Método probabilístico para identificação de zonas de acumulação de acidentes

Sara Ferreira[1] e António Couto[2]

Resumo: Neste trabalho apresenta-se um novo método de definição e identificação de zonas de acumulação (ZAA) de acidentes considerando um modelo de regressão binário. Este modelo permite calcular a probabilidade de um local ser ou não ZAA tendo em conta as características geométricas e funcionais do mesmo. Este novo método foi aplicado ao caso das interseções tendo-se, para tal, gerado uma base de dados fictícia considerando as características da sinistralidade e das interseções da cidade do Porto. Através da simulação de dados é possível conhecer *a priori* as "verdadeiras" ZAA. O desempenho do método foi analisado com base nos erros resultantes da classificação dos locais em ZAA ou não-ZAA, e comparado com dois dos métodos mais aplicados e analisados – o método de ranking pelo número de acidentes e o método Bayesiano-empírico. Desta análise verificou-se que o método binário proposto tem claramente melhor desempenho.
DOI:10.4237/transportes.v21i3.683.

Palavras-chave: zona de acumulação de acidentes; modelo binário; interseções urbanas; simulação de dados.

Abstract: This paper presents a new methodology to define and to identify hot spot based on a binary regression model. Considering the geometric and functional site characteristics, the model estimates the probability of a site being a hot spot. This method was applied to intersections using a simulation-based approach to data generated considering the characteristics of the Porto city data base. Using simulation, it is possible to establish sites that are *a priori* "true" hot spot. The performance of the new method was evaluated using the errors of the classification outcomes of the sites and compared with two commonly implemented methods that are the simple ranking of sites and the empirical Bayesian technique. This analysis has set that the proposed binary model performed better.

Keywords: hot spot; binary model; urban intersections; simulated data.

1. INTRODUÇÃO

Sob exclusivamente do ponto de vista da envolvente viária, uma zona de acumulação de acidentes (ZAA), vulgarmente designada de ponto negro, é uma zona geográfica na qual, por influência de características da infraestrutura rodoviária específicas à área, a frequência esperada de acidentes é superior ao expectável face à distribuição de acidentes nas áreas circundantes, nomeadamente em zonas aparentemente semelhantes (Cardoso, 1998). A aplicação de medidas de tratamento a locais identificados como ZAA tem demonstrado resultar numa diminuição significativa do número de acidentes, e muitas vezes associada a baixos custos em termos de investimento. Por este facto, diversos trabalhos têm vindo a ser desenvolvidos e aplicados neste âmbito, nomeadamente estudos que analisam métodos de identificação de ZAA (MIZAA) para seleção de locais a tratar. O MIZAA mais comumente utilizado corresponde ao ranking dos locais por número de acidentes ou por taxa de acidentes. O número de locais a selecionar baseia-se, em geral, num valor limite de número de acidentes ou numa condição de acordo com restrições de orçamento (Geedipally e Lord, 2010). A União Europeia estabeleceu uma diretiva (2008/969/EC) no âmbito da gestão da segurança em infraestruturas rodoviárias em que sugere que a identificação de locais com elevada concentração de acidentes se baseie no número de acidentes mortais ocorridos nos anos recentes (pelo menos 3 anos) por comprimento do segmento por volume de tráfego ou por interseção. Apesar de este ser o método mais aplicado na prática, vários trabalhos de investigação científica apontam o método Bayesian-empírico (MBE) como a melhor abordagem a considerar na identificação das ZAA, pois diminui os efeitos do fenómeno de regresso-à-média (flutuações aleatórias no número de acidentes registados ano a ano em torno de um valor médio) (Cheng e Washington, 2008, Elvik, 2008).

No entanto, este método apresenta algumas limitações tais como a necessidade de utilizar uma extensa base de dados para desenvolver um modelo de previsão de acidentes, a seleção do tipo de distribuição a considerar para os acidentes observados é pouco flexível e o facto de utilizar os mesmos dados duas vezes, isto é, por um lado, para desenvolver um modelo de previsão de acidentes e por outro lado, como dados observados (Miranda-Moreno, *et al*., 2005). Surgiu assim com alternativa a abordagem Full-Bayesian (FB) (Miranda-Moreno, *et al*., 2005, Huang e Haque, 2009, Lan e Persaud, 2011) cuja principal diferença relativamente ao MBE reside na forma como se determina a distribuição *a priori* dos parâmetros (Miranda-Moreno, *et al*., 2005). No caso da abordagem FB os hiperparâmetros são determinados com base no conhecimento *a priori* das características da base de dados (Schluter, *et al*., 1997, Carlin e Louis, 2000). No entanto, este conhecimento *a priori* não é simples de obter gerando alguma controvérsia e como tal, justifica a utilização do MBE que tem uma abordagem mais simples.

Neste sentido, vários outros métodos têm vindo a ser desenvolvidos tais como, intervalos de confiança e potencial de redução de acidentes. Recentemente, o documento designado de *Highway Safety Manual* (AASHTO, 2010) compilou alguns desses métodos de identificação de ZAA.

[1] **Sara Ferreira,** Faculdade de Engenharia da Universidade do Porto, Porto, Portugal. (e-mail: sara@fe.up.pt)

[2] **António Couto,** Faculdade de Engenharia da Universidade do Porto, Porto, Portugal. (e-mail: fcouto@fe.up.pt)

Manuscrito recebido em 04/09/2013 e aprovado para publicação em 23/09/2013. Este artigo é parte de TRANSPORTES v. 21, n. 3, 2013. ISSN: 2237-1346 (online). DOI:10.4237/transportes.v21i3.683.

Contudo, tendo em conta a dificuldade de avaliar qual destes métodos é mais preciso, vários autores têm vindo a desenvolver critérios de avaliação dos vários MIZAA (Cheng e Washington, 2008, Montella, 2010, Cafiso e Silvestro, 2011, Lan e Persaud, 2011). De notar que outras questões tais como, o volume de tráfego, o comprimento das vias e o período de observação dos acidentes, podem enfatizar a qualidade dos resultados dos MIZAA tal como Cafiso e Silvestro (2011) sugerem.

De facto, é difícil estabelecer uma clara definição de ZAA e, consequentemente uma correta identificação do local, pois uma simples observação de um número excepcionalmente elevado de acidentes pode não significar necessariamente um problema de segurança relacionado com o local mas, pelo contrário, resultar de uma flutuação aleatória no período de observação considerado (Elvik, 2008, Montella, 2010). Esta questão, entre outras, reflete-se numa errónea identificação de locais como ZAA resultando num número elevado de falsos negativos e falsos positivos tal como referido em (Montella, 2010). Locais identificados como falsos negativos são aqueles que são verdadeiramente perigosos mas foram identificados como seguros, e locais identificados como falsos positivos são aqueles que são verdadeiramente seguros mas foram identificados como perigosos. Como se pode depreender, a identificação errada de locais como ZAA produz uma ineficiente alocação de recursos com vista a melhoria da segurança dos locais, e consequentemente reduzindo a global eficácia do processo de gestão implementado.

Por este facto, alguns estudos têm-se centrado na avaliação do desempenho dos MIZAA considerando a análise dos falsos e verdadeiros positivos, e falsos e verdadeiros negativos (Geedipally e Lord, 2010, Cafiso e Silvestro, 2011, Lan e Persaud, 2011). Geralmente, estes estudos baseiam-se em dois métodos distintos - método empírico, ou um método de simulação. No primeiro caso, utiliza-se dois períodos temporais de observação de acidentes distintos em que um desses períodos define os locais como verdadeiramente perigosos ou seguros, e depois compara os MIZAA aplicados ao outro período de tempo. No caso do método de simulação existem diversas abordagens em que através, por exemplo, da simulação Monte Carlo, são gerados n locais segundo uma distribuição probabilística. O método de simulação é mais vantajoso relativamente ao método empírico na medida em que, neste último, as verdadeiras ZAA não são de facto conhecidas *a priori* (Cheng e Washington, 2005, Geedipally e Lord, 2010). Como no método de simulação os locais são gerados e identificados *a priori* como ZAA, torna-se mais fácil determinar o número de falsos positivos e negativos e consequentemente avaliar se o MIZAA identifica corretamente as ZAA. Neste caso, vários critérios, como por exemplo os critérios epidemiológicos (sensitividade e sensibilidade), podem ser aplicados baseados na classificação do tipo de erros gerados pela correta ou incorreta identificação das ZAA (Elvik, 2008, Geedipally e Lord, 2010).

Na verdade, mesmo optando pela aplicação de um MIZAA com melhor desempenho, existe sempre um grau de incerteza associado à correta identificação do local como ZAA, principalmente se o método não incluir a influência das características do local. O facto do MBE se basear, não só no número de acidentes observado, mas também num modelo de previsão de acidentes que incorpora as características do local, poderá ser a razão pela qual este método apresenta, em geral, bom desempenho.

Considerando a importância das características do local assim como o grau de incerteza geralmente associado a metodologias deste tipo, propõe-se, neste artigo, um novo MIZAA baseado num modelo discreto probabilístico, mais especificamente um modelo binário. Neste modelo a variável discreta define a classificação de um local como ZAA ou não-ZAA (isto é, local seguro). A probabilidade de um local ser ou não ZAA é determinada em função das características principais desse mesmo local. Por outro lado, a definição de ZAA baseia-se na utilização de um valor limite para o número de acidentes acima do qual se considera o local como ZAA. Este valor será determinado considerando o percentil 95. Este novo método permite identificar locais como ZAA, no entanto, estes são diferenciados pela probabilidade do local ser "verdadeiramente" uma ZAA.

Este novo MIZAA foi desenvolvido em ambiente simulado de forma a conhecer *a priori* os locais "verdadeiramente" seguros e ZAA e, a partir desta informação determinar os falsos positivos e negativos. Para gerar o ambiente simulado, considerou-se a base de dados de acidentes e de locais relativos à cidade do Porto, Portugal para o período de 2001-2005. Os locais estudados referem-se a interseções com 3 ramos com ou sem sinalização semafórica, e com 4 ramos com ou sem sinalização semafórica. Note-se que os dados gerados não têm como objetivo reproduzir a realidade, sujeita a fatores aleatórios devido à interferência do ambiente viário, sociocultural, etc. que rodeia os locais, mas, pelo contrário, gerar dados que não estão sujeitos a interferências de componente aleatória com distribuição estatística indefinida, sendo por isso, possível identificar os locais como "verdadeiramente" ZAA.

Para avaliar o desempenho desta nova metodologia proposta, foram determinados os falsos positivos e negativos e comparados com as duas metodologias mais referenciadas e aplicadas – MBE e método do número de acidentes (MNA). Nestas duas metodologias, o ranking dos locais é realizado utilizando o valor estimado para o número de acidentes e o número de acidentes determinado para cada local, respectivamente. O método de comparação utilizado neste trabalho baseia-se em indicadores que analisam o desempenho dos três métodos em identificar corretamente ou não os locais que são "verdadeiramente" ZAA.

Apresenta-se a seguir a descrição do método probabilístico, a descrição da base de dados e seu esquema de simulação, os resultados do método probabilístico, a metodologia de comparação dos três MIZAA e respetivos resultados e, por último, algumas considerações finais.

2. DESCRIÇÃO DO MÉTODO PROBABILÍSTICO

Os modelos probabilísticos têm a seguinte estrutura geral (Greene, 2008):

$$\begin{aligned} P\left(\textit{evento } j \textit{ ocorrer}\right) &= P\left(Y = j\right) \\ P\left(Y = j\right) &= F\left[\textit{ fatores relevantes, parâmetros}\right] \end{aligned} \tag{1}$$

em que um "evento" corresponde a uma categoria entre um conjunto de possíveis categorias.

O modelo binário aplica-se quando a variável de resposta tem dois resultados possíveis. Assim, tal como referido anteriormente, neste estudo considerou-se o modelo binário tendo como variável dependente a categoria Y=0 para

identificar locais seguros e a categoria *Y*=1 para identificar um local como ZAA.

As ZAA foram identificadas considerando um valor de número de acidentes limite acima do qual se considera o local como uma ZAA. Para as observações da base de dados simulada segundo o processo descrito na secção 3.2, determinou-se o percentil 95 do número de acidentes como valor limite. A escolha deste valor é flexível, e depende dos objetivos pretendidos em termos de número de locais a tratar. Outras hipóteses foram ainda analisadas tendo-se verificado que, como seria de esperar, quanto maior o valor do percentil melhor o desempenho do modelo. Assim, aos locais com número de acidentes acima desse valor, atribuiu-se a categoria 1 (ZAA) e aos restantes locais a categoria 0. Estas duas categorias definem a variável dependente do método probabilístico binário. As variáveis independentes incluídas na regressão do modelo binário correspondem ao volume de tráfego e a variáveis binárias que caracterizam a interseção quanto ao número de ramos e tipo de sinalização. O volume de tráfego é uma variável de exposição fundamental, na medida em que é a circulação dos veículos que gera os acidentes, isto é, sem tráfego não há acidentes. Esta variável tem sido referida em diversos trabalhos de modelação como a mais determinante para a ocorrência de acidentes sendo muitas vezes considerada como variável única do modelo (Fridstrom, *et al*., 1995, OCDE, 1997, Lord, 2000, Lord, 2006). Por exemplo, nos modelos desenvolvidos por Greibe, P. (2003) o tráfego (volume e suas interações) foi identificado como a variável independente mais preponderante, representando cerca de 90% e 30% da componente sistemática dos modelos para segmentos e interseções, respectivamente. Esta variável de exposição pode ser considerada de diversas formas sendo a mais utilizada o tráfego médio diário anual (TMDA). O TMDA pode assumir várias formas no modelo de regressão sendo a mais simples a que relaciona o fluxo de tráfego com os acidentes considerando o total do fluxo de entrada na interseção. No entanto, e embora esta relação tenha o mérito da simplicidade, ela não traduz o conflito do tráfego da interseção. Neste trabalho, e após uma prévia análise no âmbito de outro estudo (Ferreira, 2010), considerou-se o cálculo do TMDA que entra na interseção separando-o em ramos principais e ramos secundários. As variáveis independentes binárias que caracterizam a interseção quanto ao número de ramos e tipo de sinalização foram analisadas em diversos trabalhos, concluindo-se que são variáveis com impacto quer no número quer no tipo de acidentes aí ocorridos. Por esse facto, são geralmente utilizadas em diversos trabalhos quer como variáveis a incluir no modelo quer como critério de homogeneização da base de dados das interseções (Lord e Persaud, 2004, AASHTO, 2010).

Assim, o modelo binário é representado matematicamente por (Greene, 2008):

$$\begin{aligned} P &= \left(Y = 1 \middle| x\right) = F\left(x, \beta\right) \\ P &= \left(Y = 0 \middle| x\right) = 1 - F\left(x, \beta\right) \end{aligned} \tag{2}$$

em que x corresponde às variáveis independentes, β os respetivos parâmetros e $F(.)$ é uma função específica para assegurar que $0 \le P(Y) \le 1$.

Para o cálculo da probabilidade da componente aleatória qualquer distribuição probabilística contínua é suficiente. As mais usualmente aplicadas são, no entanto, as distribuições simétricas designadas de normal e logística (Greene, 2008). As duas distribuições são muito similares em termos de resultados. Neste trabalho optou-se pela distribuição probit, sendo a probabilidade calculada pela distribuição normal acumulada:

$$P = \left(Y = 1 \middle| x\right) = \int_{-\infty}^{\infty} \phi\left(t\right) dt = \Phi\left(x\beta\right) \tag{3}$$

em que a função $\Phi(.)$ é a notação considerada para a função da distribuição normal.

O modelo binário é construído a partir de uma regressão latente em que um conjunto de variáveis independentes x explica a decisão entre as duas alternativas, e os respetivos parâmetros β refletem o impacto na probabilidade das variações dos valores de x (Greene, 2008).

Nos modelos de resposta qualitativa como é o modelo binário, a regressão latente reflete uma variável não observada y^*, obtida por:

$$y^* = x\beta + \varepsilon \tag{4}$$

em que $x\beta$ é designado de função índex. O termo de erro aleatório não observado ε, no caso do modelo probit, segue uma distribuição normal com média zero e variância um.

A variável observada é y, tal que:

$$\begin{aligned} y &= 1 \ \text{ se } y^* > 0, \\ y &= 0 \ \text{ se } y^* \le 0. \end{aligned} \tag{5}$$

Na análise dos modelos probabilísticos não lineares, como é o caso do modelo binário, é importante ter em conta que os parâmetros β estimados não representam o efeito marginal das variáveis independentes consideradas no modelo. Por esse facto, quando se pretende analisar o efeito da variação dos valores das variáveis na probabilidade calculam-se os efeitos marginais.

O desempenho do modelo discreto binário pode ser avaliado por diferentes medidas de ajuste. Neste trabalho considerou-se para a avaliação do ajuste do modelo a percentagem de observações corretamente previstas e a curva ROC (acrónimo de *Receiver Operating Characteristics*) que corresponde a uma representação gráfica dos critérios epidemiológicos, mais concretamente a sensibilidade em função do valor de (1-especificidade), para quaisquer valores de probabilidade fronteira (P^*) entre 0 e 1. A curva da proporção de verdadeiros positivos (sensibilidade) identificados pelo modelo binário é representada no gráfico em função da proporção de falsos positivos (1-especificidade). Cada ponto da curva ROC representa o par sensibilidade/especificidade correspondente à probabilidade fronteira (P^*). Assim, o teste considerado como ideal apresenta uma curva que se aproxima do canto superior esquerdo do gráfico (100% sensibilidade, 100% especificidade). Assim, quanto maior a área do gráfico sob a curva melhor é o ajuste do modelo; geralmente considera-se que um modelo não ajustado tem uma área com valor inferior a 0,5 (Greene, 2007).

3. DESCRIÇÃO DA BASE DE DADOS – PROCESSO DE SIMULAÇÃO

Nesta secção descreve-se a seguir em 3.1., a base de dados da cidade do Porto, Portugal, relativa aos anos 2001 a 2005.

Esta base de dados será apenas utilizada para o processo de simulação de uma nova base de dados fictícia cujo esquema de procedimentos se descreve na secção 3.2.

3.1. Descrição da base de dados da cidade do Porto

Os dados utilizados neste estudo são relativos a acidentes ocorridos em interseções da cidade do Porto de 3 e 4 ramos, e com ou sem sinalização luminosa, registados ao longo de um período de 5 anos (de 1 de Janeiro de 2001 a 31 de Dezembro de 2005). De notar que os segmentos resultam de uma caracterização da rede em arcos e nós no âmbito de um estudo de doutoramento (Ferreira, 2010), sendo que os acidentes ocorridos dentro de uma área de 20 metros de raio com centro na intersecção das vias foram atribuídos aos nós/interseções. A base de dados dos acidentes foi obtida a partir de dados oficiais da Polícia de Segurança Pública e incluem todo o tipo de acidentes (com vítimas e só com danos materiais) registados com a informação do local de ocorrência. Com base nesta informação, os acidentes foram georreferenciados através de um sistema de informação geográfica. Os dados consistem em 2029 acidentes, dos quais 447 resultaram em vítimas e 1582 só com danos materiais, referenciados a 211 interseções. Estas interseções estão divididas em 48 interseções de 3 ramos e sinalização luminosa; 67 interseções de 3 ramos sem sinalização luminosa; 70 interseções de 4 ramos e sinalização luminosa; e 26 interseções de 4 ramos e sem sinalização luminosa. A Tabela 1 apresenta a descrição estatística relativa aos acidentes e ao TMDA das interseções da cidade do Porto.

Tabela 1. Descrição estatística da base de dados do Porto

Variável	*Mín*	*Máx*	*Média*	*Desvio Padrão*
Número de acidentes	0	13	1,9	2,1
$TMDA_{Princ}$	285	71.525	18.309	11.629
$TMDA_{Sec}$	0	32.882	5.286	5.232

3.2. Descrição do processo de simulação

Através da simulação é possível classificar *a priori* os locais "verdadeiramente" ZAA. A partir desta informação determina-se o número de falsos positivos e negativos para, a partir destes, avaliar o desempenho dos MIZAA em identificar corretamente estes locais.

O esquema de simulação foi elaborado utilizando o seguinte procedimento (adaptado de Geedipally e Lord (2010)):

1. Gerou-se aleatoriamente, para 1000 locais, valores do TMDA principal e secundário, e os valores 0 ou 1 como categorias das variáveis binárias que caracterizam o número de ramos (3 ou 4 ramos) e o tipo de sinalização (com ou sem sinalização luminosa) das interseções. As variáveis TMDA principal e secundário foram geradas a partir de uma distribuição log-normal cujos valores da média e do desvio padrão foram calculados considerando a base de dados das interseções da cidade do Porto e que estão apresentados na Tabela 1. As duas variáveis binárias, número de ramos e tipo de sinalização, foram geradas a partir da distribuição discreta binomial com probabilidade determinada pela frequência de ocorrência dessas características na base de dados da cidade do Porto.

2. Considerando a base de dados real dos acidentes ocorridos nas interseções da cidade do Porto, estimaram-se os parâmetros das variáveis independentes para uma regressão cuja estrutura do erro assume uma distribuição binomial negativa (BN). Estes valores estão apresentados na Tabela 2.

Tabela 2. Resultados do modelo BN para a base de dados do Porto

Parâmetro	*Valor estimado*	*Desvio padrão*	*P[Z>z]*
Constante	-3,175	0,467	0,0000
$Ln(TMDA_{Princ})$	0,303	0,049	0,0000
$Ln(TMDA_{Sec})$	0,076	0,011	0,0000
4Ramos	0,126	0,067	0,0606
Sinal. Luminosa	0,409	0,066	0,0000
Parâmetro dispersão	0,502	0,052	0,0000

A distribuição BN permite representar a sobre dispersão geralmente verificada nas bases de dados de acidentes entre locais. Com base na regressão obtida, calcularam-se, para cada um dos 1000 locais, o número de acidentes utilizando a função:

$$\mu_i = \beta_0 \cdot TDMA_{Princ}^{\beta_1} \cdot TDMA_{Sec}^{\beta_2} \cdot e^{(\beta_3 \cdot Nramos + \beta_4 \cdot Tsinal)} \quad (6)$$

em que,

μ_i: número de acidentes para cada local i (i=1 a 1000);

$TMDA_{Princ}$: tráfego médio anual de entrada na interseção segundo os ramos principais;

$TMDA_{Sec}$: tráfego médio anual de entrada na interseção segundo os ramos secundários;

Nramos: variável binária que caracteriza o número de ramos da interseção (3 ou 4 ramos);

Tsinal: variável binária que caracteriza o tipo de sinalização da interseção (com ou sem sinalização luminosa); e

β: parâmetros obtidos para a base de dados da cidade do Porto.

Este valor μ_i define, para cada local, o valor do número de acidentes assumido como "verdadeiro", permitindo assim identificar as "verdadeiras" ZAA.

3. Gerou-se o erro ε_i associado ao número de acidentes de cada local. Geralmente assume-se que $exp(\varepsilon_i)$ é independente e gamma distribuído com uma média 1 e variância $1/\phi$ para todos os i, em que o ϕ é o inverso do parâmetro de dispersão, e para o qual se considerou o valor da Tabela 2. Assim, para cada local, simulou-se os valores médios "observados" do número de acidentes, θ_i:

$$\theta_i = \mu_i \exp(\varepsilon_i) \quad (7)$$

4. Considerando o valor de θ_i de cada local, simulou-se número de acidentes para 5 anos - Y_i^{sim}. Assumiu-se para estes valores aleatórios simulados para cada local, a distribuição de Poisson:

$$Y_i^{sim} \Big| \theta_i \sim P_o(\theta_i) \quad (8)$$

Desta forma, geraram-se 5000 observações (5 anos de 1000 locais) que constituíram a base de dados considerada para a aplicação do método probabilístico e respetiva avaliação do desempenho.

4. APLICAÇÃO DO MÉTODO PROBABILISTICO – RESULTADOS

O modelo binário (MB) foi aplicado à base de dados gerada segundo o esquema descrito na secção 3.2. e considerando para as categorias (0/1) da variável dependente um valor fronteira definido pelo percentil 95 e que corresponde a 10 acidentes por interseção. Assim, aos locais com um número de acidentes igual ou superior a 10, foi atribuída a categoria 1 e classificados como ZAA, e aos restantes locais a categoria 0.

Considerando a distribuição probit, obtiveram-se os resultados para o MB apresentados na Tabela 3. Como se pode verificar, os parâmetros das variáveis independentes foram estimados com um nível de confiança de 95%. Além disso, os parâmetros estimados apresentam valores coerentes com as expectativas.

Tabela 3. Resultados do modelo binário probit para a base de dados gerada (1000 locais)

Parâmetro	*Valor estimado*	*Desvio padrão*	*P[Z>z]*
Constante	-7,269	0,507	0,0000
$Ln(TMDA_{Princ})$	0,439	0,048	0,0000
$Ln(TMDA_{Sec})$	0,111	0,010	0,0000
4Ramos	0,176	0,064	0,0059
Sinal. Luminosa	0,680	0,074	0,0000
% de observações corretamente previstas		94,4%	

A percentagem de observações corretamente previstas, isto é, observações de categoria 0 e 1 que foram previstas como sendo categorias 0 e 1, respectivamente, é de 94,4%. O gráfico que representa a curva ROC está representado na Figura 1. Como se pode observar a área sob a curva corresponde a um valor aproximado de 0,8 e como tal, claramente superior a 0,5.

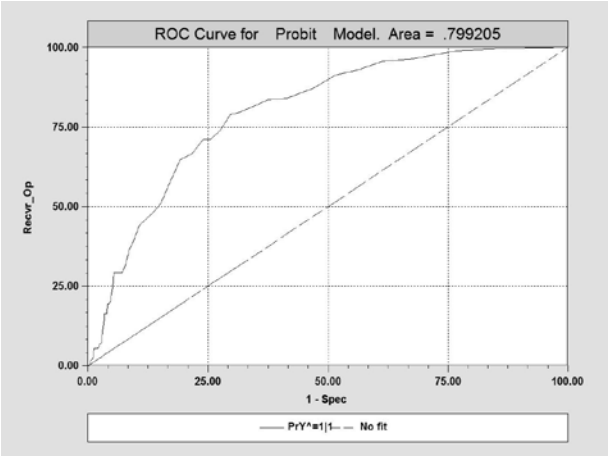

Figura 1. Gráfico da curva ROC do modelo binário probit

Destes resultados pode-se concluir que o MB apresenta um bom desempenho em termos de ajuste.

5. COMPARAÇÃO DO MODELO PROBABILÍSTICO COM OUTROS MIZAA

Para avaliar o desempenho do método probabilístico MB relativamente a outros MIZAA, aplicou-se, para a mesma base de dados gerada, o método de ranking com base no número de acidentes (MNA) e no número estimado de acidentes segundo a técnica Bayesiana-empírica (MBE). O primeiro é o método mais comumente aplicado e o segundo é o método identificado em vários trabalhos de investigação como o mais eficiente segundo diversos critérios de avaliação de desempenho (Cheng e Washington, 2008, Elvik, 2008, Montella, 2010).O ranking com base no MNA implica a disposição dos locais por ordem decrescente do número de acidentes observado no local. O MBE é calculado com base no histórico de acidentes, através do número de acidentes observado, e também com base nas características do local e os seus efeitos em locais similares através da seguinte formulação:

$$\lambda_i = wE[\lambda] + (1-w)x_i \tag{9}$$

em que ,

$E[\lambda]$: número de acidentes esperado;
x_i: número de acidentes observado no local i; e
w: fator de ponderação calculado segundo:

$$w = \frac{1}{1+E[\lambda]/k} \tag{10}$$

em que ,

k: parâmetro do modelo BN (também designado de inverso do parâmetro de dispersão).

Como se pode depreender pela equação (9) e (10) foi necessário para a aplicação do MBE aplicar um modelo BN à base de dados gerada de forma a determinar o número de acidentes esperado $E[\lambda]$. Para tal, considerou-se como variável dependente o número de acidentes observado e como variáveis independentes as mesmas referidas na descrição do modelo probabilístico binário na secção 2.

Com o objetivo de comparar o desempenho, em termos de capacidade de identificar corretamente a ZAA, do método probabilístico MB com o MNA e o MBE comumente aplicados, considerou-se critérios de avaliação baseados na classificação dos resultados segundo o tipo de erros – Tipo I (falsos positivos) e Tipo II (falsos negativos), e o número de locais detetados como ZAA ou não-ZAA (locais seguros). A Tabela 4 descreve a matriz que relaciona o número de locais identificados como ZAA ou não-ZAA por aplicação de um MIZAA com os locais "verdadeiramente" ZAA e não-ZAA e a respetiva classificação dos resultados.

A partir desta matriz de classificação dos resultados considerou-se o cálculo de cinco indicadores para avaliar e comparar o desempenho dos três métodos em identificar os "verdadeiros" locais ZAA. Os indicadores são:

Taxa de Falsos Identificados (TFI): corresponde à proporção de erros do Tipo I relativamente aos locais identificados como ZAA. O método com o menor TFI é considerado o melhor método:

$$TFI = \frac{V}{D} \tag{11}$$

Taxa de Falsos Negativos (TFN): corresponde à proporção de erros do Tipo II relativamente aos locais identificados como ZAA. Quanto menor o valor de TFN melhor o desempenho do método:

Tabela 4. Classificação dos locais identificados segundo um MIZAA

	Nº de locais identificados como não-ZAA	***Nº de locais identificados como ZAA***	
Nº de "verdadeiros" locais não-ZAA	U	V	n_0
Nº de "verdadeiros" locais ZAA	R	S	n_1
	n-D	D	n

em que:n	
N	Nº total de locais em análise
n_0	Nº de "verdadeiros" locais não-ZAA
n_1	Nº de "verdadeiros" locais ZAA
U	Nº de locais corretamente classificados como não-ZAA
V	Nº de falsos positivos ou erro Tipo I
R	Nº de falsos negativos ou erro Tipo II
S	Nº de locais corretamente classificados como ZAA
D	Nº de locais identificados como ZAA

Tabela 5. Resultados da classificação dos locais identificados segundo os três MIZAA

	Nº de locais identificados como não-ZAA			***Nº de locais identificados como ZAA***		
	MNA	MBE	MB	MNA	MBE	MB
Nº de "verdadeiros" locais não-ZAA	909	914	949	41	36	1
Nº de "verdadeiros" locais ZAA	41	36	1	9	14	49

$$TFN = \frac{R}{n - D} \tag{12}$$

Sensibilidade (SENS): é a proporção de locais que foram corretamente identificados como ZAA. Este critério epidemiológico é interpretado como a capacidade do método em identificar o "verdadeiro" local ZAA num conjunto de locais em análise. Este valor deve ser próximo de 1 se o método tiver um bom desempenho:

$$SENS = \frac{S}{n_1} \tag{13}$$

Especificidade (ESP): representa a proporção de locais não-ZAA que foram corretamente identificados como "verdadeiros" locais não-ZAA. Este critério epidemiológico avalia a capacidade do método em identificar os "verdadeiros" locais não-ZAA num conjunto de locais em análise. Este valor deve ser próximo de 1 se o método tiver um bom desempenho:

$$ESP = \frac{U}{n_0} \tag{14}$$

Risco (RISC): corresponde à proporção do total de erros (Tipo I e II) e o número de locais em análise. Quanto mais próximo de 0 for o valor do RISC, melhor é o desempenho do método:

$$RISC = \frac{(V + R)}{n} \tag{15}$$

Assim, aplicaram-se os três métodos à base de dados gerados (1000 locais) e dispuseram – se os locais por ordem decrescente dos valores considerando: número de acidentes resultante da média dos 5 anos para o MNA; número de acidentes calculado pela equação (9) e (10), considerando a média dos 5 anos para o número de acidentes observado x_i e o número estimado de acidentes $E[\lambda]$ considerando as características dos locais gerados e a função obtida pela aplicação do modelo BN à base de dados gerada, para o MBE; probabilidade do local ser ZAA (categoria 1) para o MB. Para que o número de locais selecionados como ZAA seja o mesmo na aplicação de qualquer um dos três MIZAA, considerou-se 5% dos locais, ou seja, 50 locais da lista dos 1000 locais, classificados como ZAA.

Os resultados da aplicação dos MIZAA estão apresentados na Tabela 5 elaborada segundo a Tabela 4.

A partir dos valores que constam na Tabela 5, calculou-se os cinco indicadores descritos anteriormente. Os valores relativos a esses indicadores estão apresentados na Tabela 6.

Tabela 6. Resultados dos indicadores de avaliação do desempenho dos três MIZAA

	MNA	***MBE***	***MB***
TFI	0,82	0,72	0,02
TFN	0,04	0,04	0,001
SENS	0,18	0,28	0,98
ESP	0,96	0,96	0,99
RISC	0,08	0,07	0,002

Como se pode verificar pela Tabela 6, todos os cinco indicadores indicam que o MB é o método com melhor desempenho em termos de identificação de ZAA com valores claramente melhores do que os outros dois métodos. O MBE é o segundo melhor método tendo em conta os valores da Tabela 6. É de salientar que a comparação entre o MB e os outros dois métodos é realizada entre indicadores diferentes: no caso do MNA e do MBE o indicador é o número de acidentes, observado e estimado, respectivamente; no caso do MB o indicador é uma probabilidade (valor entre 0 e 1). Este último indicador é, por este facto, mais abrangente pois calcula uma probabilidade de um determinado intervalo de frequência de acidentes ocorrer num local. De facto, ao analisar os valores

obtidos para os 50 locais selecionados como ZAAs segundo o MB, verifica-se que os valores das probabilidades não são muito elevados. Daqui concluiu-se que o modelo, apesar de atribuir valores de probabilidade relativamente baixos, numa comparação relativa entre os locais, tal como é o caso da seleção dos locais por ordem decrescente dos valores das probabilidades, não afeta o potencial do método tal como mostra a Tabela 6.

6. CONSIDERAÇÕES FINAIS

A identificação de ZAA e seu posterior tratamento através da implementação de medidas, eventualmente medidas de baixo custo, tem vindo a demonstrar em diversos países resultados muito eficientes na diminuição do número de acidentes e consequentemente melhorando a segurança rodoviária (Cardoso, 1998). Contudo, a decisão sobre qual método utilizar para a identificação das ZAA é um fator preponderante nos resultados finais em termos de eficácia do tratamento dos locais. Na verdade, um número elevado de acidentes num determinado local nem sempre representa um local perigoso, mas eventualmente uma situação aleatória normalmente explicada pelo fenómeno de regresso-à-média. Com o objetivo de considerar a possibilidade de ocorrer este fenómeno, foi desenvolvido o MBE que se baseia no número de acidentes observado mas também no número de acidentes esperado para um local com determinadas características. De facto, este método tem obtido bons resultados em termos de desempenho na identificação de ZAA em detrimento do método mais utilizado baseado apenas no número de acidentes observado. Neste contexto, desenvolveu-se um novo método baseado no cálculo da probabilidade de um local ser ZAA ou não-ZAA. Através de um modelo de regressão binário, as características do local são consideradas no cálculo da probabilidade do local ser ZAA ou não-ZAA, mitigando o fenómeno de regresso-à-média. Por outro lado, o método MB assume o grau de incerteza associado a qualquer MIZAA ao considerar uma probabilidade, o que permite também gerir a seleção dos locais a tratar de uma forma mais eficiente. Assim, quanto maior a probabilidade de um local ser ZAA tendo em conta as características do mesmo, maior a probabilidade de uma medida de tratamento reduzir os acidentes.

Para analisar o desempenho do MB considerou-se os erros na classificação dos locais como ZAA ou não-ZAA. Para tal, foi necessário identificar os locais "verdadeiramente" ZAA ou não-ZAA: Nesse sentido, utilizou-se uma base de dados simulada de forma a possibilitar o conhecimento *a priori* das "verdadeiras" ZAA. A esta base de dados simulada aplicou-se o MB bem como o MNA e o MBE, tendo-se verificado, com base nos erros de classificação dos locais e no cálculo de cinco indicadores de desempenho, que o MB tem claramente melhor desempenho.

Torna-se assim evidente o potencial do MB sendo, como tal, interessante avaliar empiricamente o seu desempenho conjuntamente com a aplicação de medidas de tratamento e correspondente eficiência. De notar que o trabalho foi desenvolvido para o caso das interseções, sendo, no entanto, extensível com eventuais adaptações ao caso dos segmentos.

Contudo, é de salientar que, tal como sugerido pelo *Highway Safety Manual* (AASHTO, 2010), é aconselhável aplicar sempre mais do que um método, pois todos têm vantagens e desvantagens, sendo que a escolha de qual método a considerar depende de vários fatores tais como dados disponíveis e objetivo final do estudo.

REFERÊNCIAS BIBLIOGRÁFICAS

AASHTO (2010) *Highway Safety Manual. American Association of State Highway and Transportation Officials*, Washington DC.
ISBN: 978-1-56051-477-0

Cafiso, S. e G. D. Silvestro (2011) Performance of Safety Indicators in Identification of Black Spots on Two-Lane Rural Roads. *Transportation Research Record: Journal of the Transportation Research Board*, n. 2237, p. 78–87. DOI: 10.3141/2237-09

Cardoso, J. L. (1998) Definição e detecção de zonas de acumulação de acidentes. *Segurança Rodoviária - Avaliação e Redução da Sinistralidade*, LNEC, Lisboa, p. 131–133.

Carlin, B. e T. Louis (2000) *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall, London. ISBN: 1439840954.

Cheng, W. e S. Washington (2008) New Criteria for Evaluating Methods of Identifying Hot Spots. *Transportation Research Record: Journal of the Transportation Research Board*, n. 2083, p. 76–85. DOI: 10.3141/2083-09

Cheng, W. e S. P. Washington (2005) Experimental evaluation of hotspot identification methods. *Accident Analysis & Prevention*, v. 37, n. 5, p. 870–881. DOI: 10.1016/j.aap.2005.04.015

Elvik, R. (2008) Comparative Analysis of Techniques for Identifying Locations of Hazardous Roads. *Transportation Research Record: Journal of the Transportation Research Board*, n. 2083, p. 72–75. DOI: 10.3141/2083-08

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

Fridstrom, L., J. Ifver, S. Ingebrigtsen, R. Kulmala e L. K. Thomsen (1995) Measuring the contribution of randomness, exposure, weather and dayligth to the variation in road accident counts. *Accident Analysis and Prevention*, v. 27, p. 1–20. DOI: 10.1016/0001-4575(94)E0023-E

Geedipally, S. R. e D. Lord (2010) Identifying Hot Spots by Modeling Single-Vehicle and Multivehicle Crashes Separately. *Transportation Research Record: Journal of the Transportation Research Board*, n. 2147, p. 97–104. DOI: 10.3141/2147-12.

Greene, W. H. (2008) *Econometric Analysis.* Sixth Edition. Pearson International Edition, New Jersey. ISBN-13:978-0-13-513740-6; ISBN-10:0-13-513740-3.

Greibe, P. (2003) Accident prediction models for urban roads. *Accident Analysis and Prevention*, v. 35, p. 273–285. DOI: 10.1016/S0001-4575(02)00005-2.

Huang, H. e H. C. C. M. M. Haque (2009) Empirical Evaluation of Alternative Approaches in Identifying Crash Hot Spots. Naive Ranking, Empirical Bayes, and Full Bayes Methods. *Transportation Research Record: Journal of the Transportation Research Board*, n. 2103, p. 32–41. DOI: 10.3141/2103-05.

Lan, B. e B. Persaud (2011) Fully Bayesian Approach to Investigate and Evaluate Ranking Criteria for BlackSpot Identification. *Transportation Research Record: Journal of the Transportation Research Board*, n. 2237, p. 117–125. DOI: 10.3141/2237-13.

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. Disponívelem: <http://hdl.handle.net/1807/14514>.

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*, v. 38, p. 751–766. DOI: 10.1016/j.aap.2006.02.001.

Lord, D. e B. N. Persaud (2004) Estimating the safety performance of urban road transportation networks. *Accident Analysis and Prevention*, v. 36, p. 609–620. DOI: 10.1016/S0001-4575(03)00069-1.

Miranda-Moreno, L. F., L. Fu, F. F. Saccomanno e A. Labbe (2005) Alternative Risk Models for Ranking Locations for Safety Improvement. *Transportation Research Record: Journal of the Transportation Research Board*, n. 1908, p. 1–8. DOI: 10.3141/1908-01.

Montella, A. (2010) A comparative analysis of hotspot identification methods. *Accident Analysis and Prevention*, v. 42, p. 571–581. DOI: 10.1016/j.aap.2009.09.025.

OCDE (1997) Road safety principles and models: review of descriptive, predictive, risk and accident consequence models. *Organisation*

for Economic Co-operation and Development, Paris. Disponível em: <http://www.oecd.org/sti/transport/roadtransportresearch/2103285.pdf>.

Schluter, P. J., J. J. Deely e A. J. Nicholson (1997) Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. *The Statistician*, v. 46, p. 293–316. DOI: 10.1111/1467-9884.00084.