ENGINEERING INTERVENTION MANAGEMENT USING A PROBABILISTIC METHODOLOGY FOR HOT SPOT IDENTIFICATION

Sara Ferreira, Porto University, Faculty of Engineering; sara@fe.up.pt
António Couto, Porto University, Faculty of Engineering; fcouto@fe.up.pt

ABSTRACT

A theoretical definition of a hotspot is any location that has a higher expected number of accidents than other similar locations as a result of local risk factors present at the location. This study presents an alternative approach to research regarding hot spot definition and identification based on a probabilistic model that defines the dependent variable as an indicator of a discrete choice. A binary choice model was used considering a binary dependent variable that differentiates a hot spot (category 1) from a safe (category 0) site set by the number of accidents per kilometer. To define these two categories, it is necessary to use a strategy that selects a hot spot as accurately as possible. Based on a threshold strategy, various hypotheses were analyzed to obtain the most appropriate value to balance sensitivity and specificity criteria (epidemiological criteria). To apply this approach, risk factors including traffic volume, the number of minor intersections per kilometer, the road function classification and land use from an urban segment data set collected over a five-year period from Porto, Portugal, were used. The probabilities were estimated by the binary choice model, and a performance evaluation was then applied. In addition, considering the probabilistic nature of this approach, it is also possible to define four classes to classify a site in terms of safety using the uncertainty in a site being a hot spot, setting for each class a range of probability values. Furthermore, a comparative analysis was considered to test the performance of the qualitative response (QR) method relative to two commonly applied methods - accident frequency (AF) and empirical Bayes (EB). Considering the probabilistic nature of this approach, two probability threshold values were determined based on the "correct proportion" of sensitivity and specificity. These two values, along with the 0.5 probability value that the model defines to separate a hot spot from a non-hot spot, were used to define four classes. Using these classes, priorities and types of engineering intervention can be determined and thus, site interventions can be efficiently managed. In addition, the QR approach was compared with the AF and EB, against four robust and informative quantitative evaluation criteria tests. These tests showed that the QR method performs better than the two other methods. In this paper, a strategy to more accurately classify hot spots from a set of sites is proposed using four classes that may define priorities

and types of engineering intervention, allowing for the efficient management of site intervention. Above all, this analysis proves that the application of a probabilistic model is an alternative approach to hot spot research.

Keywords: Qualitative response model; Binary choice model; Hot spot identification; Countermeasures; Performance evaluation

INTRODUCTION

Accident prediction models (APM) have been extensively applied in transportation safety analysis. As reported in (Lord, 2002), they are extremely useful tools for the safety analysis of transportation facilities because they have a wide range of applications, such as in finding the proper relationship between accidents and covariates, determining the long-term average number of accidents regarding transportation entities and the computation of the expected number of accidents for sites that have yet to be built or upgraded. These issues may be applied in the study of hot spot identification, countermeasures, prediction and safety evaluation. Different statistical methods have been used to develop APM by taking into account the non-negative integer nature of crash-frequency data, among which the preferred one is application of count-data regression methods. A common technique used for APM is the generalized linear model (GLM) procedure with the assumption of a negative binomial (NB) error distribution (Sawalha and Sayed, 2001, Lord and Persaud, 2004, Lord, Manar and Vizioli, 2005, Anastasopoulos and Mannering, 2009). However, methodological approaches have evolved over the years to address superior statistical fits and/or predictive capabilities. Recently, a review of the statistical analysis and its methodological alternatives has been published (Lord and Mannering, 2010). In all of these statistical techniques, the predicted outcome is the number of accidents per time period for a transportation entity (segment or intersection), i.e., a quantitative response.

Qualitative response models have been focused on the statistical analysis of accident-injury severities, where also a variety of methodological techniques have been applied. The discrete and the ordinal nature of the injury-severity data, generally represented by discrete categories such as fatal injury, incapacitating injury, non-incapacitating, possible injury, and property damage only are proper represented by a qualitative response model.

The most common techniques applied to analyze accident-injury severity were the binary, multinomial logit, nested logit and ordered probit and logit formulation (Carson and Mannering, 2001, Kockelman and Kweon, 2002, Abdel-Aty, 2003, Eluru, Bhat and Hensher, 2008, Wang and Abdel-Aty, 2008, Savolainen, Mannering, Lord and Quddus, 2011). These models can be grouped in two response mechanisms: the ordered response (ordered probit and ordered logit) and unordered response (binary, multinomial and nested). The ordered response mechanism has the advantage of being parsimonious in structure because it imposes the restriction that the regression parameters are the same for different severity levels. Hence, the adjacent severity levels are correlated. On the other hand, the unordered response mechanism is based on a utility-maximization principle hypothesis and thus the severity levels are not presumed to correspond to the successive partition of a uni-dimension latent variable (Bhat and Pulugurta, 1997).

In the context of accident frequency, the discrete choice models were seldom used. However, recently, a novel approach has been described in (Ferreira and Couto, 2011) about

the use of the ordered probit model to assess safety at the planning level. This approach based on probabilistic results has the advantage to allow to performing a risk analysis.

As may be concluded from the above briefly review, the qualitative response models have been extensively applied in determining accident-injury severity, leaving other possible applications open. One of them is hot spot (also referred to as black spots, hazard sites, high-risk sites, accident-prone sites, sites with promise or priority investigation locations) identification. Thus, this paper aims to provide an analysis of the potential of the quantitative approach as an alternative to commonly used methodologies in identifying hot spots and to determine if the novel approach improves those methodologies.

A hot spot can be theoretically defined as any location that has a higher number of accidents than other similar locations as a result of local risk factors (Elvik, 2008b, Montella, 2010). Several methods for identifying hot spots have been used; the most commonly used method is the ranking of accident frequency and the ranking of accident rate (Montella, 2010). In fact, in Europe, a recent EU directive (2008/969/EC in (2008)) on road infrastructure safety management suggests that for the identification of road sections with a high concentration of accidents, the number of fatal accidents that have occurred in previous years (at least 3 years) per unit of segment length as related to traffic volume and, in the case of intersections, the number of accidents per location should be used as a minimum. Despite these practical examples, considering the methods most commonly used by several researchers, it is generally assumed that the empirical Bayes technique is the best approach because it is more capable of accounting for random fluctuations in the record number of accidents (regression to the mean phenomenon) (Cheng and Washington, 2008, Elvik, 2008a). Other methods such as classical confidence intervals and the accident reduction potential are seldom used in practice, though they are the focus of several researchers' attention.

Assuming that it is difficult to quantitatively define the precision of estimates that are determined using different techniques, recent research has been done to develop new criteria for evaluating methods of identifying hot spots (Cheng and Washington, 2008, Montella, 2010, Cafiso and Silvestro, 2011, Lan and Persaud, 2011). Also, the use of some critical issues in the identification of hot spots, such as the traffic volume, segment length and accident observation period, reported in (Cafiso and Silvestro, 2011), can emphasize the quality of the results.

Based on the foregoing discussion, it is difficult to set a clear definition of a hot spot because a simple observation of an unusually high number of accidents does not necessarily mean that there is a safety problem on a particular section of road; it may be a random "up" fluctuation in accident count during the observation period (Elvik, 2008a, Montella, 2010). This, this difficulty arises in distinguishing sites that are truly high risk (true positives) from those that may be false positives as a result of random "up" fluctuations. As pointed out by Montella in (Montella, 2010), 'errors in hot spot identification may produce large numbers of false negatives and large numbers of false positives', where in the former, truly hazardous sites are mistakenly designated as safe and, in the latter, truly safe sites are mistakenly identified as hazardous. These errors result in the inefficient use of resources to perform safety improvements and reduce the global effectiveness of the safety management process.

Studies have focused on the assessment of true and false positives and of true and false negatives to analyze the performance of different hot spot identification methods (Geedipally and Lord, 2010, Cafiso and Silvestro, 2011, Lan and Persaud, 2011). Several criteria have been used to compare the hot spot identification methods using criteria such as sensitivity and specificity (epidemiological criteria), which are based on the classification of the type of

errors generated by correct or incorrectly hot spot detection (Elvik, 2008a, Geedipally and Lord, 2010). Although several studies have focused on the analysis of the model performance based on the error types, it is interesting to note that, to the author's knowledge, there is less attention focused on the study of hot spot selection on the basis of a threshold or according to budget constraints. The most common was considered the top percentage range between 1% and 10%, with the highest accident frequency, accident rate or other identification criterion.

In this context, we propose an alternative approach based on a qualitative response model that embraces, from a different perspective, the analysis that to date has been referenced in many studies. This approach is based on categorical modeling that defines the dependent variable as an indicator of a discrete choice, which in the case of the hot spot approach may be classified into two categories: hot spot and non-hot spot, i.e., the binary (0/1) dependent variable is set to differentiate a safe site from a hot spot site. To define these two categories, it is necessary to use a strategy that selects a hot spot as accurately as possible. Based on a threshold strategy, various hypotheses were analyzed to obtain the most appropriate value to balance sensitivity and specificity criteria (epidemiological criteria).

In this approach, a general framework of probability models is used to link the outcome to a set of factors. Because we used two categories for the dependent variable, binary choice models were used. The road attributes considered were: traffic volume, number of minor intersections per kilometer, road function classification and land use. The binary choice model also provided an analysis of the risk factor effects on a hot spot relative to those on a non-hot spot site. Finally, based on those risk factors, the model estimated the probability associated with a site, and, in turn, a category was set. With respect to the probability of a hot spot or a non-hot spot occurring, four classes of countermeasure engineering interventions are proposed. These four classes represent a tool to be used in the efficient management of engineering interventions.

We believe that the qualitative approach, as a simple alternative, may overcome the most commonly used methodology. The main characteristics of this approach are (1) the outcome of a binary model includes a matrix of predictions that compares actual values with predicted values that indicate false and true positives and predict success or failure; and (2) the outcome is a probability that determines the likelihood of a site being a hot spot, and as such, a degree of uncertainty is associated with the selection of a hot spot, allowing (3) a quick and efficient analysis of a site situation in terms of safety and, protection against misdirecting engineering efforts toward undeserving sites. In the latter, different types of countermeasure treatment approaches can be defined. Moreover, this qualitative approach provides an individual analysis of the exogenous variable effects on hot spot sites relative to those on non-hot spot sites. Furthermore, all of these characteristics of the binary model may allow for control over other issues that have not been fully addressed, such as the definition of a hot spot and random fluctuation problems.

Thus, to carry out these goals, a binary choice model was developed and applied to urban road segments using Porto (Portugal and Goldner) data covering a 5-year period. These data were used to estimate the parameters and marginal effects as well as for performance evaluation. A strategy was developed to determine whether or not an urban segment can be classified as a hot spot. Then, the same data were used to validate this strategy with respect to the studied sites. Additionally, four classes of countermeasure treatments were defined.

The remainder of this paper is organized as follows: methodology, data description, results and summary and conclusions.

METHODOLOGY

This section describes the dependent variable, the exogenous variables and the regression approach used for hot spot identification. Also, the performance evaluation criteria and possible countermeasure treatments of sites based on the likelihood of a site being a hot spot or a non-hot spot are described.

Dependent Variable

A binary dependent variable was defined to differentiate a hot spot site from a safe site (category 0 for a non-hot spot and category 1 for a hot spot site). Because the accident frequency along with some road attributes such as segment length and traffic flows are usually available in the database for hot spot identification, categories were defined using this information. To reflect the accident risk (accident frequency/exposure) per year, the segment length was used as the exposure measure to normalize the frequency of accidents. Different studies have used this variable as an offset, considering the nearly linear relationship between segment length and accident frequency (Montella, 2010, Cafiso and Silvestro, 2011). To verify this observation in the Porto segments, the relationship between this variable and accident frequency, using the segment length as a covariate, was analyzed. Applying a count-data model with a negative binomial error distribution to the Porto data, an elasticity of 0.9 was determined, which indicates that a 1% increase in segment length implies a 0.9% increase in accident frequency (Ferreira, 2010). Although this elasticity indicates a slight deviation from a linear relationship, it is not significantly nonlinear. On the contrary, as referred by Lord (Lord, 2002), traffic flow may not be a suitable measure of exposure. In fact, for traffic flow (Annual Average Daily Traffic or AADT) it was found an elasticity value of 0.5, which confirms the non-linear relationship between AADT and accident frequency. Moreover, the traffic flow is an important covariate to characterize road sections in hot spot analysis because countermeasures can be related to changing traffic flows or movements. Therefore, the accident frequency per segment length was selected to define the categories of the binary choice model.

On the basis of the threshold strategy for listing hot spots, a criterion to identify the appropriate threshold value above which a site is set as a hot spot (category 1) was defined.

Exogenous Variable

To characterize a segment, various covariates were included in the regression model. Traffic flow is considered the most determinant variable for the occurrence of accidents and as such was usually used as the only variable in the model (Fridstrom, Ifver, Ingebrigtsen, Kulmala and Thomsen, 1995, Lord, 2000, Lord, 2006). In addition to traffic flow, the number of minor intersections per length was included by taking into account the fact that the presence of an intersection with a minor road, such as access roads without traffic data, can have a significant effect on a segment with respect to accident risk (Mountain, Fawaz and Jarrett, 1996).

The geometric and functional features of the segment were typified using the road function classification. This is a broader manner of characterizing segments using single information that is easy to obtain. Although the municipal master plan defines four road classes, namely, arterial roads, principal distributor roads, local distributor roads and access roads, only principal distributor roads and local distributor roads were used because arterial roads have characteristics similar to those of a highway and a lack of traffic flow data for access roads

was encountered. Furthermore, to embrace the street environment of a segment that can describe, for example, the presence of truck vehicles or pedestrian, land use was also included in this study. Five different types of land use based on the municipal master plan were taken into account: Land Use 1 (LU1) – high density of buildings; Land Use 2 (LU2) – low density of buildings; Land Use 3 (LU3) – industrial area; Land Use 4 (LU4) – community building area (educational buildings and sports grounds); Land Use 5 (LU5) – historic center area. Although the latter's covariates are difficult to analyze from an engineering treatment perspective, it may indicate the need of changing road features and/or the street environment from a broader perspective. Also, it has the advantage of heeding the differences in risk factor effects between a hot spot located in an industrial area and a hot spot located in a central area, for example.

The regression approach

The qualitative response approach is based on categorical modeling that defines the dependent variable as an indicator of a discrete choice, i.e., the dependent variables are merely a coding for some qualitative outcome (Greene, 2008). In this approach, a general framework of probability models is used to link the outcome to a set of factors (Greene, 2008):

$$Prob(event\ j\ occurs) = Prob(Y=j) = F[effects, parameters] \quad (1)$$

where the "event" is an individual's choice among a set of alternatives.

Considering a binary response to differentiate a safe site (Y=0) from a hot spot site (Y=1) and the covariates described later, which are gathered in a vector x to define the decision between both categories:

$$\begin{aligned} Prob(Y=1|x) &= F(x,\beta) \\ Prob(Y=0|x) &= 1 - F(x,\beta) \end{aligned} \quad (2)$$

The set of parameters β reflects the impact of changes in x on the probability. A suitable model for $F(x,\beta)$ is assumed, usually considering a normal or logistic distribution. The distributions are similar in that it is difficult to justify the choice of one distribution or another on theoretical grounds (Greene, 2008). In this paper, the logistic distribution was chosen because it produced slightly better outcomes:

$$Prob(Y=1|x) = \frac{e^{x'\beta}}{1+e^{x'\beta}} \quad (3)$$

The function $\Lambda(.)$ is a commonly used notation for the logistic cumulative distribution function.

It should be noted that the parameters of the model are not necessarily the marginal effects; therefore, to compute them the following expression should be used:

$$\frac{\partial E[y|x]}{\partial x} = \Lambda(x'\beta)[1-\Lambda(x'\beta)]\beta \quad (4)$$

These values will vary with the values of *x*. Hence, the estimated model interpretation should be calculated using the means of the covariates. It should be noted that it is not appropriate to use Eq. (4) to compute the marginal effects of dummy (binary variable that takes the value 0 or 1) variables to apply the effect of a change in a dummy variable. The appropriate marginal effect of a binary independent variable, represented by *d*, would be:

$$\text{marginal effect} = E[y = 1 | \bar{x}_{(d)}, d = 1] - E[y = 1 | \bar{x}_{(d)}, d = 0] \quad (5)$$

where $\bar{x}_{(d)}$ denotes the means of all the other variables in the model.

From the latent regression – $y^*=xb+e$, the unobserved y^* variable is determined (where ε is assumed to have mean zero and a standard logistic variance $\pi^2/3$). In other words, we do not observe the predicted number of accidents; we only determine whether it is a hot spot or safe site. Therefore, our observation is:

$$\begin{cases} y = 1 \text{ if } y^* > 0 \\ y = 0 \text{ if } y^* \leq 0 \end{cases} \quad (6)$$

Evaluation criteria

As was mentioned previously, errors in hot spot identification may produce large numbers of false positives and false negatives. The former are related to sites that are detected as hot spots but are non-hot spots and the latter are those that are hot spots but are detected as non-hot spots. The number or percentage of sites that are correctly classified, and the false positives and negatives can be represented by a matrix, as in the study by Geedipally and Lord (Geedipally and Lord, 2010). The values predicted by the binary model versus the actual values provide a similar analysis of the type of errors described above; therefore, the same performance evaluation criteria, such as those described in (Elvik, 2008a, Geedipally and Lord, 2010, Lan and Persaud, 2011), were applied.

To demonstrate the performance of the binary model, a matrix with the numbers and percentages of the predicted versus actual observations will be shown. The epidemiological criteria are also presented to evaluate the diagnostic performance of this approach. These criteria are:

$$\text{[illegible]} = \frac{\text{[illegible]}}{\text{[illegible]}}$$

$$\text{[illegible]} = \frac{\text{[illegible]}}{\text{[illegible]}}$$

Assuming that the relationships between the dependent variable and covariates are correctly estimated by the model, the results related to the probabilities will be interpreted from a countermeasures treatment perspective based on the probabilities associated to each one, i.e., on the uncertainty in hot spot identification. The idea is to assume that we are in the presence of a “true” hot spot for a high value of probability of *Prob(Y=1|x),* and on the contrary, a “true” non-hot spot will be assumed for a high value of probability *Prob(Y=0|x).* By exclusion, uncertain hot spots and non-hot spots comprise in a critical intermediate range.

DATA DESCRIPTION

A binary response model was developed using accident data collected from 376 Porto segments. The data were obtained from the official Portuguese Police Security database, covering all local (accidents resulting in injury and accidents resulting in property damage only) police-recorded accidents for the years 2001 to 2005. The data consist of 5634 police-recorded accidents; of these, 1182 were personal injury accidents and 4452 were property-damage-only accidents. All accidents were related to their specific location by applying a geographic information system (GIS) tool including information on segment length, number of intersections, road function classification and land use. The AADT was estimated by the Porto "SATURN" traffic model and data provided by permanent counting stations located throughout the principal city zones belonging to the Urban Traffic Center because traffic flow values were not available for all road networks. Table I provides relevant descriptive statistics for covariates and accident data.

Table I - Statistical description of variables used in accident frequency models (5-year)

Variable	Min.	Max.	Average	S.D.
Accident frequency	0	27	3.00	3.95
AADT	142	64068	15139	116734
Segment length (in meters)	30	3343	329	36
Number of minor intersections per kilometer	0.00	10.00	1.67	2.02
High density of buildings (LU1)	0	1	0.55	0.50
Low density of buildings (LU2)	0	1	0.22	0.42
Industrial (LU3)	0	1	0.03	0.18
Community buildings (LU4)	0	1	0.06	0.23
Historic center (LU5)	0	1	0.14	0.35
Local distributor roads	0	1	0.50	0.50
Principal distributor roads	0	1	0.50	0.50

It should be noted that land use and road function classification variables are dummy variables. Correlations among the variables presented in Table I were analyzed using a correlation matrix, allowing us to assume that the explanatory variables are not correlated ($\rho \leq 0.3$) (Ferreira, 2010).

The accident frequency per segment length values for each observation were distributed by one of the two categories using a threshold value of 6.88 for the number of accidents per km achieved by the methodology described below. Hence, it was found that 51.9% of observations were related to category 0 and 48.1% to category 1.

RESULTS

This section is divided into three subsections. First, a strategy to determine a threshold number to list hot spots is described. Second, estimation results and performance evaluation are presented. Finally, a methodology is applied to the case of Porto data to define countermeasure treatments.

Threshold strategy for hot spot definition

On the basis of the threshold strategy used to list hot spots, a criterion to identify the appropriate threshold value above which a site is classified as a hot spot (category 1) was defined using the epidemiological evaluation criteria, namely, the sensitivity and the specificity, whose definition was presented above.

Taking into account the fact that the binary choice model set an observation as category 1 when the probability was above 0.5 ($Prob(Y=1|x)>0.5$), this value was used to allocate the point where the sensitivity and specificity lines plotted against cut-off (P^*) probability may cross each other. The idea was to balance the values of those criteria, ensuring roughly the same number of false positives (Type I errors) and false negatives (Type II errors) at the same point that the model used to separate a hot spot from a non-hot spot.

Under this constraint and for the case of Porto data, the percentile 52, corresponding to 6.88 accidents per kilometer, was obtained. Thus, by setting the threshold above this value, the observation was considered a hot spot by being identified as belonging to category 1. Figure 1 shows the sensitivity and specificity lines plotted against the cut-off (P^*) probabilities estimated by the binary model. As can be seen, the cross point of the two lines roughly matches the probability value of 0.5, above which the observation is classified as belonging to category 1.

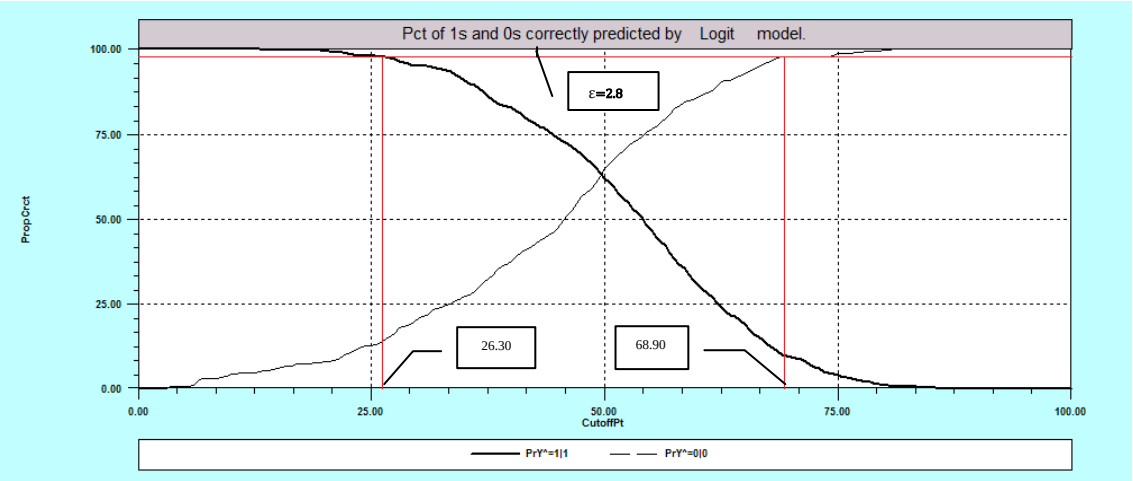

Figure 1 – Sensitivity and specificity against cut-off (P*) probability

Figure 1 shows that the range of probability values approximately between 0.25 and 0.75 correspond to a category choice near the threshold value of either being category 1 or 0; therefore, this region is defined as a critical zone, where there is no certainty about the hot spot identification. A deep analysis of this critical zone must be performed using an approach we propose and describe later in this paper.

Estimation results and performance evaluation

The binary choice model (BM) was estimated for Porto database, and the marginal effects for each variable were computed as discussed above (see results shown in Table II). It should

be noted that the parameter estimates in the BM represent the effects of variables on category 1 relative to category 0.

The parameter estimates are significant at a 95% confidence level, indicating a possible relationship between the covariates and the probability of the categories used in the BM occurring, with the exception of the local distributor road function classification and the land use classified as LU5. This latter observation indicates that there are no differences in the marginal effects between a local distributor road and a principal distributor road. Also, no differences were found between a site located in a historic center area and a site located in an area with a high density of buildings, at least within the categories used in this approach.

Considering the marginal effects shown in Table II, it can be seen that when the AADT increases, there is an increase in the probability of a site being a hot spot (category = 1). Also, when the number of minor intersections per km increases, the probability of a site being a hot spot increases as well, although to a lesser extent. In the case of the land use variables, Table II shows that a site located in an industrial area has a higher probability of being a hot spot than in areas with a high density of buildings, as may be expected. Additionally, a site located in an area with a low density of buildings or community buildings is associated with a decrease in the probability of being a hot spot, with no major differences in the marginal effects between the dummy variables in question.

Table I - Binary choice model results

Parameter	Prob[Y=1]			Marginal effects
	Estimated value	Standard error	P[Z>z]	
Constant	-5.580	0.619	0.0000	-1.391
LnAADT	0.583	0.064	0.0000	0.145
Minor intersections per km	0.047	0.010	0.0000	0.012
Low density of buildings (LU2)	-0.648	0.125	0.0000	-0.157
Industrial (LU3)	0.695	0.285	0.0147	0.170
Community buildings (LU4)	-0.507	0.210	0.0159	-0.122
Historic center (LU5)	0.221	0.146	0.1315	0.055
Local distributor road	-0.067	0.106	0.5246	-0.017
Number of observations	1880			
Restricted log-likelihood	-1301.738			
Log-likelihood function	-1207.145			

Table III shows value predicted by the BM versus the actual values for the category classification. Out of 904 true hot spot observations, the BM detected 557 (30%). In other words, there were 344 type I errors (false positives) and 347 type II errors (false negatives). In total, there were 691 observations that were incorrectly classified by the BM. Table IV gives the percentage of the prediction successes including the sensitivity and specificity percentage values. As expected, because of the constraint (threshold criterion) mentioned above, those values are quite similar. Figure 2 plots Sensitivity(*Prob/P**) against 1-Specificity(*Prob/P**) to provide a 'fit measure'. This 'fit measure' is computed as the area under the ROC curve (Greene, 2007). A greater area implies a greater model fit. A model with no fit has an area of 0.5. As shown in Figure 2, the area has a value of 0.672.

Table II - Classification of the outcomes[a]

	Predicted value		
Actual value	0	1	Total
0	632(34%)	344(18%)	976(52%)
1	347(19%)	557(30%)	904(48%)
Total	979(52%)	901(48%)	1880(100%)

a) Note that column or row total percentages may not sum to 100% because of rounding

Table III - Successful BM predictions

Prediction Success	%
Sensitivity	61.6%
Specificity	64.8%
Positive predicted value (predicted 1s that were actual 1s)	61.8%
Negative predicted value (predicted 0s that were actual 0s)	64.6%
Correct prediction = actual 1s and 0s correctly predicted	63.2%

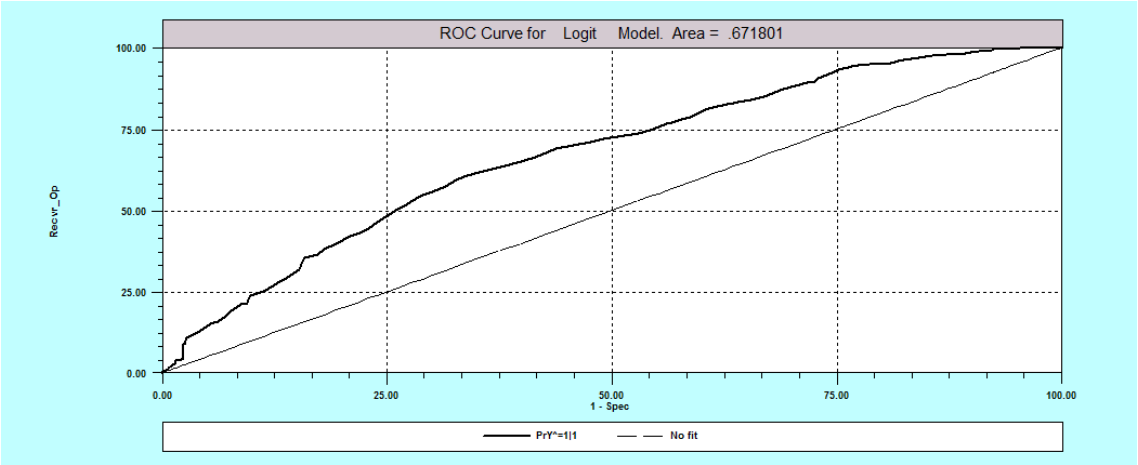

Figure 2 - Sensitivity(*Prob/P**) against 1-Specificity(*Prob/P**)

Classification of countermeasure treatments based on uncertainty in hot spot identification

This section focuses on the two values located in the middle of what is defined as a critical zone (Figure 1) and the classification of four different types of engineering interventions. The first step is to establish the two values. To do so, a comparison between the category classifications using the accident frequency per kilometer observed as well as the probability estimated by the BM ($Prob(Y=1|x)>0.5$) was performed. Assuming a "correct proportion" of sensitivity and specificity of 97.2% (100%-ε) (see Figure 1), the two probability values are obtained: 0.263 and 0.689 cut-off probability.

Thus, for efficient hot spot management, we defined the classification of sites in terms of safety so that when the probability of a site being a hot spot is lower than 26.3%, we assume that the site is safe; therefore, no intervention is needed (Class A). When the probability estimated is between 26.3% and 50%, despite the fact that the site could be identified as

being safe, we cannot ensure that some prevention countermeasures are not required (Class B). When the probability estimated between 50% and 68.9%, the site is a prone hot spot, and engineering interventions of the low-cost-measure type should be implemented (Class C). When the estimated hot spot probability is higher than 68.9%, we can safely assume that the site is a hot spot (Class D), and a deep analysis based on engineering interventions should be performed.

Through the probabilistic approach described above, a more efficient tool to manage the engineering site interventions can be provided by considering four classes of priority of intervention: Class D – high priority of intervention; Class A – no intervention; Class B – intervention with few resources; Class C – intervention with low-cost resources. Thus, for a high-priority intervention (Class D), given the covariates included in the BM, strong measures should be used, such as changing the traffic flow and/or movements, placing traffic signs in minor intersections, parking features changes, and diverting access to the adjacent buildings. Sites classified as Class C require traffic-calming measures with a low-cost investment. For Class B, weak interventions such as the implementation of traffic signs and street illumination may be considered to prevent unsafe situations.

Finally, to validate the classification procedure described above, the procedure was applied to all 376 segments (assuming for each covariate value the average site value of the five-year observations). To do so, the percentage error related to the site classifications Class A and Class D was analyzed and compared to the percentage error assumed to define the threshold values of the 4 classes – ε = 2.8%. Under this condition, 1.6% of sites were found not to be hot spots despite being classified as Class D. Also, 0.5% of Class A sites were incorrectly classified. Thus, it was confirmed that the percentage errors of Class A and Class D were less than the previous limit value of 2.8% set for ε.

SUMMARY AND CONCLUSIONS

Hot spot analysis, including identification methodologies, performance evaluation criteria and so on, has been extensively studied for a long time. However, some issues still merit attention, particularly the use of alternative methodologies that may improve the current ones. Thus, this paper presents an alternative approach based on a qualitative response model that, given those specific properties, may improve the currently used methodologies while being capable of providing useful tools for the efficient treatment of hot spots.

This novel approach is based on a probabilistic model that defines the dependent variable as an indicator of a discrete choice used to, in this case, differentiate a hot spot from a non-hot spot considering a binary (0/1) dependent variable set by the number of accidents per kilometer. To define the two categories, it is necessary to use a strategy that selects a hot spot from a set of sites as accurately as possible. Based on a threshold strategy, various hypotheses were evaluated to obtain the most appropriate value toward more balanced values of sensitivity and specificity criteria (epidemiological criteria).

In this approach, binary choice models were used to link the outcome to a set of factors. The factors considered were: traffic volume, number of minor intersections per kilometer, road function classification and land use. Based on these risk factors, the model estimates the probability of a site being a hot spot ($Prob(Y=1|x)$). To apply this approach, data gathered from Porto, Portugal, over a five-year period was used. A performance evaluation was applied based on the epidemiological criteria, and a matrix with the predicted values versus the actual values was provided.

Finally, considering the probabilistic nature of this approach, it was possible to define four classes of sites using the estimated probability. Two probability threshold values were determined assuming a value of 97.2% for the "correct proportion" of sensitivity and specificity. These two values, along with the 0.5 probability value that the model defines to separate a hot spot from a non-hot spot, were used to determine threshold values for four classes. Using these classes, priorities and types of engineering intervention can be determined and thus, site interventions can be efficiently managed.

From the application briefly described above, various advantages can be pointed out related to the probabilistic approach that may improve the current hot spot identification and intervention process. First, the outcome is a probability that determines the likelihood of a site being a hot spot and, as such, allows for different engineering intervention measures, depending on the degree of uncertainty that is associated with the selection of a hot spot. Moreover, this qualitative approach provides a relative probabilistic analysis of the exogenous variable effects on hot spot sites. In fact, it is expected that a site in an industrial area is likely to be more hazardous than a site in a residential area. With a probabilistic model, the risk effects of various road attributes are differentiated between categories.

Furthermore, the outcome of a binary model includes a matrix of predictions that compares the actual values with the predicted values thus indicating the percentage of false and true positive and prediction success and failure. Also, in this paper, a strategy to more accurately classify hot spots from a set of sites is proposed using threshold values.

Above all, this analysis proves that the application of a probabilistic model is an alternative approach to hot spot research. Nevertheless, further applications to other data are required to fully assess the utility of this approach.

REFERENCES

Abdel-Aty, M. (2003) Analysis of driver injury severity levels at multiple locations using ordered probit models. *Journal of Safety Research,* 34**,** 597-603.

Anastasopoulos, P. C. and Mannering, F. L. (2009) A note on modeling vehicle accident frequencies with random-parameters. *Accident Analysis and Prevention,* 41**,** 153-159.

Bhat, C. R. and Pulugurta, V. (1997) A comparison of two alternative behavioral choice mechanism for household auto ownership decisions. *Transportation Research Part B,* 32**,** 61-75.

Cafiso, S. and Silvestro, G. D. (2011) Safety-Indicator Performance in the identification of Black Spots on 2-Lane Rural Roads. Performance Evaluation of Black Spot Identification Methods. *Transportation Research Record*.

Carson, J. and Mannering, F. (2001) The effect of ice warning signs on ice-accident frequencies and severities. *Accident Analysis and Prevention,* 33**,** 99-109.

Cheng, W. and Washington, S. (2008) New Criteria for Evaluating Methods of Identifying Hot Spots. *Transportation Research Record: Journal of the Transportation Research Board,* 2083**,** 76-85.

Eluru, N., Bhat, C. R. and Hensher, D. A. (2008) A mixed generalized ordered responde model for examining pedestrian and byciclist injury severity level in traffic crashes. *Accident Analysis and Prevention,* 40**,** 1033-1054.

Elvik, R. (2008a) Comparative Analysis of Thechniques for Identifying Locations of Hazardous Roads. *Transportation Research Record: Journal of the Transportation Research Board,* 2083**,** 72-75.

Elvik, R. (2008b) A survey of operational definitions of hazardous road locations in some European countries. *Accident Analysis and Prevention,* 40**,** 1830-1835.

Ferreira, S. (2010) *A Segurança Rodoviária no processo de planeamento de redes de transportes.* Ph.D. dissertation, University of Porto.

Ferreira, S. and Couto, A. (2011) Categorical Modeling to Evaluate Road Safety at the Planning Level. *3rd International Conference on Road Safety and Simulation*, Transportation Research Board, Indianapolis, Indiana, USA.

Fridstrom, L., Ifver, J., Ingebrigtsen, S., Kulmala, R. and Thomsen, L. K. (1995) Measuring the contribution of randomness, exposure, weather and dayligth to the variation in road accident counts. *Accident Analysis and Prevention,* 27**,** 1-20.

Geedipally, S. R. and Lord, D. (2010) Hot Spot Identification by Modeling Single-Vehicle and Multi-Vehicle Crashes Separately. *Transportation Research Record,* 2147**,** 97-104.

Greene, W. H. (2008) *Econometric Analysis.* Sixth Edition. Pearson International Edition, New Jersey.

Kockelman, K. M. and Kweon, Y.-J. (2002) Driver injury severity: an application of ordered probit models. *Accident Analysis and Prevention,* 34**,** 313-321.

Lan, B. and Persaud, B. (2011) Investigation and Evaluation of Ranking Criteria for Hot Spot Identification. *Transportation Research Record*.

Lord, D. (2000) *The prediction of accidents on digital networks: characteristics and issues related to the application of accident prediction models.* Ph.D. dissertation, University of Toronto.

Lord, D. (2002) Application of Accident Prediction Models for Computation of Accident Risk on Transportation Networks. *Transportation Research Record,* 1784.

Lord, D. (2006) Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis and Prevention,* 38**,** 751-766.

Lord, D., Manar, A. and Vizioli, A. (2005) Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. *Accident Analysis and Prevention,* 37**,** 185-199.

Lord, D. and Mannering, F. (2010) The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A,* 44**,** 291-305.

Lord, D. and Persaud, B. N. (2004) Estimating the safety performance of urban road transportation networks. *Accident Analysis and Prevention,* 36**,** 609-620.

Mountain, L., Fawaz, B. and Jarrett, D. (1996) Accident prediction models for roads with minor junctions. *Accident Analysis and Prevention,* 28**,** 695-707.

Portugal, L. d. S. and Goldner, L. G. (2003) *Estudo de pólos geradores de tráfego e de seus impactos nos sistemas viários e de transportes.* Editora Edgard Blucher Ltda, São Paulo.

Savolainen, P. T., Mannering, F., Lord, D. and Quddus, M. (2011) The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. *Accident Analysis and Prevention,* 43**,** 1666-1676.

Sawalha, Z. and Sayed, T. (2001) Evaluating safety of urban arterial roadways. *Journal of Transportation Engineering,* 127**,** 151-158.

Wang, X. and Abdel-Aty, M. (2008) Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models. *Accident Analysis and Prevention,* 40**,** 1674-1682.