 2018). Such high statistics are also observed in the rising number of road fatalities in the state. In 2015, São Paulo registered more than six thousand deaths caused by road traffic accidents (DATASUS, 2018).

In order to use the geostatistical modeling software – geoMS which requires data on Universal Transverse Mercator (UTM) projection, firstly a ground coordinate system was generated. Then, analyses were conducted based on the information of car occupant fatalities on public roads of 2010. In spite of the challenges to obtain the punctual spatial location of road fatalities

and explanatory variables related to crash occurrence, especially those related to road features, aggregated fatality statistics are easily obtained on the Mortality Information System (*SISTEMA DE INFORMAÇÕES DE MORTALIDADE* – SIM), where our data was gathered. SIM is a Brazilian public source created by DATASUS (SUS Department of informatics), the Ministry of Health's institution responsible for gathering data provided by the official sources of information in Brazil, for instance the State and Municipal Healthy Departments (DATASUS, 2018). Hence, this information was associated with the geo-referenced centroid for each of the 645 municipalities of the state. Figure 1 shows the thematic map obtained for the number of road crash fatalities.

Figure 1 – Distribution of road crash fatalities in the state of São Paulo, Brazil

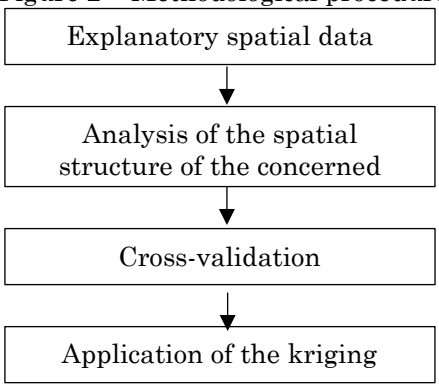


Source: Prepared by the authors.

2.3 Methodological procedure

In order to apply both estimators, IK and OK, the methodological procedure followed four main steps, synthetized on Figure 2, and described on the following sections.

Figure 2 – Methodological procedure



Source: Prepared by the authors.

2.3.1 Explanatory spatial data analysis

Initially, the spatial distribution of car occupant fatalities and its descriptive statistics were assessed. This practice enabled us to detect the main features of the concerned variable, identify outliers, and adequate the original data to the application of geostatistics, as it includes in its assumptions the spatial continuity of the event.

In order to minimize the variance and obtain a better spatial distribution of the data, before applying the estimator OK, detected outliers were omitted during the variographic analyses. This exercise was held along with the calculation of the experimental variograms and adjustment of the theoretical ones. In this sense, we restricted car occupant fatalities to a maximum value, which was determined based on the descriptive statistics and the observed spatial distribution of the sample for different values tested. Based on this, several variograms were generated. Hence, the best-fitted variograms were chosen according to their better spatial structures.

Thereafter, in order to enable the IK, binary variables (0 and 1) were generated from the original data. We assumed three classes of binary variables, produced according to the occurrence, or not, of fatalities (IK – 1) and based on the 1st and 3rd quartile values (IK – 3 and IK – 5 respectively), where:

a) IK – 1: (1 for occurrence ≥ 1; 0 for non-occurrence);
b) IK – 3: (1 for occurrence ≥ 3; 0 for occurrence < 3);
c) IK – 5: (1 for occurrence ≥ 5; 0 for occurrence < 5).

Such quartiles were chosen as cut off levels as they granted symmetrical data distribution, thus producing the best fitted variograms, in terms of spatial structure, in relation to the median and other percentiles, for example.

2.3.2 Analysis of the spatial structure of the concerned variable

At this stage, variographic analysis and the adjustment of the experimental variograms were held. The primary tool in geostatistical modeling is the variogram, which graphically represents a regionalized variable. The variogram function is given by Equation 1, where N(h) is the set of all pairwise, and z (xi) and z (xi + h) are data values at spatial locations i and i+h, respectively (MATHERON, 1963).

$$y\,(h) = \frac{1}{2N}\sum_{i=1}^{n(k)}[Z\,(xi) - Z\,(xi+h)]^2 \qquad (1)$$

The representation of an experimental variogram requires a further understanding of graphical aspects. Some measures include lag distance, tolerance, cut distance and angle direction. The next step is to model a theoretical variogram based on the experimental one. From the various theoretical models for adjustment of variograms, the most frequently used are Spherical, Gaussian and Exponential. In this sense, the experimental variogram is replaced by a theoretical variogram function, from which the main parameters for spatial modelling can be obtained: nugget effect (Co), Range (a) and Sill (C+ Co) (MATHERON, 1963, 1971; WACKERNAGEL, 2003).

2.3.3 Cross validation

Cross Validation (CV) compares various assumptions either concerning the model (e.g. type of function to be adjusted, variogram parameters) or data. In the cross validation procedure, each sample value Z(x1) is removed in turn from the dataset and a value Z*(x1) at the location is estimated using the remaining n-1 samples. The difference between a data value and the estimated value (Z(x1) - Z*(x1)) gives an indication of how well the data value fits into the neighborhood of the surrounding data values (JOURNEL and HUIJBREGTS, 1978; WACKERNAGEL, 2003).

2.3.4 Application of the kriging estimator

Regarding the CV, the geostatistical modeling general method follows the kriging estimation. Kriging technique is able to provide a best linear unbiased estimator (BLUE) for variables that have tendency to vary over space (JOURNEL and HUIJBREGTS, 1978; MATHERON, 1963). The idea behind the technique is to conduct estimates with minimum error and variance using the parameters defined in the theoretical variogram. The name kriging comes from the pioneering work of a mining called Daniel Krige (KRIGE, 1951). The most usual univariate kriging methods are simple kriging, universal kriging and ordinary kriging. In this paper, OK and IK are compared in the estimation of road crash fatalities. OK is the most widely used kriging method. The main aim behind OK is to estimate a value at a point of a region, for which the correspondent variogram is known, using data in neighborhoods (WACKERNAGEL, 2003).

IK is an estimation technique with the same basis as linear kriging, but it is applied to attributes with non-Gaussian distribution, which are transformed according to non-linear mapping and codification by means of indication. Applied on a set of numeric sample values, Z(u=ua), codification by means of an indicator, for a cutting value zk, a sample set by means of

indicator I (u=ua; zk) has the formulation expressed in Equation 2 (ISAAKS and SRVASTAVA, 1989).

$$I(u; z_k) = \begin{cases} 1 \xrightarrow{if} Z(u) \leq z_k \\ 0 \xrightarrow{if} Z(u) > z_k \end{cases} \quad (2)$$

As previous mentioned, in this paper, the original variable of interest was codified considering the 1st and 3rd quartile values, besides the occurrence, or not, of fatalities.

3 Results and Discussion

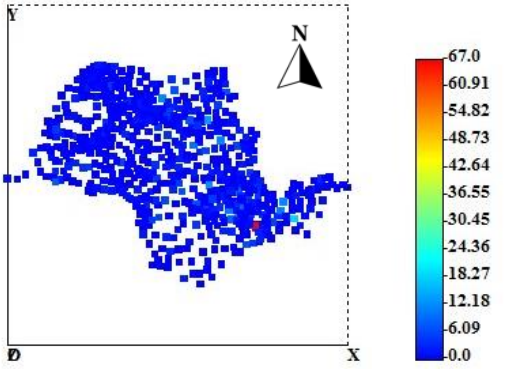
3.1Ordinary kriging application

In the next subsections, results produced at each methodological step with OK are presented, followed by their discussion.

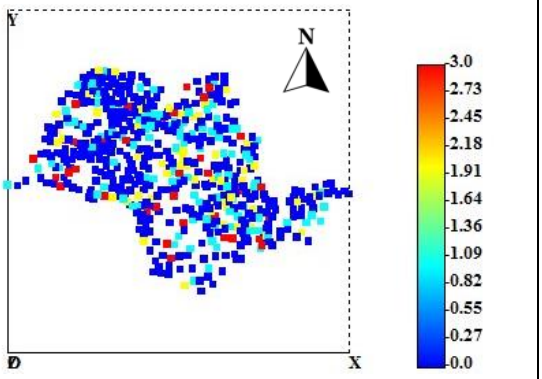
3.1.1 Spatial data analysis

At this stage, the visualization of the spatial distribution of the data, illustrated in Figure 3, and their descriptive statistics (see Table 1) enabled us to detect high variance in the data. One explanation could be the presence of many outliers in the dataset. For instance, the city of São Paulo registered a much higher number of fatalities compared to the other municipalities. Besides, there were many repetitions of the value "zero" in the dataset, meaning that many regions, registered in that year, any fatality involving car occupants. Therefore, such positive asymmetry was a criterion and the major motivation for the application of IK.

Figure 3 – Maps of points



Map of points including outliers



Map of points excluding outliers

Source: Prepared by the authors.

Table 1 – Descriptive statistics of the variable of study

Statistics	Including Outliers	Excluding Outliers
Observations	645	585
Mean	1.20	0.55
Min.	0	0
Max.	67	3
Variance	11.38	0.81
Std. Dev.	3.38	0.9
1st quartile	0.00	0.00
3rd quartile	1.00	1.00

Source: Prepared by the authors.

3.1.2 Analysis of the spatial structure of the concerned Variable

At this stage, the recommended steps for Geostatistical modeling were followed. Initially, the experimental variograms were generated, characterizing the spatial structure of the variable of interest. Afterwards, the variograms were adjusted and the main direction, in which the variability is higher compared to the other directions (CLARK, 1979), and its respective orthogonal direction were verified, including important parameters for the kriging step (range, sill and nugget effect).

The development of the experimental variograms and their main and orthogonal directions proceeded from the angle 0° (North to South) to 90° (East to West), according to the standardization of axes of geoMS. Initially,

experimental variograms were generated with test angles ranging from 15° to 15°, and angular tolerance of 1°. 100 lags were used (maximum allowed by the software). The size of the lag (h) to the Omni-directional case was obtained according to the average of Euclidean distances between the nearest neighbors considering all angular directions. For the others, the mean of the Euclidean distances for each of the studied angular directions was considered. The cutting distance adopted was 550 kilometers based on half of the state's length. Table 2 shows the parameters that best described the spatial structure of the samples for the variable of study. Subsequently, Figure 4 presents the obtained adjusted theoretical variograms for both directions. The spherical model was the one that best suited.

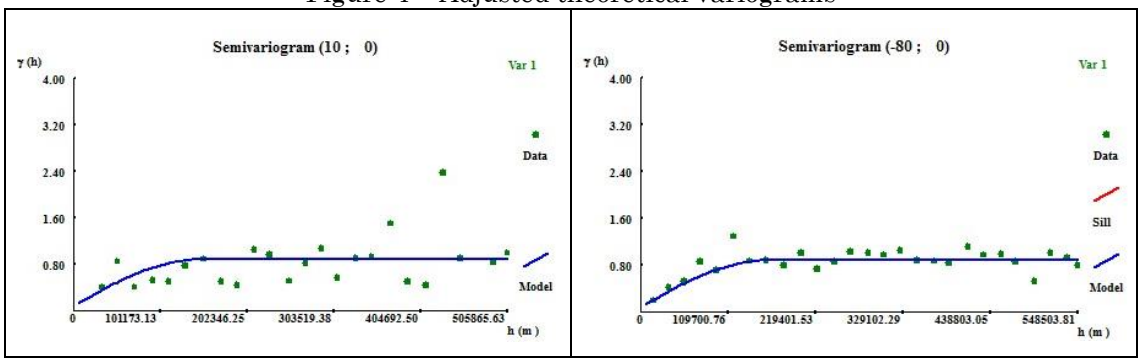
Table 2 – Final parameters (experimental variograms)

Parameters	Direction: 10	Direction: -80*
Lag (m)	$20x10^3$	$21x10^3$
Tolerance	1º	1º
Number of lags	100	100
Cut distance (m)	$510x10^3$	$510x10^3$

*Main direction

Source: Prepared by the authors.

Figure 4 – Adjusted theoretical variograms



Source: Prepared by the authors.

The variable of study showed an anisotropic structure, which implies in different spatial variability in different directions, and a main direction.

3.1.3 Cross validation

In order to carry out the validation considering observed and predicted values, goodness of fit measures were calculated, and are shown in Table 3.

Table 3 – Performance measures for variogram modeling

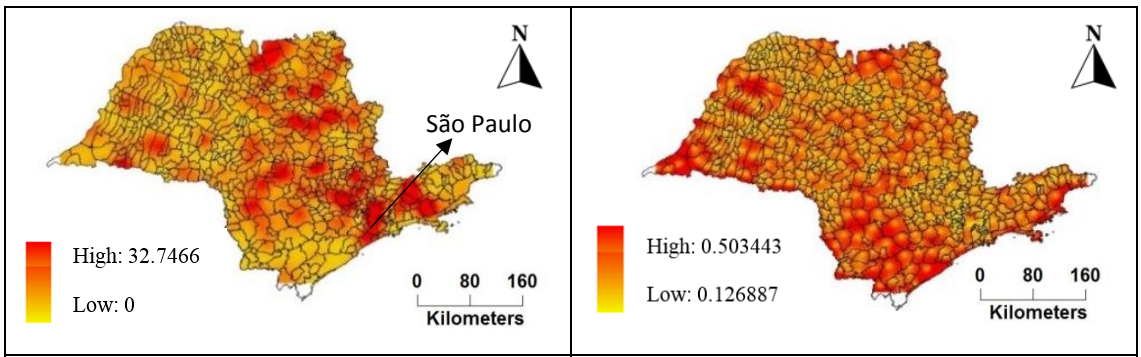
Pearson's Correlation Coefficient	0.081
Mean Error	-0.03
Mean Absolute Error	1.61
Mean Relative Error	-0.05

Source: Prepared by the authors.

3.1.4 Ordinary kriging

The definition of the main direction and the necessary parameters for adjusting the variograms enabled the composition of the spatial surface. Figure 5 presents the kriging map obtained with the distribution of estimated values of deaths on a continuous surface, and its corresponding variance estimation map.

Figure 5 – Kriging and variance of estimation maps



Kriging map | Variance of estimation map

Source: Prepared by the authors.

The obtained kriging maps enabled us to associate car occupant fatalities to local aspects of areas where estimations were higher. For

instance, the majority of fatalities were estimated in municipalities where important highways of the country are also located. Moreover, in addition to the steady high flow of cars and motorcycles on these highways, they play an important role in the transport of agricultural and industrial products, thus resulting in a great amount of trucks, and consequently greater accident risks.

High estimates were also observed to the east of the state, where highways often connect Sao Paulo city to other rich economic poles. Furthermore, the most important export corridors of Brazil are located in this axis, including its most important port: Port of Santos. Hence, import and export goods transported in big and heavy trucks eventually cross the state through these highways. In most cases, drivers are expected to drive long distances of more than 400km, and journeys that often take more than 6 hours. Hence, more effort and attention is required from these drivers, as they have greater physical and physiological burdens, thus more prone to the risk of crash occurrence.

In addition, in São Paulo city and other many municipalities in the state, a great amount of highways cut through urban areas, thus combining the flow of active (pedestrians and cyclists) and motorized transports, and unfortunately increasing the risk of road crashes. Furthermore, high estimates were observed in municipalities with a great amount of highlands. This leads to many heavy trucks with reduced speeds on the highways. As a result, vehicle traffic is slower, increasing the exposure to risk of accidents. Regarding the estimation variance map, lower accuracy of inference was seen mostly in areas where the estimation of incidences of fatalities were also low.

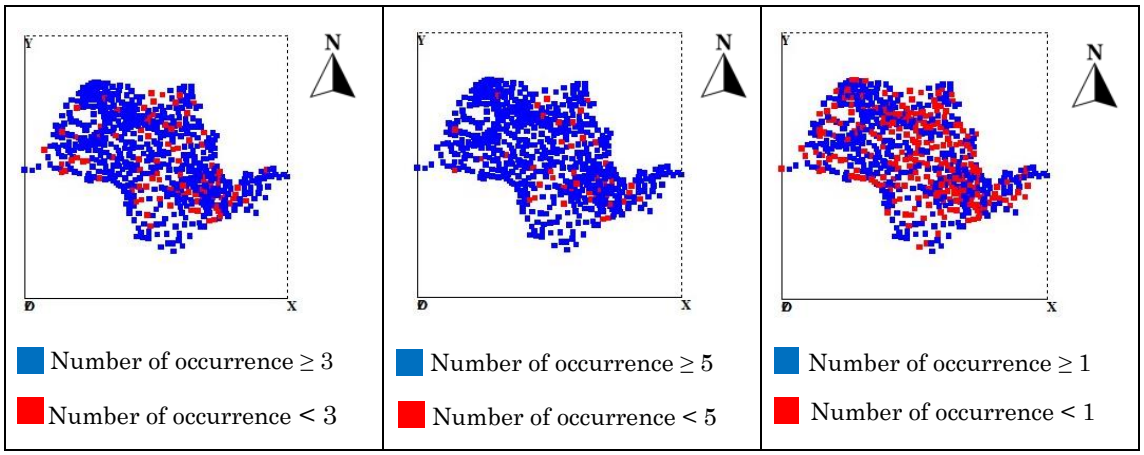
3.2 Indicador kriging application

In the next subsections, results produced at each methodological step with IK are presented followed by their discussion.

3.2.1 Spatial data analysis

After the binary transformation of the variable, according to the three investigated classes, maps of points were obtained, and are shown in Figure 6.

Figure 6 – Map of points for the produced binary variables



Source: Prepared by the authors.

3.2.2 Analysis of the spatial structure of the concerned variable

Experimental variograms and their respective main and orthogonal directions were generated likewise for OK. Table 4 shows the parameters that better described the spatial behavior of the samples for the variable of interest.

Table 4 – Final parameters for the calculation of experimental variograms

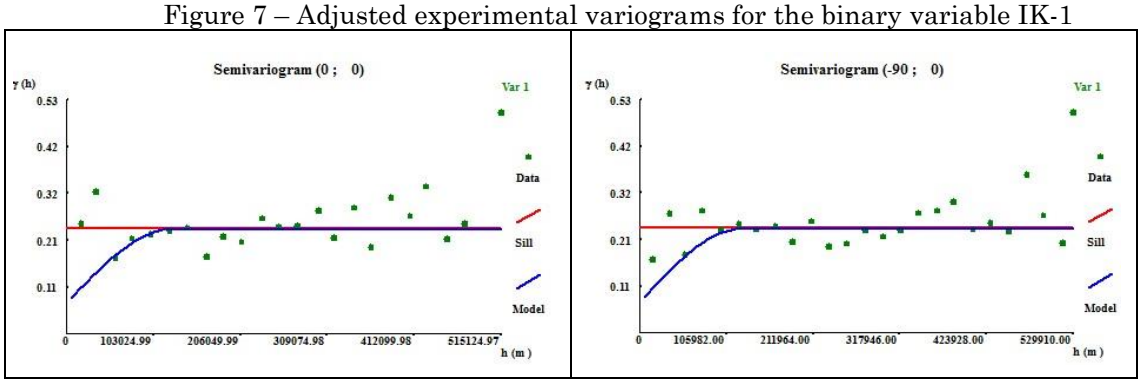
Var.	Dir.	Lag (m)	Tol.	Nº Lags	Cut dist. (m)
IK – 1	0	$22x10^3$	1º	100	$550x10^3$
	-90*	$22x10^3$			
IK - 3	10	$23x10^3$	1º	100	$550x10^3$
	-80*	$22x10^3$			
IK - 5	10	$26x10^3$	1º	100	$550x10^3$
	-80*	$28x10^3$			

*Main direction

Source: Prepared by the authors.

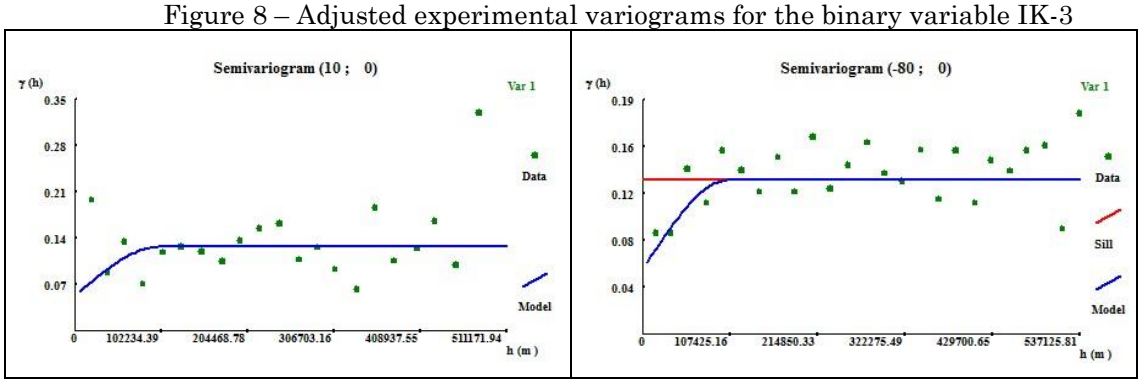
Figures 7, 8 and 9 present the theoretical variograms for the binary variables, which showed an anisotropic structure for each binary condition. The spherical model was the one that best suited variograms IK -1 and IK - 3, and the Gaussian model to IK - 5.

Figure 7 – Adjusted experimental variograms for the binary variable IK-1



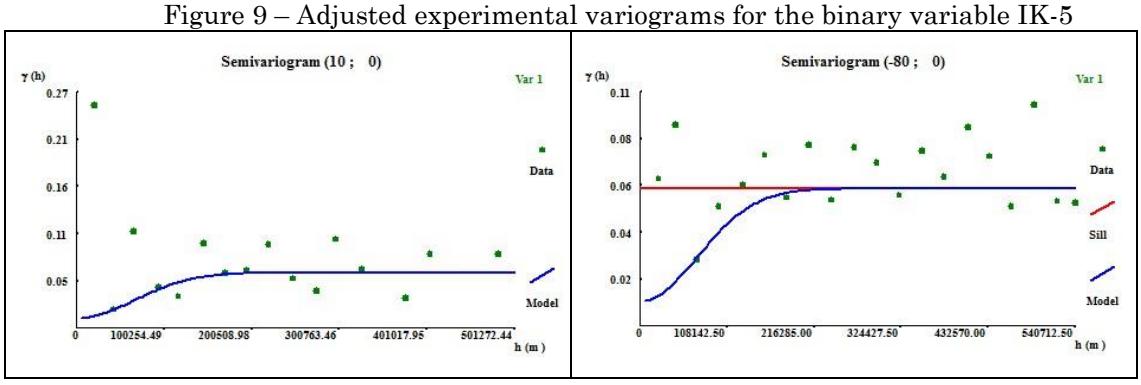
Source: Prepared by the authors.

Figure 8 – Adjusted experimental variograms for the binary variable IK-3



Source: Prepared by the authors.

Figure 9 – Adjusted experimental variograms for the binary variable IK-5



Source: Prepared by the authors.

3.2.3 Cross validation

Given the categorical nature of the observed and estimated variables, the associations between their binary variables were observed by means of the chi-squared test.

a) Estimated binary variable: the transformed variable of the estimated value, based on the median (i.e. if the estimated value was greater or equal to the median value of the values estimated for the sample, then 1, if not 0). The median was chosen for the discretization of the estimated values, as the estimated results are continuous, ranging between zero and one;
b) Observed binary variable: defined based on the indicators, previously created (IK-1, IK-3 and IK-5).

Thereafter, based on the two qualitative variables, the assumptions of the chi-square test were formulated as H0: the two variables are independent; H1: there is an association between the two variables. Table 5 shows the obtained results for the test.

Table 5 – Cross validation results

Variable	Chi-Square	Significance	Hit Rates
IK – 1	22.75	0.00	59.38%
IK – 3	10.51	0.00	56.43%
IK – 5	5.60	0.01	54.88%

Source: Prepared by the authors.

Tables 6, 7 and 8 present the percentage of hit rates obtained for each binary variable. In the three cases, the number of hit rates was higher compared to the number of errors. The hit rate obtained from IK was around 60%. Moreover, the calculated chi-square values were found to be high, which indicates the significant associations between the observed and estimated values of the variables.

Table 6 – Percentage of hit rates with IK-1 in geographic coordinates of known values

		Estimated value		Total
		0	1	
Observed value	0	178 (59.93%)	119 (40.07%)	297 (100%)
	1	143 (41.09%)	205 (58.91%)	348 (100%)
Total		321 (49.77%)	324 (50.23%)	645 (100%)

Source: Prepared by the authors.

Table 7 – Percentage of hit rates with IK-3 in geographic coordinates of known values

		Estimated value		Total
		0	1	
Observed value	0	26 (8.75%)	271 (81.25%)	297 (100%)
	1	10 (2.87%)	338 (97.13%)	348 (100%)
Total		36 (5.58%)	609 (94.42%)	645 (100%)

Source: Prepared by the authors.

Table 8 – Percentage of hit rates with IK-5 in geographic coordinates of known values

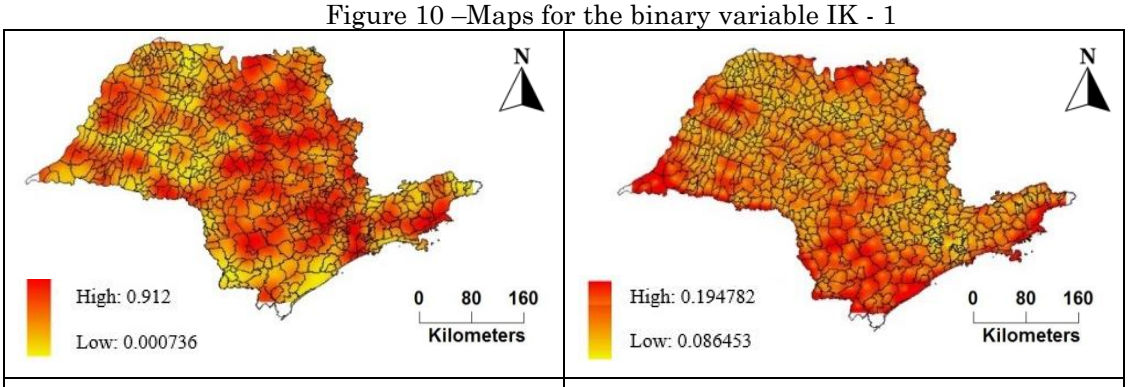
		Estimated value		Total
		0	1	
Observed value	0	7 (2.36%)	290 (97.64%)	297 (100%)
	1	1 (0.29%)	347 (99.71%)	348 (100%)
Total		8 (1.24%)	637 (98.76%)	645 (100%)

Source: Prepared by the authors.

3.2.4 Indicator kriging

Kriging maps with areas that were assigned higher probabilities of road crash fatalities on public roads were produced after setting the main and orthogonal direction and its parameters for the adjustment of variograms (see Figures 10 to 12).

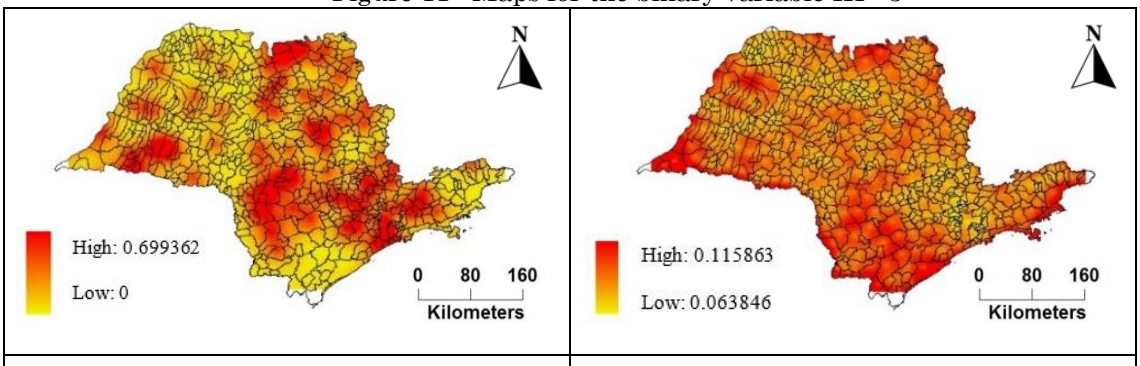
Figure 10 –Maps for the binary variable IK - 1



Kriging map	Variance of estimation map

Source: Prepared by the authors.

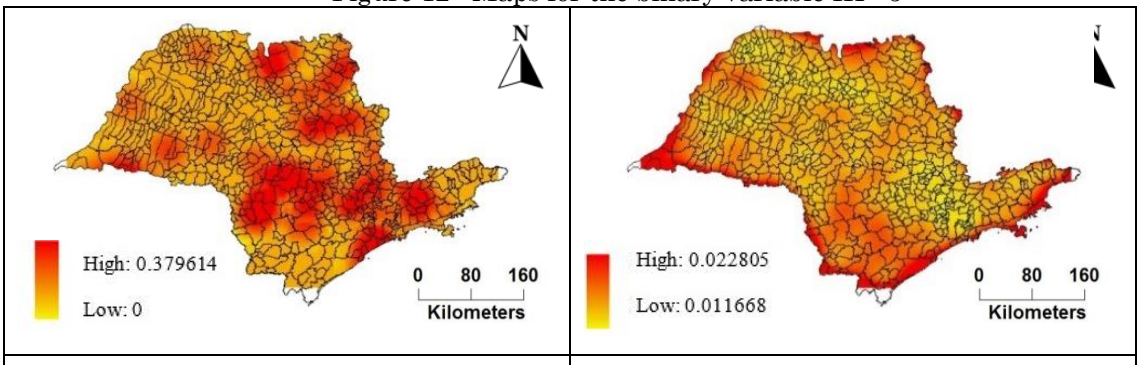
Figure 11 –Maps for the binary variable IK - 3



Kriging map	Variance of estimation map

Source: Prepared by the authors.

Figure 12 –Maps for the binary variable IK - 5



Kriging map	Variance of estimation map

Source: Prepared by the authors.

Based on the obtained results, initially high probability of occurrence of accidents in the entire state (IK-1) was inferred, but mainly in the center and southeast where the probabilities were higher (IK-5). Considering the kriging map obtained for the binary variable IK – 5, higher probabilities of

road crash fatalities were assigned in the same municipalities as observed from OK. Thus, the same considerations and assumptions by means of influence factors are now repeated for the IK results.

4 Conclusions

The unavailability of essential information related to road crashes in Brazil and its negative effects on crash modelling, were the major motivation to carry out this study. Although some data can be found for instance in transportation, health and census websites in Brazil, they are often unreliable or restricted to socio-economic variables. In this context, researchers and planners are forced to either manipulate the data or try to overcome at least part of this lack of information by contacting the official transport departments of the country. Considering that efforts to explore these strategies were not successful, this study aimed to present geostatistics as a suitable tool to help estimate the number of deaths in areas where there is no information. To this end, ordinary kriging and indicator kriging were performed, and analyses were conducted based on aggregated information of road crash fatalities in São Paulo, Brazil.

The results revealed a statistical outperformance in favor of IK, although spatial patterns found on kriging maps obtained from both techniques were similar in terms of the identified hot spots. Moreover, they were found coherent with local aspects observed in the state, for instance those related to the features of the roads and vehicles. This exploratory investigation enabled us to make the following assumptions:

a) Higher road death probability and estimates in the east of the state, where characteristics of the road and vehicles, as those from the municipalities in this axis play an important role in the high values estimated in these areas;
b) Higher road death probability and estimates in municipalities close to highlands;

c) Fatality figures are more likely to happen in areas where highways go through urban areas.

Concerning the analysis, the data dispersion was solved using artifices such as removing atypical values for modeling the variograms and using different parameters (size and number of lags, directions, tolerance, etc.). These artifacts contributed to the geostatistical modeling and, consequently, to obtaining better variograms. However, they were not sufficient if we observe the prediction errors from the estimations made through OK. Only by the transformation of the variables with IK did the variograms have a better structure, and greater effectiveness in the obtained results.

Results obtained with both estimators were considered reasonable. Moreover, the visualization and geostatistical modeling of the spatial distribution of the data enabled us to estimate fatalities at unsampled sites, and identify areas where fatal crash estimates and probability of occurrence were higher. This information could be useful for authorities responsible for traffic safety management as they could contribute to the development of road safety studies, as well as ensuring that safety measures are directed first in areas where the number of deaths is greater. This could be considered as the main contribution of this study.

Despite some recent studies into the transportation field have explored geostatistical tools, for instance on travel demand forecasting (PITOMBO et al., 2015; LINDNER et al., 2016; GOMES et al., 2016), to the best of our knowledge, its application using crash data, is recent and has not yet been fairly explored, especially considering IK, and in Brazil, where these studies are even more incipient.

Unfortunately, obtaining reliable data has been a challenge in developing countries, as Brazil, and its unavailability an obstacle to road safety promotion. In this context, this study presented IK as an alternative and effective tool to predict road death probability and concluded that kriging can be used as a potential tool to estimate road fatalities in areas where there is no information. Besides the obtained results, this statement is supported

by the fact that only the spatial location of the centroid of the municipalities and the values of the variable of interest were considered. However, in order to better understand the phenomenon of road accidents, it is essential the inclusion of appropriate explanatory variables, as well as spatial dependence.

Therefore, for further studies and analysis, the authors recommend validating the identified inferences, as well as using multivariate spatial models, including in the analysis variables associated to the characteristics of the roads and municipalities, generation and attraction of trips and goods, etc. Furthermore, in spite of geostatistics being indicated as an adequate method for the intended purposes in this research, for future studies, we suggest the exploration of the concept of "semivariogram deconvolution", which has been recently applied in health studies aiming to adequate the data, spatially discrete, to the application of geostatistics (GOOVAERTS, 2008).

Acknowledgments

This research was supported by the Brazilian National Council for Scientific and Technological Development - CNPq.

References

CIUFFO, B. F.; PUNZO, V. and QUAGLIETTA, E. 2011. Kriging meta-modelling to verify traffic micro-simulation calibration methods. **TRB 90th Annual Meeting Compendium of Papers**, Washington, 2011.

CLARK, I. **Practical geostatistics**. London: Applied Science Publishers, 1979. 126p.

DATASUS, Ministério Da Saúde - Departamento de Informática do SUS. Sistema de Informações sobre Mortalidade - Estatísticas Vitais. In: http://www2.datasus.gov.br/DATASUS, accessed in April, 2018.

GOMES, V. A.; PITOMBO, C. S.; ROCHA, S. S.; SALGUEIRO, A. R. Kriging geostatistical methods for travel mode choice: a spatial data analysis to travel

demand forecasting. **Open Journal of Statistics**, vol. 6, 2016. pp. 514-527. doi: 10.4236/ojs.2016.63044

GOOVAERTS, P. and JACQUEZ G.M. Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. **International Journal of Health Geographics**, vol. 3, n. 14, 2004. doi: 10.1186/1476-072X-3-14

GOOVAERTS, P. Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. International **Journal of Health Geographics**, vol. 4, n. 31, 2005. doi: 10.1186/1476-072X-4-31

GOOVAERTS, P. Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. **International Journal of Health Geographics**, vol. 5, n. 52, 2006. doi: 10.1186/1476-072X-5-52

GOOVAERTS, P. Kriging and semivariogram deconvolution in the presence of irregular geographical units. **Mathematical Geosciences**, vol. 40, n. 1, 2008. pp. 101-128. doi: 10.1007/s11004-007-9129-1

GOOVAERTS, P. Medical geography: a promising field of application for geostatistics. **Mathematical Geosciences**, vol. 41, 2009. pp. 243-264. doi: 10.1007/s11004-008-9211-3

GUNDOGDU, I. B. Risk governance for traffic accidents by Geostatistical Analyst methods. **International Journal of Research in Engineering and Science,** vol. 2, 2014. pp. 35-40.

IBGE, Instituto Brasileiro de Geografia e Estatística. In: <https://ww2.ibge.gov.br/estadosat>, accessed in April 2018.

ISAAKS, E. H. and SRVASTAVA, R. M. **An introduction to applied Geostatistics**. Oxford University Press, 1989.

JOURNEL, A. G. and HUIJBREGTS, C. J. **Mining geostatistics**. New York: Academic Press, 1978.

KASSTEELE, J. van de. and VELDERS, G. J. M. Uncertainty assessment of local NO2 concentrations derived from error-in-variable external drift kriging and its relationship to the 2010 air quality standard. **Atmospheric**

Environment, vol. 40, 2006. pp. 2583–2595. doi: 10.1016/j.atmosenv.2005.12.023

KASSTEELE J. van de. and STEIN, A. A model for external drift kriging with uncertain covariates applied to air quality measurements and dispersion model output. **Environmetrics**, vol. 17, 2006. pp. 309–322. doi: 10.1002/env.771

KRIGE, D. A statistical approach to some basic mine valuation problems on the Witwatersrand. **Journal of the Chemical, Metallurgical and Mining Society of South Africa**, vol. 52, 1951. pp. 119–139.

LASCALA, E. A.; JOHNSON, F. W. and GRUENEWALD, P. J. N. Neighborhood characteristics of alcohol-related pedestrian injury collisions: a geostatistical analysis. **Prevention Science**, vol. 2, 2001. pp. 123-134.

LINDNER, A.; PITOMBO, C. S.; ROCHA, S. S.; QUINTANILHA, J. A. Estimation of transit trip production using Factorial Kriging with External Drift: an aggregated data case study. **Geospatial Information Science**, vol. 19, n. 4, 2016. pp.245-254. doi: 10.1080/10095020.2016.1260811.

MAJUMDAR, A.; NOLAND, R. B. and OCHIENG, W. Y. A spatial and temporal analysis of safety-belt usage and safety-belt laws. **Accident Analysis & Prevention,** vol. 36, 2004. pp. 551-560. doi: 10.1016/S0001-4575(03)00061-7

MANEPALLI, U. R. R. and BHAM, G.H. Crash prediction: evaluation of empirical bayes and kriging methods. **Proc., 3rd International Conference on Road Safety and Simulation**, Indianapolis, 2011.

MATHERON, G. Principles of Geostatistics. **Economic Geology**, vol. 58, 1963. pp. 1246–1266.

MATHERON, G. **The theory of regionalized variables and its applications**. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau. École Nationale Supérieure dês Mines de Paris, 1971.

MATSUMOTO, P. S. S. and FLORES, E. F. Aplicações de Geoestatística na saúde: acidentes de trânsito em Presidente Prudente SP. **Anais do III Simpósio de Geoestatística Aplicada em Ciências Agrárias**, Botucatu, 2013.

MAZZELLA, A.; PIRAS, C. and PINNA, F. Use of Kriging Technique to Study Roundabout Performance. **Transportation Research Record, Journal of the Transportation Research Board**. N. 2241, 2011. pp. 78-86. doi: 10.3141/2241-09, 2011

MCMILLAN, G. P.; HANSON, T. E. and LAPHAM, S. C. Geographic variability in alcohol-related crashes in response to legalized Sunday packaged alcohol sales in New Mexico. **Accident Analysis & Prevention,** vol. 39, 2007. pp. 252-257. doi: 10.1016/j.aap2006.07.012

MOLLA, M. M.; STONE, M. L. and LEE, E. Geostatistical Approach to Detect Traffic Accident Hot Spots and Clusters in North Dakota. **Upper Great Plains Transportation Institute**, n. 276, 2014.

PEARCE, J.L.; RATHBUN, S.L.; AGUILAR-VILLALOBOS, M.; NAEHER, L.P. Characterizing the spatiotemporal variability of PM2.5 in Cusco, Peru using kriging with external drift. **Atmospheric Environment,** vol. 43, 2009. pp. 2060-2069. doi: 10.1016/j.atmosenv.2008.10.060

PITOMBO, C. S.; SALGUEIRO, A. R..; COSTA, A. S. G.; ISLER, C. A. A two-step method for mode choice estimation with socioeconomic and spatial information. **Spatial Statistics**, vol. 11, 2015. pp. 45-64. doi: 10.1016/j.spasta.2014.12.002

SOARES, A. **Geoestatística para as ciências da terra e do meio ambiente**. 2nd ed. Lisboa: IST PRESS, 2006.

WACKERNAGEL, H. **Multivariate Geostatistics**. 3rd ed. Berlin: Springer, 2003. 388p.

WHO, World Health Organization. Global Status Report on Road Safety. In:<http://www.who.int/violence_injury_prevention/road_safety_status/2015>, accessed in April 2016.

YAMAMOTO, J. K. and LANDIM, P. M, B. **Geoestatística: conceitos e aplicações**. São Paulo: Oficina de textos, 2013. 215p.

ZHANG, D. and WANG, X. **Traffic volume estimation using network interpolation techniques: an application on transit ridership in NYC Subway System**. Final Report, New York, 2013.

ZOU, H.; YUE, Y.; LI, Q; YEH, AGO. An improved distance metric for the interpolation of link-based traffic data using kriging: a case study of a large-scale urban road network. **International Journal of Geographical Information Science,** vol. 26, 2012. pp. 667–689. doi: 10.1080/13658816.2011.609488