A Comparison of Techniques for the Identification of Hazardous Locations

JULIA L. HIGLE AND MARI B. HECHT

Techniques for the identification of hazardous locations, based on both classical and Bayesian statistical analyses, are evaluated and compared in terms of their ability to identify hazardous locations correctly. A simulation experiment, which is described in detail, is used. One classically based technique exhibits a tendency to err in the direction of false negatives. Another classically based technique yields relatively few false negative errors and produces results that are virtually indistinguishable from the results obtained from the Bayesian techniques. A variation of the Bayesian method proposed by Higle and Witkowski exhibits a tendency to perform well, producing low numbers of both false negative and false positive errors. Observed sensitivities of the various procedures are discussed.

The identification of hazardous locations is an important first step in any highway safety plan. The technique used to identify these locations should be sufficiently accurate to instill a high degree of confidence in the reported results. Hauer and Persaud compare the identification process to a sieve that should "catch" the hazardous sites while allowing the nonhazardous sites to "pass through" (*1*). A technique that tends to catch hazardous sites while allowing nonhazardous sites to pass through works well. Similarly, a technique that tends to allow a large fraction of hazardous sites to pass through while catching a number of nonhazardous sites is probably flawed. In this paper, the results of an experiment designed to evaluate empirically the relative performance of various techniques for the identification of hazardous locations are reported.

From the Hauer and Persaud analogy, it can be seen that an identification procedure partitions the sites under consideration into four categories, depending on whether or not they are truly hazardous (labeled H and NH, respectively) and whether or not they are identified, or flagged, by the identification technique (labeled F and NF, respectively). If a technique works perfectly, all sites that are hazardous are flagged and all sites that are not hazardous are not flagged. That is, in the absence of error, $F = H$ and $NF = NH$. Unfortunately, sites are flagged on the basis of data that are subject to random variation. It follows that some sites that are not hazardous are flagged and some sites that are hazardous are not flagged. The former event is a "false positive" identification, and the latter is a "false negative" identification. If the number of false negative identifications tends to be low, one can be reasonably assured that the set of sites that is flagged contains most of the truly hazardous sites. Similarly, if the number of false positive identifications tends to be high, then one suspects that many of the sites that are flagged are not truly hazardous. In evaluating the effectiveness of an identification technique, the relative severity of these two types of errors must be considered. It is generally agreed that a false negative error is far more serious than a false positive error. Thus, one might conclude that a technique that tends to yield a low number of false negatives is a good technique, as long as the low number of false negatives is not accompanied by an excessively large number of false positives.

Higle and Witkowski propose an empirical Bayesian method for the identification of hazardous locations (*2*). In comparing the output of their procedure to that of procedures based on classical statistical techniques, they speculate that at least one of the classically based techniques may be prone to err in the direction of false negatives. This observation is based on the authors' interpretation of the empirical results presented in the paper. On the basis of their study, it is impossible to know which sites are actually hazardous (H); thus their observations cannot be considered conclusive. In an effort to compare the ability of various identification techniques to distinguish correctly between hazardous and nonhazardous locations, the performance of the various techniques discussed by Higle and Witkowski (*2*) are investigated. In particular, the performance of identification techniques that are derived from classical statistical procedures is compared with those that are derived from the Bayesian procedure defined by Higle and Witkowski.

This paper is organized as follows. The next section contains a detailed explanation of the experimental method used. Salient points regarding the experimental method are discussed next. The results of the experiment are tabulated and presented, and the following section contains a discussion of the sensitivities of the various techniques observed during the course of the experiment. Conclusions are presented last.

EXPERIMENTAL METHOD

The experiment is based on a simple design. It is assumed that we are given a collection of accident rates (i.e., the number of accidents per million vehicles entering an intersection) for m locations, $\{\Lambda_i\}_{i=1}^m$, which represents a state of "perfect knowledge." That is, it is assumed that Λ_i represents the long-term expected accident rate at location i under its current configuration. With these accident rates, three years' worth of accidents are randomly generated, thereby simulating the accident data that might be available to a safety analyst. The simulated data are then analyzed using the various techniques. The set of sites identified as hazardous by each technique is compared with a set of sites that is known to be hazardous based on the true rates, $\{\Lambda_i\}_{i=1}^m$. Because the output of the

Department of Systems and Industrial Engineering, University of Arizona, Tucson 85721.

experiment can be expected to vary with the simulated data, 30 independent repetitions of the experiment are performed to allow for the observation of general trends that might emerge.

In its most basic form, the experiment consists of the following phases:

1. The selection of the "true" accident rates, $\{\Lambda_i\}_{i=1}^m$,
2. The identification of the set of sites that is "truly" hazardous,
3. The generation of the simulated accident data,
4. The analysis of the simulated data and the identification of hazardous sites, and
5. The comparison of the sites identified as hazardous with those that are truly hazardous.

Each of these phases is discussed in turn.

True Accident Rates

The goal of this experiment is to glean an understanding of the performance characteristics of the various techniques as they are commonly implemented. Thus a hypothetical collection of "true accident rates" that can capture some of the idiosyncrasies associated with true rate distributions with a reasonable degree of accuracy is required. To do this, observed accident data were used as a hypothetical collection of true rates. Four data sets, containing accident and traffic volume data for signalized intersections from communities under various jurisdictions within the state of Arizona, are used. By using data sets that differ in size and underlying characteristics, a variety of test scenarios is achieved. Consistent results across all scenarios considered should provide evidence for and against the various techniques.

The two data sets included by Higle and Witkowski (*2*) are used in this study and are labeled DS1 and DS2. In addition, two remaining data sets, DS3 and DS4, are also used. DS3 corresponds to the set of accident rates associated with signalized intersections under the jurisdiction of the City of Tempe Traffic Engineering Division, Tempe Arizona, during 1982 and 1984, whereas DS4 corresponds to accident histories associated with signalized intersections under the jurisdiction of the City of Tucson Department of Transportation, Tucson Arizona, during 1983–1986. Within each data set, each location's accident data are combined to represent a single annual accident rate (i.e., average number of accidents per million vehicles entering the intersection). The cumulative distributions for these preliminary sets of rates are depicted in Figure 1.

DS1 consists of 33 intersections and yields an approximately s-shaped curve. DS2 consists of 35 intersections and is characterized by an approximately linear curve with three rates that are significantly higher than all other rates. In Figure 1b, these outliers can be seen to affect the tail of the distribution. DS3 consists of 28 sites and exhibits the most nearly linear curve, with no obvious outliers. DS4 consists of 96 intersections and is characterized by a gently shaped s-curve. For the purposes of this experiment, the accident rates taken from the data sets DS1–DS4 are thought of as representing perfect information, subject to no randomness whatsoever, and thus are considered to be the true accident rates (i.e., those that the identification procedures attempt to estimate).

Truly Hazardous Locations

To assess the performance of the various techniques, it is necessary to identify a set of locations that are truly hazardous. This identification is based on the collection of true rates, $\{\Lambda_i\}_{i=1}^m$, as it represents a state of perfect knowledge. To be consistent with current practice, a location is considered hazardous if the true accident rate is sufficiently higher than the population mean. That is, if

$$\Lambda_i > \mu + k\sigma \tag{1}$$

where μ and σ are the mean and standard deviation associated with the true distribution of rates (as depicted in Figure 1), then location i is identified as truly hazardous. The symbol H is used to represent those sites that are truly hazardous.

For the purpose of this experiment, three values of k are considered: 1.282, 1.645, and 2.327. These values are arbitrarily selected to correspond to the critical values associated with a classical statistical test of hypothesis at the 0.90, 0.95, and 0.99 confidence levels, respectively. These values are obtained from the normal distribution function and thus have no particular meaning in terms of the distribution of true rates associated with the four preliminary data sets DS1–DS4. Note that they are used simply because they coincide with the commonly used classical statistical procedures discussed elsewhere (*2*, *3*) and thus provide for a convenient comparison of the set of truly hazardous sites and the sets of sites that are flagged by the various techniques.

Recognizing that Equation 1 may be an unsatisfactory method for determining those locations that are truly hazardous, a second method in which a threshold value is used is also considered. That is, one may alternatively state that if

$$\Lambda_i > \Lambda_T, \tag{2}$$

where Λ_T represents an upper limit on the acceptable accident rates, then location i is identified as truly hazardous (H). Note that Equations 1 and 2 are equivalent when

$$k = \frac{\Lambda_T - \mu}{\sigma} \tag{3}$$

Thus one may equivalently think of Λ_T as implying a critical value for the multiplier k.

It should be noted that because $\{\Lambda_i\}_{i=1}^m$ represents the true accident rates, locations are identified as truly hazardous (H) independent of both the simulated data and the technique being tested. Thus, all techniques attempt to identify the same set of locations for all simulated data sets. Naturally, the set of truly hazardous locations will vary with the four preliminary data sets, DS1–DS4.

Simulation of Data

Given a collection of true rates, accident data comparable with what would be available to a safety analyst are randomly generated using a computer simulation. The number of accidents at intersection i is generated according to a Poisson distribution with a mean of $\Lambda_i V_i$, where V_i represents the traffic volume at the intersection. The Poisson is a logical

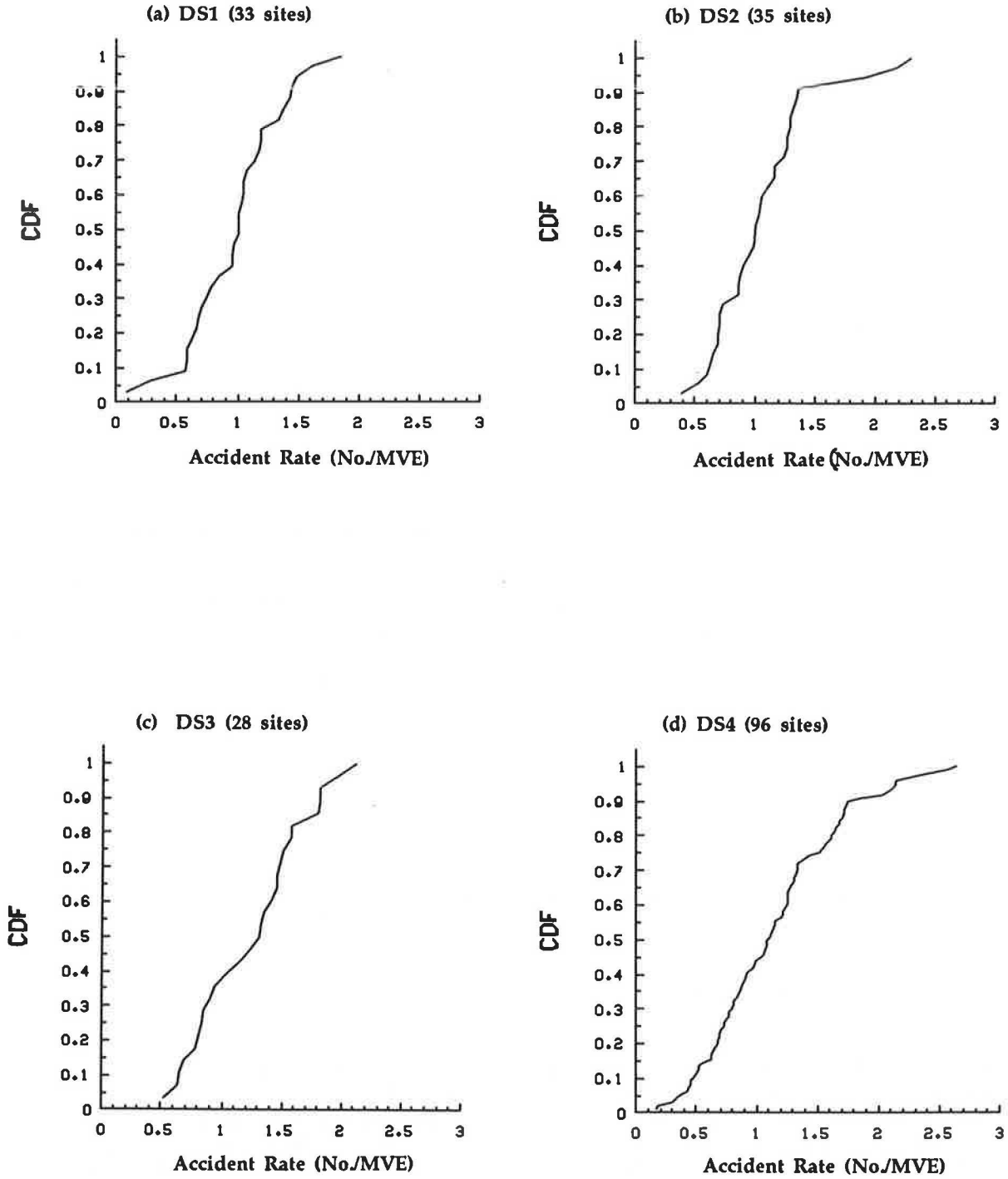

FIGURE 1 Cumulative distributions.

choice for a probabilistic model for numerous reasons, such as those presented by Ross (*4*). In general, if the numbers of events (e.g., accidents) occurring in disjoint time intervals are independent random variables with distributions varying with the length of the interval and the probability of multiple events occurring in a small interval of time is low, the Poisson distribution is an appropriate model for the random process. Similarly, if one envisions the relationship between traffic volume and accidents at a particular intersection in such a fashion that each vehicle entering the intersection has some probability of being involved in an accident, then again the Poisson distribution is an appropriate choice for modeling the number of accidents at the intersection (*5*, p. 66).

Analysis of Simulated Data

The simulated data are analyzed using the techniques discussed by Higle and Witkowski (*2*). The term "flag" and the symbol *F* are used to distinguish between those sites that are

identified as hazardous by the techniques being tested (i.e., using the simulated data) and those sites that are truly hazardous (i.e., "H," those based on the true rates).

The techniques being tested vary somewhat, depending on whether Equation 1 or 2 is used to determine the set of truly hazardous locations. In what follows, $\hat{\lambda}_i$ represents the accident rate observed at location i, based on the traffic volume at the intersection and the simulated number of accidents, $\bar{x}$ and s represent the sample mean and standard deviation of the observed (simulated) data, and x_R is the observed system rate. The symbol $\tilde{\lambda}_i$ is used to represent the true accident rate at location i, which some techniques explicitly treat as a random variable. When truly hazardous locations are determined by Equation 1, the identifcation techniques are defined as follows:

C1 Site i is flagged as hazardous if

$$\hat{\lambda}_i > \bar{x} + k_\delta s \tag{4}$$

C2 Site i is flagged as hazardous if

$$\hat{\lambda}_i > x_R + k_\delta \left(\frac{x_R}{V_i}\right)^{\frac{1}{2}} + \frac{1}{2V_i} \tag{5}$$

B1 Site i is flagged as hazardous if

$$P(\tilde{\lambda}_i > \bar{x}) > \delta \tag{6}$$

B2 Site i is flagged as hazardous if

$$P(\tilde{\lambda}_i > x_R) > \delta. \tag{7}$$

The values of $\bar{x}$ and x_R are computed as described by Higle and Witkowski (*2*), and s is computed as described by Quaye (*2*, p. 33). The probabilities presented in Equations 6 and 7 are computed as done by Higle and Witkowski (*2*), with the modification for the estimation of the parameters of the regional distribution of the accident rates provided in the discussion by Morris (*2*). Note that in each of the preceding techniques, the values of δ used to determine whether or not a site is flagged are allowed to vary freely and are generally supplied by the safety analyst. In making comparisons between the various techniques, δ is held constant so that the four inequalities have analogous interpretations. That is, with δ fixed, C1 and C2 are interpreted as indicating that a flagged site has a true rate that exceeds $\bar{x}$ and x_R, respectively, with a confidence of $\delta \times 100$ percent, whereas B1 and B2 are interpreted as indicating that a flagged site has a true rate that exceeds $\bar{x}$ and x_R, respectively, with a probability of δ. The values of δ tested are 0.90, 0.95, and 0.99. Correspondingly, k_δ takes on the values 1.282, 1.645, and 2.327, respectively. These values are intentionally selected to agree with the values used when identifying the truly hazardous locations, thereby facilitating the comparison of H and F.

C1 is based on the concept of confidence intervals associated with classical statistical tests of hypothesis. C2 is the rate-quality technique defined by Norden, Orlansky, and Jacobs (*6*) and Morin (*7*) and can be thought of as a refinement of C1. B1 and B2 differ only in their respective use of $\bar{x}$ and x_R and are simply two variations of the Bayesian technique defined by Higle and Witkowski (*2*). Both C1 and C2 are computationally straightforward, whereas B1 and B2 require the numerical integration of a gamma probability density function and are thus computationally intensive.

When truly hazardous locations are identified via a threshold value, as in Equation 2, the values of δ used in the Bayesian techniques will vary. As B1 is intended to be analogous to C1 (in a Bayesian fashion), the appropriate value of δ is taken to depend on $\bar{x}$, s, and Λ_T. That is, building from the observation leading to Equation 3, let

$$k_T = \frac{\Lambda_T - \bar{x}}{s} \tag{8}$$

and δ_T represent the confidence level whose critical value corresponds to k_T. For example, if $\frac{\Lambda_T - \bar{x}}{s} = 1.58$ the value of δ_T used in Equation 7 is 0.9429 [see Ross (*4*, p. 73)]. On the other hand, B2 is intended to be analogous to C2, which differs from C1; thus the value of δ used will differ from the value associated with Equation 8. Using logic similar to that which yields Equation 3, let

$$k_{Ti} = \frac{\Lambda_T - (x_R + 1/2V_i)}{\left(\frac{x_R}{V_i}\right)^{1/2}} \tag{9}$$

and let δ_{Ti} represent the confidence level whose critical value corresponds to k_{Ti}. Note that because k_{Ti} depends on V_i, δ_{Ti} will vary with the location. This is not the case for B1 when Equation 2 is used to define the truly hazardous locations. When truly hazardous locations are identified using Equation 2, the techniques being tested are defined as follows:

B1T Site i is flagged as hazardous if

$$P(\tilde{\lambda}_i > \bar{x}) > \delta_T \tag{10}$$

B2T Site i is flagged as hazardous if

$$P(\tilde{\lambda}_i > x_R) > \delta_{Ti} \tag{11}$$

where the threshold probabilities δ_T and δ_{Ti} are computed in accordance with Equations 8 and 9, respectively. The letter T is appended to the names given to the techniques to indicate that a threshold rate has been used to determine the set of truly hazardous locations.

Comparison of Techniques

As previously mentioned, a procedure that works perfectly will flag every site that is truly hazardous and no others, yielding $F = H$. It is reasonable, however, to expect that none of the techniques being tested work perfectly. Thus, errors will be made, and the challenge lies in determining how to describe these errors quantitatively.

As discussed in earlier work (*2*), a simple accounting of the two types of mistakes, false negative and false positive identifications, yields an inadequate summary of the accuracy of the procedures. For example, consider a site that is deemed

truly hazardous based on Equation 1 with k corresponding to $\delta = 0.95$. If the probability computed in Equation 6 (criterion B1) is at least 0.95, the site is correctly flagged as hazardous. Otherwise, B1 yields a false negative identification for this location. A false negative identification based on a computed probability of 0.949 (representing a "near miss") might be judged less harshly than one based on a computed probability of 0.70. Hence, there is some value in knowing the magnitude of the error associated with a misidentification.

Consider a hazardous site i that is not flagged (NF) by the technique under consideration (i.e., a false negative). Let δ represent the value corresponding to k for which the site is hazardous according to Equation 1 (i.e. $\delta = 0.90$, 0.95, or 0.99). Under C1 and C2, one can easily determine $\bar{\delta}_i$, the maximum level for which the site would be flagged. For example, using C1 (Equation 4), set $\bar{k}_i = \hat{\lambda}_i - \bar{x}/s$. Then $\bar{\delta}_i$ can easily be determined. The values of $\underline{k}_i$ and $\underline{\delta}_i$ are analogously defined using C2 (Equation 5). In this case, because i is a false negative identification, $\bar{\delta}_i < \delta$, and thus,

$$\rho_i = \delta - \bar{\delta}_i \tag{12}$$

represents the magnitude of the error associated with the misidentification. With the Bayesian methods B1 and B2, $\bar{\delta}_i$ is equal to the probabilities computed via Equations 6 and 7, respectively. The magnitude of the error associated with a false positive identification is similarly obtained. In this case, a site that is not hazardous is flagged at some value of δ. Using logic similar to that above, let Δ_i represent the value corresponding to $\Lambda_i - \mu/\sigma$ (from Equation 1). Then Δ_i can be interpreted as representing the appropriate level at which the site would be identified as truly hazardous. In this case, because i is a false positive, $\Delta_i < \delta$, and thus

$$\tau_i = \delta - \Delta_i \tag{13}$$

represents the magnitude of the error. Because a site that is flagged when $\delta = 0.95$ will also be flagged when $\delta = 0.90$, a site need not be uniquely flagged (or identified as hazardous) at one level. In cases where a site is misidentified for multiple values of δ, the maximum value of the error is used for comparative purposes.

Finally, the consequences of false negative and false positive identifications are distinct. As such, the magnitudes of the two types of errors are considered separately. That is, letting $NF \cap H$ represent the set of sites receiving false negative identifications and n represent the number of sites in this category, then

$$\frac{1}{n} \sum_{i \in NF \cap H} \rho_i \tag{14}$$

represents the average error per false negative identification. Similar values can be computed using τ_i for the false positive identifications. To summarize, two statistics—(I) number of false identifications and (II) average error per misidentification—can be computed for each of the techniques identified by Equations 4 to 6, for each type of error. Statistic I is averaged over the 30 repetitions of the experiment, whereas statistic II is averaged only over those repetitions in which a misidentification occurs. In addition, two other statistics are of interest: (III) maximum number of misidentifications (over the 30 repetitions) and (IV) number of repetitions on which there are no misidentifications.

When Equation 2, the threshold technique, is used to determine the set of truly hazardous sites, comparisons are made between B1T and B2T (Equations 10 and 11, respectively). Because B1T and B2T involve the direct computation of a probability, the same four statistics can be computed for these methods (i.e., by comparing the computed probabilities to δ_T for B1T and δ_{Ti} for B2T). One might be interested in knowing which observed rates exceed Λ_T. Unfortunately, this simple "hit or miss" decision precludes an opportunity for comparison with other methods. As such, when Equation 2 is used, comparisons are made only between the Bayesian techniques B1T and B2T, affording an opportunity to investigate the relative value of allowing the threshold probability to vary with the site.

DISCUSSION

As discussed earlier, four distinct data sets are used to provide four sets of "true" accident rates. The reasons for this are twofold. First, to be able to make useful conclusions regarding the various techniques, it is necessary to use data that somehow capture some of the idiosyncrasies of a collection of real accident rates (such as accident rate/traffic volume pairings, central or outlying tendencies, etc.). Although the literature suggests a tendency toward using a gamma distribution to represent the true rates (e.g. *1,2,8,9*) it is possible that, within this experiment, such a choice might bias the results toward the Bayesian techniques, as the gamma distribution figures heavily in its development. Second, it is possible that a given set of rates might, on average, favor a particular technique. Note that the variance of this hypothetical collection of true rates is (on average) higher than the variance of those rates from which they have been generated. To overcome these potential shortcomings, multiple data sets with differing characteristics are used in an effort to establish sensitivities to the characteristics of the preliminary data set. Consistent results across all data sets would suggest that any such sensitivities that might exist are probably negligible.

The second phase of the experiment, the identification of truly hazardous locations, proves to be by far the most elusive part of the experimental design. To complete this task, it is necessary to decide how to interpret perfect information. The method corresponding to Equation 1 is used simply because it provides a convenient comparison of H and F (i.e., via δ). Thus, the reader is urged to interpret Statistic II, the average deviation per misidentification, as nothing more than a "ballpark estimate." Note that the use of Equation 1 is likely to bias the experiment in favor of C1. To allow for other interpretations of perfect information, Equation 2 is also used as a second method for determining a set of truly hazardous intersections. As an added benefit, Equation 2 allows for an assessment of the relative value of allowing the critical probabilities used in the Bayesian techniques to vary from one site to another.

The multiple statistics presented earlier are necessary to gain a full appreciation of the differences between the various techniques. For example, a large value of Statistic I (number of errors, averaged over the 30 repetitions) combined with low values of Statistic II (average deviation per error) suggests

that a large number of "near misses" are observed. Because an increase in the statistics associated with the false positives may be tolerable if a substantial decrease in the statistics associated with the false negatives ensues, the statistics are computed separately for false negative and false positive identifications.

RESULTS

The results of the study are separated on the basis of the method used to determine the set of truly hazardous locations. Presented first are the results of those repetitions associated with the method identified by Equation 1.

In Table 1 the distribution of the sites among the four categories determined by the combinations of *H*, *NH*, *F*, and *NF* are summarized. The values reported represent average values over the 30 repetitions of the experiment. As throughout this paper, within these figures, *H* corresponds to those sites that are truly hazardous, and *F* corresponds to those sites that are flagged as hazardous by the technique under consideration. The prefix *N* is used to represent the complement of the set. For the purposes of consistent comparison, the δ levels at which a site is deemed truly hazardous and flagged by the technique under consideration must agree to obtain a correct identification.

In reading Table 1, note that when $\delta = 0.90$, C1 correctly identifies (on average) 2.17 of the truly hazardous sites (*H-F*) and 28.67 of the truly nonhazardous sites (*NH-NF*) for DS1, while incorrectly identifying 0.83 of the truly hazardous sites (*H-NF*, false negatives) and 1.33 of the truly nonhazardous sites (*NH-F*, false positives). The corresponding values for B1, on the other hand, are 2.50 and 24.93 for the correct identification of hazardous and nonhazardous sites, respectively, and 0.50 and 5.07 for the incorrect identifications. Note that for DS1, when $\delta = 0.90$, 3 sites (2.17 + 0.83) are truly hazardous but 30 are not.

A number of trends are immediately visible in Table 1. First, note that in nearly every instance, C2, B1, and B2 correctly identify a significantly higher fraction of the truly hazardous sites than does C1. Consequently, it follows that, of the four techniques tested, C1 consistently yields the highest number of false negative identifications. The obvious exception to this trend is found in DS3 when $\delta = 0.99$. In this case, there is no site that is truly hazardous, and none of the techniques flag any sites. Further, for all techniques, the false positive identifications appear to be lowest for DS2. This is the data set that contains the outliers, true rates that are

TABLE 1 DISTRIBUTION SUMMARY

		Technique							
		C1		C2		B1		B2	
		F	NF	F	NF	F	NF	F	NF
DS1 (33 Sites)									
δ=0.90	H	2.17	0.83	2.43	0.57	2.50	0.50	2.43	0.57
	NH	1.33	28.67	4.63	25.37	5.07	24.93	4.60	25.40
δ=0.95	H	1.13	0.87	1.30	0.70	1.33	0.67	1.23	0.77
	NH	0.73	30.27	4.47	26.53	4.93	26.07	4.20	26.80
δ=0.99	H	0.20	0.80	0.63	0.37	0.57	0.43	0.57	0.43
	NH	0.17	31.83	2.90	29.10	2.93	29.07	2.73	29.27
DS2 (35 Sites)									
δ=0.90	H	2.70	0.30	3.00	0.00	3.00	0.00	3.00	0.00
	NH	0.40	31.60	4.50	27.50	4.60	27.40	4.33	27.67
δ=0.95	H	2.50	0.50	2.97	0.03	2.97	0.03	2.97	0.03
	NH	0.23	31.77	3.07	28.93	3.10	28.90	2.83	29.17
δ=0.99	H	1.27	0.73	2.00	0.00	2.00	0.00	2.00	0.00
	NH	0.27	32.73	2.23	30.77	2.20	30.80	2.03	30.97
DS3 (28 Sites)									
δ=0.90	H	3.07	1.93	4.60	0.40	4.67	0.33	4.60	0.40
	NH	0.33	22.67	3.20	19.80	3.63	19.37	3.13	19.87
δ=0.95	H	0.70	0.30	1.00	0.00	1.00	0.00	1.00	0.00
	NH	1.03	25.97	5.83	21.17	6.13	20.87	5.70	21.30
δ=0.99	H	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	NH	0.17	27.83	5.00	23.00	5.13	22.87	4.83	23.17
DS4 (96 Sites)									
δ=0.90	H	8.83	1.17	9.93	0.07	9.97	0.03	9.93	0.07
	NH	2.10	83.90	18.87	67.13	20.63	65.37	18.77	67.23
δ=0.95	H	6.40	2.60	8.97	0.03	8.97	0.03	8.97	0.03
	NH	0.40	86.60	16.70	70.30	18.50	68.50	16.33	70.67
δ=0.99	H	1.37	0.63	2.00	0.00	2.00	0.00	2.00	0.00
	NH	1.03	92.97	18.70	75.30	20.13	73.87	17.77	76.23

substantially higher than all other rates within the set. One should expect that, in general, the observed rates associated with these outliers will be correspondingly higher, and it follows that any reasonable identification technique should perform well in such a situation. Of course, note also that for a fixed value of δ, C1 tends to flag fewer sites than the remaining techniques. Consequently, it yields a larger number of false negative errors and a smaller of false positive errors than the other techniques.

To appreciate the magnitudes of the errors associated with the various techniques, the four summary statistics are presented in Tables 2 and 3.

Statistic I, the average number of false identifications, is related to the data presented in Table 1. In deriving the statistics presented in these tables, a site that is incorrectly identified for two or more values of δ is counted only once when computing Statistic I. As previously stated, the maximum magnitude of the error associated with the misidentification is used when computing Statistic II.

Again, a number of trends are immediately visible in these tables. Note again that C1 consistently yields a significantly higher number of false negative identifications (Statistic I, Table 2) than the other techniques. Moreover, its corresponding error per false negative identification (Statistic II, Table 2) tends to be comparable with that of the other techniques, although in most situations it is somewhat lower than the others. Note that for all but DS1, a substantial majority of the 30 repetitions of the experiment yield no false negative identifications for C2, B1, and B2. By looking at Statistics III and IV, note that only one false negative identification is ever made by C2, B1, and B2 over the 30 iterations with DS2 (Table 2). Similar comments hold for DS4 and, to a lesser extent, for DS3. As previously mentioned, this trend among the false negative identifications is reversed when the false positive identifications (Table 3) are considered, where all four statistics are generally better for C1 than for C2, B1, and B2.

TABLE 3 FALSE POSITIVE SUMMARY STATISTICS

Data Set	Statistic	Criterion C1	C2	B1	B2
DS1 (33 Sites)	I	1.70	5.97	6.37	5.90
	II	0.111	0.183	0.192	0.182
	III	3	8	8	8
	IV	6	0	0	0
DS2 (35 Sites)	I	0.63	5.40	5.50	5.20
	II	0.154	0.234	0.235	0.231
	III	2	9	9	9
	IV	15	0	0	0
DS3 (28 Sites)	I	1.47	7.57	8.10	7.50
	II	0.068	0.141	0.149	0.139
	III	3	10	11	10
	IV	2	0	0	0
DS4 (96 Sites)	I	3.30	26.70	28.50	26.53
	II	0.069	0.167	0.179	0.167
	III	6	30	32	30
	IV	1	0	0	0

A final trend emerging from these tables warrants observation. There is virtually no difference between the performances of C2, B1, and B2. Close agreement between B1 and B2 is to be expected, as there is generally little difference between $\bar{x}$ and x_R in Equations 6 and 7. However, that C2 so closely agrees with the Bayesian techniques may follow from the fact that the Poisson distribution plays so heavily in the derivations of Equations 5, 6, and 7.

We now turn to the presentation of the results of the experiment when Equation 2, the threshold criterion, is used to define the set of truly hazardous intersections. In Table 4 the distribution of the sites among the four categories determined by the combinations of *H*, *NH*, *F*, and *NF* is summarized.

TABLE 2 FALSE NEGATIVE SUMMARY STATISTICS

Data Set	Statistic	Criterion C1	C2	B1	B2
DS1 (33 Sites)	I	1.70	0.90	0.97	1.00
	II	0.101	0.186	0.165	0.184
	III	3	2	2	2
	IV	0	11	9	8
DS2 (35 Sites)	I	1.13	0.03	0.03	0.03
	II	0.063	0.059	0.061	0.063
	III	3	1	1	1
	IV	6	29	29	29
DS3 (28 Sites)	I	2.17	0.40	0.33	0.40
	II	0.115	0.149	0.155	0.170
	III	4	1	1	1
	IV	0	18	20	18
DS4 (96 Sites)	I	3.70	0.07	0.03	0.07
	II	0.049	0.081	0.087	0.083
	III	6	1	1	1
	IV	1	28	29	28

TABLE 4 DISTRIBUTION SUMMARY, THRESHOLD VALUES

	Technique B1T		B2T	
	F	NF	F	NF
DS1 (33 Sites) Λ_T=1.5/*MVE*				
H	1.57	0.43	1.27	0.73
NH	6.40	24.60	1.97	29.03
DS2 (35 Sites) Λ_T=1.5/*MVE*				
H	3.00	0.00	2.93	0.07
NH	6.30	25.70	1.47	30.53
DS3 (28 Sites) Λ_T=2.0/*MVE*				
H	1.00	0.00	0.73	0.27
NH	6.17	20.83	0.93	26.07
DS4 (96 Sites) Λ_T=2.0/*MVE*				
H	8.97	0.03	8.17	0.83
NH	19.63	67.37	3.37	83.63

Again, the values reported represent average values over the 30 iterations. The threshold values, Λ_T were arbitrarily chosen in the range of the 90th to 95th percentile of the true distributions presented in Figure 1 and are comparable with, but not identical to, the critical values identified in Equation 1.

As with the results presented in Table 1, B1T and B2T appear to identify successfully most sites that are truly hazardous (i.e., $H \cap NF$ tends to be small, relative to H). Although B2T results in a slightly higher number of false negative identifications than does B1T, it is generally accompanied by a substantial decrease in the number of false positive identifications.

In Table 5, the average fractions of truly hazardous sites that are not flagged by the various techniques, the false negative identifications, are presented. Similar information regarding the false positive identifications is contained in Table 6. Note that this information is directly obtained from Tables 1 and 4. It is included here to facilitate qualitative comparisons between the various techniques.

In evaluating the false negative identifications as a fraction of the number of sites that are truly hazardous, Table 5 indicates that in 8 of the 11 cases in which there are truly hazardous locations, C2, B1, and B2 yield fractions that are, at most, 0.08 (in 7 cases, this fraction is, at most, 0.01). The corresponding fractions for C1 are much higher, with values above 0.25 in 7 of the 11 cases and one fraction as high as 0.8 (although this corresponds to a situation in which only one site in the data set is hazardous). This can be taken to be encouraging evidence that C2, B1, and B2 are generally successful in flagging sites that are truly hazardous. Unfortunately, this success comes at a cost of a substantial increase in the number of false positive identifications over the number associated with C1. In reviewing Table 6, note that of the sites flagged, C1 yields a consistently lower fraction of false positive identifications than the other three. This false positive trend, which opposes the false negative trend, is to be expected because false positive and false negative errors are inversely related.

As one might expect, B1T exhibits both false negative and false positive profiles that are roughly comparable to those exhibited by B1 (and hence, by C2 and B2 as well). Of course, differences between B1T and B1 are to be expected because the true rates do not follow a normal distribution; thus, for example, the 95th percentile of the true distribution does not necessarily correspond to $\mu + 1.645\sigma$. B2T, on the other hand, appears to exhibit false negative profiles that are comparable with those exhibited by B2 (and hence, by B1 and C2) while exhibiting false positive profiles that are comparable with those exhibited by C1. The most notable difference occurs with DS3, where an average of 27 percent of the truly hazardous sites receive a false negative identification under B2T (note that there is only one such site).

Tables 7 and 8 contain the summary statistics associated with the portion of the experiment pertaining to the threshold values. In reviewing Tables 7 and 8, note again that B2T suffers from a higher number of false negative identifications than does B1T. In looking at Statistic II, however, it appears that with the exception of DS1, the false negative identifications associated with B2T are "near misses." This is true even for DS3, for which the large fraction of false negative identifications was previously reported. This "near miss" phenomenon is observed to an even greater extent with the false positives. Thus, it seems that in general, B2T yields more false negatives than B1T, although it tends to come very close to identifying all hazardous sites correctly. In addition, B2T identifies fewer false positives than B1T and tends to come closer to identifying them correctly than does B1T, as indi-

TABLE 5 FALSE NEGATIVE FRACTIONS, $NF \cap H/H$

	δ	C1	C2	B1	B2	B1T	B2T
DS1							
	0.90	0.28	0.19	0.17	0.19		
	0.95	0.43	0.35	0.34	0.39	0.215	0.365
	0.99	0.80	0.37	0.43	0.43		
DS2							
	0.90	0.10	0.	0.	0.		
	0.95	0.17	0.01	0.01	0.01	0.	0.023
	0.99	0.37	0.	0.	0.		
DS3							
	0.90	0.39	0.08	0.07	0.08		
	0.95	0.30	0.	0.	0.	0.	0.27
	0.99	n/a	n/a	n/a	n/a		
DS4							
	0.90	0.12	0.007	0.007	0.007		
	0.95	0.29	0.003	0.003	0.003	0.033	0.092
	0.99	0.32	0.	0.	0.		

TABLE 6 FALSE POSITIVE FRACTIONS, $F \cap NH/NH$

	δ	C1	C2	B1	B2	B1T	B2T
DS1							
	0.90	0.044	0.154	0.169	0.153		
	0.95	0.024	0.144	0.159	0.135	0.206	0.063
	0.99	0.005	0.091	0.092	0.085		
DS2							
	0.90	0.012	0.141	0.144	0.135		
	0.95	0.007	0.096	0.097	0.088	0.197	0.046
	0.99	0.008	0.067	0.067	0.062		
DS3							
	0.90	0.014	0.139	0.158	0.136		
	0.95	0.038	0.216	0.227	0.211	0.228	0.034
	0.99	0.006	0.178	0.183	0.173		
DS4							
	0.90	0.024	0.219	0.240	0.218		
	0.95	0.005	0.192	0.213	0.188	0.226	0.039
	0.99	0.011	0.199	0.214	0.189		

TABLE 7 FALSE NEGATIVE SUMMARY STATISTICS, THRESHOLD VALUES

Data Set	Criterion Statistic	B1T	B2T
DS1 (33 Sites)	I	0.43	0.73
	II	0.2533	0.2352
	III	2	2
	IV	19	13
DS2 (35 Sites)	I	0.00	0.07
	II	n/a	0.0306
	III	0	1
	IV	30	28
DS3 (28 Sites)	I	0.00	0.27
	II	n/a	0.0027
	III	0	1
	IV	30	22
DS4 (96 Sites)	I	0.03	0.83
	II	0.0627	0.0119
	III	1	2
	IV	29	9

TABLE 8 FALSE POSITIVE SUMMARY STATISTICS, THRESHOLD VALUES

Data Set	Criterion Statistic	B1T	B2T
DS1 (33 Sites)	I	6.40	1.97
	II	0.0823	0.0113
	III	12	4
	IV	0	3
DS2 (35 Sites)	I	6.30	1.47
	II	0.1082	0.0133
	III	11	5
	IV	0	10
DS3 (28 Sites)	I	6.17	0.93
	II	0.0449	0.0001
	III	9	3
	IV	0	8
DS4 (96 Sites)	I	19.63	3.37
	II	0.0539	0.0001
	III	22	7
	IV	0	0

cated by Statistic II. Thus, with the Bayesian technique, it seems that there is some merit in allowing the probability that must be exceeded before a site is flagged to vary with the site.

DISCUSSION OF SENSITIVITY

Given that the procedures being tested represent statistical analyses of data that are subject to random variations, one must expect that some sites will be incorrectly categorized. One should expect sites with true accident rates that are close to the critical rate, $\mu + k\sigma$ or Λ_T, to be prone to misidentification. In fact, *all* of the false negative errors associated with DS1 for $\delta = 0.99$ are attributed to this phenomenon. There is one site with a true rate only slightly higher than the critical rate suggested by Equation 1. Similarly, in DS3 there is one rate that is slightly higher than the critical rate for $\delta = 0.90$. This particular site accounts for nearly 60 percent of the C2, B1, and B2 false negative errors and for nearly 30 percent of the C1 false negative errors. This phenomenon, which is largely unavoidable, accounts for approximately one-half of all misidentifications. Several other phenomena cause false negative and false positive identifications among locations that do not have this characteristic. These phenomena, which might not be expected, are discussed next.

Many of the false negatives produced by C1 can be directly attributed to the dependence of the C1 critical rates (defined by Equation 4) on the sample mean and variance of the simulated data. A high sample mean and/or a high sample variance yields a high critical value that must be exceeded before C1 flags a site as hazardous. When these sample statistics are high, the generated accident rate at a given location must be high enough to exceed the critical rate. Such an occurrence will naturally result in a lower number of sites flagged as hazardous and a correspondingly higher number of false negative identifications. Additionally, cases were observed in which a single site having the same observed accident rate in two different repetitions of the experiment was flagged for only one of the repetitions, because of differences in the sample statistics.

The false positive identifications are much more dramatic for C2, B1, and B2 than for C1, as indicated by Tables 3 and 6. This appears to result from a sensitivity of these criteria to large variations in the traffic volume at the locations involved, as alluded to in the discussion by Morris (*2*). Recall that in the Bayesian procedure, the parameters associated with the gamma distributions from which the desired probabilities are computed are updated using the number of accidents observed at the site and the traffic volume at the site [see Higle and Witkowski (*2*)]. In general, a low traffic volume will result in a distribution with a high variance, a distribution that is fairly "spread out." For a site with a low traffic volume, the computed probabilities tend to be lower than for a site with a high traffic volume, whose accident rate distribution has a lower variance and is more "peaked." Thus, when using B1 and B2, a site with a low rate and a high volume will be sensitive to the observation of a "higher than expected" number of accidents. Such an occurrence leads to a false positive identification. Most of the false positive identifications associated with B1, B2, and C2 are due to this sensitivity to the traffic volumes. Because traffic volumes tend to be high, relative to the number of accidents, C2 and the Bayesian criteria used in the first phase of the experiment, indicated by Equations 6 and 7, can be expected to be plagued by a high number of false positive identifications.

Although C2, B1, and B2 tend to perform well in terms of the number of false negative identifications, Tables 2 and 5 indicate that they yield the largest number of false negative identifications for preliminary data set DS1. This increase over the remaining preliminary data sets appears to be directly attributed to a single site that is truly hazardous for $\delta = 0.95$ and, consequently, for $\delta = 0.90$ as well. This particular site has a low volume, and the updated distribution used in the Bayesian procedure has a correspondingly high variance. As a result it is prone to false negative identification, particularly when the observed accident rate is lower than the true accident rate.

The second phase of the experiment allows for the comparison of two variations of the Bayesian technique presented elsewhere (*2*). With B1T, a single "critical value" is applied to all sites, whereas with B2T, the critical value varies among the sites as a function of the traffic volume at the site. Based on Table 4, the difference between these two variations is dramatic. B1T is essentially the same as the procedures B1 and B2 and thus exhibits the previously discussed sensitivity to the traffic volume, resulting in a number of false positive identifications. With B2T, a large traffic volume results in a correspondingly large value that the computed probability must exceed to be flagged as hazardous. In addition, these values are generally higher than the value used in B1T, and it follows that fewer sites are flagged by B2T than by B1T. The dramatic reduction in the number of false positive identifications follows. As might be expected, this reduction is accompanied by an increase in the number of false negative identifications made by B2T. With the exception of preliminary data set DS1 (See Tables 4 and 5), however, B2T still exhibits a tendency to identify the truly hazardous locations correctly. As with C2, B1, and B2, B2T is plagued by the truly hazardous site in DS1 that is subject to low traffic vol-

umes. This site still receives false negative identifications when the observed accident rate is low.

CONCLUSIONS

In evaluating the performance of the various criteria, one must weigh the relative severity of false negative and false positive errors. A false negative error is likely to result in a failure to improve a truly hazardous location, and may lead to various forms of catastrophic loss. False positive errors may or may not result in the unnecessary improvement of a location that is not truly hazardous, depending on the judgment of the safety analyst and budgetary constraints. Thus, false negative errors are considered far more serious than false positive errors.

Because all techniques tested within the confines of this experiment behave consistently across all data sets, it does not appear that the underlying characteristics of the set of true accident rates influence the performance of a technique. Thus, these techniques can be expected to perform in a manner similar to that observed within this experiment, independent of regional data characteristics that might exist for a given jurisdiction. In reviewing cases in which the various techniques tested yield incorrect identifications (either false negatives or false positives), a number of trends become clear. Many of the errors are associated with locations whose true accident rates are close to the critical rates used to define the set of sites that are truly hazardous. This phenomenon, which accounts for a large fraction of the errors, is observed among all techniques tested; it is to be expected and is probably unavoidable.

C1, the classically based statistical technique, flags a smaller number of sites and consequently yields a greater number of false negative identifications and larger magnitudes of false negative error than the other techniques. Surprisingly, many of the false negative identifications associated with C1 result from its apparent sensitivity to the sample mean and standard deviation of the observed accident rates. It is disconcerting to note that even when two sets of data yield the same observed accident rate at a given location, differing sample statistics can result in differing identifications. This sensitivity, which is not observed among the other techniques, casts doubt on the reliability of C1 as an appropriate technique for the identification of hazardous locations.

The Bayesian techniques B1 and B2 and the classically based rate-quality technique C2 perform in a similar fashion. Each yields low numbers of false negative identifications and correspondingly low false negative errors. This comes at the expense of an increase in the number of false positive identifications, which may be the result of a sensitivity to the volume of traffic at the sites. The relatively large number of false positive errors is disconcerting but may not be serious, given that false negative identifications are to be avoided. It appears that B1, B2, and C2 exhibit a sensitivity to the volume of traffic at the sites that can result in a large number of false positive identifications. This may present difficulties in that many of the sites identified as hazardous are often not truly hazardous. Nonetheless, the rate-quality technique, C2, yields results that are virtually indistinguishable from those of the Bayesian techniques and is computationally straightforward.

In an effort to counteract the apparent sensitivity of C2, B1, and B2 to large variations in traffic volume, one can use a variation of the Bayesian method in which the value the computed probability must exceed before a site is identified as hazardous is allowed to vary as a function of the traffic volume at the site. B2T is an example of such a technique. With this technique, a small (but noticeable) increase in the number of false negative identifications is accompanied by a dramatic decrease in the number of false positive identifications. Thus, B2T tends to identify correctly sites that are truly hazardous without additionally identifying as hazardous a large number that are not truly hazardous. On the basis of the results obtained when Equation 2 is used in the identification of truly hazardous locations, it appears that there is some merit in allowing the probability used to flag locations in the Bayesian technique to vary among the sites, as suggested by Equations 8–11. Continued investigation into the potential advantage of this observation is encouraged. In particular, in using the observation leading to Equation 3, it will be interesting to note whether or not such a refinement of the Bayesian technique will yield sufficient improvement over the rate-quality technique to justify the computational burden associated with the technique.

REFERENCES

1. E. Hauer and B. 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.
2. J. L Higle 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.
3. J. C. Laughlin, 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.
4. S. M. Ross. *Introduction to Probability Models*. Academic Press, Inc., Orlando, Fla., 1985.
5. P. G. Hoel. *Introduction to Mathematical Statistics*. John Wiley and Sons, New York, 1984.
6. M. Norden, J. Orlansky, and H. Jacobs. Application of Statistical Quality-Control Techniques to Analysis of Highway-Accident Data. In *Highway Research Bulletin 117*, HRB, National Research Council, Washington, D.C., 1956, pp. 17–31.
7. D. A. Morin. Application of Statistical Concepts to Accident Data. In *Highway Research Record 188*, HRB, National Research Council, Washington, D.C., 1967, pp. 72–79.
8. W. D. Glauz, K. M. Bauer, and D. J. Migletz. Expected Traffic Conflict Rates and Their Use in Predicting Accidents. In *Transportation Research Record 1026*, TRB, National Research Council, Washington, D.C., 1985, pp. 1–12.
9. B. N. Persaud and E. Hauer. Comparison of Two Methods for De-Biasing Before- and After Accident Studies. In *Transportation Research Record 975*, TRB, National Research Council, Washington, D.C., 1984, pp. 43–49.

Publication of this paper sponsored by Committee on Methodology for Evaluating Highway Improvements.