Accident Analysis and Prevention 42 (2010) 571–581

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

A comparative analysis of hotspot identification methods

Alfonso Montella*

University of Naples Federico II, Department of Transportation Engineering "Luigi Tocchetti", Via Claudio 21, 80125 Naples, Italy

ARTICLE INFO

Article history:
Received 28 May 2009
Received in revised form
26 September 2009
Accepted 29 September 2009

Keywords:
Hotspots
Ranking criteria
Empirical Bayes method
Crash prediction models
Quantitative evaluation criteria
Engineering countermeasures

ABSTRACT

The identification of crash hotspots is the first step of the highway safety management process. Errors in hotspot identification may result in the inefficient use of resources for safety improvements and may reduce the global effectiveness of the safety management process. Despite the importance of using effective hotspot identification (HSID) methods, only a few researchers have compared the performance of various methods. In this research, seven commonly applied HSID methods were compared against four robust and informative quantitative evaluation criteria. The following HSID methods were compared: crash frequency (CF), equivalent property damage only (EPDO) crash frequency, crash rate (CR), proportion method (P), empirical Bayes estimate of total-crash frequency (EB), empirical Bayes estimate of severe-crash frequency (EBs), and potential for improvement (PFI). The HSID methods were compared using the site consistency test, the method consistency test, the total rank differences test, and the total score test. These tests evaluate each HSID method's performance in a variety of areas, such as efficiency in identifying sites that show consistently poor safety performance, reliability in identifying the same hotspots in subsequent time periods, and ranking consistency. To evaluate the HSID methods, five years of crash data from the Italian motorway A16 were used.

The quantitative evaluation tests showed that the EB method performs better than the other HSID methods. Test results highlight that the EB method is the most consistent and reliable method for identifying priority investigation locations. The EB expected frequency of total-crashes (EB) performed better than the EB expected frequency of severe-crashes (EBs), although the results differed only slightly when the number of identified hotspots increased. The CF method performed better than other HSID methods with more appealing theoretical arguments. In particular, the CF method performed better than the CR method. This result is quite alarming, since many agencies use the CR method. The PFI and EPDO methods were largely inconsistent. The proportion method performed worst in all of the tests. Overall, these results are consistent with the results of previous studies.

The identification of engineering countermeasures that may reduce crashes was successful in all of the hotspots identified with the EB method; this finding shows that the identified hotspots can also be corrected.

The advantages associated with the EB method were based on crash data from one Italian motorway, and the relative performances of HSID methods may change when using other crash data. However, the study results are very significant and are consistent with earlier findings. To further clarify the benefits of the EB method, this study should be replicated in other countries. Nevertheless, the study results, combined with previous research results, strongly suggest that the EB method should be the standard in the identification of hotspots.

1. Introduction

The identification of crash hotspots, also referred to as hazardous road locations, high-risk locations, accident-prone locations, black spots, sites with promise, or priority investigation locations, is the first step of the highway safety management process. Crash hotspot identification results in a list of sites that are prioritized for detailed engineering studies that can identify crash patterns, contributing factors, and potential countermeasures (Hauer et al., 2002a, 2004). The most cost effective projects are selected in order to ensure that the best use is made of the limited funds available (Montella, 2001, 2005). A crash hotspot can be theoretically defined as any location that has a higher number of crashes than other similar locations as a result of local risk factors (Elvik, 2007). This definition implies that true crash hotspots are sites at which local risk factors related to road design and/or traffic control make a substantial contribution to crashes; as a result, engineering improvements can reduce crashes. Errors in hotspot identification

* Tel.: +39 0817683941; fax: +39 0817683946.
E-mail address: alfonso.montella@unina.it.

0001-4575/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2009.09.025

may produce large numbers of false negatives (i.e., truly hazardous sites mistakenly designated as safe) and large numbers of false positives (i.e., truly safe sites that are wrongly identified as hazardous). These errors result in an inefficient use of the resources applied to safety improvements and reduce the global effectiveness of the safety management process. Therefore, the correct identification of crash hotspots is essential for the successful implementation of any highway safety plan.

Highway agencies use different hotspot identification (HSID) methods. The HSID methods most commonly used in practice are the ranking of crash frequency and the ranking of crash rate (DTLR, 2001; Persaud, 2001; PIARC, 2004; ROSPA, 2002; TAC, 2004; Tarko and Kanodia, 2004). To take crash costs into account, the equivalent property damage only (EPDO) crash frequency method is sometimes used (PIARC, 2004; TAC, 2004). Based on the assumption that only excess crashes over those expected from similar sites can be prevented by applying appropriate treatments, the potential for improvement, measured as the difference between the observed (or estimated with the empirical Bayes method) and the expected crash frequency, is also used as an HSID method (Austroads, 2009; El-Basyouny and Sayed, 2006; Persaud et al., 1999; PIARC, 2004; ROSPA, 2002). Since the existence of crash patterns susceptible to correction may be accompanied by an overrepresentation of crash frequency (Kononov, 2002), the proportion method is sometimes used; this method prioritizes sites according to the probability that the observed proportion of a crash type at a site is truly above a given critical proportion (Lyon et al., 2007; PIARC, 2004). Crash count does not always give an unbiased estimate of the long-term expected number of crashes, because crash counts can randomly fluctuate during the observation period. This observation has generated interest in techniques that control for random fluctuations in the recorded number of crashes; the empirical Bayes (EB) technique is one example (Hauer, 1997; Hauer et al., 2002b; Persaud and Lyon, 2007).

Compared with the large number of studies focused on the development of various HSID methods, considerably less research has been dedicated to comparing the performance of various methods (Cheng and Washington, 2005). Central to the comparison of HSID methods is the identification and development of robust and informative quantitative and qualitative criteria that can be used to evaluate the methods. Like the selection of statistical and econometric models, numerous assessment criteria are needed to assess HSID methods. Cheng and Washington (2008) proposed the use of five (four new) quantitative evaluation tests aimed at comparing alternative HSID methods. These five tests evaluate a variety of aspects of each HSID method's performance. The site consistency test measures each method's efficiency in identifying sites that show consistently poor safety performance. The method consistency test and the total rank differences test measure the reliability and consistency of HSID methods in terms of the number of hotspots with consistent underlying safety problems that are identified in a relatively short period of time. The false identification test and the Poisson mean differences test measure each method's performance in terms of the false identification of hotspots and the corresponding consequences that arise from erroneous identifications. The latter two tests require that truth be known *a priori*; they are useful only when HSID methods are being compared in a simulated environment, where the truth is known by design. Cheng and Washington demonstrated the new criteria using three years of crash data collected in Arizona and four HSID methods: the ranking of crash frequency, the ranking of crash rate, the ranking of potential for improvement, and the EB method. The EB HSID method was the superior method in most of the evaluation tests. In contrast, the crash rate method consistently performed the worst.

Elvik (2007, 2008a) used data from Norwegian roads to compare five HSID techniques: crash frequency, crash rate, combination of crash frequency and crash rate, EB crash estimate, and contribution of local risk factors to the EB estimate (equivalent to the potential for improvement method). The diagnostic performance of the five techniques was assessed in terms of epidemiological criteria (sensitivity and specificity). Confirming the results of Cheng and Washington (2008), the EB technique performed the best and the crash rate method performed the worst.

Persaud et al. (1999) compared the crash frequency method, the crash rate method, the PFI method, and the EB method; they used both the crash frequency and the difference between observed and predicted crashes in a subsequent time period as measures of effectiveness. The results differed in relation to the measure of effectiveness.

Despite recent research showing that the EB approach is the best available HSID method, most highway agencies continue to use less effective methods. For example, Elvik (2007, 2008b) surveyed HSID methods in eight European countries and found that most of the approaches are primitive and are likely to involve substantial inaccuracies. The European Parliament's (2008) directive on road infrastructure safety management does not make any reference to the EB method. Furthermore, existing highway safety manuals (e.g., Austroads, 2009; DTLR, 2001; PIARC, 2004) do not adequately highlight the problems associated with the use of older HSID methods instead of the EB method. In addition, the forthcoming Highway Safety Manual of the Transportation Research Board will suggest that many HSID methods can be used; practitioners who are skeptical of new techniques may misinterpret this information. There is an urgent need for research studies that use robust and quantitative evaluation criteria to compare HSID methods.

The aim of this paper is to compare of a set of commonly applied HSID methods against quantitative evaluation criteria. Five years of crash data from the Italian motorway A16 were used in this comparison. Furthermore, in order to test whether the best method identifies correctible hotspot sites, diagnostic evaluations were performed and the existence of potential effective safety treatments was verified. The remainder of the paper is organized as follows: Section 2 describes the HSID methods that were evaluated; Section 3 explains the analytics of the quantitative evaluation criteria; Section 4 describes the case study data; Section 5 describes the crash prediction models used in the evaluations; Section 6 reports and discusses the results of the comparison of HSID methods according to the evaluation criteria; Section 7 describes the diagnostic evaluations performed in the hotspots identified with the selected best method; and finally, conclusions are drawn in Section 8.

2. Alternative hotspot identification methods used in comparison

The following HSID methods were compared: the ranking of crash frequency (CF), the ranking of equivalent property damage only (EPDO) crash frequency, the ranking of crash rate (CR), the proportion method (P), the ranking of EB estimate of total-crash frequency (EB), the ranking of EB estimate of severe-crash frequency (EBs), and the potential for improvement method (PFI). Each of these HSID methods are described below. Since the methods were compared with reference to highway segments, only explanations relative to the segments are given.

2.1. Crash frequency

Crash frequency is the simplest identification criterion. Applying this method, sites are ranked in descending order of observed crash frequencies. In order to compare segments of different lengths, the total number of crashes is divided by the segment length and the time period. The result is a value showing crashes per year per kilometer.

2.2. Equivalent property damage only crash frequency

The EPDO crash frequency measure weights crashes according to severity (fatal, injury, and property damage only) to develop a combined frequency and severity score for each site. The weighting factors are based on property damage only (PDO) crash costs. An EPDO value summarizes the crash costs and severity.

In the calculations, weighting factors were assessed from the crash cost estimates developed by the UK Department for Transport (2007) based on the willingness to pay approach. Using average crash costs for motorways, fatal crashes (1,751,150 £) have a weight factor equal to 771, injury crashes (78,930 £) have a weight factor equal to 35, and PDO crashes (2,270 £) have a weight factor equal to 1.

2.3. Crash rate

The crash rate normalizes the frequency of crashes with exposure (measured by traffic volume). Roadway segment traffic volume is measured as vehicle-kilometers traveled for the study period. This method reflects crash risk for the individual road user.

2.4. Proportion method

This method is frequently applied as a diagnostic tool to identify crash patterns; it is sometimes applied as an HSID method. Sites are prioritized based on the probability that the proportion of a particular crash type is greater than the threshold proportion. A threshold proportion is selected based on the proportion of the target crash type in the reference population. Since each crash can be viewed as an independent Bernoulli trial with the probability of a particular crash type equal to the proportion of the crash type in the comparison group (p), the probability of observing less than x crashes out of n trials can be computed as follows:

$$P(X \leq x-1, n; p) = B(x-1, n; p) = \sum_{i=0}^{x-1} \frac{n!}{(n-i)! \times i!} \times p^{i} \times (1-p)^{n-i} \tag{1}$$

In the calculations, the following crash types were considered: run-off-the-road (ROR), rear-end, nighttime, wet road, and rainy. Since crash patterns are different in the tangents and in the curves, reference groups were different for the segments located in the curves and in the tangents.

2.5. EB method

In the EB procedure, a crash prediction model (CPM) is used to first estimate the number of crashes that would be expected in the analysis period at locations with traffic volumes and other characteristics similar to the one being analyzed. The crash estimates are then combined with the count of crashes to obtain a better estimate of the expected number of crashes according to the following equations:

$$EB = w \times \hat{E}(Y) + (1-w) \times \text{count} \tag{2}$$

$$w = \frac{1}{1 + (\hat{E}(Y)/k)} \tag{3}$$

where k is the negative binomial parameter (also called the inverse dispersion parameter), constant for a given CPM, and is estimated from the CPM calibration process with the use of a maximum likelihood procedure; count is the observed crash frequency; and $\hat{E}(Y)$ is the predicted crash frequency.

The expected number of crashes is divided by the segment length and the analysis period length to obtain the crashes per year per kilometer.

2.6. EBs method

In this method, the EB expected frequency of severe crashes (fatal plus all injury) is used.

2.7. Potential for improvement method

The potential for improvement (PFI) is the difference between the EB expected crash frequency and the crash frequency predicted from a CPM, as follows:

$$PFI = EB - \hat{E}(Y) = w \times \hat{E}(Y) + (1-w) \times \text{count} - \hat{E}(Y) \tag{4}$$

When the potential for improvement value is greater than zero, a site experiences more crashes than expected. When the potential for improvement value is less than zero, a site experiences fewer crashes than expected.

3. Hotspot identification methods evaluation criteria

The various HSID methods were compared using four quantitative evaluation tests: the site consistency test, the method consistency test, the total rank differences test, and the total score test. The first three tests have been recently introduced by Cheng and Washington (2008); the fourth test is new and combines the results of the previous tests in order to give a synthetic and easily readable index. Each of these tests are described below. Since the methods are not being compared in a simulated environment, there is no known set of correct hotspots. Thus, the false identification test was not performed.

3.1. Site consistency test

The site consistency test (SCT) measures the ability of an HSID method to consistently identify a high-risk site over repeated observation periods. The test rests on the premise that a site identified as high risk during time period i should also reveal an inferior safety performance in a subsequent time period $i+1$, given that the site is in fact high risk and no significant changes have occurred at the site. The method that identifies sites in a future period with the highest crash frequency is the most consistent. The test statistic is given as:

$$SCT_j = \frac{\sum_{k=n-n\alpha+1}^{n} C_{k,j,i+1}}{\sum_{k=n-n\alpha+1}^{n} L_{k,j} \times y_{i+1}} \tag{5}$$

where j is the HSID method being compared, n is the total number of sites, α is the threshold of identified hotspots (e.g., $\alpha = 0.01$ corresponds with top 1% of n sites identified as hotspots, and $n\alpha$ is the number of identified hotspots), C is the crash count for site ranked k in the time period $i+1$, L is the length of the site ranked k in the time period $i+1$ (km), and y_{i+1} is the length of the time period $i+1$ (years).

3.2. Method consistency test

The method consistency test (MCT) evaluates a method's performance by measuring the number of the same hotspots identified in both time periods. It is assumed that road sections are in the same or similar underlying operational state and their expected safety performance remains virtually unaltered over the two analysis periods. With this assumption of homogeneity, the greater the number of hotspots identified in both periods the more consistent the performance of the HSID method. The test statistic is given as:

$$\mathrm{MCT}_j = \left\{k_{n-na+1}, k_{n-na}, \ldots, k_n\right\}_{j,i} \cap \times \left\{k_{n-na+1}, k_{n-na}, \ldots, k_n\right\}_{j,i+1} \quad (6)$$

3.3. Total rank differences test

The total rank differences test (TRDT) takes into account the safety performance rankings of the road sections in the two periods. The test is conducted by calculating the sum of the total rank differences of the hotspots identified across the two periods. The smaller the total rank difference, the more consistent the HSID method. The test statistic is given as:

$$\mathrm{TRDT}_j = \sum_{k=n-n\alpha+1}^{n} \left|R(k_{j,i}) - R(k_{j,i+1})\right| \quad (7)$$

3.4. Total score test

The total score test (TST) combines the site consistency test, the method consistency test, and the total rank difference test in order to provide a synthetic index. The test statistic is given as:

$$\mathrm{TST}_j = \frac{100}{3} \times \left[\left(\frac{\mathrm{SCT}_j}{\max_j \mathrm{SCT}}\right) + \left(\frac{\mathrm{MCT}_j}{\max_j \mathrm{MCT}}\right) + \left(1 - \frac{\mathrm{TRDT}_j - \min_j \mathrm{TRDT}}{\max_j \mathrm{TRDT}}\right)\right] \quad (8)$$

The test assumes that the SCT, MCT, and TRDT have the same weight. The former three tests provide absolute measures of effectiveness, whereas the total score test gives an effectiveness measure relative to the methods being compared. If method j performed best in all of the previous tests, the TST value is equal to 100. If method j performed worst in all of the tests, the TST value is positive since all three components of the test have a positive value. Indeed, SCT and MCT, which should be maximized by the HSID methods, are weighted in relation to the maximum values in the tests, whereas TRDT, which should minimized by the HSID methods, is weighted in relation to its difference from the minimum value in the test.

4. Case study data

Geometric data, traffic volumes, and crash records were collected for motorway A16 between Naples and Candela in Italy (Montella et al., 2008). Horizontal alignment characteristics and traffic flow volumes were used to divide the study section into 646 homogeneous segments (343 for each direction of the motorway), with a mean length of 395 meters. Crash data from 2001–2005 were collected by analyzing police reports and were integrated with detailed site inspections. In the analysis period, 2245 crashes occurred. The number of severe crashes (fatal plus all injury) was 728.

To compare the HSID methods, the five-year crash data were separated into two time periods, Period 1 (2001 and 2002) and Period 2 (2003, 2004, and 2005). In the study period, horizontal curve delineation improvements have been carried out for 15 curves (Montella, 2009). These sites were excluded from the comparison because the evaluation tests are based on the hypothesis that no significant changes have occurred at the sites.

5. Crash prediction models

The EB, EBs, and PFI methods require the use of pertinent crash prediction models. For the purpose of the study, a refinement of the CPMs developed by Montella et al. (2008) was used.

5.1. Model description

Generalized linear modeling techniques were used to fit the models, and a negative binomial distribution error structure was assumed. The selected model form is as follows:

$$\hat{E}(Y) = L \times e^{a_0 + a_1 \times \ln(\mathrm{AADT})} \times e^{\sum_{i=1}^{n} b_i \times x_i} \quad (9)$$

where $\hat{E}(Y)$ is the predicted crash frequency, L is the segment length (m), AADT is the average annual daily traffic (vpd) on the segment, a_i and b_i are the model parameters, and x_i are the explanatory variables in addition to L and AADT.

In the model, segment length is an offset variable. It captures the logical requirement that if N crashes are expected to occur on 1 km of road, $2N$ crashes should be expected to occur on an identical road that is 2 km long (Hauer, 2004).

Explanatory variables related to traffic composition, horizontal alignment, design consistency, vertical alignment, sight distance, roadside context, cross-section, and yearly effects were considered. Table 1 shows descriptions and analytical details of the variables included in the final models.

The model parameters and the dispersion parameter of the negative binomial distribution were estimated by the maximum

Table 1
CPMs: Significant explanatory variables.

Variable	Description
Traffic volume and composition	
ln(AADT)[vpd]	Natural logarithm of the average annual daily traffic
P_{hv} [%]	Percentage of heavy vehicles
Horizontal alignment	
$1/R^2$ [1/km^2]	Square of horizontal curvature
Def [gon]	Deflection angle
Design consistency[a]	
ΔV_{85} [km/h][b]	85th percentile of speed reduction
Δf_r[c]	Difference between assumed and demanded side friction
L_t [km]	Length of the tangent preceding the curve
Vertical alignment	
G_u [%][d]	Equivalent upgrade
Roadside context	
Bridge	Binary variable, equal to 1 if the segment is on a bridge
Cross-section	
Median	Binary variable, equal to 1 if a concrete barrier is present in the median (0 if a steel barrier is present)
Yearly effects	
Yr02	Binary variable, equal to 1 in the year 2002

[a] Evaluated only on curves.
[b] $V_{85} = 155 - \frac{1352}{R} - 0.42 \times \frac{\mathrm{def}_2}{2} - 4.1 \times |g|$ where def_2 is the total deflection angle in the 2 km before the end of the segment (gon), and $|g|$ is the absolute value of the longitudinal grade (%).
[c] $\Delta f_r = f_{ra} - f_{rd} = 0.6 \times 0.925 \times (0.59 - 4.85 \times 10^{-3} \times V_d + 1.51 \times 10^{-5} \times V_d^2) - \left(\frac{V_{85}^2}{127 \times R} - e\right)$ where f_{ra} is the side friction assumed with respect to the design speed, f_{rd} is the side friction demanded at the 85th percentile speed, R is the radius of horizontal curve (m), and e is the curve superelevation.
[d] $G_u = \frac{\sum_i G_{iu} \times L_{iu}}{\sum_i L_i}$ where G_u is the equivalent upgrade, G_{iu} is the grade of the upgrade sub-segment i, L_{iu} is the length of the upgrade sub-segment i, and L_i is the length of the sub-segments (both upgrade and downgrade) which make up the segment under consideration.

likelihood method using the GENMOD procedure in SAS. Separate models were developed for total-crashes and severe (fatal plus all injury) crashes. The models were developed by the stepwise forward procedure. Variables were kept in the model based on two criteria. The first was whether the *t*-ratio of the variable's estimated coefficient was significant at the 5% level. The second criterion was related to improvement in the goodness-of-fit measures of the model including the variable.

5.2. Measuring goodness-of-fit

Two goodness-of-fit measures were used: R^2_α (Miaou, 1996) and the Akaike information criterion (AIC) (Akaike, 1987).

R^2_α, a dispersion parameter-based R^2, is calculated as follows:

$$R^2_\alpha = 1 - \left(\frac{k_{\min}}{k}\right) \tag{10}$$

where $k_{\min}$ is the smallest possible inverse dispersion parameter that is obtained by having no covariates in the model (by assuming that all sites have an identical prediction estimate equal to the mean over all sites) and k is the inverse dispersion parameter for the calibrated model. This measure is bound between 0 (when no covariate is included) and 1 (when covariates are perfectly specified).

The AIC value is calculated as follows:

$$\mathrm{AIC} = -2 \times \mathrm{ML} + 2 \times p \tag{11}$$

where ML is the maximum log-likelihood of the fitted model, and p is the number of parameters in the model. A smaller AIC value reflects a better model. The first term in the AIC equation measures the badness-of-fit, or bias, when the maximum likelihood estimates of the parameters are used. The second term measures the complexity of the model, thus penalizing the model for using more parameters. The goal for selecting the best model is to choose the best fit with the least complexity.

Table 2
CPMs: parameter estimates and goodness-of-fit measures.

Variable	Total-crashes		Severe-crashes	
	Estimate	Standard deviation	Estimate	Standard deviation
$1/k$	0.9192	0.0636	1.2752	0.1221
Constant	−12.7360	1.0580	−16.7040	2.7594
Traffic volume and composition				
ln(AADT)[vpd]	0.6511	0.1143	0.9738	0.2844
P_{hv} [%]			−0.0305	0.0148
Horizontal alignment				
$1/R^2$ [1/km^2]	0.1290	0.0185	0.1236	0.0188
Def [gon]	−0.0112	0.0019	−0.0099	0.0023
Design consistency				
ΔV_{85} [km/h]	0.0323	0.0071	0.0514	0.0088
Δf_r	−4.3016	0.7402	−4.6086	0.8565
L_t [km]	0.2974	0.0660	0.2996	0.0802
Vertical alignment				
G_u [%]	−0.0745	0.0166	−0.1046	0.0218
Roadside context				
Bridge			−0.3869	0.1526
Cross-section				
Median			0.1805	0.0879
Yearly effects				
Yr02	0.1960	0.0639	0.3039	0.0813
R^2_α	0.562		0.663	
AIC	4739.2		5474.8	

5.3. Modeling results

Table 2 presents parameter estimates and goodness-of-fit measures for both the total-crashes and severe-crashes models. Non-significant parameters are not reported. All the parame-

Table 3
Test results: top 1% of hotspots.

HSID method		Site consistency test [crashes/(km × year)]	Method consistency test	Total rank differences test	Total score test
EB	Test value	11.1[a]	2 (33.3%)	137	98.3
	Test ranking	*1*	*1*	*2*	*1*
CF	Test value	11.0	2 (33.3%)	497	88.2
	Test ranking	*2*	*1*	*4*	*2*
CR	Test value	10.5	2 (33.3%)	618	83.3
	Test ranking	*3*	*1*	*6*	*3*
EB_s	Test value	9.0[b]	1 (16.7%)	76	77.0
	Test ranking	*4*	*6*	*1*	*4*
PFI	Test value	8.2	2 (33.3%)	877	69.5
	Test ranking	*5*	*1*	*10*	*5*
EPDO	Test value	7.6	1 (16.7%)	792	53.2
	Test ranking	*6*	*6*	*7*	6
$\text{Proportion}_{\text{ROR}}$	Test value	2.9	2 (33.3%)	872	53.4
	Test ranking	*8*	*1*	*9*	*7*
$\text{Proportion}_{\text{Rear-end}}$	Test value	2.5	1 (16.7%)	499	46.1
	Test ranking	*9*	*6*	*5*	*8*
$\text{Proportion}_{\text{Rainy}}$	Test value	2.5	0	415	31.7
	Test ranking	*9*	*10*	*3*	*9*
$\text{Proportion}_{\text{Wet}}$	Test value	4.6	0	870	25.3
	Test ranking	*7*	*10*	*8*	*10*
$\text{Proportion}_{\text{Night}}$	Test value	2.5	1 (16.7%)	1213	26.4
	Test ranking	*9*	*6*	*11*	*11*

The percentage shown in parenthesis stands for the percentage of hotspots identified in period 1 that were also identified in period 2.

[a] The modified site consistency test is equal to 2.43 severe crashes/(km × year).

[b] The modified site consistency test is equal to 2.90 severe crashes/(km × year).

ters have a logical and expected sign. The difference from the previous models developed by Montella et al. (2008) was the inclusion of the variable $1/R^2$ (square of horizontal curvature) as a substitute for the variable $1/R$. The new models had lower AIC values and greater R_α^2 values. For the total-crashes model, the AIC and R_α^2 values were 4739.2 and 56.2%, respectively. For the severe-crashes model, the AIC and R_α^2 values were 5474.8 and 66.3%, respectively. Overall, the models showed a reasonable goodness-of-fit.

6. Test results and discussion

The four quantitative tests described in Section 3 were used to assess the relative performance of the seven commonly applied HSID methods described in Section 2. Segments with the highest rankings were flagged as hotspots. In this evaluation, the top 1% (see Table 3), 5% (see Table 4), and 10% (see Table 5) sites were selected as hotspots.

6.1. Site consistency test results

The EB method outperformed other HSID methods in identifying the top 1%, 5%, and 10% of hotspots with the highest crash frequency in period 2, equal respectively to 11.1, 5.2, and 4.3 crashes per km per year. In contrast, the proportion method consistently performed the worst. In period 2, the top 1%, 5%, and 10% of hotspots identified with the EB method showed crash frequencies that were 334%, 158%, and 140% greater than the crash frequencies of the hotspots identified when the proportion method was applied to nighttime crashes. That is, the proportion method is highly ineffective at identifying sites that will exhibit higher crash frequencies in the future. On the site consistency test, the EB method was followed by the following methods: CF, CR, EBs, PFI, and EPDO. The EPDO method performed better than the PFI and EBs methods only with reference to the top 10% of hotspots.

The EBs method is aimed at identifying sites with a high frequency of severe-crashes and it is therefore not expected to perform well in identifying sites with a high frequency of total-crashes. The comparison between the EB method and the EBs method was thus also performed with reference to the modified site consistency test, in which the total-crashes count (Eq. (5)) was substituted with the severe-crashes count. Surprisingly, the EB method (except the top 1% of sites) was more effective than the EBs method in identifying sites that will exhibit a higher frequency of severe-crashes in the future.

6.2. Method consistency test results

The EB method was superior in this test by identifying the largest number of the same hotspots (2 and 15, respectively) in both the top 1% and 5% of sites. In the case of the top 10% of sites, the EBs method performed slightly better than the EB method, identifying 33 sites (vs. 31) in period 1 that were also identified as hotspots in period 2. In this case, the CF and EB methods gave the same results. The CR, PFI, and EPDO methods performed significantly worse than the EB method. The proportion method performed the worst.

6.3. Total rank differences test results

In this test, the EBs method performed the best and was closely followed by the EB method. The difference between the EB method (referring to both total and severe-crashes) and the other HSID methods was very impressive and was consistent with the results reported by Cheng and Washington (2008). Indeed, the other methods showed significantly greater summed ranked differences. In the

Table 4
Test results: top 5% of hotspots.

HSID method		Site consistency test [crashes/(km × year)]	Method consistency test	Total rank differences test	Total score test
EB	Test value	5.2[a]	15 (46.9%)	2180	98.1
	Test ranking	*1*	*1*	*2*	*1*
CF	Test value	4.9	13 (40.6%)	3965	83.5
	Test ranking	*2*	*2*	*3*	*2*
EB_s	Test value	3.9[b]	11 (34.4%)	1772	82.7
	Test ranking	*4*	*3*	*1*	*3*
CR	Test value	4.7	11 (34.4%)	4016	77.1
	Test ranking	*3*	*3*	*4*	*4*
PFI	Test value	3.5	8 (25.0%)	6745	50.4
	Test ranking	*5*	*5*	*9*	*5*
EPDO	Test value	3.5	4 (12.5%)	5810	45.3
	Test ranking	*6*	*7*	*6*	*6*
$\text{Proportion}_{\text{ROR}}$	Test value	1.9	5 (12.5%)	5302	39.7
	Test ranking	*10*	*6*	*5*	*7*
$\text{Proportion}_{\text{Rear-end}}$	Test value	2.0	4 (15.6%)	5934	35.4
	Test ranking	*7*	*8*	*7*	*8*
$\text{Proportion}_{\text{Night}}$	Test value	2.0	4 (12.5%)	6220	34.0
	Test ranking	*7*	*8*	*8*	*9*
$\text{Proportion}_{\text{Rainy}}$	Test value	2.0	4 (12.5%)	6925	30.7
	Test ranking	*7*	*8*	*10*	*10*
$\text{Proportion}_{\text{Wet}}$	Test value	1.2	4 (12.5%)	7029	25.3
	Test ranking	*11*	*8*	*11*	*11*

The percentage shown in parenthesis stands for the percentage of hotspots identified in period 1 that were also identified in period 2.

[a] The modified site consistency test is equal to 1.37 severe crashes/(km × year).

[b] The modified site consistency test is equal to 1.06 severe crashes/(km × year).

Table 5
Test results: top 10% of hotspots.

HSID method		Site consistency test [crashes/(km × year)]	Method consistency test	Total rank differences test	Total score test
EB	Test value	4.3[a]	31 (48.4%)	6291	95.8
	Test ranking	*1*	*2*	*2*	*1*
EB_s	Test value	2.9[b]	33 (51.6%)	5336	89.2
	Test ranking	*6*	*1*	*1*	*2*
CF	Test value	4.1	24 (37.5%)	8910	81.5
	Test ranking	*2*	*3*	*4*	*3*
CR	Test value	4.1	22 (34.4%)	8865	79.5
	Test ranking	*3*	*4*	*3*	*4*
PFI	Test value	2.9	22 (34.4%)	12820	61.0
	Test ranking	*5*	*4*	*9*	*5*
EPDO	Test value	3.0	16 (25.0%)	11553	58.5
	Test ranking	*4*	*8*	*7*	*6*
$Proportion_{Rainy}$	Test value	1.8	19 (29.7%)	11283	52.7
	Test ranking	*9*	*6*	*6*	*7*
$Proportion_{Rear\text{-}end}$	Test value	1.8	17 (26.6%)	10965	51.6
	Test ranking	*8*	*7*	*5*	*8*
$Proportion_{ROR}$	Test value	1.9	16 (25.0%)	13382	46.0
	Test ranking	*7*	*8*	*10*	*9*
$Proportion_{Wet}$	Test value	1.7	14 (21.9%)	12349	44.7
	Test ranking	*11*	*10*	*8*	*10*
$Proportion_{Night}$	Test value	1.8	5 (7.8%)	14429	31.2
	Test ranking	*9*	*11*	*11*	*11*

The percentage shown in parenthesis stands for the percentage of hotspots identified in period 1 that were also identified in period 2.
[a] The modified site consistency test is equal to 1.21 severe crashes/(km × year).
[b] The modified site consistency test is equal to 0.97 severe crashes/(km × year).

case of the top 5% of sites, differences with the EBs in summed ranked differences were as follows: 124% for the CF method, 127% for the CR method, 228% for the EPDO method, 281% for the PFI method, and between 199% and 297% for the proportion method.

The CF and CR methods, which were only slightly worse than the EB method in the site consistency test, revealed large inconsistencies in the total rank differences test, mainly because they do not control for random fluctuation in crashes over time.

6.4. *Total score test results*

In the total score test, which combines the results of the previous three tests, the EB method performed better than the other HSID methods in all of the case studies (the top 1%, 5%, and 10% of the sites), reaching values very close to 100 (98.3, 98.1, and 95.8, respectively).

The EB expected frequency of severe-crashes (EBs) performed worse than the EB expected frequency of total-crashes (EB) (77.0 vs. 98.3 for $\alpha = 0.01$, 82.7 vs. 98.1 for $\alpha = 0.05$, 89.2 vs. 95.8 for $\alpha = 0.10$). The largest difference was observed in the site consistency test, where the EB method performed better with reference to both total-crashes and severe-crashes.

The CF and CR methods, which were only slightly worse than the EB method in the site consistency test, revealed large inconsistencies in the method consistency test and in the total rank differences test. The CF method performed worse than the EB method in all of the tests and for any threshold of hotspots identified (total test score equal to 88.2, 83.5, and 81.5). The CR method, which is used by many agencies based on the assumption that it performs better than the CF method, showed even worse results (83.3, 77.1, and 79.5).

The PFI (69.5, 50.4, and 61.0) and EPDO (53.2, 45.3, and 58.5) methods were largely inconsistent. The proportion method performed worst in all of the tests.

6.5. *Discussion of test results*

Overall, our study results are consistent with the results of the previous quantitative evaluations carried out by Cheng and Washington (2008) and Elvik (2007, 2008a).

The test results highlight that the EB method is the most consistent and reliable method for identifying priority investigation locations. The EB estimate of total-crashes performed better than the EB estimate of severe-crashes, even if the resulting difference was insignificant when the number of identified hotspots increased. This result is supported by a strong theoretical basis, because the EB method increases the precision of the safety estimation and corrects for the regression-to-mean bias. Indeed, EB estimates of safety allow the relative contributions of random variation, general factors, and local factors to the observed number of crashes to be identified.

From a practical point of view, an important result is that the CF method performed better than other HSID methods with more appealing theoretical arguments. In particular, the CF method performed better than the CR method in this and other studies (Cheng and Washington, 2008; Elvik, 2007, 2008a; Persaud et al., 1999). In fact, the crash rate method is biased towards low-volume sites and incorrectly assumes a linear relationship between traffic volume and crash frequency. This result is quite alarming, as many agencies use this method.

The PFI method, which is data-intensive like the EB method, did not perform well. The main drawback of the PFI method is that larger values of predicted crashes decrease the corresponding probability that a site is selected as a hotspot (because the difference between the EB estimate and predicted crashes is diminished), whereas in the EB method larger values of predicted crashes increase the ranking. Elvik (2007, 2008a) found that the major problem associated with the PFI method appears to be a large number of false positives that result from a temporary instability in the contribution of local risk factors to crashes. In previous

Table 6
Top 5% of hotspots identified using the EB method: crash patterns.

Segment	Length (m)	Crash count	EB rank	EB estimate [crashes/(km × year)]	P_{ROR} (%)	$P_{rear\text{-}end}$ (%)	P_{night} (%)	P_{wet} (%)	P_{rainy} (%)	Overrepresented crashes
224	319	64	1	39.4	79.7	4.7	9.4	84.4	54.7	ROR, Wet, Rainy
545	255	52	2	36.7	96.2	0.0	42.3	92.3	57.7	ROR, Wet, Rainy
108	309	55	3	34.5	74.5	21.8	38.2	72.7	40.0	ROR, Wet, Rainy
160	86	21	4	26.0	66.7	19.0	47.6	61.9	33.3	Wet, Rainy
546	255	33	5	22.5	87.9	6.1	33.3	60.6	27.3	ROR, Wet, Rainy
223	319	35	6	21.3	65.7	17.1	17.1	77.1	40.0	ROR, Wet, Rainy
533	188	22	7	21.1	95.5	0.0	13.6	77.3	22.7	ROR, Wet
4	225	29	8	19.0	93.1	0.0	58.6	37.9	20.7	ROR, Night
84	181	18	9	18.0	66.7	16.7	33.3	61.1	44.4	Wet, Rainy
353	248	20	10	15.0	75.0	20.0	30.0	90.0	45.0	ROR, Wet, Rainy
221	236	17	11	14.6	82.4	0.0	35.3	64.7	23.5	ROR, Wet
135	171	11	12	13.7	72.7	27.3	9.1	54.5	45.5	Wet, Rainy
373	347	26	13	12.6	92.3	7.7	34.6	96.2	34.6	ROR, Wet, Rainy
354	248	16	14	12.1	75.0	6.3	31.3	50.0	18.8	ROR
24	476	28	15	10.9	75.0	7.1	32.1	25.0	17.9	ROR
222	236	10	16	8.9	90.0	0.0	30.0	70.0	30.0	ROR, Wet
298	206	10	17	8.8	50.0	20.0	10.0	50.0	30.0	
159	86	6	18	8.3	33.3	0.0	16.7	0.0	0.0	
103	514	21	19	8.1	61.9	23.8	38.1	71.4	42.9	Wet, Rainy
326	160	8	20	7.8	75.0	12.5	12.5	50.0	25.0	
531	159	6	21	7.5	100.0	0.0	16.7	83.3	66.7	ROR, Wet, Rainy
426	474	18	22	7.4	94.4	0.0	16.7	94.4	27.8	ROR, Wet
487	196	7	23	7.4	14.3	0.0	28.6	28.6	0.0	
139	159	5	24	7.4	80.0	20.0	40.0	80.0	40.0	Wet
246	434	16	25	7.1	75.0	25.0	31.3	62.5	6.3	ROR, Wet
335	240	9	26	7.0	33.3	22.2	66.7	66.7	55.6	Night, Wet, Rainy
131	177	6	27	6.1	33.3	33.3	16.7	0.0	0.0	
457	501	16	28	5.8	50.0	0.0	25.0	18.8	6.3	
187	292	12	29	5.5	16.7	50.0	33.3	8.3	8.3	Rear-end
458	501	14	30	5.5	85.7	0.0	64.3	78.6	42.9	ROR, Night, Wet, Rainy
488	196	5	31	5.2	100.0	0.0	0.0	100.0	40.0	ROR, Wet
329	284	7	32	5.2	71.4	0.0	71.4	71.4	42.9	Night, Wet

Underline indicates crash types overrepresented at the 10% level of significance. All other crash types were overrepresented at the 5% level of significance.

studies (Cheng and Washington, 2008; Elvik, 2008a; Persaud et al., 1999), the PFI method performed worse than the EB and CF methods just as it did in this study, but performed better than the CR method, which was not found in this study. Persaud et al. (1999) used both the crash frequency and the difference between observed and predicted crashes as measures of the effectiveness of the HSID methods; the results of the second measure, which was not used in the evaluation tests of this study, indicated that the PFI was the best HSID method. The difference between our results and those of previous studies might depend both on the different range of AADTs and on the different CPMs used in the evaluations. In the present study, the AADT ranged between 7266 and 16,000 vehicles per day, whereas in the previous studies greater variations in traffic volume were observed. Cheng and Washington (2008) used only segment length and average annual daily traffic as explanatory variables (Persaud et al. added total width). Elvik (2008a) fitted one model for more highway types, using average annual daily traffic, speed limit, number of lanes, density of intersections, and a dummy for trunk highways as explanatory variables. In this study, more detailed variables related to traffic volume and composition, such as horizontal alignment, design consistency, vertical alignment, roadside context, cross-section, and yearly effects were used. As a result, the contribution to the EB estimate of safety of the general factors included in the CPMs is greater and the contribution of local factors, measured by the PFI method, is diminished.

The EPDO method was inconsistent in all of the tests, since it overemphasizes sites with a low frequency of fatal or severe crashes. To date, other quantitative evaluations of the EPDO method have not been performed.

The proportion method may be appealing because it is based on the hypothesis that dominant crash patterns are associated with risk factors; it is not, however, based on safety estimates. Dominant crash patterns might also be observed in sites with low frequencies of total-crashes or with unusually low frequencies of other crash types. Lyon et al. (2007) compared the proportion method with the EB method and the PFI method based on the fraction of the same hotspots ranked by the different methods. Using as reference the hotspots ranked by the EB method, they found that between 45% and 80% of the same hotspots were ranked by the PFI method and between 45% and 70% of the same hotspots were ranked by the proportion method. Based on this result, they concluded that the proportion method appears to be a workable alternative to the preferred PFI and EB methods. In the present study, smaller fractions of the same hotspots were ranked by the PFI (50–61%) and proportion (22–50%) methods. Furthermore, the quantitative evaluation tests clearly showed the inconsistency of the proportion method.

7. Correctability of hotspots identified with the EB method

In order to the verify whether the method classified as the best according to the four quantitative tests also ensures the correctability of the hotspot sites (i.e., the sites are not false hotspots), diagnostic evaluations were performed and the existence of potential effective safety treatments was verified.

Based on the 646 homogeneous segments and five years of crash data, hotspots were identified with the EB method, which gave the best results according to the quantitative evaluation criteria. The top 5% of sites were selected as hotspots. With the exception of one site, the identified hotspots were located on horizontal curves. In each site, the following analyses were performed: (a) diagnosis of safety concerns by means of comparison between crash patterns in the site and in the reference sites; (b) identification of local risk factors by means of assessment of site characteristics and field investigations; (c) identification of countermeasures which have proven to be effective in situations with the identified crash patterns and local risk factors.

Table 7
Top 5% of hotspots identified using the EB method: local risk factors.

Segment	Local risk factors
224	Inadequate coordination between horizontal and vertical alignment: sag before the curve. Polished pavement. Faded markings.
545	Inadequate delineation due to chevrons missing in the curve beginning. Polished pavement.
108	Curve following long downhill with steep grade. Polished pavement. Faded markings.
160	Inadequate coordination between horizontal and vertical alignment: horizontal and vertical curves with different centers. Poor curve perception due also to small deflection angle (6°) and short length (86 m). Polished pavement. Faded markings.
546	Polished pavement.
223	Polished pavement. Faded horizontal markings.
533	Short tunnel with inadequate lighting. Inadequate delineation due to faded chevrons and markings. Polished pavement.
4	Poor curve perception in the night due to an exit ramp in the curve. Polished pavement.
84	Inadequate delineation on a steep downgrade due to chevrons missing. Curve warning signs.
353	Inadequate delineation due to missing chevrons and faded markings. Curve warning signs missing. Polished pavement. Pavement unevenness.
221	High pavement unevenness due to discontinuities along the buried bridge joints. Inadequate delineation due to local discontinuities of the chevrons and faded markings. Polished pavement.
135	Design inconsistency due to the presence of a small radius curve (R = 250 m) on a steep downgrade which follows a larger radius curve (R = 500 m) after a short tangent (L = 173 m). Polished pavement.
373	Inadequate coordination between horizontal and vertical alignment: sag before the curve. Polished pavement. Faded markings.
354	Design inconsistency due to a small radius curve (R = 320 m) which follows a larger radius curve (R = 760 m) after a short tangent (L = 121 m). Inadequate delineation due to missing chevrons in the curve beginning, irregular spacing and height of the chevrons, and faded markings.
24	Narrow left shoulder.
222	Inadequate delineation due to local discontinuities of the chevrons. Polished pavement. Faded markings.
298	Inadequate coordination between horizontal and vertical alignment: sag before the curve. Inadequate delineation due to chevrons missing.
159	Inadequate coordination between horizontal and vertical alignment: horizontal and vertical curves with different centres. Poor curve perception due also to small deflection angle (6°) and short length (86 m). Faded markings.
103	High pavement unevenness due to discontinuities along the buried bridge joints. Inadequate delineation due to chevrons missing in the curve beginning. Polished pavement.
326	Inadequate curve perception due to a tunnel with inadequate lighting preceding the curve, dazzling at the sunset in the tunnel exit, inadequate delineation, and missing curve warning sign. Shoulder narrowing in the final part of the curve.
531	Polished pavement.
426	Inadequate perception of the curvature change due to the presence of a small radius curve (R = 305 m) which follows a larger radius curve (R = 785 m) after a short tangent (L = 93 m). Chevrons installed in the tangent between the curves increase the perception problem. Polished pavement.
487	Entry ramp in the curve with short length. Faded markings.
139	High pavement unevenness due to discontinuities along the buried bridge joints before the curve beginning. Polished pavement.
246	Inadequate coordination between horizontal and vertical alignment: crest before the curve. Inadequate perception of the curve due to the transition along the curve between a tunnel with inadequate lighting and a bridge. Inadequate delineation due to chevrons missing and faded markings. Curve warning signs missing. High pavement unevenness due to discontinuities along the buried bridge joints. Polished pavement.
335	Inadequate delineation due to chevrons missing. Curve warning signs missing. Polished pavement.
131	Inadequate coordination between horizontal and vertical alignment: crest before the curve. Inadequate delineation due to chevrons missing.
457	Inadequate delineation due to chevrons missing in the curve beginning and faded markings. Pavement unevenness due to discontinuities along the buried bridge joints.
187	Sag with short radius. Faded horizontal markings.
458	High pavement unevenness due to discontinuities along the buried bridge joints. Inadequate delineation due to chevrons missing. Curve warning sign missing. Polished pavement.
488	Polished pavement.
329	Inadequate lighting of the tunnel preceding the curve. Curve warning sign missing. Pavement unevenness due to discontinuities along the buried bridge joints. Inadequate delineation due to chevrons missing and faded markings. Polished pavement.

7.1. Crash patterns

Since crash patterns in the curves and in the tangents showed statistically significant differences, two different reference groups were used for the segments located in the curves and in the tangents. For each curve, the reference group was composed of all the curves in the study sample except the segment under consideration (325 reference curves). The same methodology was used for the hotspot located in the tangent (segment 187, with 319 reference tangents). Statistical analyses were performed with the binomial test (see Eq. (1) in Section 2.4). The following crash types were considered: ROR, rear-end, nighttime, wet road, and rainy.

In 25 out of 32 sites, crash types that were overrepresented at the 5% level of significance were found (see Table 6). In one site, an overrepresentation at the 10% level of significance was found. Only six sites did not exhibit an overrepresentation of crash patterns. In all of the top 16 hotspots, crash types overrepresented at the 5% level of significance were found.

In the hotspots, the most frequent crash patterns were wet road crashes, which were overrepresented at the 5% level of significance in 21 sites (and at the 10% level of significance in one site). ROR crashes were overrepresented in 18 sites, rainy crashes were overrepresented in 14 sites, and nighttime crashes were overrepresented in three sites. Rear-end crashes were overrepresented (at the 10% level of significance) only in the hotspot located in the tangent (segment 187, see Table 6 for further information).

Hypotheses regarding crash contributory factors were drawn based on the identified crash patterns.

7.2. Local risk factors

The safety diagnosis was complemented by an assessment of site characteristics and field investigations; these studies examined how the road environment was perceived and ultimately utilized by different road users. Furthermore, detailed site features were observed. As a result, many local risk factors that could be corrected were identified (see Table 7). The most common local risk factors which were not detectable from the analysis of crash data and crash patterns were (a) inadequate coordination between horizontal and vertical alignment and (b) high pavement unevenness due to discontinuities along the buried bridge joints. Segment 426 was a special case, where the curvature change was inadequately perceived due to the presence of a small radius curve (R = 305 m) following a larger radius curve (R = 785 m) after a short tangent (L = 93 m). The perception problem was greatly increased by the installation of chevrons along the tangent connecting the curves with different radii. Furthermore, chevron spacing was the same along the two curves and along the tangent. This is a misleading

Table 8
Top 5% of hotspots identified using the EB method: potential countermeasures.

Segment	Potential countermeasures
224	Resurfacing. Installation of high performance markings. Sag displacement.
545	Resurfacing. Curve delineation improvement.
108	Resurfacing. Installation of high performance markings. Installation of optical speed bars.
160	Resurfacing. Horizontal and vertical alignment improvement: increase of the horizontal curve radius, introduction of spiral transitions, and displacement of the sag curve. Installation of high performance markings.
546	Resurfacing.
223	Resurfacing. Installation of high performance markings.
533	Resurfacing. Tunnel lighting upgrading. Curve delineation improvement. Installation of high performance markings.
4	Lengthening of the right turn lane. Curve delineation improvement: chevrons and sequential flashing beacons along the curve.
84	Resurfacing. Installation of curve warning signs. Curve delineation improvement: chevrons and sequential flashing beacons along the curve.
353	Resurfacing. Installation of curve warning signs. Curve delineation improvement: chevrons and sequential flashing beacons along the curve. Installation of high performance markings.
221	Resurfacing. Buried bridge joints upgrading. Curve delineation improvement. Installation of high performance markings.
135	Resurfacing.
373	Resurfacing. Displacement of the sag curve. Installation of high performance markings.
354	Resurfacing. Installation of curve warning signs. Curve delineation improvement. Installation of high performance markings.
24	Widening of left shoulder.
222	Resurfacing. Curve delineation improvement. Installation of high performance markings.
298	Displacement of the sag curve and increase of the horizontal curve radius. Installation of chevrons and sequential flashing beacons along the curve.
159	Horizontal and vertical alignment improvement: increase of the horizontal curve radius, introduction of spiral transitions, and displacement of the sag curve. Installation of high performance markings.
103	Resurfacing. Curve delineation improvement: chevrons and sequential flashing beacons along the curve. Buried bridge joints upgrading.
326	Tunnel lighting improvement. Installation of curve warning signs. Curve delineation improvement: chevrons and sequential flashing beacons along the curve.
531	Resurfacing.
426	Resurfacing. Removal of the chevrons in the tangent preceding the curve. Change of chevrons spacing in the successive curves with different radius.
487	Lengthening of the entry ramp. Installation of high performance markings.
139	Resurfacing. Buried bridge joints upgrading.
246	Tunnel lighting improvement. Resurfacing. Installation of curve warning signs. Curve delineation improvement. Installation of high performance markings. Buried bridge joints upgrading.
335	Resurfacing. Installation of curve warning signs. Installation of chevrons and sequential flashing beacons along the curve.
131	Increase of the radius of the crest curve. Installation of chevrons and sequential flashing beacons along the curve.
457	Curve delineation improvement. Installation of high performance markings. Buried bridge joints upgrading.
187	Increase of the radius of the crest curve. Resurfacing. Installation of high performance markings
458	Resurfacing. Installation of chevrons and sequential flashing beacons along the curve. Buried bridge joints upgrading.
488	Resurfacing.
329	Resurfacing. Tunnel lighting upgrading. Installation of curve warning signs. Buried bridge joints upgrading. Curve delineation improvement. Installation of high performance markings.

clue that gives road users a dangerous message about the presence of only one curve with a constant radius.

7.3. Countermeasures

Based on the results of previous investigations, identification of countermeasures that may reduce crashes was successful in all of the hotspots (see Table 8). Engineering improvements that had been evaluated by properly conducted empirical Bayes observational before-and-after studies were preferred. For example, in the sites where wet road crashes were overrepresented, resurfacing was identified as a potential countermeasure. Indeed, Persaud and Lyon (2008) conducted an EB observational before-and-after study and found statistically significant reductions of 31.6% in total crashes and 65.4% in wet road crashes for rural segments; these segments required skid resistance improvement because of a high frequency of wet road crashes and low friction numbers.

8. Conclusions

Seven commonly applied HSID methods were compared against four robust and informative quantitative evaluation criteria: the site consistency test, the method consistency test, the total rank differences test, and the total score test. These tests evaluated different aspects of each HSID method's performance. The site consistency test measures the method's efficiency in identifying sites that show a consistently poor safety performance. The method consistency test measures the efficiency of the HSID methods in terms of the number of the same hotspots identified in subsequent time periods. The total rank differences test measures the methods' efficiency in identifying sites that have a small rank difference in subsequent time periods. The total score test combines the previous tests in order to provide a synthetic and easily readable index. For this evaluation, five years of crash data from the Italian motorway A16 were used.

The quantitative evaluation tests showed that the EB method performs better than the other HSID methods. The test results highlight that the EB method is the most consistent and reliable method for identifying priority investigation locations. The EB expected frequency of total-crashes (EB) performed better than the EB expected frequency of severe-crashes (EBs), although the difference was insignificant when the number of identified hotspots increased. The CF method performed better than other HSID methods with more appealing theoretical arguments. In particular, the CF method performed better than the CR method. This result is quite alarming, as many highway agencies use the CR method. The PFI and EPDO methods were largely inconsistent. The proportion method performed worst in all of the tests. Overall, these results are consistent with the results of previous studies.

The identification of engineering countermeasures that have the potential for crash reduction was successful in all of the hotspots identified with the EB method; this shows that the method identified hotspots that can also be corrected.

We recognize that the advantages associated with the EB method were obtained based on data from only one Italian motorway and that the relative performances of the HSID methods may change when other crash data are used. However, our study results are very significant and are consistent with earlier findings. To gain more confidence in the benefits of the EB method, the study should be replicated in other countries. Nevertheless, these study results, combined with previous research results, strongly suggest that the EB method should be the standard approach in the identification of crash hotspots.

Acknowledgments

The comments and suggestions of Wen Cheng, Tommaso Esposito, and Dominique Lord are gratefully acknowledged.

References

Akaike, H., 1987. Factor analysis and AIC. Psycometrika 52, 317–332.
Austroads, 2009. Guide to Road Safety PART 8: Treatment of Crash Locations. Austroads Publication AGRS08/09, Sydney, New South Wales.
Cheng, W., Washington, S., 2005. Experimental evaluation of hotspot identification methods. Accident Analysis and Prevention 37, 870–881.
Cheng, W., Washington, S., 2008. New Criteria for Evaluating Methods of Identifying Hot Spots. Transportation Research Record, 2083. TRB, National Research Council, Washington, DC, pp. 76–85.
DTLR, Department for Transport, Local Government and the Regions, 2001. A Road Safety Good Practice Guide. London.
El-Basyouny, K., Sayed, T., 2006. Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models. Transportation Research Record, 1950. TRB, National Research Council, Washington, DC, pp. 9–16.
Elvik, R., 2007. State-of-the-Art Approaches to Road Accident Black Spot Management and Safety Analysis of Road Networks. Report 883. Institute of Transport Economics, Oslo.
Elvik, R., 2008a. Comparative Analysis of Techniques for Identifying Hazardous Road Locations. Transportation Research Record, 2083. TRB, National Research Council, Washington, DC, pp. 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.
European Parliament, 2008. Road Infrastructure Safety Management. Directive 2008/96/EC.
Hauer, E., 1997. Observational Before–After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Pergamon Press (Elsevier Science), Oxford.
Hauer, E., 2004. Statistical Road Safety Modeling. Transportation Research Record, 1897. TRB, National Research Council, Washington, DC, pp. 81–87.
Hauer, E., Allery, B.K., Kononov, J., Griffith, M.S., 2002a. Screening the road Network for Sites with Promise. Transportation Research Record, 1784. TRB, National Research Council, Washington, DC, pp. 27-32.
Hauer, E., Harwood, D.W., Council, F.M., Griffith, M.S., 2002b. Estimating Safety by the Empirical Bayes Method: a Tutorial. Transportation Research Record, 1784. TRB, National Research Council, Washington, DC, pp. 126–131.
Hauer, E., Allery, B.K., Kononov, J., Griffith, M.S., 2004. How Best to Rank Sites with Promise. Transportation Research Record, 1897. TRB, National Research Council, Washington, DC, pp. 48–54.
Kononov, J., 2002. Identifying Locations with Potential for Accident Reductions: Use Of Direct Diagnostics and Pattern Recognition Methodologies. Transportation Research Record, 1784. TRB, National Research Council, Washington, DC, pp. 153–158.
Lyon, C., Gotts, B., Wong, W.K.F., Persaud, B., 2007. Comparison of Alternative Methods for Identifying Sites with High Proportion of Specific Accident Types. Transportation Research Record, 2019. TRB, National Research Council, Washington, DC, pp. 212–218.
Lyon, C., Persaud, B., 2008. Safety Effects of Targeted Program to Improve Skid Resistance. Transportation Research Record, 2068. TRB, National Research Council, Washington, DC, pp. 135–140.
Miaou, S.P., 1996. Measuring the goodness of fit of accident prediction models. Publication FHWA-RD- 96-040. FHWA, US Department of Transportation.
Montella, A., 2001. Selection of Roadside Safety Barrier Containment Level According to European Union Standards. Transportation Research Record, 1743. TRB, National Research Council, Washington, DC, pp. 104–110.
Montella, A., 2005. Safety Reviews of Existing Roads: Quantitative Safety Assessment Methodology. Transportation Research Record, 1922. TRB, National Research Council, Washington, DC, pp. 62–72.
Montella, A., 2009. Safety Evaluation of Curve Delineation Improvements: an Empirical Bayes Observational Before-and-After Study. Transportation Research Record, 2103. TRB, National Research Council, Washington, DC, pp. 69–79.
Montella, A., Colantuoni, L., Lamberti, R., 2008. Crash Prediction Models for Rural Motorways. Transportation Research Record, 2083. TRB, National Research Council, Washington, DC, pp. 180–189.
Persaud, B., 2001. Statistical Methods in Highway Safety Analysis. NCHRP Synthesis 295. Transportation Research Board, Washington, DC.
Persaud, B., Lyon, C., Nguyen, T., 1999. Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Improvement. Transportation Research Record, 1665. TRB, National Research Council, Washington, DC, pp. 7–12.
Persaud, B., Lyon, C., 2007. Empirical Bayes Before–After Studies: Lessons Learned from Two Decades of Experience and Future Directions. Accident Analysis and Prevention 39, 546–555.
PIARC, World Road Association, Technical Committee on Road Safety C13, 2004. Road Safety Manual.
ROSPA, The Royal Society for Prevention of Accidents, 2002. Road Safety Engineering Manual. Birmingham.
TAC, Transportation Association of Canada, 2004. The Canadian Guide to In-Service Road Safety Reviews. Ottawa.
Tarko, A.P., Kanodia, M., 2004. Hazard Elimination Program. Manual on Improving Safety of Indiana Road Intersections and Sections. Report FHWA/IN/JTRP-2003/19, West Lafayette, Indiana.
UK Department for Transport, 2007. 2005 Valuation of the Benefits of Prevention of Road Accidents and Casualties. Highways Economic Note No. 1, London.