

New Criteria for Evaluating Methods of Identifying Hot Spots

Wen Cheng and Simon Washington

Identification of hot spots, also known as the sites with promise, black spots, accident-prone locations, or priority investigation locations, is an important and routine activity for improving the overall safety of roadway networks. Extensive literature focuses on methods for hot spot identification (HSID). A subset of this considerable literature is dedicated to conducting performance assessments of various HSID methods. A central issue in comparing HSID methods is the development and selection of quantitative and qualitative performance measures or criteria. The authors contend that currently employed HSID assessment criteria—namely false positives and false negatives—are necessary but not sufficient, and additional criteria are needed to exploit the ordinal nature of site ranking data. With the intent to equip road safety professionals and researchers with more useful tools to compare the performances of various HSID methods and to improve the level of HSID assessments, this paper proposes four quantitative HSID evaluation tests that are, to the authors' knowledge, new and unique. These tests evaluate different aspects of HSID method performance, including reliability of results, ranking consistency, and false identification consistency and reliability. It is intended that road safety professionals apply these different evaluation tests in addition to existing tests to compare the performances of various HSID methods, and then select the most appropriate HSID method to screen road networks to identify sites that require further analysis. This work demonstrates four new criteria using 3 years of Arizona road section accident data and four commonly applied HSID methods [accident frequency ranking, accident rate ranking, accident reduction potential, and empirical Bayes (EB)]. The EB HSID method reveals itself as the superior method in most of the evaluation tests. In contrast, identifying hot spots using accident rate rankings performs the least well among the tests. The accident frequency and accident reduction potential methods perform similarly, with slight differences explained. The authors believe that the four new evaluation tests offer insight into HSID performance heretofore unavailable to analysts and researchers.

Identifying crash hot spots (also known as sites with promise, black spots, accident-prone locations, and priority investigation locations) is an imperative step to improve the safety performance of roadway network. Its great significance is echoed in two federal transportation bills, the Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991 and the subsequent Transportation Equity Act for

the 21st Century, which brought hot spot identification (HSID) squarely into transportation planning activities. In particular, ISTEA requires each state to develop a work plan outlining strategies to implement safety management systems (1). One of the objectives outlined in this system requires that metropolitan planning organizations, departments of transportation (DOTs), or both to undertake the adoption of a defensible methodology for identifying safety deficiencies within a region. In addition to the legislation requirement, there are also public pressures to spend safety funds wisely and professional pressures to conduct rigorous analyses and be held accountable for generation of reliable data.

There is a fairly extensive literature focused on methods for HSID. There are papers that discuss methods based on accident count or frequency (2), papers that employ both accident rate (AR) and rate quality control (3–6), and others that adopt the joint use of accident frequency and rate to flag sites with promise (7). To correct for the regression-to-the-mean bias associated with typical HSID methods (8), some researchers have suggested using the empirical Bayes (EB) techniques (9–14). This method combines clues from both the accident history of a specific site and expected safety of similar sites, and has the advantage of revealing underlying safety problems which otherwise would not be detected. Applications of the EB techniques include estimating the safety of railway–highway crossings, signalized intersections, drivers, etc.

Rather than using overall accident frequencies at sites, some researchers have suggested using accident reduction potential (ARP) to identify hot spots (15–18). ARP-based methods rest on the premise that only “excess” accidents over those expected from similar sites can be prevented by applying appropriate treatments, and thus the potential for reduction is a better method for identifying sites with promise. Yet other researchers (19–22), using sites with large traffic volumes, using total crash counts, or both have focused on crash outcomes and pattern recognition to identify sites with promise (e.g., a site has greater than expected numbers of run-off-road crashes). Finally, there are papers that emphasize the importance of crash severity and costs (23, 24).

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. Persaud and Hauer (25) compared and evaluated the performances of an EB and a nonparametric method for debiasing before-and-after analyses. The analysis of several data sets show that the Bayesian methods in most cases yield superior estimates. Hauer and Persaud (26) also proposed the use of false identifications to measure the performances of various methods for HSID. In response to this study, Hagle and Hecht (27) conducted a simulation experiment to evaluate and compare techniques for the identification of hazardous locations in relation to crash rates. Maher and Mountain (28) also used a simulation-based approach to compare methods, including ranking

W. Cheng, Tetra Tech, Inc., 16241 Laguna Canyon Road, Suite 200, Irvine, CA 92618. S. Washington, Department of Civil and Environmental Engineering, Arizona State University, University Drive and Mill Avenue, Tempe, AZ 85287. Corresponding author: S. Washington, Simon.Washington@asu.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2083, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 76–85.
DOI: 10.3141/2083-09

of sites on the basis of annual accident totals and ARP. Subsequent work by Cheng and Washington (29) employed several practical empirical crash distributions from the state of Arizona and several degrees of crash heterogeneity to compare HSID method performances and explore the optimal accident history duration in an experiment design.

Central to comparisons of HSID is the identification and development of robust and informative quantitative and qualitative criteria for evaluating such methods. To the authors' knowledge, no paper has reported on the criterion used for assessing HSID methods. Without robust and meaningful criteria for comparing HSID methods, the results of such comparisons rest on a shaky foundation. Similar to the selection of statistical and econometric models, numerous assessment criteria are needed to assess HSID methods. For example, in linear regression, numerous goodness-of-fit statistics are used to assess the appeal of a model, including adjusted R^2 , F -ratio, t -statistics of model variables, signs and magnitudes of coefficients, and mean square error.

This paper describes research that aims to equip road safety professionals with more useful tools for comparing and assessing the performances of various HSID methods and to aid in the selection of preferred methods with confidence. Five tests are proposed (four new) for conducting performance assessments of alternative HSID methods. These five tests evaluate the HSID methods' performances in different aspects: the site consistency test measures the 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 relation to the number of hot spots that have consistent underlying safety being identified in a relatively short period. The false identification test and the Poisson mean differences test measure the performance in the basis of false identification of hot spots and the corresponding consequences that arise from erroneous identifications.

The remainder of this paper first describes the HSID methods to be compared in the analysis. Then, it describes and develops the analytics of the five performance criteria, four of which are new. The data, consisting of safety performance functions of nine functional classifications of road sections, are then described. Then presented are the results of a comprehensive test of the HSID methods against the five criteria described. The paper ends with conclusions and recommendations.

HSID METHODS USED IN COMPARISON

Four HSID methods are commonly used in practice. These methods include the ranking of accident frequency (AF), the ranking of ARs, the ranking of ARP, and the EB method and ARP method. The details of each of these HSID methods are now described.

The AF HSID method is straightforward. Applying this method, a set of "similar" locations (e.g., all four-lane signalized intersections in a jurisdiction) is ranked in descending order of observed AFs (or counts, x) during a given observation period (e.g., 1 year). Top-ranked sites are identified as sites with promise for further examination and possible treatment. Typically, resources are invested to improve correctable sites from the top down until allocated funds are expended. One disadvantage of this method is that it does not account for the underlying (and unknown) safety of locations. For example, it cannot distinguish between a random versus a "real" up fluctuation in crash counts.

The AR is the accident count divided by traffic volume or entering vehicle volume. Analysis methods based on AR have also enjoyed broad applications. The AR method simply ranks road sections (inter-

sections, etc.) according to their calculated ARs. As discussed by Hauer (30), the use of ARs makes an implicit assumption that safety performance functions between accidents and exposure are linear for a collection of sites. The possible nonlinearity of safety performance functions and the lack of an estimable coefficient for exposure (even if the effect is linear) are the two primary arguments against using ARs. Moreover, if exposure is treated as an independent variable and a linear (fixed-effect) coefficient is estimated (say, in a predictive model of crashes), then the difference between a rate-based model and a frequency-based model is simply a multiplicative constant.

The EB method for HSID has been developed and applied more recently than the AF and AR methods. This method accounts for both accident history of a specific site and the performance or similar reference sites. It follows that safety of a site is affected by not only some common measurable factors shared by a corresponding reference population (generally captured in a multivariate regression model) but also some unique characteristics associated with the site (reflected in its accident history). The expected safety of a site is then expressed as follows:

$$\lambda_i = wE[\lambda] + (1-w)x_i \quad (1)$$

where w is a weighting factor that is calculated through the following equation:

$$w = \frac{E[\lambda]}{\{E[\lambda] + \text{Var}[\lambda]\}} \quad (2)$$

In Equations 1 and 2, $E[\lambda]$ represents the expected safety of a reference population, $\text{Var}[\lambda]$ is the corresponding variance, and x_i is the observed count history for site i . Hauer (31) presented two ways to calculate the $E[\lambda]$ and $\text{Var}[\lambda]$, that is, the method of sample moments and the multivariate regression method. The former one is relatively straightforward but produces less precise results, whereas the latter one is more complicated and requires relatively large amounts of accident data. In the multivariate regression method, w can be rewritten as follows (32):

$$w = \left[\frac{1 + (\mu * Y)}{\phi} \right] \quad (3)$$

where

- μ = expected number of accidents per kilometer per year on similar segments or accidents per year expected on similar intersections,
- Y = number of years of accident count data used, and
- ϕ = overdispersion parameter, which is a constant for a given model and is derived during the regression calibration process.

Finally, the ARP was first introduced as the difference between the observed accident count of a site and the expected count estimated using a collection of reference sites, $\text{ARP}_i = \lambda_i - E[\lambda]$ (16, 17). Subsequently, to account for the random fluctuations associated with observed data, Persaud (18) proposed that the EB-estimated AF replace the observed accident counts of the specific site. In so doing, the ARP associated with each site is calculated and sites ranked to identify sites with promise. Appropriately substituting Equation 1, the ARP for site i is expressed as follow:

$$\text{ARP}_i = wE[\lambda] + (1-w)x_i - E[\lambda] \quad (4)$$

Comparing Equations 1 and 4, one can see that the higher crash counts in history (x_i) will increase the priority for further investigation

of each individual site in both EB and ARP methods, whereas the value of $E[\lambda]$ has the reverse impact on selection of hot spots. In the EB method, larger values of $E[\lambda]$ increase the probability that a specific location makes it on the hot spot list (because it increases the estimate at a site), whereas in the ARP method, larger values of $E[\lambda]$ decrease the corresponding probability that a site is selected as a site with promise (because the potential for improvement is diminished).

EVALUATION TESTS OF PERFORMANCE OF HSID METHODS

With the HSID methods and safety performance functions described and ready for supporting the EB and ARP methods, it is now possible to describe new evaluation criteria for evaluating HSID methods. To evaluate HSID methods, prior research has relied solely on the percentage of false positives (FPs), claiming a site is unsafe when it is not; the percentage of false negatives (FNs), failing to claim that a site is unsafe when it is; and the percentage of incorrect classifications, the sum of the FPs and the FNs divided by the total. This paper contends that the FP and FN measures are too coarse and uninformative to provide adequate insight into the relative performances of HSID methods. Specifically, these measures are binary—a site is either an FP or not, or an FN or not—when in fact HSID methods rank sites and provide much richer information than can be

accurately captured by the count of FPs and FNs. HSID methods produce ordinal data that to date have been ignored in their evaluation. The additional measures described in this paper make use of the ranking data produced using HSID (which is ignored when FPs and FNs are used), providing greater insight into HSID method performance.

The four new evaluation tests developed and described in this section include the site consistency test, the method consistency test, the total rank differences test, and the Poisson mean differences test; in addition, the false identification test, which is not new, is also described. To facilitate the understanding of and illustrate the mechanics of the new tests, a random sample of 20 principal arterial road sections, shown in Table 1, is used. The sites are sorted with respect to increasing crash counts in 2000. Furthermore, the table shows the results of the calculations for AF, AR, EB, and ARP (by means of Persaud's correction) methods. For example, Site ID 20 recorded 14 crashes per kilometer per year (in 2000) and had an accident rate of 2.197 crashes per million vehicles per kilometer per year, an EB estimate of crashes of 4.02 crashes per kilometer per year, and an ARP of 5.07 crashes per kilometer per year. The 20 sites are used for illustration only: a more comprehensive and extensive analysis on all sites in the Arizona database is summarized later in the paper, after the new criteria have been illustrated. These later results produce results that do not suffer from small-sample bias, as is evident in the sample of 20 described here.

TABLE 1 Accident Information of Sample of 20 Principal Arterial Road Sections

Site ID	Period 1 (2000)				Period 2 (2001–2002)				3-Year Crash Mean
	Crash Frequency (AF)	Crash Rate ^a (AR)	EB Estimated Crashes	ARP ^b Estimated Crashes	Crash Frequency (AF)	Crash Rate ^a (AR)	EB Estimated Crashes	ARP ^b Estimated Crashes	
1	1	0.16	1.21	-0.03	4	0.63	2.34	1.54	1.67
2	1	0.15	0.82	0.01	2	0.31	1.84	0.12	1.00
3	1	0.16	0.76	0.09	1	0.16	1.14	-0.08	0.67
4	2	0.31	0.83	0.44	2	0.31	1.52	0.06	1.33
5	2	0.32	0.79	0.96	1	0.16	0.84	0.02	1.00
6	4	0.52	1.71	1.42	5	0.65	2.68	1.16	3.00
7	4	0.63	1.81	1.62	10	1.58	2.94	4.3	4.67
8	5	0.77	1.86	1.50	13	2.00	3.94	2.98	6.00
9	5	0.79	1.79	3.20	4	0.63	2.98	0.56	3.00
10	6	0.74	1.92	0.20	17	2.1	4.06	6.78	7.67
11	6	0.93	1.82	1.49	11	1.71	3.9	0.22	5.67
12	6	0.94	1.89	3.75	16	2.51	3.62	7.98	7.33
13	7	1.11	1.95	0.80	15	2.39	4.24	1.12	7.33
14	7	1.10	2.01	1.82	17	2.68	4.86	8.36	8.00
15	8	1.28	1.96	1.09	20	2.40	3.78	7.56	9.33
16	8	1.26	1.99	4.37	15	2.37	4.54	5.24	7.67
17	11	1.38	3.56	1.08	16	2.01	6.36	0.16	9.00
18	11	1.73	4.33	4.06	31	3.31	6.82	13.98	14.00*
19	12	1.84	3.58	4.15	28	3.38	6.86	11.78	13.33
20	14	2.20	4.02	5.07	32	3.14	8.9	18.52	15.33*

NOTE: Sites 18 and 20 are considered as truly top 10% hazardous locations based on the 3-year accident mean, which is assumed as the true Poisson mean of each road section. Shaded cells represent the top 10% sections of the 20 samples in the 2 time periods according to HSID method.

^aAR: Accidents/million vehicles/km/year.

^bARP: Accident reduction potential. The negative values result from sections whose Bayesian estimators are less than regression values of similar sections.

Test T1. Site Consistency Test

The site consistency test (T1) is used to measure the ability of a HSID method to identify consistently a site as high risk over subsequent observation periods. The test rests on the premise that a site identified as high risk during Period 1 should also reveal inferior safety performance in a subsequent Period 2, given that no significant changes have occurred at the site and given that the site is in fact high risk. Test T1 simply requires a comparison of the sum of observed crashes occurring on $n\alpha$ high-risk sites during future time period $i + 1$ identified by method j (during time period i) to crashes occurring at high-risk sites (in time period $i + 1$) identified by other possible HSID methods. In equation form, this comparison is given as

$$T1 = \sum_{k=n-n\alpha}^n C_{k, \text{method}=j(i), i+1} > \sum_{k=n-n\alpha}^n C_{k, \text{method} \neq j, i+1} \quad (5)$$

where

- n = total number of sites being compared (e.g., crashes occurring on $n = 100$ rural signalized intersections),
- C = crash count for site ranked site k ,
- α = threshold of identified high-risk sites (e.g., $\alpha = .05$ corresponds with top 5% of n sites identified as high risk, and $n\alpha$ = number of identified high-risk sites),
- j = HSID method being compared (e.g., $j = 1$ could be the AF method, $j = 2$ the AR method, etc.), and
- i = observation period (e.g., $i = 2000$, $i + 1 = 2001$). In this comparison, the method j that identifies sites in a future period with the highest crash count is the most consistent method for identifying underlying safety problems.

To illustrate this test, consider the small sample of data presented in Table 1. The AF and AR methods identify the same top 10% ($\alpha = .10$) of the sections during 2000, Sites 19 and 20. The crash counts of these same sites during 2001 to 2002 are 60 (28 + 32). The EB method identifies Sites 18 and 20 as high risk in 2000, while the total crashes in Period 2 at these sites are 63 (31 + 32). The ARP method identifies Sites 16 and 20, whose total crashes during 2001 and 2002 are 47 (15 + 32). Thus, comparing the performance of all methods with this test, the EB method performs best (based on this small sample used for illustrative purposes). In other words, the EB method identified sites in Period 1 with the highest crash counts also observed in Period 2, or sites with the most consistent poor safety record.

Test T2. Method Consistency Test

In Test T1 just discussed, crash counts in Period 2 are used as a benchmark to compare different HSID methods. The underlying assumption is that the identification performance of the HSID methods is revealed through the safety performance of the corresponding identified hot spots. The method consistency test (T2), in contrast, is designed to evaluate a method's performance by measuring the number of the same hot spots identified in both periods. Inasmuch as the two periods are close in time, it is assumed that road sections are in the same or similar underlying operational states (similar traffic volumes, driver populations, geometric designs, weather fluctuations, etc.) and that their expected safety performance remains virtually unaltered over the two periods. Under this homogeneity assumption, a good HSID method will identify the same set of hot spots across two periods. The greater the number of hot spots that are identified in both periods, the more reliable and consistent is the performance of HSID method.

Analytically, Test T2 is simply the intersection of ranked sites k identified in subsequent time periods i and $i + 1$ that are high risk, or

$$T2_j = \{k_{n-n\alpha}, k_{n-n\alpha+1}, \dots, k_n\}_{j,i} \cap \{k_{n-n\alpha}, k_{n-n\alpha+1}, \dots, k_n\}_{j,i+1} \quad (6)$$

Only sites identified in the top threshold α are compared.

In this test, the intersection of sites identified as high risk in two subsequent periods is compared over methods $j = 1$ to J , and the method yielding the largest intersection of sites is said to be the most consistent.

In a review of Table 1 for the AF method, Sites 19 and 20 and Sites 18 and 20, respectively, are selected as the top 10% of the hot spots in the two periods; the number of consistent sites is thus one, or, Site 20. For the ARP method, the hazardous sections are Sites 19 and 20 and Sites 18 and 19, respectively. In similar fashion, the number of hot spots identified by the EB and the ARP methods is one as well.

Test T3. Total Rank Differences Test

In Test T2, the number of consistently identified hot spots common to both periods is used to measure a method's reliability. Test T3, the total rank differences test, building on Test T2, takes into account the rankings of safety performances of road sections in the two periods. To illustrate, a group of 100 road sections is considered; among them there is one site whose AR ranks first in Period 1 and ranks 10th in Period 2. If the top 10 sites are identified as hot spots, then the AR method has identified a site in both periods even though the rankings have changed between periods. Test T3 is proposed to account for the rankings. The test is conducted by calculating the sum of total rank differences of the hazardous road sections identified across the two periods. The smaller is the total rank difference, the more consistent is the HSID method on this test, reflecting consistent ranking of sites across periods. In comparison with the other tests, this test is based on the assumptions that no safety treatments are implemented on the road sections and that the underlying safety of the road section during the two periods is constant. Hence, it is of great importance to ensure that all data outliers (i.e., road sections treated during Period 2) have been identified and removed when conducting this test.

In equation form, this test statistic is given as

$$T3_j = \sum_{k=n-n\alpha}^n (\mathfrak{R}(k_{j,i}) - \mathfrak{R}(k_{j,i+1})) \quad (7)$$

where \mathfrak{R} is the rank of site k in period i for method j . The difference in ranks is summed over all identified sites for threshold level α for period i .

In Table 1, the top 10% of the sections—Sites 19 and 20—identified by the frequency method (based on data in Period 1) possess the 19th and 20th rankings of AF in Period 1 and the 18th and 20th rankings in Period 2; the total ranking difference of AF as a result is 1 (19 – 18 + 20 – 20). In contrast, for the ARP method, the hazardous sections, Sites 16 and 20, have the rankings of 19th and 20th in Period 1 and 13th and 20th in Period 2, and the resultant total ranking difference is 6.

False Identification Test

False identifications consist of FNs (i.e., the truly hazardous sites mistakenly considered as safe) and FPs (i.e., the truly safe sites

wrongly viewed as hazardous). As the name implies, this test counts false identifications to assess the performances of various HSID methods. The false identification test is not new and has been used in previous research (27, 33). An important concern of this test is how to identify the “truly” hazardous and “truly” safe locations. In other words, calculation of false identification test statistics requires that truth be known a priori. It is this fact that justifies the use of simulated data for comparing HSID methods, and in fact historical data cannot be used with reliability. To satisfy this requirement in this analysis, the mean of 3 years of accident data is used as the true Poisson mean (TPM) of the road section; thereby, road sections with higher TPMs are assumed to be truly hazardous locations in this simulation (in real data, the sites that are truly hazardous are unknown). The relatively small sample of data (3 years) is a caveat for the results presented in this paper, and more accurate results may be obtained with longer crash histories. (However, there is a risk with longer crash histories in that the homogeneity assumption becomes threatened due to changes in pavement conditions, vehicle fleet, surrounding land uses, etc.)

The false identification test consists of a sequence of steps similar to those used in previous research (29). First, the TPM cumulative distributions of each roadway classification are divided into truly hazardous locations and nonhazardous locations by using various specified separation thresholds. Generally, if the road sections within the top $x\%$ of TPMs are viewed as truly hazardous, then the separation threshold of $1 - x\%$ is used. Second, the four HSID methods are applied as usual to identify hot spots. For the truly hazardous sites (the road sections within the top $x\%$ of TPMs), if their associated safety performance (i.e., AF, AR, EB-estimated accidents, and ARP) rank lower than the top $x\%$, then FNs are produced. Similarly, for the collection of truly nonhazardous locations, when the corresponding safety performance ranks higher than the top $x\%$, FPs result. The number of false identifications is the sum of the number of FNs and FPs. Finally, false identifications across classifications of road sections in both periods are tallied and reported.

In Table 1, the column titled 3-Year Crash Mean represents the TPM of each road section. From the ranking of these TPMs, it is known that Sites 18 and 20 are hazardous locations, and the rest are truly safe sites. For the AF, in Period 1, one of the truly hazardous sites (i.e., Site 18) is wrongly identified as safe because its recorded accident count of 11 ranks lower than the top 10% among the 20 accident counts in Period 1. In Period 2, the two sites are correctly identified. Thus, the number of FNs for the AF in both periods is 1. Similarly, the truly safe Site 19 in Period 1 is mistakenly identified as hazardous, and the number of FPs for the AF in both periods is also 1. The corresponding number of total false identifications is then 2. In comparison, for the AR, Site 18 in Period 1 (the associated rate is 1.73) and Site 20 in Period 2 (the associated rate is 3.14) are wrongly identified as safe, so the total number of FNs in the two periods is then 2. It is easy to calculate that the number of FPs and identifications are 2 and 4, respectively. The EB and ARP methods follow the same logic to compute the erroneous identifications.

Test T4. Poisson Mean Differences Test

The false identification test just discussed uses the number of FNs and FPs to assess the performances of various HSID methods. One disadvantage of this test is that each false identification is weighted equally. This disadvantage can be seen through a simple example. If a site with a TPM (underlying true safety) of 15.6 is wrongly

selected for treatment instead of one with a TPM of 15.7, the error is really rather small, whereas, if a site with a TPM of 5.7 is mistakenly selected instead of one with a TPM of 15.7, the error is much more significant. An HSID method that makes the first error might be preferred over one that makes the second. The TPM differences associated with the two false identifications are 0.1 and 10, respectively, a relatively large difference. To obviate this drawback, a Poisson mean differences test (T4) is developed to differentiate various false identifications and to quantify more clearly the adverse consequences that result from erroneous identifications of unequal importance. The test statistic is the sum of the absolute difference of TPMs associated with the falsely identified sites and critical TPMs, which separate the truly safe and hazardous locations. In numerical form, the test is

$$T4_j = \sum_{k=n-n\alpha}^n (|TPM_{k,j}| - |TPM_{k,j+1}|) \quad (8)$$

This test requires that “truth” be known, so it is useful only when HSID methods are being compared in a simulated environment. Some comparisons in the literature have been done on field data and so could not apply this test; in those cases, the previous tests should be used. In this simulation experiment, the truth is known by design and so this test can be applied.

Table 1 shows that the critical TPM of the top 10% truly hazardous road sections is 14. The false identification test results show that the AF results in one FN (Site 18) and one FP (Site 19) in the two periods. Thus, the TPM differences of FNs and FPs of the AF are 0 and 0.67 (absolute value of $13.33 - 14$), respectively. A value of 0 is obtained because the TPM of the FN (Site 18) is used as the critical TPM. This effect diminishes quickly as the number of FNs increases. The total TPM differences of false identifications are then 0.67 ($0 + 0.67$), whereas, for the ARP, Site 18 in Period 1 and Site 20 in Period 2 are FNs and Site 19 in both periods are FPs, so the relative TPM differences of FNs and FPs are 1.33 ($15.33 - 14 + 0$) and 1.34 ($0.67 * 2$), respectively, and the total TPM difference of false identifications is 2.67 ($1.33 + 1.34$). With the same procedure, the TPM differences of FNs, FPs, and false identifications for the EB and the ARP methods are 0, 0.67, and 0.67, respectively.

DATA DESCRIPTION

Data used for comparing HSID methods by means of the new criteria and for developing safety performance functions (SPFs) needed to apply the EB and ARP methods were obtained from the Arizona DOT. The data consisted of both highway and accident data. The highway data contained information on functional classification, speed limit, section length, pavement type, and the like. The accident data contained detailed accident information, such as accident locations, accident times, prior events, involved vehicles, and so on. For each of the 3 years from 2000 to 2002, accident data were extracted for road segments, and intersection-related accidents were excluded.

For the purpose of comparing alternate HSID approaches, the 3-year accident data were separated into two periods, Period 1 (Year 2000) and Period 2 (Years 2001 and 2002). Within the state of Arizona, the highway system is divided into reference sections generally delimited by intersections; thus, the road section lengths vary. To make road sections comparable, the four HSID methods were applied on the basis of accident density (i.e., accidents per kilometer)—equivalent to an offset effect for length. Thus, the AF method uses accidents

TABLE 2 Functional Classification Codes

Rural		Urban	
Code	Description	Code	Description
1	Principal arterial–Interstate	11	Principal arterial–Interstate
2	Principal arterial–other	12	Principal arterial–other
6	Minor arterial		freeways and expressways
7	Major collector	14	Principal arterial–other
8	Minor collector	16	Minor arterial
9	Local	17	Collector
		19	Local

per kilometer-year to identify road sections, the AR employs accidents per million kilometers to flag dangerous sections, and EB and ARP methods rely on accidents per kilometer.

As discussed earlier, safety performance functions are required when the EB and the ARP methods are being assessed. Because design criteria and level of service vary according to the functional classification of the highway, the SPFs (based on Year 2000 accident data) are developed by roadway classification. The functional classification of roadways and corresponding codes used in the database are shown in Table 2.

Summary information for roads of various functional classifications is shown in Table 3 and includes total section lengths of each classification, total accidents, and the like. Shown is the number of road sections—2, 10, and 3, respectively—for the three classifications, rural local (9), urban collector (17), and urban local (19). Due to small sample sizes, SPFs for these three classifications of road sections are not modeled in this paper.

DEVELOPMENT OF SPFS

SPFs should not be judged on their ability to explain the causal factors related to accident occurrence. The main purpose for developing these functions is to provide the expected accident counts for spe-

TABLE 3 Statistics for Roads of Various Functional Classifications

Functional Classification	Number of Sections	Total Lengths (km)	Accidents (2000–2002)	Average AADT
1	403	996.1	8,122	23,810
2	441	1,115.1	7,012	7,603
6	436	1,132.1	5,261	5,483
7	628	1,856.1	5,285	2,637
8	100	365.4	416	684
9	12	22.1	29	3,194
11	207	171.6	11,999	106,338
12	165	199.5	9,557	95,931
14	429	270.6	9,685	17,407
16	164	139.7	2,282	11,499
17	10	7.5	136	4,144
19	3	4.4	13	433

NOTE: Functional classification codes are provided in Table 1.

cific types of road sections that are required to apply the EB and ARP methods; thus, the focus is on crash prediction, not explanation. Due to overdispersion of accidents observed on various classifications of road segments in Arizona, negative binomial regression models were fit to generate SPFs.

To remind the reader, the advantage of using a negative binomial model (34) to model the distribution of crash frequencies is that the Poisson distribution restricts the mean and the variance to be equal [$E(y_i) = \text{Var}(y_i)$]. When this equality does not hold (statistically), the data are said to be underdispersed [$E(y_i) > \text{Var}(y_i)$] or overdispersed [$E(y_i) < \text{Var}(y_i)$]. The negative binomial model allows for overdispersion in that the mean of Poisson counts over sites i is itself gamma distributed and has the following form:

$$\lambda_i = \text{EXP}(\beta x_i + \epsilon_i) = \text{EXP}(\beta x_i) \text{EXP}(\epsilon_i) \tag{9}$$

where $\text{EXP}(\epsilon_i)$ is a gamma-distributed error term with mean 1 and variance $\alpha 2$. The addition of this term allows the variance to differ from the mean as shown in Equation 10:

$$\text{Var}[y_i] = E[y_i][1 + \alpha E[y_i]] = E[y_i] + \alpha E[y_i]^2 \tag{10}$$

It is known that many highway elements (e.g., speed limit, geometry, and surrounding land uses) vary significantly across functional classifications. Accordingly, SPFs were created for nine functional classifications of roads. The model functional form used for the SPFs is given as

$$\lambda_i = \alpha * \text{EXP}(\text{SL}) \beta * (\text{AADT})^\gamma \tag{11}$$

where SL is the independent variable representing the road section length and α , β , and γ are estimated parameters. SL is never zero, and so it is not technically problematic to have a nonzero prediction for the case when SL = 0. Hence, the model form including an intercept is used to develop the SPFs.

Two goodness-of-fit methods were used to assess fit of SPFs (see 34). The R_p^2 statistic (similar to R^2 in linear regression) based on standardized residuals is computed as

$$R_p^2 = 1 - \frac{\sum_{i=1}^n \left[\frac{y_i - \lambda_i}{\sqrt{\lambda_i}} \right]^2}{\sum_{i=1}^n \left[\frac{y_i - \bar{y}}{\sqrt{\bar{y}}} \right]^2} \tag{12}$$

where the numerator is similar to a sum of square errors and the denominator is similar to a total sum of squares. A second method for assessing model fit is the G^2 statistic, for which a better-fitting model will yield smaller G^2 values. G^2 is calculated as

$$G_i^2 = 2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{\lambda_i} \right) \right] \tag{13}$$

The results of SPFs for the nine classifications of road sections are shown in Table 4, including the estimated coefficient values, the associated t -statistics, the overdispersion parameter, and measures of goodness of fit. It is known that all the independent variables are significant with 95% confidence and that the sign and magnitude of the modeled effect agrees with theoretical expectations of the accident process. Some important factors not considered in the models

TABLE 4 Summary of Safety Performance Function Model Results

Functional Classification	Estimated Parameters				Goodness of Fit	
	Intercept	SL	Ln (AADT)	Dispersion	R_p^2	G^2
1	-9.5065 (-11.5175)	0.2583 (18.1505)	1.0427 (13.1583)	3.1840	0.5396	1,230.6
2	-3.9588 (-5.1830)	0.1912 (14.5189)	0.5493 (6.5847)	4.5573	0.3947	1,873.3
6	-6.2632 (-11.5879)	0.2297 (18.2387)	0.7995 (13.3474)	3.2221	0.6319	1,380.6
7	-2.5743 (-5.3246)	0.1955 (16.9021)	0.3518 (5.7967)	4.1135	0.4026	2,058.6
8	-2.3758 (-2.1318)	0.2201 (6.6748)	0.2041 (2.2062)	1.7539	0.3950	141.5
11	-15.0635 (-11.9517)	0.6413 (11.0918)	1.4893 (14.4197)	9.8268	0.7196	1,739.5
12	-16.3219 (-5.6998)	0.1155 (8.5783)	1.6289 (6.7093)	9.7150	0.3374	1,946.3
14	-10.1915 (-7.8556)	0.8251 (8.9443)	1.1243 (8.7093)	10.3190	0.5672	3,646.5
16	-7.7549 (-5.2006)	0.7031 (8.1219)	0.8642 (5.6913)	4.9392	0.4711	681.4

NOTE: SL: Variable of road section length (km). Ln(AADT): Logarithm transformation on variable AADT. Functional classification code is illustrated in Table 1. Values shown in parentheses represent *t*-statistic values associated with various independent variable coefficients.

(e.g., human factors, number of lanes, weather, human error, etc.) probably result in the relatively small R_p^2 statistic values associated with some SPFs.

TEST RESULTS

The five tests described previously (new Tests T1 through T4 and the existing false identification test) are used to assess the relative performance of four commonly applied HSID methods described in the first section.

The evaluation experiment uses the following procedure, which closely mimics how reactive safety management programs are conducted in practice:

1. Road sections (intersections, ramps, two-lane rural roads, etc.) are segregated so that the safety of similar sites can be fairly compared. In this evaluation, the analysis is based on the analysis of nine functional classifications of road sections.
2. For each HSID method, similar road sections are sorted in descending order of estimated safety (noting that the four HSID methods rank sites according to different criteria).
3. Sections with the highest rankings are flagged as hot spots (in practice these sites will be further scrutinized). Typically, a threshold is assigned according to safety funds available for improvement, such as the top 10% of sites. In this evaluation, both the top 10% and 5% of the locations are used as experimental values.

The test results described in the following sections and shown in the associated tables are the accumulated results based on assessment of four HSID methods using five test statistics for nine functional classes of roads in Arizona. The reader wishing more detail on this analysis can refer to the full report (33). The results are discussed in relation to the HSID methods performance on the five statistical tests.

T1 Site Consistency Test Results

From Test T1, it is shown in Table 5 that the AF method outperforms other HSID methods in identifying the top 10% of hot spots with highest average accidents, 9,611, in Period 2, followed closely by the EB method. The EB method performs best in identifying the top 5% of hot spots, with 6,377 accidents occurring at high-risk sites

identified in 2000. The ARP performs the worst in both cases, with the identified hot spots experiencing the lowest number of accidents, say, 7,556 and 4,313, respectively (although the ARP is based on reduction potential, so the total count can be misleading). The AR method also performs poorly, as the sites identified by this method in 2000 produce quite a few accidents in the 2001-to-2003 period.

T2 Method Consistency Test Results

Table 6 shows the number of similarly identified hot spots identified by alternate HSID methods over the two periods. The EB method is superior in this test by identifying the largest number of the same hot spots in both cases of $\alpha = .90$ and $\alpha = .95$, with 167 and 71 sites, respectively. In other words, the EB method identified 167 sites in 2000 that were also identified as hot spots in 2001 to 2003. The AF, which performs slightly better than the ARP method, places 2nd with identifying 148 consistent hot spots (in the case of $\alpha = .90$) and 63 consistent hot spots (in the case of $\alpha = .95$). The ARP performs last, with the lowest number of consistent hot spots identified in the two periods. Again, the EB method outperforms the other HSID methods.

Also shown in Table 6 are differences between percentages (shown in the parentheses) of Column 2 and Column 3 for the four methods. There is a consistent drop in percentages as threshold values drop. The explanation is that the top sites suffer from greater random fluctuations in crashes, and thus the higher is the threshold, the larger are

TABLE 5 Accumulated Results of Site Consistency Test (Test T1) of Various Methods for All Classifications of Highways

Method	$\alpha = 0.90$		$\alpha = 0.95$	
	Accidents 2000	Average Accidents 2001–2002	Accidents 2000	Average Accidents 2001–2002
AF	8,276	9,611	5,639	6,303
AR	6,899	7,556	4,257	4,313
EB	8,123	9,603	5,399	6,377
ARP	7,314	8,450	5,611	6,260

NOTE: ARP—Method of Accident Reduction Potential. $\alpha = 0.90$ and $\alpha = 0.95$ represent the cases of top 10% and 5% hazardous sections.

TABLE 6 Accumulated Results of Method Consistency Test (Test T2) of Various Methods for All Classifications of Highways

Method	$\alpha = 0.90$	$\alpha = 0.95$
EB	167 (56.0%)	71 (47.3%)
AF	148 (50.0%)	63 (42.0%)
AR	131 (44.0%)	49 (32.7%)
ARP	139 (46.6%)	59 (39.3%)

NOTE: $\alpha = 0.90$ and $\alpha = 0.95$ represent the cases of top 10% and 5% hazardous sections. The number represents locations identified by methods in both periods, the percent shown in parenthesis stands for the percentage of consistent hot spots, or the percentage of hot spots identified in 2000 that were also identified in 2001, 2002, and 2003.

the random fluctuations and the likelihood of not being identified in a prior year.

T3 Total Rank Differences Test Results

Table 7 illustrates that the EB method is vastly superior in Test T3. In both the $\alpha = .90$ and $\alpha = .95$ cases, the EB method has significantly smaller-summed ranked differences, by about 50% compared with the AF, 80% compared with the ARP, and by more than 100% compared with the AR. This result suggests that the EB method is the best HSID method (of the four evaluated here) for ranking sites consistently from period to period.

False Identification Test Results

Review of Table 8 reveals the results of the false identification test, which has been presented elsewhere (29). First, the percentage of FNs and FPs show an inverse relationship with α for the four HSID methods. The former decreases with increasing α , whereas the latter increases. This phenomenon suggests that stricter identification criteria (higher α) will result in fewer failures to select truly hazardous sites for remediation but will lead to a larger number of nonhazardous locations identified as hazardous. Second, the false identifications interestingly tend in the same direction as the FNs with the increase of the value of α . The most probable explanation for this phenomenon is that the relatively small number of FNs can lead to more FPs and then reduce the efficiency of the investment of local governments.

TABLE 7 Accumulated Results of Total Rank Differences Tests (Test T3) of Various Methods for All Classifications of Highways

Methods	$\alpha = 0.90$	$\alpha = 0.95$
EB	17,851	10,349
AF	29,602	15,357
AR	34,869	21,212
ARP	32,601	18,787

TABLE 8 Accumulated Results of False Identification Test of Various Methods for All Classifications of Highways

Method		$\alpha = 0.90$	$\alpha = 0.95$
EB	FN	153 (2.9%)	94 (1.7%)
	FP	153 (25.6%)	94 (31.3%)
	FI	314 (5.1%)	188 (3.2%)
AF	FN	152 (2.8%)	91 (1.6%)
	FP	152 (25.5%)	91 (30.3%)
	FI	304 (5.0%)	182 (3.1%)
AR	FN	333 (6.2%)	199 (3.5%)
	FP	333 (55.9%)	199 (66.3%)
	FI	666 (11.2%)	398 (6.7%)
ARP	FN	226 (3.4%)	174 (2.7%)
	FP	226 (30.9%)	174 (50.7%)
	FI	452 (6.2%)	348 (5.1%)

NOTE: FN = false negatives; FP = false positives; FI = false identifications. The number shown in the table is false identifications; the percent in the parenthesis is the corresponding percentage. FN% is defined as the number of FNs divided by the number of sites viewed as safe in various periods; FP% is defined as the number of FPs divided by the number of sites viewed as safe in various periods; FI% is defined as the number of FIs divided by all the number of sites.

In Table 8, the AF appears to top other alternative methods by yielding both the lowest number and rate of FPs and FNs. The EB method trails closely behind the AF, with slightly more (about 1% to 3%) false identifications. The ARP method remains in third place, followed by the AR, which generates the highest false identifications in both the $\alpha = 0.90$ and $\alpha = 0.95$ cases.

As mentioned earlier, the results shown in this test should be interpreted with care due to the relatively small accident history (3 years). Fortunately, a former study (29) used a long accident history by simulating 30 years of data. The results of this prior analysis showed that the EB method consistently outperformed the AF in almost every situation. In consideration of this prior result, it is plausible to expect that the advantage of the AF in this test would diminish with longer accident histories and would be surpassed by the EB method. Regardless, the EB and AF methods perform similarly with this test.

T4 Poisson Mean Differences Test Results

It should be recalled that Test T4 measures the absolute difference in TPM across periods. Large differences in Test T4 represent large differences in ranking across periods and are less desirable (than small differences). Table 9 illustrates that the AF performs best, followed by the EB method, followed by the ARP method, and last the AR method (which is significantly worse than all other methods). The results agree with intuition and logic that total TPM differences are positively related to the number of false identifications. The greater is the number of false identifications, the greater are the total TPM differences.

DISCUSSION AND CONCLUSIONS

Three years of road section accident data from the State of Arizona were used to compare alternate HSID methods, including the AF, the AR, the ARP, and the EB methods for identifying hot spots. Five

TABLE 9 Results of Poisson Mean Differences Test (T4) of Various Methods for All Classifications of Highways

Method		$\alpha = 0.90$	$\alpha = 0.95$
EB	FN	1,141.5 (7.46)	1,097.3 (11.67)
	FP	1,271.3 (8.31)	1,169.7 (12.44)
	FI	2,412.8 (7.68)	2,267.0 (12.06)
AF	FN	1,041.4 (6.85)	858.9 (9.44)
	FP	1,279.0 (8.41)	1,168.8 (12.84)
	FI	2,320.4 (7.63)	2,027.7 (11.14)
AR	FN	4,164.7 (12.54)	3,175.5 (15.96)
	FP	2,549.4 (7.66)	2,378.2 (11.95)
	FI	6,714.1 (10.08)	5,553.7 (13.95)
ARP	FN	1,047.5 (4.63)	1,017.0 (5.84)
	FP	1,465.6 (6.48)	1,358.5 (7.81)
	FI	2,513.1 (5.56)	2,375.5 (6.83)

NOTE: Values in parentheses represent average true Poisson mean differences per erroneous identification.

evaluation tests—four new ones and one that has been used in past research efforts—are applied to evaluate the performance of the four HSID methods. The tests applied include the site consistency test (T1), the method consistency test (T2), the total rank differences test (T3), the false identification test, and the Poisson mean differences test (T4). The intended use of these methods is akin to the selection of statistical models by which multiple criteria are used to select the “best” model, including adjusted R^2 , F -ratio, t -statistics of model variables, signs and magnitudes of coefficients, and mean square error. As in statistical modeling, a model will not be best among all criteria, and the analyst must compare models against a set of criteria and subjectively choose the most appealing model.

These four new tests are argued to offer improved ability (a) to evaluate HSID methods versus comparing FPs and FNs and (b) to provide additional criterion on which to gauge the performance of HSID methods. The four new tests evaluate dimensions of HSID methods heretofore not acknowledged and provide tools for evaluating future HSID methods against one another. The four new methods are unique and represent a contribution to the understanding and evaluation of HSID methods that are integral to safety management programs. All the methods rest on the assumption that in adjacent periods safety has not changed as a result of interventions or significant road user changes, traffic, and like. Tests T1 through T3 can be used to evaluate field or simulated data (where truth is known), whereas Test T4 can be used only on simulated data (where truth is known).

From a practical perspective, four commonly applied HSID methods were evaluated by using the new criteria against a subset of Arizona crash data for nine functional classifications of roads. After evaluation of these four methods, the following conclusions are drawn:

- The AF method outperforms the other three HSID methods on the Test T1 (site consistency), followed very closely by the EB method. That is, the AF and EB methods identify sites in Period 1 that produced the highest number of crashes in Period 2, demonstrating good consistency. The AR method performed the worst.
- The EB method is superior to the other three methods for the Test T2 (method consistency). That is, the EB method consistently identified a larger intersection of sites across observation periods. The AF and the AR method follow the EB method in second and in third place, respectively, while the ARP method performed worst.

- Compared with the Test T2 (method consistency), the Test T3 (total rank differences) revealed pronounced benefits associated with the EB method. The EB method outperformed all competing HSID methods by a wide margin on this criterion, showing great consistency in ranking sites across observation periods. The AR method performed the worst by a large margin.

- In the false identification test, the AF slightly outperforms the EB method, with fewer false identifications, although the differences are small. Again, the AR method performed the worst against this criterion.

- The results of Test T4 (Poisson mean differences) are generally positive relative to the number of false identifications examined in the false identification test. The greater is the number of false identifications, the greater are the false TPM differences. The best-performing HSID method on this criterion is the AF, followed by the EB, and then the ARP, which differed in performance by at most 10%. The AR method performed significantly worse, with a differences as large as 300%.

- Overall, the five tests reveal that the EB method is the most consistent and reliable method for identifying hot spots. Although it was not the best on all measures, in the cases in which it was not the best, it was always similar in performance to the best. On other criteria, the EB outperforms other methods by a wide margin. This evaluation suggests that the EB method (of the methods compared) should remain or become the industry standard.

- In contrast, the AR method performed consistently the worst. This result is quite alarming, as many agencies use this method due to its simplicity for ranking hot spots. In fact, the AF method is better than the AR method in many cases and would be a better choice than the AR. These results discourage the use of the AR method for identifying hot spots.

The results observed in this paper require some caveats. First, only two independent variables are included in the safety performance functions, and as a result, the functional forms might not be appropriate in some cases. Second, the advantages associated with the EB method are obtained based on Arizona accident data, and the relative performances of HSID methods may change when other accident data is used (this result is possible but not expected).

It is intended that road safety professionals can apply the five tests described here to compare the performance of current and future proposed HSID methods. The tests are meant to provide insight into dimensions of hot spot identification that have heretofore not been exploited.

ACKNOWLEDGMENT

This paper is based on research performed through contract with the Arizona Department of Transportation, whose supply of the Arizona Local Government Safety Project model and accident data is gratefully acknowledged, along with support of the research.

REFERENCES

1. *NCHRP Synthesis 322: Safety Management Systems, A Synthesis of Highway Practice*. Transportation Research Board of the National Academies, Washington, D.C., 2003.
2. Deacon, J. A., C. V. Zegeer, and R. C. Deen. Identification of Hazardous Rural Highway Locations. In *Transportation Research Record 543*, TRB, National Research Council, Washington, D.C., 1975, pp. 16–33.

3. Norden, M., J. Orlansky, and H. Jacobs. Application of Statistical Quality-Control Techniques to Analysis of Highway-Accident Data. *Bulletin 117*, HRB, National Research Council, Washington, D.C., 1956, pp. 17–31.
4. Rudy, B. M. Operational Route Analysis. *Bulletin 341*, HRB, National Research Council, Washington, D.C., 1962, pp. 1–17.
5. Morin, D. A. Application of Statistical Concepts to Accident Data. In *Highway Research Record 188*, HRB, National Research Council, Washington, D.C., 1967, pp. 72–79.
6. Stokes, R. W., and M. I. Mutabazi. Rate-Quality Control Method of Identifying Hazardous Road Locations. In *Transportation Research Record 1542*, TRB, National Research Council, Washington, D.C., 1996, pp. 44–48.
7. Laughland, J. C., L. E. Haefner, J. W. Hall, and D. R. Clough. *NCHRP Report 162: Methods for Evaluating Highway Safety Improvements*. TRB, National Research Council, Washington, D.C., 1975.
8. Hauer, E. Bias-by-Selection: Overestimation of the Effectiveness of Safety Countermeasures Caused by the Process of Selection for Treatment. *Accident Analysis and Prevention*, Vol. 12, No. 2, 1980, pp. 113–117.
9. Abbess, C., D. Jarret, and C. C. Wright. Accidents at Blackspots: Estimating the Effectiveness of Remedial Treatment, with Special Reference to the “Regression-to-the-Mean” Effect. *Traffic Engineering and Control*, Vol. 22, No. 10, 1981, pp. 535–542.
10. Hauer, E. On the Estimation of the Expected Number of Accidents. *Accident Analysis and Prevention*, Vol. 18, No. 1, 1986, pp. 1–12.
11. Hauer, E., and B. N. Persaud. How to Estimate the Safety of Rail-Highway Grade Crossings and the Safety Effects of Warning Devices. In *Transportation Research Record 1114*, TRB, National Research Council, Washington, D.C., 1987, pp. 131–140.
12. Hauer, E., J. C. N. Ng, and J. Lovell. Estimation of Safety at Signalized Intersections. In *Transportation Research Record 1185*, TRB, National Research Council, Washington, D.C., 1988, pp. 48–61.
13. Hauer, E., B. N. Persaud, A. Smiley, and D. Duncan. Estimating the Accident Potential of an Ontario Driver. *Accident Analysis and Prevention*, Vol. 23 No. 2–3, 1991, pp. 133–152.
14. Hagle, J. L., and J. M. Witkowski. Bayesian Identification of Hazardous Locations. In *Transportation Research Record 1185*, TRB, National Research Council, Washington, D.C., 1988, pp. 24–36.
15. Hakkert, A. S., and D. Mahalel. Estimating the Number of Accidents at Intersections from a Known Traffic Flow on the Approaches. *Accident Analysis and Prevention*, Vol. 10, No. 1, 1978, pp. 69–79.
16. McGuigan, D. R. D. The Use of Relationships Between Road Accidents and Traffic Flow in “Black-Spot” Identification. *Traffic Engineering and Control*, Aug.–Sept. 1981, pp. 448–453.
17. McGuigan, D. R. D. Nonjunction Accident Rates and their Use in “Black-Spot” Identification. *Traffic Engineering and Control*, Feb. 1982, pp. 45–56.
18. Persaud, B. N., C. Lyon, and T. Nguyen. Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1665*, TRB, National Research Council, Washington, D.C., 1999, pp. 7–12.
19. Heydecker, B. G., and J. Wu. Using the Information in Road Accident Records. *Planning and Transport Research and Computation*, Vol. 35, 1991, pp. 241–255.
20. Kononov, J., and B. N. Jason. Diagnostic Methodology for the Detection of Safety Problems at Intersections. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1784*, Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 51–56.
21. Shankar, V., F. Mannering, and W. Barfield. Effect of Roadway Geometric and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis and Prevention*, Vol. 27, No. 3, 1995, pp. 371–389.
22. Kim, D., S. Washington, and J. Oh. Modeling Crash Outcomes: New Insights into the Effects of Covariates on Crashes at Rural Intersections. *Journal of Transportation Engineering*, Vol. 132, No. 4, 2004, pp. 505–513.
23. Tamburri, T. N., R. N. Smith, J. P. Mills, and V. J. Perini. The Safety Index: Method of Evaluating and Rating Safety Benefits. In *Highway Research Record 332*, Highway Research Board, National Research Council, Washington, D.C., 1970, pp. 28–43.
24. Taylor, J. L., and H. T. Thompson. Identification of Hazardous Locations. FHWA-RD-77-81. FHWA, U.S. Department of Transportation, 1997.
25. Persaud, B. N., and E. Hauer. Comparison of Two Methods for Debiasing Before-and-After Accident Studies. In *Transportation Research Record 975*, TRB, National Research Council, Washington, D.C., 1984, pp. 43–49.
26. Hauer, E., and B. N. Persaud. Problem of Identifying Hazardous Locations Using Accident Data. In *Transportation Research Record 975*, TRB, National Research Council, Washington, D.C., 1984, pp. 36–43.
27. Hagle, J. L., and M. B. Hecht. A Comparison of Techniques for the Identification of Hazardous Locations. In *Transportation Research Record 1238*, TRB, National Research Council, Washington, D.C., 1989, pp. 10–19.
28. Maher, M. J., and L. J. Mountain. The Identification of Accident Blackspots: A Comparison of Current Methods. *Accident Analysis and Prevention*, Vol. 20, No. 2, 1988, pp. 143–151.
29. Cheng, W., and S. Washington. Experimental Evaluation of Hotspot Identification Methods. *Accident Analysis and Prevention*, Vol. 37, No. 5, 2005, pp. 870–881.
30. Hauer, E. *Observational Before-After Studies in Road Safety*. Pergamon Press, Tarrytown, N.Y., 1997.
31. Hauer, E. Empirical Bayes Approach to the Estimation of “Unsafety”: The Multivariate Regression Method. *Accident Analysis and Prevention*, Vol. 24, No. 5, 1992, pp. 457–477.
32. Hauer, E., D. W. Harwood, F. M. Council, and M. S. Griffith. Estimating Safety by the Empirical Bayes Method: A Tutorial. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1784*, Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 126–131.
33. Washington, S., and W. Cheng. *High Risk Crash Analysis*. FHWA-AZ-05-558. Arizona Department of Transportation, Phoenix, Dec. 2005.
34. Washington, S., M. G. Karlaftis, and F. L. Mannering. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall, Washington, D.C., 2003.

The contents are the responsibility of the authors.

The Safety Data, Analysis, and Evaluation Committee sponsored publication of this paper.