Microscopic Accident Potential Models for Two-Lane Rural Roads

BHAGWANT N. PERSAUD AND KORNEL MUCSI

Fundamental to the present research is the use of hourly traffic volumes in regression models for estimating accident potential on two-lane rural roads. By using data from Ontario, Canada, a simple model form, and a regression package that allows the assumption of a negative binomial error structure, regression models were calibrated for the different combinations of time periods (24 hr, day hours, and night hours) and geometric (roadway and shoulder width) characteristics. It is shown that the effect of day/night conditions is different for single-vehicle and multivehicle accidents. For single-vehicle accidents the accident potential is higher during the night, whereas for multivehicle accidents the opposite is true. This indicates the importance of differentiating between single-vehicle and multivehicle accidents and day/night conditions. The refinement of the regression predictions by the empirical Bayesian (EB) estimation procedure for individual road sections is illustrated. It is shown through a validation exercise that the EB procedure provides better estimates of accident potential than the conventional method only on the basis of the short-term accident count for a section.

Estimation of the accident potentials of road sections usually requires the use of a relationship between accidents and a measure of traffic volume, traditionally, the average daily traffic (ADT). If this relationship is nonlinear, ADT-based models would be unsuitable for use in estimating safety during portions of a day, for example, specific hours, peak periods, and nighttime. Such estimates might be required to evaluate strategies that affect traffic volumes or safety during certain parts of the day and to identify potentially hazardous traffic operating conditions. The fundamental premise of the research on which this paper is based is that for these estimates it is preferable to use microscopic models with hourly volumes as the measure of traffic intensity. The increasing availability of hourly traffic volume data, taken together with recent advances in accident modeling, prompted a fresh look at developing models for estimating the accident potential of two-lane rural roads. These constitute a substantial portion of the North American road network.

The procedure for developing these microscopic models follows on that used in other recent work (*1,2*) and differs from the early attempts at accident modeling in two important aspects. First, like the early work, regression models are developed to relate accident occurrence and traffic volume, but the models are developed by using a generalized linear modeling package (GLIM) (*3*) that allows errors in accident counts to be more properly described by the negative binomial distribution rather than the traditional normal distribution assumed in conventional regression packages. Second, the procedure allows for the refinement of the regression estimate of accident potential by using an empirical Bayesian (EB) procedure.

B. N. Persaud, Department of Civil Engineering, Ryerson Polytechnic University, 350 Victoria Street, Toronto, Ontario M5B 2K3, Canada. K. Mucsi, Department of Civil Engineering, University of Toronto, Ontario M5S 1A4, Canada.

The remainder of the paper presents the results of research aimed at estimating microscopic accident prediction models for two-lane rural roads in Ontario, Canada. A summary of the theory, data, results, and validation along with a detailed example application are presented. Every attempt has been made to minimize repetition from related earlier publications, but a certain amount has been unavoidable in the interest of making this a paper that can reasonably stand alone.

THEORETICAL FOUNDATIONS

The fundamental estimator for $E(m)$, the expected number of accidents during T hours on a section of length L km, is given by Equation 1, where F represents the traffic volume, and a and b are parameters to be estimated in a regression model.

$$E(m) = a\,L\,T\,F^{b} \tag{1}$$

Reviews by Satterthwaite (*4*) and Hauer (*5*) indicate that when geometric and environmental conditions are appropriately controlled, this model form is quite common. Other reasons for its selection are its simplicity, the parsimony of the independent parameters, and the correspondence with logic that the predicted number of accidents would be zero for a traffic volume of zero. In a special case in which b is equal to 1, a linear relationship is indicated, but this need not be assumed a priori as has been done in several studies.

Regarding the error distribution for accident counts, several studies (*2,6*) have indicated that it is more appropriate to describe the accident count in a population of entities with the negative binomial distribution than with the Poisson or normal distributions.

Following recent work by Persaud (*1*) and others the generalized linear interactive modeling (GLIM) software package (*3*) was applied for the parameter estimation. GLIM allows the specification of different error structures, including the negative binomial, and it uses an algorithm for the parameter estimation by the method of maximum likelihood.

According to the theory the variance of m is related to $E(m)$ as follows:

$$\mathrm{VAR}(m) = E(m)^2/k \tag{2}$$

where k can be estimated by using a maximum likelihood procedure that assumes that each squared residual of the regression model is an estimate of $\mathrm{VAR}(m)$ and that each count comes from a negative binomial distribution with mean $E(m)$ and variance given by Equation 2.

$E(m)$ is an average over sites that are similar in the values of their independent variables, and if the variance of the m values is large,

TABLE 1 Parameter Estimates for Level 1 Models

Code	Model Group: Acc. Type	Severity Group	Time Period	Model Parameters: ln(a) (std. error)	b (std. error)	k	Sample Size: Number of Accidents
101	S.V.	F.& I.	24 Hr	-12.97 (0.03)	0.430 (0.006)	1.5	2340
102	M.V.	F.& I.	24 Hr	-16.88 (0.05)	1.137 (0.009)	0.7	1677
103	All	F.& I.	24 Hr	-13.59 (0.03)	0.674 (0.007)	1.3	4018
104	S.V.	F.& I.	Day	-13.49 (0.05)	0.502 (0.010)	2.1	1124
105	M.V.	F.& I.	Day	-17.12 (0.69)	1.180 (0.012)	0.7	1109
106	All	F.& I.	Day	-14.30 (0.05)	0.793 (0.010)	1.2	2234
107	S.V.	F.& I.	Night	-13.02 (0.04)	0.491 (0.009)	1.1	1028
108	M.V.	F.& I.	Night	-16.61 (0.06)	1.080 (0.011)	1.1	430
109	All	F.& I.	Night	-13.33 (0.04)	0.643 (0.009)	1.5	1459
110	S.V.	Total	24 Hr	-11.80 (0.03)	0.444 (0.005)	2.2	8033
111	M.V.	Total	24 Hr	-15.93 (0.04)	1.123 (0.008)	0.8	3992
112	All	Total	24 Hr	-12.26 (0.03)	0.627 (0.005)	1.8	12026
113	S.V.	Total	Day	-12.30 (0.04)	0.490 (0.008)	1.7	3470
114	M.V.	Total	Day	-16.21 (0.06)	1.173 (0.011)	0.8	2663
115	All	Total	Day	-13.02 (0.04)	0.741 (0.008)	1.5	6134
116	S.V.	Total	Night	-11.97 (0.03)	0.557 (0.007)	3.6	3878
117	M.V.	Total	Night	-15.70 (0.05)	1.071 (0.010)	0.7	1007
118	All	Total	Night	-12.14 (0.03)	0.650 (0.007)	2.7	4886

the value of $E(m)$ by itself is not very useful, particularly if it is applied to a specific site. Therefore, VAR(m) must also be estimated and used in the refinement of estimates for a specific site. For such a case and in general, a refined estimate of accident potential can be obtained from an EB procedure used by others in recent years (*1,2*). This procedure combines the regression estimate, $E(m)$, of sites similar in all independent variables and the short-term accident count (x) of the site.

For reasonable assumptions on the distributions of x and m (*7*), the EB estimate of accident potential is

$$E(m|x) = wE(m) + (1 - w)x \quad (3)$$

where

$$w = [1 + \text{VAR}(m)/E(m)]^{-1} = [1 + E(m)/k]^{-1} \quad (4)$$

It can also be shown (*2,7*) that the variation in (m/x) can be estimated by

$$\text{VAR}(m|x) = (x + k)/[1 + (k/E(m)]^2 \quad (5)$$

As discussed in some of the earlier work, a large value of k is consistent with a sound regression estimate, which is given a relatively large weight w. Conversely, for a small k, indicated by a regression estimate with a large variance, substantial weight is given to the accident count.

DATA

Three groups of data pertaining to traffic flow, geometry, and accidents were obtained from the Ministry of Transportation, Ontario (MTO), for each of 2,014 two-lane rural road sections. Hourly traffic volumes on these sections were estimated on the basis of the Annual ADT and the seasonal and hourly variation factors at permanent counting stations (PCSs). It was assumed that the variations in traffic volume (both seasonal and hourly) at a section were similar to those at the associated PCS. The accident data file provided information on all reported accidents on two-lane roads in Ontario for 1988 and 1989. Each accident record was associated with a unique road section and was characterized by the date and time of occurrence, class, severity, number of vehicles involved, and other relevant information. The raw data were used to assemble a regression data set stratified by section, hour, weekday/weekend, season, and day/night condition.

MODEL CALIBRATION AND ANALYSIS

Two levels of models were developed. In Level 1 models all two-lane roads are placed in the same category; Level 2 models

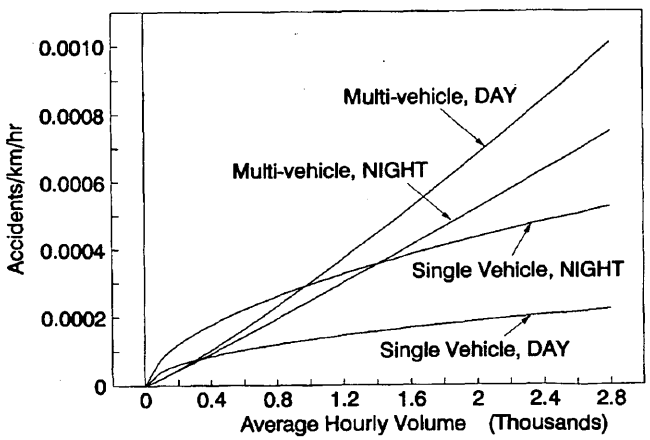

FIGURE 1 Plots of selected Level 1 models (all severity levels combined).

were calibrated separately for roads grouped according to geometric and other factors that may be associated with accident causation.

Level 1 Models

The availability of estimated hourly volumes for each road section made it possible to estimate models for various time periods. Three different periods were identified: 24 hr, daytime hours only, and nighttime hours only. For each period the models were calibrated separately for groupings on the basis of severity and whether the accident involved a single vehicle or multiple vehicles. The values of k and the model parameters are given in Table 1, and the graphs of selected models are depicted in Figure 1. Note that GLIM uses the linear form of Equation 1, resulting in estimates of $\ln(a)$ rather than a and that the model, as calibrated, is for estimating the number of accidents per hour per kilometer.

Examination of the curves in Figure 1 and the model parameters in Table 1 reveals two main points of interest. First, the relationship for single-vehicle accidents is described with a convex curve ($b < 1$), whereas that for multivehicle accidents is characterized by a concave curve ($b > 1$). Second, it appears that the potential for single-vehicle accidents is higher during nighttime hours, whereas for multivehicle accidents the opposite is true. These observations serve to emphasize the importance of analyzing single-vehicle and multivehicle accidents separately and also the usefulness of differentiating between nighttime and daytime accidents.

Level 2 Models

As indicated earlier the Level 1 regression models apply for all two-lane rural roads in the province of Ontario. It is natural that within this group there are sections with different geometric characteristics. Level 2 models are the result of an attempt to create more homogeneous groups of road sections. The importance of second-level models is twofold. First, models built for subgroups provide better predictions for member sections; second, comparison of the accident potentials on sections with different geometric features is facilitated.

To facilitate the development of the Level 2 models, subgroups of road sections with similar lane and shoulder width combinations were formed after exploratory analysis revealed that these features were important in explaining differences in accident occurrence on sections with similar traffic volumes. Tables 2 and 3 contain the estimated Level 2 model parameters and estimates of the coefficient k for four different geometric groups, whereas Figures 2 and 3 are plots of selected models.

In analyzing these results it is important to stress that accident prediction models do not necessarily explain accident causation; they only represent an estimate of the accident potential by taking into consideration factors associated with accident occurrence. Nevertheless, the following points are worth noting:

1. For the 24-hr period the highest accident potential for both single-vehicle and multivehicle accidents is on roads with total lane widths of 6.7 m and shoulders with widths of 2.4 m (narrow lanes and wide shoulders).
2. The lowest potential for single-vehicle accidents is on roads with total lane widths of 7.3 m and shoulders with widths of 3.0 m (wide lanes, wide shoulders).
3. By contrast, the lowest potential for multivehicle accidents is on roads with total lane widths of 6.7 m and shoulders with widths of 1.8 m (narrow lane, narrow shoulders).
4. Similar to what has been observed for Level 1 models, parameter b for multivehicle accidents is almost always greater than 1, indicating a convex relationship between accidents and traffic volume, whereas for single-vehicle accidents it is always less than 1.
5. By comparing the accident potential for day and night hours on a homogeneous group of sections in terms of lane and shoulder widths, it can be observed that the accident potential for single-vehicle accidents is higher during nighttime hours (Figure 3), whereas for multivehicle accidents the accident potential is higher during daytime hours.

TABLE 2 Parameter Estimates for Level 2 Models (Fatal and Injury Accidents)

Code	Model Group					Model Parameters			Sample Size	
	Lane Width	Shldr. Width	Acc. Type	Severity Group	Time Period	ln(a) (std. error)	b (std. error)	k	# of Accs.	Total km
201	6.7	1.8	S.V.	F.& I.	24 Hr	-12.40 (.12)	0.291 (.027)	2.0	225	1899
202	6.7	1.8	M.V.	F.& I.	24 Hr	-17.36 (.15)	1.212 (.029)	0.8	116	1899
203	6.7	1.8	All	F.& I.	24 Hr	-13.29 (.12)	0.589 (.027)	4.0	342	1899
204	6.7	2.4	S.V.	F.& I.	24 Hr	-13.30 (.12)	0.491 (.026)	1.8	274	1953
205	6.7	2.4	M.V.	F.& I.	24 Hr	-17.85 (.17)	1.321 (.033)	1.1	201	1953
206	6.7	2.4	All	F.& I.	24 Hr	-14.29 (.13)	0.786 (.026)	1.9	476	1953
207	7.3	2.4	S.V.	F.& I.	24 Hr	-11.63 (.14)	0.214 (.027)	2.1	376	1694
208	7.3	2.4	M.V.	F.& I.	24 Hr	-17.47 (.21)	1.236 (.036)	1.5	259	1694
209	7.3	2.4	All	F.& I.	24 Hr	-12.92 (.15)	0.568 (.027)	1.7	636	1694
210	7.3	3.0	S.V.	F.& I.	24 Hr	-11.75 (.13)	0.201 (.023)	0.9	330	1649
211	7.3	3.0	M.V.	F.& I.	24 Hr	-15.72 (.16)	0.956 (.027)	0.8	441	1649
212	7.3	3.0	All	F.& I.	24 Hr	-13.06 (.13)	0.592 (.023)	1.3	772	1649

TABLE 3 Parameter Estimates for Level 2 Models (All Severity Groups Combined)

Code	Model Group					Model Parameters			Sample Size	
	Lane Width	Shld. Width	Acc. Type	Severity Group	Time Period	ln(a) (std. error)	b (std. error)	k	# of Accs.	Total km
301	6.7	1.8	S.V.	Total	24 Hr	-11.27 (.09)	0.342 (.022)	1.6	861	1899
302	6.7	1.8	M.V.	Total	24 Hr	-15.95 (.14)	1.101 (.027)	1.2	271	1899
303	6.7	1.8	All	Total	24 Hr	-11.72 (.10)	0.509 (.022)	2.0	1133	1899
304	6.7	2.4	S.V.	Total	24 Hr	-11.86 (.11)	0.448 (.022)	4.9	932	1953
305	6.7	2.4	M.V.	Total	24 Hr	-16.92 (.16)	1.323 (.029)	1.0	515	1953
306	6.7	2.4	All	Total	24 Hr	-12.67 (.11)	0.704 (.021)	2.6	1448	1953
307	7.3	2.4	S.V.	Total	24 Hr	-10.82 (.12)	0.284 (.023)	2.4	1203	1694
308	7.3	2.4	M.V.	Total	24 Hr	-16.41 (.18)	1.206 (.032)	1.2	622	1694
309	7.3	2.4	All	Total	24 Hr	-11.78 (.12)	0.552 (.022)	2.6	1826	1694
310	7.3	3.0	S.V.	Total	24 Hr	-10.59 (.11)	0.221 (.020)	3.0	1175	1649
311	7.3	3.0	M.V.	Total	24 Hr	-15.36 (.14)	1.038 (.024)	0.8	1020	1649
312	7.3	3.0	All	Total	24 Hr	-11.81 (.11)	0.558 (.019)	1.5	2196	1649
313	6.7	1.8	S.V.	Total	Day	-11.28 (.18)	0.292 (.037)	1.4	369	1899
314	6.7	1.8	M.V.	Total	Day	-16.20 (.20)	1.140 (.038)	3.3	173	1899
315	6.7	1.8	All	Total	Day	-12.24 (.17)	0.577 (.035)	2.4	543	1899
316	6.7	2.4	S.V.	Total	Day	-13.48 (.21)	0.707 (.041)	16	407	1953
317	6.7	2.4	M.V.	Total	Day	-17.47 (.24)	1.423 (.043)	1.0	359	1953
318	6.7	2.4	All	Total	Day	-14.53 (.20)	1.023 (.037)	2.3	767	1953
319	7.3	2.4	S.V.	Total	Day	-11.39 (.23)	0.349 (.042)	1.5	520	1694
320	7.3	2.4	M.V.	Total	Day	-17.56 (.28)	1.403 (.049)	1.1	409	1694
321	7.3	2.4	All	Total	Day	-13.26 (.23)	0.794 (.040)	1.9	930	1694
322	7.3	3.0	S.V.	Total	Day	-10.18 (.22)	0.105 (.038)	2.4	470	1649
323	7.3	3.0	M.V.	Total	Day	-15.96 (.23)	1.139 (.037)	0.8	677	1649
324	7.3	3.0	All	Total	Day	-12.83 (.21)	0.709 (.034)	1.2	1148	1649
325	6.7	1.8	S.V.	Total	Night	-11.80 (.12)	0.538 (.029)	3.1	405	1899
326	6.7	1.8	M.V.	Total	Night	-15.48 (.17)	0.993 (.037)	2.0	69	1899
327	6.7	1.8	All	Total	Night	-11.89 (.12)	0.600 (.028)	2.7	475	1899
328	6.7	2.4	S.V.	Total	Night	-12.18 (.12)	0.600 (.027)	20	452	1953
329	6.7	2.4	M.V.	Total	Night	-16.19 (.18)	1.179 (.036)	22	119	1953
330	6.7	2.4	All	Total	Night	-12.43 (.12)	0.711 (.026)	25	572	1953
331	7.3	2.4	S.V.	Total	Night	-11.66 (.14)	0.513 (.030)	4.2	594	1694
332	7.3	2.4	M.V.	Total	Night	-15.70 (.21)	1.068 (.040)	3.4	164	1694
333	7.3	2.4	All	Total	Night	-11.96 (.14)	0.627 (.029)	6.2	759	1694
334	7.3	3.0	S.V.	Total	Night	-11.71 (.14)	0.489 (.026)	2.5	583	1649
335	7.3	3.0	M.V.	Total	Night	-14.66 (.17)	0.903 (.030)	0.5	263	1649
336	7.3	3.0	All	Total	Night	-12.08 (.13)	0.619 (.025)	1.5	847	1649

EXAMPLE OF ACCIDENT PREDICTION BY EB METHOD

The following example illustrates how the regression prediction can be refined by the EB procedure to estimate the accident potential of an individual road section.

Suppose that it is desired to estimate the expected number of single-vehicle accidents during an a.m. peak hour (8:00 to 9:00 a.m.) for a 6-month period in 1989 for a 9-km section ($T = 183$ hr and $L = 9$ km in Equation 1) having 1.8-m shoulders and a roadway width of 6.7 m. The section had an average volume of 171 vehicles per hr ($F = 171$ in Equation 1) and recorded two single-vehicle accidents from 8:00 to 9:00 a.m. during the 6-month period of interest. First, the expected number of accidents is calculated by using Equation 1 and the appropriate regression parameters, in this case, for Model 301 in Table 3, for which $\ln(a)$ is equal to -11.27, b is equal to 0.342, and k is equal to 1.6. This yields

$$E_{89}(m) = 9 \times 183 \times 0.00007395 = 0.1218 \text{ accidents}$$

with variance given by Equation 2,

$$\text{VAR}_{89}(m) = 3 \times 10^{-9} \times 9^2 \times 183^2 = 0.00927 \text{ accidents}^2$$

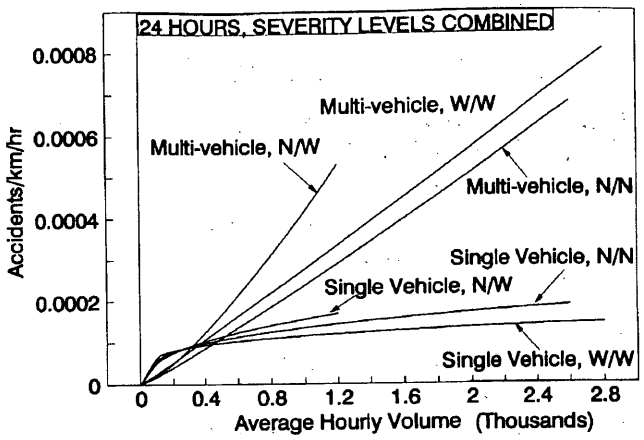

FIGURE 2 Plots of Level 2 models for various lane and shoulder width combinations (n = narrow; w = wide).

In refining this estimate, the weight w to be applied is (by Equation 4)

$$w = 1/[1 + E_{89}(m)] = 1/(1 + 0.1218/1.6) = 0.9293$$

The refined estimate (by Equation 3) is

$$E_{89}(m|x) = w[E_{89}(m)] + (1 - w)x_{89}$$

$$= 0.9293 \times 0.1218 + (1 - 0.9293) \times 2 = 0.2546 \text{ accidents}$$

with variance (Equation 5) given by

$$VAR_{89}(m|x) = (x_{89} + k)/[1 + k/E_{89}(m)]^2$$

$$= (2 + 1.6)/(1 + 1.6/0.1218)^2 = 0.0180 \text{ accident}^2$$

MODEL VALIDATION

This section shows that the EB method provides estimates of accident potential that are better than the results that would be obtained by a method that is based only on accident counts and that assumes that the number of accidents in one period is a good estimate of the accident potential for another period, after adjusting for traffic volume differences.

The method involved a comparison of the prediction accuracy resulting from the use of 1989 accident counts as estimates of the 1988 counts, as opposed to using the EB procedure on the basis of 1989 data. In either case the relative measure of the discrepancy between the predicted and observed numbers of accidents was the weighted squared difference between the two values. The weight is the squared product of the section length, hours of exposure, and the average hourly volume. This calculation was done for each hour for each section in the data set, and a mean value was calculated. It is assumed that the method that provides a lower value of this mean weighted squared difference is better. It can be seen from the values in Table 4 that the EB estimate provides lower values for the measure of accuracy than those obtained by using the accident count method, thus confirming its superiority.

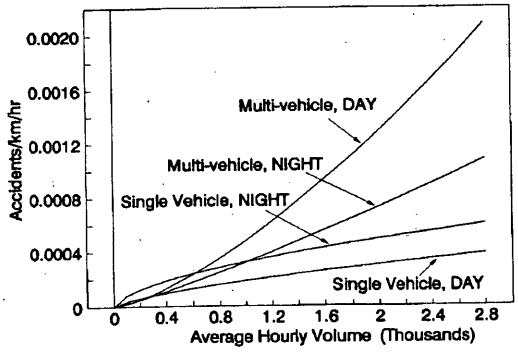

FIGURE 3 Plots of selected Level 2 models illustrating differences between daytime and nighttime accident frequency (severity levels combined: narrow lanes, wide shoulders).

ACKNOWLEDGMENTS

The research for this paper resulted from a project with MTO and was supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada. These sources of support are gratefully acknowledged.

TABLE 4 Mean* Squared Differences Between Predicted and Observed Numbers of Accidents Per Kilometer Per Hour

Model	Accident Count Method	EB Estimate
112	8.555	5.775
115	8.200	6.070
118	15.869	0.118
301	10.419	1.389
302	2.358	0.046
303	10.721	1.436
313	5.536	0.910
314	0.605	0.030
315	6.140	0.950
325	11.386	1.602
326	0.621	0.103
327	12.151	1.713

* Sum of weighted squared differences divided by the number of observations.

Values in Table are to be multiplied by 10^{-10}.

REFERENCES

1. Persaud, B. N. *Estimating the Accident Potential of Ontario Road Sections.* Project Report. Department of Civil Engineering, Ryerson Polytechnic University, Ontario, Canada, 1993.
2. Hauer, E. Empirical Bayes Approach to the Estimation of "Unsafety." The Multivariate Regression Approach. *Accident Analysis and Prevention,* Vol. 24, 1992, pp. 457–477.
3. Baker, R. J., and J. A. Nelder. *The GLIM System—Release 3.* Rothamsted Experimental Station, Harpenden, United Kingdom, 1978.
4. Satterthwaite, S. P. A Survey of Research into Relationships Between Traffic Accidents and Traffic Volumes. Supplementary Report SR 692. Transport and Road Research Laboratory, Crowthorne, United Kingdom, 1981.
5. Hauer, E. *Traffic Flow and Safety.* TRB, National Research Council, Washington, D.C., forthcoming.
6. Maycock, G., and R. D. Hall. *Accidents at 4 Arm Roundabouts.* Laboratory Report LR 1120. Transport and Road Research Laboratory, Crowthorne, United Kingdom, 1984.
7. Hauer, E. On the Estimation of the Expected Number of Accidents. *Accident Analysis and Prevention,* Vol. 18, 1986, pp. 1–12.

Publication of this paper sponsored by Committee on Traffic Records and Accident Analysis.