Accident Analysis and Prevention 145 (2020) 105684

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Generalized criteria for evaluating hotspot identification methods

Xiaoyu Guo[a], Lingtao Wu[b,*], Dominique Lord[a]

[a] Zachry Department of Civil and Environmental Engineering, Texas A&M University, 3136 TAMU, College Station, TX, 77843-3136, United States
[b] Center for Transportation Safety, Texas A&M Transportation Institute, Texas A&M University System, 3135 TAMU, College Station, TX, 77843-3135, United States

ARTICLE INFO

Keywords:
Network screening
Hotspot identification method
Performance evaluation
Multiple periods

ABSTRACT

Hotspot identification (HSID) is one of the most important components in the highway safety management process. Previous research has found that hazardous sites identified with different methods are not consistent. It is therefore necessary to evaluate the performance of various HSID methods. The existing evaluation criteria are limited to two consecutive periods, and do not consider the temporal instability of crashes. In addition, one existing criterion does not precisely evaluate HSID method under given circumstances. This paper proposed three generalized criteria to evaluate the performance of HSID methods: (1) High Crashes Consistency Test (HCCT) is proposed to evaluate HSID methods in terms of their reliabilities of identifying sites with high crash counts; (2) Common Sites Consistency Test (CSCT) is proposed to gauge HSID methods in consistently identifying a set of common sites as hazardous sites; and, (3) Absolute Rank Differences Test (ARDT) is proposed to measure the consistency of HSID methods in measuring the absolute differences in rankings. Further, three commonly used HSID methods are applied to estimate crashes on Texas rural two-lane roadway segments with eight years of crash data. The performance of these three HSID methods were evaluated to validate the proposed criteria. Comparisons between the existing criteria and the generalized criteria revealed that: (1) the generalized criteria are capable of evaluating different HSID methods over multiple periods; and (2) the generalized criteria are enhanced with a consistent result and with less discrepancy in scores of the best identified HSID method.

1. Introduction

Network screening is the first step in the process of roadway safety management recommended by the American Association of State Highway and Transportation Officials' (AASHTO) *Highway Safety Manual (HSM)*. The primary purpose of network screening is to review the roadway network, rank sites and identify those with higher potentials to be improved with the implementation of a countermeasure or a set of countermeasures. Network screening is also known as hotspot identification (HSID), identification of sites with promise, hazardous locations or high-risk sites (sites that experience greater risk than expected). It is one of most important parts to improve the safety performance of roadway network (Persaud et al., 2010).

Transportation management agencies use various HSID methods. The first edition of the *HSM* provides 13 types of network screening methods, e.g., average crash frequency, crash rate, Equivalent Property Damage Only (EPDO) average crash frequency, relative severity index (RSI), and the empirical Bayes (EB) estimate (AASHTO, 2010). Previous studies have shown that observed crash frequency and crash rate based HSID methods are not reliable because they do not account for the regression-to-the-mean (RTM) bias (Montella, 2010; Wu et al., 2014). In addition to the methods introduced in the *HSM*, safety analysts have proposed new statistical approaches for identifying hotspots. Unfortunately, sites identified with different methods are not consistent (Hauer and Persaud, 1984; Hauer, 1996). In other words, some methods result in inaccurate hotspot identifications, thus leading to two types of errors: (1) false positive (FP, identifying a safe site as hazardous) and (2) false negative (FN, identifying an unsafe site as safe) (Cheng and Washington, 2005; Wu et al., 2014). These two errors can be the results of inaccurate methods and/or random fluctuation of crashes over time. Either of them in identifying hotspots can result in inefficient use of limited investments and additional loss of lives (Persaud et al., 1999; Cheng and Washington, 2005; Wu et al., 2014; Washington et al., 2018; Meng et al., 2020). It is therefore necessary to evaluate the performance of the various HSID methods.

Traditionally, safety analysts have been using statistical measures to assess safety models, such as R^2, Akaike information criterion (AIC), Bayesian information criterion (BIC), and the p-value. However, higher goodness-of-fitting (GOF) measures in crash models do not always guarantee the accuracy in prediction and the performance of practical HSID methods (Geedipally et al., 2014). Hence, there is a need to develop robust, informative, quantitative and qualitative criteria for

* Corresponding author.
E-mail addresses: xiaoyuguo@tamu.edu (X. Guo), l-wu@tti.tamu.edu (L. Wu), d-lord@tamu.edu (D. Lord).

https://doi.org/10.1016/j.aap.2020.105684
Received 20 February 2020; Received in revised form 3 July 2020; Accepted 6 July 2020
Available online 13 August 2020
0001-4575/

evaluating the performance of HSID methods. To the authors' knowledge, the first criteria for assessing the performance of HSID method were developed by Cheng and Washington (2005) and Cheng and Washington (2008). Cheng and Washington (2008) proposed three tests for conducting performance assessments of alternative HSID methods based on field data: (1) site consistency test; (2) method consistency test; and, (3) total rank differences test. The tests have been heavily used by safety researchers in the identification of hotspots and the evaluation of the performance (Montella, 2010; Wu et al., 2014). However, researchers have reported that one of the criteria (i.e., total rank differences test) does not precisely evaluate HSID method under certain circumstances (Guo et al., 2019; Meng et al., 2020). In addition, the approach is only capable of assessing the performance of HSID methods in two periods. Recent studies have reported that temporal instability could negatively affect traditional safety analyses (Mannering, 2018). Without accounting for the instability of crashes, the HSID results may be inconsistent and/or inaccurate. Aggregating crash data into only two periods can potentially reduce the accuracy of HSID performance evaluation. Considering the importance of both network screening and HSID method performance assessment, it is necessary to generalize the criteria (i.e., three tests) to handle multiple periods. Thus, the primary objective of this paper is to improve those tests initially proposed by Cheng and Washington (2008).

To achieve the objectives, this paper first describes the criteria and test methods developed by Cheng and Washington (2008). Then, it documents the proposed improvements and generalization. To validate the improvements, this paper analyses eight years of crash data that were collected on 18,154 rural two-lane roadway segments with three HISD methods (i.e., crash rate, EB, and ratio between EB and prediction). In conclusion, this study evaluates the performance of HSID methods using the generalized criteria, and compares them to the performance evaluated by the existing criteria documented in Cheng and Washington (2008).

2. Methodology

This section first introduces the existing criteria developed by Cheng and Washington (2008). Then, the section describes the generalized criteria, which include three quantitative tests: (1) High Crashes Consistency Test; (2) Common Sites Consistency Test; and, (3) Absolute Rank Differences Test. Finally, it briefly documents three representative HSID methods that are evaluated by the existing and proposed criteria in this study.

2.1. Existing evaluation criteria

Cheng and Washington (2008) introduced three criteria (i.e., site consistency test, method consistency test, and total rank differences test) to gauge the performance of HSID methods on field data. The site consistency test (T1) was designed to measure the ability of a HSID method to identify a site as high-risked consistently over two adjacent periods. In a total of n sites, a Method m identifies the high-risk sites S_j by ordering the estimated crash rates or means from $\hat{y}_{m_{(n)}}$ to $\hat{y}_{m_{(n-\alpha n+1)}}$ during Period 1 (i.e., $P = 1$). The expressions of $\hat{y}_{m_{(n)}}$, $\hat{y}_{m_{(n-1)}}$, ...,$\hat{y}_{m_{(n-\alpha n+1)}}$ are following the notation of the order statistics. For example, $x_{(3)} = \max(x_1, x_2, x_3)$, $x_{(1)} = \min(x_1, x_2, x_3)$ for some x. The subscript of m indicates that the estimated crash rates or means are from the Method m. The notation α is the threshold of identified high-risk sites. Then, T1 compares the summation of $C_{S_j, P=2}$, the crash count for identified high-risk sites S_j at the subsequent observation period (i.e., $P = 2$), among HSID methods. For example, a Method 1 (i.e., $m = 1$) is better than any other methods in T1, when $T1_{m=1} > T1_{m\neq 1}$. The mathematical expression of this test is given as Eq. 1,

$$T1_m = \sum_{j=1}^{j=\alpha n} C_{S_j, P=2};$$
$$for\ S_j = \{S_1|\hat{y}_{m_{(n)}}, S_2|\hat{y}_{m_{(n-1)}}, ...,S_{\alpha n}|\hat{y}_{m_{(n-\alpha n+1)}}\}_{P=1} \quad (1)$$

In T1, observed crash counts in Period 2 are used as the score to compare different HSID methods. In contrast, the method consistency test (T2) considered the consistency of the method on selecting the same numbers of high-sites over two adjacent periods. Let a Method m identify a set of high-risk sites in Period 1 by ordering the estimated crash rates or means, $\{S_1|\hat{y}_{m_{(n)}}, S_2|\hat{y}_{m_{(n-1)}}, ...,S_{\alpha n}|\hat{y}_{m_{(n-\alpha n+1)}}\}_{P=1}$, or simply $\{S_1, S_2, ..., S_{\alpha n}\}_{P=1}$, and identify another set of high-risk sites in Period 2, $\{S_1, S_2, ..., S_{\alpha n}\}_{P=2}$. The numbers of elements in the intersection set between $\{S_1, S_2, ..., S_{\alpha n}\}_{P=1}$ and $\{S_1, S_2, ..., S_{\alpha n}\}_{P=2}$, is the score of T2. A Method 1 (i.e., $m = 1$) is better than any other methods in T2, when $T2_{m=1} > T2_{m\neq 1}$. The mathematical expression of T2 is shown in Eq. (2).

$$T2_m = |\{S_1, S_2, ..., S_{\alpha n}\}_{P=1} \bigcap \{S_1, S_2, ..., S_{\alpha n}\}_{P=2}|;$$
$$for S_1|\hat{y}_{m_{(n)}}, S_2|\hat{y}_{m_{(n-1)}}, ...,S_{\alpha n}|\hat{y}_{m_{(n-\alpha n+1)}} \quad (2)$$

The third test, the total rank differences test (T3), was introduced to compare the summation of ranking differences in high-risked sites identified between Period 1 and Period 2. For example, in a Method m, there are identified high-risk sites S_j in Period 1, or explicitly $\{S_1|\hat{y}_{m_{(n)}}, S_2|\hat{y}_{m_{(n-1)}}, ...,S_{\alpha n}|\hat{y}_{m_{(n-\alpha n+1)}}\}_{P=1}$. They are with ranks from 1 to αn in Period 1. With these identified high-risk sites S_j, their corresponding ranks in Period 2 are $\Re(S_j)_{P=2}$. The variations of the ranks are then summed as the output of this method as shown in Eq. (3):

$$T3_m = \sum_{j=1}^{j=n\alpha} (j - \Re(S_j)_{P=2});$$
$$for\ S_j = \left\{S_1|\hat{y}_{m_{(n)}}, S_2|\hat{y}_{m_{(n-1)}}, ..., S_{\alpha n}|\hat{y}_{m_{(n-\alpha n+1)}}\right\}_{P=1} \quad (3)$$

Because it is a measure on variations of ranking, a smaller output indicates a better performance. For example, a Method 1 (i.e., $m = 1$) is better than any other methods in T3, when $T3_{m=1} < T3_{m\neq 1}$.

These three tests have been used by many safety researchers in the identification of hotspots and the evaluation of the performance (Montella, 2010; Wu et al., 2014; Ferreira and Couto, 2015). However, some researchers have reported that one of the criteria (i.e., total rank differences test) does not precisely evaluate HSID method under certain circumstances (Guo et al., 2019). For instance, the existing T3 criterion assigns a score of 0 for both of the following two cases:

- Case 1:
  - o Site ID #1 ranks as top 1 in Period 1 and Period 2,
  - o Site ID #3 ranks as top 2 in Period 1 and Period 2;
- Case 2:
  - o Site ID #3 ranks as top 1 in Period 1, and as top 2 in Period 2,
  - o Site ID #1 ranks as top 2 in Period 1, and as top 1 in Period 2.

In Case 1, the top two hazardous sites' ID identified in Period 1 are site ID 1 and site ID 3 (i.e., $\Re(Site_{ID_1}) = 1$, $\Re(Site_{ID_3}) = 2$), which rank exactly the same in Period 2 as $\Re(Site_{ID_1}) = 1$, $\Re(Site_{ID_3}) = 2$. The existing T3 criterion scores 0 for Case 1, by $(1 - 1) + (2 - 2) = 0$. In Case 2, the top two hazardous sites' ID identified in Period 1 are site ID 3 and site ID 1 (i.e., $\Re(Site_{ID_3}) = 1$, $\Re(Site_{ID_1}) = 2$), which rank in Period 2 as $\Re(Site_{ID_3}) = 2$, $\Re(Site_{ID_1}) = 1$. But the existing T3 criterion scores the same for Case 2, as 0, by $(2 - 1) + (1 - 2) = 0$. This difference in scores is crucial. When a practitioner implements the criterion T3 with a score of 0, it is unable to precisely evaluate the ranking differences occur in which of the above 2 cases. Thus, in order to overcome this shortcoming, a generalized criterion (Absolute Rank Differences Test) is inspired from the existing T3 and is introduced in a following subsection.

Moreover, all three existing criteria are only capable of assessing the performance of HSID methods in two periods (i.e., Period 1 and Period

2). Aggregating crash data into only two periods can potentially reduce the accuracy of HSID performance evaluation. For instance, Cheng and Washington (2008) utilized three years (Years 2000–2002) crash data and aggregated them into two periods: Period 1 (Year 2000) and Period 2 (Year 2001-Year 2002); Ferreira and Couto, 2015 applied five years (Years 2001–2005) crash data and aggregated them into Period 1 (Year 2001-Year 2003) and Period 2 (Year 2004 -Year 2005). Both the HSID estimated crashes and HSID performance measurement results may vary, depending on how a study aggregates multiple years of crash data into two periods. Hence, it is necessary to generalize all three criteria (i.e., T1, T2, and T3) to directly handle multiple periods, instead of aggregating into two periods.

It is worth mentioning that Cheng and Washington (2008) proposed another test named "T4 Poisson Mean Differences Test," which only applies to simulated data (also known as artificial realistic data). Although a few researchers have utilized this test with real-world data (Lan and Persaud, 2011) and with certain assumptions, the application is not recommended. The mechanism and algorithm for evaluating HSID methods are different between real-world data and simulated data, as the "true" safety (i.e., Poisson mean) is unknown with the former dataset. This study focuses on real-world datasets, hence T4 was not used in the analyses.

2.2. High crashes consistency test (HCCT)

The proposed High Crashes Consistency Test (HCCT) is inspired by the T1 in Cheng and Washington (2008). It overcomes the limitations of T1, that is generalizing the capability of the existing test in literature by handling multiple HSID methods and multiple periods at the same time. It is proposed to evaluate HSID methods by its ability to consistently capture high-risk sites associated with high crash counts. The test first identifies high-risk sites based on the estimates in each HSID method, then sums the crash counts associated with these identified high-risk sites at each observation period; finally, it evaluates the HSID methods by comparing the averages of these summations over all observation periods. The mathematical expression, Eq. (4), can be used to calculate the score for the HCCT:

$$HCCT_m = \frac{\sum_{d=i+1}^{d=f}\left[\sum_{j=1}^{j=\alpha n} C_{S_j,\,P=d}\right]}{f-i};$$
$$for\ S_j \in \{S_1|\hat{y}_{m(n)},\ S_2|\hat{y}_{m(n-1)},\ ...,S_{\alpha n}|\hat{y}_{m(n-\alpha n+1)}\}_{P=i},$$
$$i \in \{1,\ 2,...,f-1\},\ m \in \{1,\ 2,\ ...\}, \quad (4)$$

Where $HCCT_m$ is the HCCT score of a Method m; S_j is in a set of identified high-risk sites by ordering the estimated crash rates or means $\hat{y}$ during the initial period (i.e., $P = i$); $C_{S_j,\,P}$ is the crash count for an identified high-risk site S_j at a future observation period (i.e., $P = d$). In Eq. (4), P is an index for observation period starting with an initial Period i and ending with a Period f; i is an index for the initial observation period, which can be any period between Period 1 and Period $f-1$; d is an index for the future observation period, which counts from Period $i+1$ to Period f; j is the count for the high-risk sites, from 1 to αn, which the notation n indicates the total number of the sites, the notation α is the threshold of identified high-risk sites within the total of n sites; and m is an index for HSID methods.

To better understand how this test works, the authors created a small artificial dataset (i.e., a total of 7 sites), shown in Table 1, with observed crash count per site over three consecutive periods (i.e., Period 1, Period 2, and Period 3). In addition, the authors also made-up values for three HSID methods (i.e., M1, M2, and M3) with an estimated crash value associated at each site over every period. A two-step process is described to guide readers to understand the implementation of HCCT.

2.2.1. Step 1: Computing a HCCT score for each HSID method over multiple periods

Step 1.1: Rank sites based on the estimates of one method (e.g., M1) from the highest to the lowest in a period (e.g., Period 1), and identify the top two hazardous sites' ID. The highest two sites in the rank of M1 in Period 1 are site 3 and site 1, with safety estimates of 0.75 and 0.72, respectively.

Step 1.2: Sum the observed crash counts in Period 2 and Period 3 for those identified high-risk sites in Period 1. For example, the summation of crash counts observed at site ID 1 and ID 3 during Period 2 is 31 (i.e., 17 plus 14). Then, the HCCT score for M1 in Period 2 is 31.

Step 1.3: Repeat the above two sub-steps for all observation periods. The scores are 31 and 18 for M1 as in Period 2 and Period 3, as illustrated in Table 2. The last sub-step is to take the average of the scores over all the periods, and round the average to the whole number as the score of HCCT for a HSID method. Taking M1 as an example, its HCCT score is 25 (please see the last row of Table 2).

2.2.2. Step 2: Comparing HCCT scores across all HSID methods

Repeat the HCCT score (i.e., step 1) for all HSID methods. In the illustration example, the HCCT score for the three methods (i.e., M1, M2 and M3) are 25, 16, and 25, respectively. Since the HCCT is designed to evaluate HSID methods by its ability to capture high-risk sites associated with high crash counts, a higher HCCT score indicates a better HSID method performance. In this example, M1 and M3 are better than M2.

2.3. Common sites consistency test (CSCT)

The Common Sites Consistency Test (CSCT) looks at the consistency of high-risk sites identified by a HSID method over multiple time periods, rather than being site-based. This test was developed from the T2 in Cheng and Washington (2008). This paper further generalizes the T2 in the capability of handling multiple HSID methods and periods at the same time. CSCT is developed to evaluate HSID methods by its ability to consistently identify a number of common sites as high-risk sites. The test first identifies a set of high-risk sites per evaluation period based on the estimates of each HSID method, then it determines the common sites among the sets. Finally, it counts the number (i.e., cardinality) of common sites contained and evaluates HSID methods by comparing cardinalities among all methods. The mathematical expression, Eq. (5), is used to calculate the score of the CSCT:

$$CSCT_m = \frac{\sum_{d=i+1}^{d=f} |\{S_1,\ S_2,...,\ S_{\alpha n}\}_{P=i} \cap \{K_1,\ K_2,...,\ K_{\alpha n}\}_{P=d}|}{f-i};\{K_1|\hat{y}_{m(n)},\ K_2|\hat{y}_{m(n-1)},\ ...,K_{\alpha n}|\hat{y}_{m(n-\alpha n+1)}\}_{P=d},$$
$$for\ \{S_1|\hat{y}_{m(n)},\ S_2|\hat{y}_{m(n-1)},\ ...,S_{\alpha n}|\hat{y}_{m(n-\alpha n+1)}\}_{P=i},\quad i \in \{1,\ 2,...,f-1\},\ m \in \{1,\ 2,\ ...\},$$

Where $CSCT_m$ is the CSCT score of a Method m; $\{S_1|\hat{y}_{m(n)},\ S_2|\hat{y}_{m(n-1)},\ ...,S_{\alpha n}|\hat{y}_{m(n-\alpha n+1)}\}_{P=i}$, or simply $\{S_1,\ S_2,...,S_{\alpha n}\}_{P=i}$, is a set of identified high-risk sites by ordering the estimated crash rates or means $\hat{y}$ during the initial period (i.e., $P = i$); $\{K_1,\ K_2,...,K_{\alpha n}\}_{P=d}$ is another set of identified high-risk sites by ordering $\hat{y}$ in a future observation period (i.e., $P = d$). In Eq. (5), P is an index for observation period starting with an initial Period i and ending with a Period f; i is an index for the initial observation period, which can be any period between Period 1 and Period $f-1$; d is an index for the future observation period, which counts from Period $i+1$ to Period f; n

Table 1
An Example with a Need to Identify Top 2 Hazardous Sites.

Site ID*	Observed Crash Count			Estimates by M1			Estimates by M2			Estimates by M3		
	Period 1	Period 2	Period 3	Period 1	Period 2	Period 3	Period 1	Period 2	Period 3	Period 1	Period 2	Period 3
1	11	14	11	0.72	0.96	0.72	1.81	3.84	2.88	1.37	1.74	1.31
2	7	5	12	0.33	0.44	0.33	1.83	2.44	1.83	0.73	0.97	0.67
3	10	17	7	0.75	0.56	0.14	6.12	1.5	0.01	0.75	2.4	0.01
4	5	7	5	0.27	0.45	0.18	1.35	2.34	0.81	0.64	1.88	0.36
5	5	9	3	0.18	0.24	0.18	1.29	1.72	1.29	0.51	0.78	0.61
6	4	5	4	0.21	0.14	0.35	1.17	1.56	1.17	0.46	0.6	0.45
7	3	4	3	0.18	0.24	0.18	1.85	1.48	0.37	0.77	0.59	0.14

Note: for illustration purpose, the numbers are artificially created.
* Ranked by site ID.

Table 2
Scores of High Crashes Consistency Test (HCCT) for M1.

Period 1			Period 2			Period 3		
Site ID*	Observed Crash Count	Estimates by M1	Site ID*	Observed Crash Count	Estimates by M1	Site ID*	Observed Crash Count	Estimates by M1
3	10	0.75	3	17	0.56	3	7	0.14
1	11	0.72	1	14	0.96	1	11	0.72
2	7	0.33	2	5	0.44	2	12	0.33
4	5	0.27	4	7	0.45	4	5	0.18
6	4	0.21	6	5	0.14	6	4	0.35
5	5	0.18	5	9	0.24	5	3	0.18
7	3	0.18	7	4	0.24	7	3	0.18
			17 + 14 = 31			7 + 11 = 18		

HCCT Score for HSID Method M1 $= \frac{31+18}{3-1} \approx 25$

Note: for illustration purpose, the numbers are artificially created.
* Ranked by the value of estimates by M1 in Period 1.

indicates the total number of the sites; α is the threshold of identified high-risk sites within the total of n sites; and m is an index for HSID methods. Further, the vertical bar in the right-hand side of the Eq. (5) is denoted for the cardinality of all sets.

To better apprehend how the CSCT works, this paper again showcases a two-step process using the made-up example in Table 1 with observed crash count per site over three consecutive periods (i.e., Period 1, Period 2 and Period 3).

2.3.1. Step 1: Computing a CSCT score for each HSID method over multiple periods

Step 1.1: Rank the sites by estimates of a method (e.g., M1) from the highest to lowest in one period (e.g., Period 1), and find the top two hazardous sites' ID with the highest estimates. The highest two sites in the rank of M1 in Period 1 are site ID 3 and ID 1, which creates a set for M1 in Period 1, $\{Site_{ID_3}, Site_{ID_1}\}$.

Step 1.2: Repeat Step 1.1 for all other evaluation periods, the set for M1 in Period 2 is $\{Site_{ID_1}, Site_{ID_3}\}$ and the set for M1 in Period 3 is $\{Site_{ID_1}, Site_{ID_6}\}$. Then, Sub-Step 2 is used to determine the common sites of all sets associated with all evaluation periods. For example, for M1, the common sites are in $\{Site_{ID_3}, Site_{ID_1}\} = \{Site_{ID_3}, Site_{ID_1}\} \cap \{Site_{ID_1}, Site_{ID_3}\}$. The cardinality of $\{Site_{ID_3}, Site_{ID_1}\}$ is 2.

Step 1.3: Repeat Step 1.2 for all remaining periods. The scores are obtained for M1 in Period 2 and Period 3, as illustrated in Table 3. The last sub step is to take an average of the CSCT scores over all the periods, and round the average to the whole number as the score of CSCT for a HSID method. Taking M1 as an example, its CSCT score is 2 (please see the last row of Table 3). Details of the calculation are shown in Table 3.

2.3.2. Step 2: Comparing CSCT scores with all HSID methods

Calculate the CSCT scores for all the HSID methods. In the example, the CSCT scores for the three methods are 2, 0, and 1. Because the CSCT is designed to evaluate HSID methods by its ability to consistently have some common sites identified as high-risk sites, a higher CSCT score indicates better HSID method performance. In the example, M1 is the best performed HSID method.

2.4. Absolute rank differences test (ARDT)

The Absolute Rank Differences Test (ARDT) evaluates the ability of HSID methods to rank sites steadily by summing the absolute rank differences (regardless differences are positive or negative) over multiple periods. This test is inspired by the T3 proposed by Cheng and Washington (2008). More than what T3 does, the ARDT considers the absolute rank differences and improves the capability of handling multiple periods at once. The ARDT first identifies the ranks with their associated site ID for an initial period by a HSID method. Then, it determines the ranks of those site with the same ID in a future period, and calculates absolute differences between ranks in the two periods. Finally, it evaluates HSID methods by comparing the averages of these summations over all absolute rank differences. The mathematical expression, Eq. (6), is used to calculate the score for the ARDT:

$$ARDT_m = \frac{\sum_{d=i+1}^{d=f} \left[\sum_{j=1}^{j=\alpha n} abs\left(j - \Re(K_l|_{K_l=S_j})\right)\right]}{f-i};$$
$$for\ S_j \in \left\{S_1|\hat{y}_{m(n)}, S_2|\hat{y}_{m(n-1)}, \ldots, S_{\alpha n}|\hat{y}_{m(n-\alpha n+1)}\right\}_{P=i},$$
$$K_l \in \left\{K_1|\hat{y}_{m(n)}, K_2|\hat{y}_{m(n-1)}, \ldots, K_{\alpha n}|\hat{y}_{m(n-\alpha n+1)}, \ldots, K_n|\hat{y}_{(1)}\right\}_{P=d},\ i$$
$$\in \{1,2,\ldots,f-1\},\ m \in \{1,2,\ldots\}, \quad (6)$$

Where $ARDT_m$ is the ARDT of a Method m; S_j is in $\{S_1|\hat{y}_{m(n)}, S_2|\hat{y}_{m(n-1)}, \ldots, S_{\alpha n}|\hat{y}_{m(n-\alpha n+1)}\}_{P=i}$, a set of identified high-risk sites by ordering the estimated crash rates or means $\hat{y}$ during the initial period (i.e., $P = i$); K_l is in a set including all number of sites, $\{K_1|\hat{y}_{m(n)}, K_2|\hat{y}_{m(n-1)}, \ldots, K_{\alpha n}|\hat{y}_{m(n-\alpha n+1)}, \ldots, K_n|\hat{y}_{(1)}\}_{P=d}$ ordered by the

Table 3
Scores of Common Site Consistency Test (CSCT) for M1.

Period 1			Period 2			Period 3		
Site ID*	Observed Crash Count	Estimates by M1	Site ID**	Observed Crash Count	Estimates by M1	Site ID***	Observed Crash Count	Estimates by M1
3	10	0.75	1	14	0.96	1	11	0.72
1	11	0.72	3	17	0.56	6	4	0.35
2	3	0.33	4	7	0.45	2	12	0.33
4	7	0.27	2	5	0.44	4	5	0.18
6	5	0.21	7	4	0.24	7	3	0.18
7	5	0.18	5	9	0.24	5	3	0.18
5	4	0.18	6	5	0.14	3	7	0.14
$\{Site_{ID_3}, Site_{ID_1}\}$			$\{Site_{ID_1}, Site_{ID_3}\}$			$\{Site_{ID_1}, Site_{ID_6}\}$		
			$\lvert\{Site_{ID_3}, Site_{ID_1}\} \cap \{Site_{ID_1}, Site_{ID_3}\}\rvert = 2$			$\lvert\{Site_{ID_3}, Site_{ID_1}\} \cap \{Site_{ID_1}, Site_{ID_6}\}\rvert = 1$		
CSCT Score for HSID Method M1 $= \frac{2+1}{3-1} \approx 2$								

Note: for illustration purpose, the numbers are artificially created.
* Ranked by the value of estimates by M1 in Period 1.
** Ranked by the value of estimates by M1 in Period 2.
*** Ranked by the value of estimates by M1 in Period 3.

value of $\hat{y}$ in a future observation period (i.e., $P = d$). In Eq. (6), P is an index for observation period starting with an initial Period i and ending with a Period f; i is an index for the initial observation period, which can be any period between Period 1 and Period $f - 1$; d is an index for the future observation period, which counts from Period $i + 1$ to Period f; j is the count for the high-risk sites, from 1 to αn, which the notation n indicates the total number of the sites, the notation α is the threshold of identified high-risk sites within the total of n sites; l is the count for the sites, from 1 to n ; and m is an index for HSID methods. The notation $abs(\,)$ indicates the absolute value of a number. The notation $\Re$ indicates the rank of a site. The notation $K_l|_{Kl=S_j}$ indicates the site K_l given that the site ID of K_l is equal to the site ID of S_j.

To better understand how ARDT is applied, this paper showcases a two-step process using artificial data described in Table 1 with observed crash counts per site over three consecutive periods (i.e., Period 1, Period 2 and Period 3).

2.4.1. Step 1: Computing an ARDT score for each HSID method over multiple periods

Step 1.1: Rank the sites by estimates of a method (e.g., M1) from the highest to lowest in the initial period (e.g., Period 1), and find the corresponding ranks in Period 2 of the top two hazardous sites' ID identified in Period 1. The top two hazardous sites' ID identified in Period 1 are site ID 3 and site ID 1, which rank in Period 2 as $\Re(Site_{ID_3}) = 2$, $\Re(Site_{ID_1}) = 1$, respectively.

Step 1.2: Subtract the ranks between initial period (Period 1) and the following period (Period 2), obtain the absolute differences, and sum these differences. That is, $|2 - 1| + |1 - 2| = 2$.

Step 1.3: Repeat the above two steps for all evaluation periods, the corresponding ranks in Period 3 of the top two hazardous sites' ID identified in Period 1, which is $\Re(Site_{ID_3}) = 7$, $\Re(Site_{ID_1}) = 1$, respectively. The sum of the absolute differences is $|7 - 1| + |1 - 2| = 7$. Thus, the ARDT score for M1 is the average of the two scores (i.e., $(2 + 7)/2 \approx 5$). Details are illustrated in Table 4.

2.4.2. Step 2: Comparing ARDT scores across all HSID method

Repeat step 1 over all HSID methods to compute their ARDT scores. In the example, the ARDT test scores for M1, M2, M3 are 5, 10, and 6 respectively. Because this ARDT is designed to evaluate HSID methods by its ability to have high-risk sites identified with consistent ranks, a smaller ARDT score indicates a better HSID performance. In the example, M1 is the best performed HSID method.

Moreover, this paper considers that the proposed ARDT with an absolute value of the differences is more precise on measuring the ranking differences than the existing criterion proposed by Cheng and Washington (2008), the total rank differences test (i.e., T3).

Using the same two cases example (i.e., Table 5) given in the Existing Evaluation Criteria subsection, the existing criterion (i.e., T3) scores 0 for both Case 1 and Case 2; while the generalized criterion ARDT provides a score of 0 for Case 1 and a score of 2 for Case 2. This shows that, the proposed criterion, ARDT overcomes the shortcoming of T3. ARDT not only precisely identifies Case 1 from Case 2, but also correctly assigns a smaller ARDT score for Case 2, because Case 1 contains a smaller rank difference than Case 2.

2.5. HSID methods

The previous three subsections have proposed three generalized criteria for evaluating hotspot identification methods. This subsection introduces some widely implemented HSID methods, which serve as examples to be evaluated with the criteria. It is important to note that safety analysts and practitioners have proposed a number of network screening and HSID methods. For example, in the first edition of *HSM*, there are 13 HSID approaches (curious readers are referred to *HSM* Chapter 4 Section 4.4.2). Each approach has its own assumptions, advantages and limitations. It is worth noting that most previous HSID studies (e.g., Montella, 2010) assumed that there is an underlying assumption that the safety levels of investigated sites being investigated remain unchanged during the study period. This assumption is in line with the existing evaluation criteria proposed by Cheng and Washington (2008). In the case when the safety level of some site(s) has changed substantially (for example, a certain improvement projects are implemented), the common HSID methods are not applicable (because a hotspot may become a safer site after the improvement). The purpose of this study is not to describe or assess the performance of all the HSID methods. Instead, the primary objective is to apply the three proposed HSID performance evaluation tests (i.e., generalized criteria) in evaluating a few HSID methods. Thus, this section mainly focuses on three frequently used HSID methods: crash rate based, EB based, and ratio based HSID methods. The former two are adopted from the *HSM*. Although previous studies have pointed out that crash rate based HSID method is not reliable, it is still included in this study for comparison purposes. The last HSID method (i.e., ratio based) has been recently used by practitioners and showed superiority (Wunderlich et al., 2019). The three methods are briefly discussed below.

2.5.1. Crash rate based HSID approach

The crash rate based method calculates the rate of crashes at each site, and ranks the sites by their rates. It is usually computed by dividing the observed crash number by exposure (e.g., traffic volume or vehicle

Table 4
Scores of Absolute Rank Differences Test (ARDT) for M1.

Period 1			Period 2			Period 3		
Site ID*	Observed Crash Count	Estimates by M1	Site ID**	Observed Crash Count	Estimates by M1	Site ID***	Observed Crash Count	Estimates by M1
3	10	0.75	1	14	0.96	1	11	0.72
1	11	0.72	3	17	0.56	6	4	0.35
2	3	0.33	4	7	0.45	2	12	0.33
4	7	0.27	2	5	0.44	4	5	0.18
6	5	0.21	7	4	0.24	7	3	0.18
7	5	0.18	5	9	0.24	5	3	0.18
5	4	0.18	6	5	0.14	3	7	0.14
$\Re(Site_{ID_3}) = 1,\ \Re(Site_{ID_1}) = 2$			$\Re(Site_{ID_3}) = 2,\ \Re(Site_{ID_1}) = 1$			$\Re(Site_{ID_3}) = 7,\ \Re(Site_{ID_1}) = 1$		
			$\lvert 2-1\rvert + \lvert 1-2\rvert = 2$			$\lvert 7-1\rvert + \lvert 1-2\rvert = 7$		
ARDT Score for HSID Method M1 $= \frac{2+7}{3-1} \approx 5$								

Note: for illustration purpose, the numbers are artificially created.
* Ranked by the value of estimates by M1 in Period 1.
** Ranked by the value of estimates by M1 in Period 2.
*** Ranked by the value of estimates by M1 in Period 3.

Table 5
Scores Comparison between Existing T3 and Absolute Rank Differences Test (ARDT).

Case 1			Case 2		
	Period 1	Period 2		Period 1	Period 2
Rank #1	Site ID #1	Site ID #1	Rank #1	Site ID #3	Site ID #1
Rank #2	Site ID #3	Site ID #3	Rank #2	Site ID #1	Site ID #3
…	…	..	..	..	..
T3 Score $= (1-1) + (2-2) = 0$			T3 Score $= (1-2) + (2-1) = 0$		
ARDT Score $= \lvert 1-1\rvert + \lvert 2-2\rvert = 0$			ARDT Score $= \lvert 1-2\rvert + \lvert 2-1\rvert = 2$		

miles of travelling). This study considers the number of total crashes, and the exposure as traffic volume travelling through a site. Therefore, the crash rate for a site is calculated as:

$$\text{Crash Rate}_{i,j} = \frac{Y_{i,j}}{\text{Exposure}_{i,j} \times N} \times 10^5 \tag{7}$$

Where $\text{Crash Rate}_{i,j}$ indicates crash rate at site i in period j; $Y_{i,j}$ is total number of observed crashes at site i in period j; N means number of years in period j; and, $\text{Exposure}_{i,j}$ indicates the exposure (traffic volume or vehicle miles of travelling) at site i in period j.

Since the crash rate relies simply on the number of observed crashes and exposure at the sites, the randomness of crashes and RTM bias are not well addressed. Furthermore, the crash rate does not account for the non-linear relationship between crashes and traffic flow (Hauer, 1997). It has been shown that the HSID results using crash rates are not reliable (Cheng and Washington, 2008; Elvik, 2008; Montella, 2010).

2.5.2. EB based HSID approach

The EB approach is proposed by safety researchers to correct for the RTM bias (Hauer, 1992, 1997; Hauer et al., 2002); note, as discussed by Lord and Kuo (2012), the EB method may not properly address the site selection bias. With the EB approach, the long-term safety estimate of a site is obtained from two sources: observed number of crashes and predicted number of crashes. Let Y be the observed number of crashes, and assume it is Poisson distributed; let k be the expected crash number; the EB estimate of k is given as:

$$\hat{k} = w \times \mu + (1-w) \times Y \tag{8}$$

Where $\hat{k}$ denotes the EB estimate (i.e., expected number of crashes); w is the weight factor, which will be discussed in detail later; and, μ is the predicted number of crashes, which is calculated by a safety performance function (SPF). Many statistical models have been proposed to develop SPFs (Lord and Mannering, 2010; Zou et al., 2013; Park et al., 2016; Wu et al., 2017), for instance, the negative binomial (NB), the Poisson-lognormal (Miranda-Moreno et al., 2005; Lord and Miranda-Moreno, 2008), the Conway-Maxwell-Poisson (Lord et al., 2008, 2010), the gamma (Oh et al., 2006), the Sichel (Wu et al., 2014), the Poisson-Tweedie (Saha et al., 2020), the negative binomial-Lindley models (Geedipally et al., 2012), the survival model (Wu et al., 2020), and machine learning methods (Das et al., 2018, 2019; Yang et al., 2020). Among these models, the NB model has been the most frequently used. It has the following structure: the number of crashes y during a given time period is assumed to be Poisson-distributed, and the probability mass function (PMF) for which is given by:

$$p(y|\lambda) = \frac{\lambda^y \times \exp(-\lambda)}{y!},\ \lambda > 0,\ y = 0,\ 1,\ 2,\ ..., \tag{9}$$

Where λ is the mean response of the observed crash counts during a given period.

The rate parameter λ is assumed to be gamma-distributed with $E(\lambda) = \mu$ and $Var(\lambda) = \mu^2/\phi$ (ϕ is the shape parameter of the gamma distribution). Eq. (10) shows the probability density function (PDF) for λ :

$$p(\lambda|\mu,\ \phi) = \frac{1}{(\mu/\phi)^\phi} \times \frac{\lambda^{\phi-1} \times \exp(-\lambda\phi/\mu)}{\Gamma(\phi)} \tag{10}$$

The NB distribution is actually a mixture of Poisson distribution, where the Poisson mean λ is gamma distributed. The PDF of the NB (Eq. (11)) can be obtained by summing out λ in Eq. (10) (readers are referred to (Hilbe, 2007) for complete derivation of the NB model):

$$p(y|\mu,\ \phi) = \frac{\Gamma(y+\phi)}{\Gamma(y+1) \times \Gamma(\phi)} \times \left(\frac{\mu}{\phi+\mu}\right)^y \times \left(\frac{\phi}{\phi+\mu}\right)^\phi \tag{11}$$

Where y is the response variable (i.e., number of crashes); μ indicates mean response of the observation; and, ϕ is the shape parameter of the gamma distribution, also known as the inverse dispersion parameter in the context of Eq. (11).

With the NB model structure, the weight factor w in the EB method (i.e., Eq. (8)) is given as:

$$w = \frac{1}{1+\mu/\phi} \tag{12}$$

For the detailed procedures of estimating roadway safety and ranking sites using the EB method, readers are referred to Hauer (1997), Persaud et al. (1999), and Hauer et al. (2002).

2.5.3. Ratio based HSID approach

Although previous studies have shown that the EB method outperforms other approaches in HSID, it is not without any limitation. One of

the concerns is that the exposure is not well addressed while ranking sites. In other words, sites with higher traffic volume and/or longer length will be more likely ranked higher than the sites having similar roadway feature but lower traffic volume and/or shorter length. The ranking results are biased toward sites having higher exposure. To overcome this issue, safety analysts proposed using the ratio between the expected number of crashes and the predicted number of crashes (Wunderlich et al., 2019). The ratio is calculated using Eq. (13).

$$Ratio = \frac{\widehat{k}}{\mu} \tag{13}$$

Where *Ratio* is the ratio between expected and predicted number of crashes; $\widehat{k}$ is the expected number of crashes (i.e., EB estimate); and, μ means predicted number of crashes. As can be seen from Eq. (13), the ratio addresses the exposure issue of the classic EB method. Two sites with the same roadway feature and traffic characteristics but different lengths will be ranked similar using the ratio based HSID method. This method has recently been used by safety researchers in practical projects (Wunderlich et al., 2019).

3. Data description

In order to validate the proposed criteria with field measured data, the authors prepared roadway, traffic, and crash data on rural two-lane highways in Texas. The roadway and traffic data were extracted from Texas roadway inventory database (RHiNO). A few filters have been used while selecting roadway segments: (1) Average daily traffic should be between 1000 and 12,000; (2) Segment length should be between 0.1 and 2.0 miles; and, (3) No special lanes (e.g., two-way left turn lane, passing lane). These three filters were used such that the selected roadway segments share similar roadway characteristics. Eight years (2011–2018) of crash data were collected from the Crash Records Information System (CRIS) database managed by the Texas Department of Transportation (TxDOT). The analyzed years were re-organized into four periods, with each period containing two years. The purpose is to make sure every site contained with enough observed crashes in each period to represent its characteristic. Intersection, intersection-related, and pedestrian crashes were excluded from the analyses.

In total, 18,154 rural two-lane roadway segments with corresponding roadway information (i.e., segment length, lane width, left shoulder width, right shoulder width, and speed limit) and crash counts were prepared. A statistical summary of the dataset is documented in Table 6.

4. Results

Two comparisons are tabulated and discussed in this section. One compares the evaluation results determined from the scoring system.

Table 6
Summary Statistics of Sample Roadway Segments.

Variable	Min	Max	Mean (SD)
Segment Length (mile)	0.1	2.0	0.363 (0.23)
Lane Width (ft)	10	14	11.9 (0.31)
Right Shoulder Width (ft)	0	12	6.6 (3.5)
Left Shoulder Width (ft)	0	12	6.5 (3.5)
Posted Speed Limit (mph)	20	75	60.9 (11.2)
ADT (P1: 2011 – 2012)	1,002	11,991	3,538.3 (2,110.7)
ADT (P2: 2013 – 2014)	1,024	11,626	3,471.4 (2,018.4)
ADT (P3: 2015 – 2016)	1,016	11,741	3,485.4 (2,001.3)
ADT (P4: 2017 – 2018)	1,000	11,750	3,236.3 (1,848.9)
Crash (P1: 2011 – 2012)	0	36	1.349 (2.413)
Crash (P2: 2013 – 2014)	0	36	1.255 (2.243)
Crash (P3: 2015 – 2016)	0	48	1.162 (2.085)
Crash (P4: 2017 – 2018)	0	35	1.072 (1.916)

SD: standard deviation; P#: period number.

The primary purpose is to examine the consistency of generalized criteria proposed in this paper versus the existing criteria introduced by Cheng and Washington (2008) in identifying the best performed HSID method. The other criteria comparison is in terms of the dispersions in scores. The purpose is to assess the stability of generalized criteria versus the existing criteria in measuring the performance of HSID methods. Recall that the abbreviations of the existing and generalized criteria are:

- Existing criteria introduced by Cheng and Washington (2008),
  - o Site Consistency Test (T1),
  - o Method Consistency Test (T2),
  - o Total Rank Differences Test (T3),
- Generalized criteria proposed in this study,
  - o High Crashes Consistency Test (HCCT),
  - o Common Sites Consistency Test (CSCT),
  - o Absolute Rank Differences Test (ARDT).

Table 7 shows evidence that the existing criteria are a special case of the generalized criteria when considering a comparison period between P3 and P4, (i.e., when only considering two periods). For example, when identifying the top 3.0 % hazardous sites, the scores of crash rate, EB and ratio based HSID methods are 2,624, 3427 and 2924 from the existing criterion T1, and 225, 264 and 275 from the existing criterion T2. These scores are exactly the same in their corresponding generalized criteria HCCT and CSCT. Although the existing T3 is a special case of the ARDT, the scores are not exactly the same, because ARDT is enhanced by counting the changes in the rankings with an absolute value over the rank differences, instead of just counting with the rank differences in T3.

With the existing criteria, each test is limited to compare two consecutive periods. Although there are multiple periods of crash data, the existing criteria require to group into only two time periods. For instance, in the case when practitioners need to evaluate HSID methods with crash data in three periods (i.e., P2, P3 and P4), in order to evaluate HSID methods with existing criteria, it is required to aggregate them into two periods. That is, if practitioners consider P2 as the initial period, then P3 is ought to aggregated with P4 (i.e., P2, P3 – P4); alternatively, practitioners could consider the aggregation of values from P2 and P3 as the initial period (i.e., P2 – P3, P4). Both are acceptable combinations. As there was no standard on aggregation, Cheng and Washington (2008) aggregated a three-year crash information into two periods as Year 2000, and Year 2001 – 2002; while Montella (2010) aggregated a five-year data into two periods as Year 2001 – 2002, and Year 2003 – 2005. In the example, there are two possible combinations of aggregation with P2, P3 and P4. Their evaluation outcomes using the existing criteria are compared in Table 8, when identifying top 3.0 % hazardous sites. Under T1, T2 and T3, the performance evaluations vary with different combinations of aggregation. Furthermore, considering the temporal instability of crashes (Mannering, 2018), aggregating crash data can potentially reduce the accuracy of HSID performance evaluation. However, with the generalized criteria, crash data in each period are treated separately (since each test is designed to compare in multiple periods). As a result, the evaluation results by existing criteria for the most appropriate HSID method may be inconsistent. For example, in Table 7 when identifying top 2.5 % hazardous sites, the criterion T3 evaluates ratio based HSID as the best performance through P1 and P2 – P4, and evaluates EB as the best performance through P2 and P3 – P4. However, the generalized criterion ARDT consistently evaluates EB as the most appropriated HSID. Similar inconsistency exists in using the criterion T3 when identifying top 0.5 % hazardous sites.

The enhancement of ARDT over T3 shows the improvements in terms of measuring the performance of HSID methods as well. Table 9 compares the existing and generalized criteria by the dispersions in scores. Considering that practitioners need to evaluate HSID methods

Table 7
Criteria Comparisons: Performance of HSID Methods (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.).

Top 0.5% Hazardous Sites																			
	$P1^+$, P2 – P4			$P2^+$, P3 – P4			$P3^+$, P4				$P1^+$, P2, P3, P4			$P2^+$, P3, P4			$P3^+$, P4		
	CR	EB	R	CR	EB	R	CR	EB	R		CR	EB	R	CR	EB	R	CR	EB	R
T1	632	1,013	746	627	1,062	755	642	1,001	775	HCCT	632	1,013	746	627	1,062	755	642	1,001	775
T2	34	43	38	34	49	37	30	38	37	CSCT	28	38	35	31	43	33	30	38	37
T3	33,130	24,426	20,723	49,552	21,983	21,697	116,178	23,624	42,872	ARDT	84,480	50,349	61,110	88,704	31,595	51,686	116,758	24,564	43,704
Top 2.5% Hazardous Sites																			
	$P1^+$, P2 – P4			$P2^+$, P3 – P4			$P3^+$, P4				$P1^+$, P2, P3, P4			$P2^+$, P3, P4			$P3^+$, P4		
	CR	EB	R	CR	EB	R	CR	EB	R		CR	EB	R	CR	EB	R	CR	EB	R
T1	2,302	3,265	2,715	2,345	3,088	2,542	2,300	3,023	2,588	HCCT	2,302	3,265	2,715	2,345	3,088	2,542	2,300	3,023	2,588
T2	208	236	257	214	232	249	189	222	228	CSCT	173	204	229	184	208	224	189	222	228
T3	544,158	264,092	258,711	585,996	292,200	325,934	710,839	414,685	553,827	ARDT	856,347	490,063	604,588	792,506	464,691	618,875	732,995	439,173	578,227
Top 3.0% Hazardous Sites																			
	$P1^+$, P2 – P4			$P2^+$, P3 – P4			$P3^+$, P4				$P1^+$, P2, P3, P4			$P2^+$, P3, P4			$P3^+$, P4		
	CR	EB	R	CR	EB	R	CR	EB	R		CR	EB	R	CR	EB	R	CR	EB	R
T1	2,658	3,667	3,074	2,682	3,514	2,993	2,624	3,427	2,924	HCCT	2,658	3,667	3,074	2,682	3,514	2,993	2,624	3,427	2,924
T2	258	296	321	262	284	306	225	264	275	CSCT	211	254	279	228	258	286	225	264	275
T3	689,538	350,349	371,801	754,254	383,145	463,995	926,256	522,211	762,186	ARDT	1,073,843	649,734	830,108	1,015,810	612,585	827,760	955,546	559,812	798,102

CR: crash rate based HSID approach; EB: empirical Bayes based HSID approach; R: ratio based HSID approach.
P#: Period #; $^+$: Initial Period.
Deep Orange Color: a HSID method with the best identification performance.
Orange Color: a HSID method with the second-best identification performance.
Yellow: a HSID method with the worst identification performance.

with crash data in four periods (i.e., P1, P2, P3 and P4). With the existing criteria, the practitioner can evaluate three possible pairs/sets, P1 and P2-P4, P2 and P3-P4, P3 and P4, due to the limited two-period constraint. There is no constraint with the generalized criteria, but for comparison purpose, three corresponding sets are used. A dispersion of scores is measured by the standard deviation among the three sets in each HSID method. For instance, for the ratio-based HSID method, the scores from the three sets for the criterion T3 are 371,801, 463,995 and 762,186 with a standard deviation of 204,050 (ratio between standard deviation and mean is 0.38); while the scores for the generalized criterion ARDT are 830,108, 827,760 and 798,102 with a standard deviation of 17,840 (ratio between standard deviation and mean is 0.02). The dispersion in scores of ARDT is significantly smaller than that of the criterion T3. This reveals that the scores quantified by ARDT are more stable and reliable.

The scores quantified by CSCT are also more stable and consistent than those of T2, as shown in Table 9. For example, the scores identified the ratio based HSID method as the best HSID under criterion T2 are 321, 306, and 275, respectively. With a standard deviation of 23 (ratio between standard deviation and mean is 0.08); while the scores identified the ratio based approach as the best HSID under criterion CSCT are 279, 286, and 275, respectively. With a standard deviation of 6 (ratio between standard deviation and mean is 0.02). There is no difference between criterion T1 and generalized criterion HCCT, because the comparisons listed in Table 9 are based on the same initial period. However, the scores quantified by T1 are subject to the selection of initial period as well as aggregation of periods. In other words, the scores vary depending on the combination of periods. The proposed HCCT overcomes this issue through assessing the performance during disaggregated periods (i.e., no period aggregation).

5. Conclusions

The purpose of HSID is to analytically rank sites and identify those with higher potentials to be improved in the roadway network. This is an important part in the roadway safety management process. Safety analysts have proposed various methods and criteria for better identifying hotspots. Unfortunately, the ranking results are not consistent

Table 8
Existing Criteria in Different Combinations of Aggregation (Top 3.0 % Hazardous Sites) (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.).

	T1			T2			T3	
	$P2^+$,P3– P4	P2-$P3^+$,P4		$P2^+$,P3– P4	P2-$P3^+$,P4		$P2^+$,P3– P4	P2-$P3^+$, P4
CR	2,682	2,823	CR	262	258	CR	754,254	826,683
EB	3,514	3,519	EB	284	126	EB	383,145	1,222,994
R	2,993	4,211	R	306	345	R	463,995	2,054,206

CR: crash rate based HSID approach; EB: empirical Bayes based HSID approach; R: ratio based HSID approach.
P#: Period #; +: Initial Period.
Deep Orange Color: a HSID method with the best identification performance.
Orange Color: a HSID method with the second-best identification performance.
Yellow: a HSID method with the worst identification performance.

Table 9
Criteria Comparisons: Dispersions in Scores (Top 3.0 % Hazardous Sites) (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.).

T1				
	$P1^+$, P2 – P4	$P2^+$, P3 – P4	$P3^+$, P4	SD
CR	2,658	2,682	2,624	29
EB	3,667	3,514	3,427	122
R	3,074	2,993	2,924	75
T2				
	$P1^+$, P2 – P4	$P2^+$, P3 – P4	$P3^+$, P4	SD
CR	258	262	225	20
EB	296	284	264	16
R	321	306	275	23
T3				
	$P1^+$, P2 – P4	$P2^+$, P3 – P4	$P3^+$, P4	SD
CR	689,538	754,254	926,256	122,344
EB	350,349	383,145	522,211	91,243
R	371,801	463,995	762,186	204,050

HCCT				
	$P1^+$,P2, P3, P4	$P2^+$, P3, P4	$P3^+$, P4	SD
CR	2,658	2,682	2,624	29
EB	3,667	3,514	3,427	122
R	3,074	2,993	2,924	75
CSCT				
	$P1^+$, P2, P3, P4	$P2^+$, P3, P4	$P3^+$, P4	SD
CR	211	228	225	9
EB	254	258	264	5
R	279	286	275	6
ARDT				
	$P1^+$,P2, P3, P4	$P2^+$, P3, P4	$P3^+$, P4	SD
CR	1,073,843	1,015,810	955,546	59,152
EB	649,734	612,585	559,812	45,187
R	830,108	827,760	798,102	17,840

CR: crash rate based HSID approach; EB: empirical Bayes based HSID approach; R: ratio based HSID approach.
SD: Standard deviation.
P#: Period #; +: Initial Period.
Deep Orange Color: a HSID method with the best identification performance.
Orange Color: a HSID method with the second-best identification performance.
Yellow: a HSID method with the worst identification performance.

when using different methods and/or criteria. It is therefore necessary to evaluate the performance of these methods. However, to the best of the authors' knowledge, there are very limited criteria for assessing the HSID method performance. Although existing criteria (i.e., site consistency test, method consistency test, total rank differences test) developed by Cheng and Washington (2005) and Cheng and Washington (2008) have been heavily used, researchers have recently reported that the criteria are not sensitive for certain conditions (Guo et al., 2019). In addition, the existed tests are only capable of assessing the performance of HSID methods over two consecutive periods (i.e., before and after). This paper improved and generalized the criteria proposed by Cheng and Washington (2008) with higher consistency and ability to handle multi-period hotspot analyses. Specifically, three tests are proposed: (1) High Crashes Consistency Test (HCCT); (2) Common Sites Consistency Test (CSCT); and (3) Absolute Rank Differences Test (ARDT).

To validate the improvements and generalizations, this paper has documented a comparative analysis for three HISD methods (crash rate, EB, and ratio between EB and prediction) using eight years of crash data that occurred on 18,154 rural two-lane roadways. The paper first assessed the performance of the three HSID methods using existing criteria on two-period basis (i.e., aggregated). Then, it evaluated the performance using the proposed generalized criteria for multiple periods. Comparative analyses between the two results (i.e., HSID method performance and testing scores) were conducted to verify the improvements. In the two-period comparison with the HSID methods in identifying both top 0.5 % and 2.5 % hazardous sites, the results showed that the proposed tests are more consistent in detecting the best performed HSID method, and remain at least the same level of correctness as those existing criteria when identifying top 3.0 % hazardous sites. Inconsistent assessing results were found when using existing criteria in identifying top 0.5 % and 2.5 % hazardous sites. At the same time, each enhanced test showed its ability to handle multiple periods with the multiple-period comparison. The comparison also showed that the testing scores using the proposed criteria are always less dispersed than those using the existing criteria, suggesting the stability and consistency of the proposed generalized criteria.

To compare the performance of HSID methods, this study has used 18,154 roadway segments, which should be considered as an extremely large dataset. In practice, however, safety analysts may have limited number of sites, which could influence the assessment results significantly. It is suggested that HSID methods with varying numbers of sites should be assessed with the existing and generalized criteria in the future. Further, there is another test (i.e., Test 4: Poisson Mean Differences Test) in Cheng and Washington (2008), which only applies when the true Poisson means are known. Finally, it may be necessary to generalize the test with artificial realistic data (Wu et al., 2014; Zou et al., 2015; Wu and Lord, 2017).

A free online tool to conduct the three tests has been developed and is hosted on this website: https://ceprofs.civil.tamu.edu/dlord/HSID_Evaluation/. The authors will maintain and update the tool periodically.

CRediT authorship contribution statement

Xiaoyu Guo: Methodology, Formal analysis, Visualization, Writing - original draft. **Lingtao Wu:** Conceptualization, Data curation, Investigation, Writing - original draft. **Dominique Lord:** Supervision, Writing - review & editing.

Acknowledgement

The authors thank for valuable comments and insights from two anonymous reviewers.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aap.2020.105684.

References

AASHTO, 2010. Highway Safety Manual, 1st edition ed. American Association of State

Highway and Transportation Officials, Washington, D.C.
Cheng, W., Washington, S.P., 2005. Experimental evaluation of hotspot identification methods. Accid. Anal. Prev. 37 (5), 870–881.
Cheng, W., Washington, S., 2008. New criteria for evaluating methods of identifying hot spots. Transp. Res. Rec. 2083, 76–85.
Das, S., Dutta, A., Jalayer, M., Bibeka, A., Wu, L., 2018. Factors influencing the patterns of wrong-way driving crashes on freeway exit ramps and median crossovers: exploration using 'Eclat'association rules to promote safety. Int. J. Transp. Sci. Technol. 7 (2), 114–123.
Das, S., Minjares-Kyle, L., Wu, L., Henk, R.H., 2019. Understanding crash potential associated with teen driving: survey analysis using multivariate graphical method. J. Safety Res. 70, 213–222.
Elvik, R., 2008. Comparative analysis of techniques for identifying locations of hazardous roads. Transp. Res. Rec. (2083), 72–75.
Ferreira, S., Couto, A., 2015. A probabilistic approach towards a crash risk assessment of urban segments. Transp. Res. Part C Emerg. Technol. 50, 97–105.
Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-lindley generalized linear model: characteristics and application using crash data. Accid. Anal. Prev. 45, 258–265.
Geedipally, S.R., Lord, D., Dhavala, S.S., 2014. A caution about using deviance information criterion while modeling traffic crashes. Saf. Sci. 62, 495–498.
Guo, X., Wu, L., Zou, Y., Fawcett, L., 2019. Comparative analysis of empirical bayes and bayesian hierarchical models in hotspot identification. Transp. Res. Rec. 2673, 111–121.
Hauer, E., 1992. Empirical bayes approach to the estimation of "unsafety": the multivariate regression method. Accid. Anal. Prev. 24 (5), 457–477.
Hauer, E., 1996. Identification of sites with promise. Transp. Res. Rec. 1542 (1), 54–60.
Hauer, E., 1997. Observational before-after studies in road safety: estimating the effect of highway and traffic engineering measures on road safety Pergamon, Tarrytown, N.Y. U.S.A.
Hauer, E., Persaud, B.N., 1984. Problem of identifying hazardous locations using accident data. Transp. Res. Rec. 975, 36–43.
Hauer, E., Harwood, D.W., Council, F.M., Griffith, M.S., 2002. Estimating safety by the empirical bayes method - a tutorial. Transportation Research Record: Journal of the Transportation Research Board 1784, 126–131.
Hilbe, J., 2007. Negative Binomial Regression Cambridge. Cambridge University Press, Cambridge.
Lan, B., Persaud, B., 2011. Fully Bayesian approach to investigate and evaluate ranking criteria for black spot identification. Transp. Res. Rec. 2237 (1), 117–125.
Lord, D., Kuo, P.-F., 2012. Examining the effects of site selection criteria for evaluating the effectiveness of traffic safety countermeasures. Accid. Anal. Prev. 47, 52–63.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44 (5), 291–305.
Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of poisson-gamma models for modeling motor vehicle crashes: a bayesian perspective. Saf. Sci. 46 (5), 751–770.
Lord, D., Guikema, S.D., Geedipally, S.R., 2008. Application of the conway-maxwell-poisson generalized linear model for analyzing motor vehicle crashes. Accid. Anal. Prev. 40 (3), 1123–1134.
Lord, D., Geedipally, S.R., Guikema, S.D., 2010. Extension of the application of conway-maxwell-poisson models: analyzing traffic crash data exhibiting underdispersion. Risk Anal. 30 (8), 1268–1276.
Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Anal. Methods Accid. Res. 17, 1–13.
Meng, Y., Wu, L., Ma, C., Guo, X., Wang, X., 2020. A comparative analysis of intersection hotspot identification: fixed vs. Varying dispersion parameters in negative binomial models. J. Transp. Saf. Secur. 1–18.
Miranda-Moreno, L.F., Fu, L.P., Saccomanno, F.F., Labbe, A., 2005. Alternative risk models for ranking locations for safety improvement. Transp. Res. Rec. (1908), 1–8.
Montella, A., 2010. A comparative analysis of hotspot identification methods. Accid. Anal. Prev. 42 (2), 571–581.
Oh, J., Washington, S.P., Nam, D., 2006. Accident prediction model for railway-highway interfaces. Accid. Anal. Prev. 38 (2), 346–356.
Park, B.-J., Lord, D., Wu, L., 2016. Finite mixture modeling approach for developing crash modification factors in highway safety analysis. Accid. Anal. Prev. 97, 274–287.
Persaud, B., Lyon, C., Nguyen, T., 1999. Empirical bayes procedure for ranking sites for safety investigation by potential for safety improvement. Transportation Research Record: Journal of the Transportation Research Board 1665 (1), 7–12.
Persaud, B., Lan, B., Lyon, C., Bhim, R., 2010. Comparison of empirical bayes and full bayes approaches for before-after road safety evaluations. Accid. Anal. Prev. 42 (1), 38–43.
Saha, D., Alluri, P., Dumbaugh, E., Gan, A., 2020. Application of the poisson-tweedie distribution in analyzing crash frequency data. Accid. Anal. Prev. 137, 105456.
Washington, S., Afghari, A.P., Haque, M., 2018. Detecting high-risk accident locations. In: Lord, D., Washington, S. (Eds.), Safe Mobility: Challenges, Methodology and Solutions. Emerald Publishing Limited, Bingley, UK, pp. 351–382.
Wu, L., Lord, D., 2017. Examining the influence of link function misspecification in conventional regression models for developing crash modification factors. Accid. Anal. Prev. 102, 123–135.
Wu, L., Zou, Y., Lord, D., 2014. Comparison of sichel and negative binomial models in hot spot identification. Transportation Research Record: Journal of the Transportation Research Board 2460, 107–116.
Wu, L., Lord, D., Geedipally, S.R., 2017. Developing crash modification factors for horizontal curves on rural two-lane undivided highways using a cross-sectional study. Transp. Res. Rec. 2636 (1), 53–61.
Wu, L., Meng, Y., Kong, X., Zou, Y., 2020. Incorporating survival analysis into the safety effectiveness evaluation of treatments: jointly modeling crash counts and time intervals between crashes. J. Transp. Saf. Secur. 2020.
Wunderlich, R., Dixon, K., Wu, L., Geedipally, S., Dadashova, B., Shipp, E., 2019. A Data-driven Safety Analysis (DDSA) Framework for the Beaumont District. Report Number: 5-9052-01. Texas A&M Transportation Institute.
Yang, X., Zou, Y., Tang, J., Liang, J., Ijaz, M., 2020. Evaluation of short-term freeway speed prediction based on periodic analysis using statistical models and machine learning models. J. Adv. Transp. 2020.
Zou, Y., Lord, D., Zhang, Y., Peng, Y., 2013. Comparison of sichel and negative binomial models in estimating empirical bayes estimates. Transp. Res. Rec. 2392, 11–21.
Zou, Y., Wu, L., Lord, D., 2015. Modeling over-dispersed crash data with a long tail: examining the accuracy of the dispersion parameter in negative binomial models. Anal. Methods Accid. Res. 5–6, 1–16.