Paper No. 03-3363

Modeling Traffic Crash–Flow Relationships for Intersections

Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes Methods

Shaw-Pin Miaou and Dominique Lord

Statistical relationships between traffic crashes and traffic flows at roadway intersections have been extensively modeled and evaluated in recent years. The underlying assumptions adopted in the popular models for intersections are challenged. First, the assumption that the dispersion parameter is a fixed parameter across sites and time periods is challenged. Second, the mathematical limitations of some functional forms used in these models, particularly their properties at the boundaries, are examined. It is also demonstrated that, for a given data set, a large number of plausible functional forms with almost the same overall statistical goodness of fit (GOF) is possible, and an alternative class of logical formulations that may enable a richer interpretation of the data is introduced. A comparison of site estimates from the empirical Bayes and full Bayes methods is also presented. All discussions and comparisons are illustrated with a set of data collected for an urban four-legged signalized intersection in Toronto, Ontario, Canada, from 1990 to 1995. In discussing functional forms, the need for some goodness-of-logic measures, in addition to the GOF measure, is emphasized and demonstrated. Finally, analysts are advised to be mindful of the underlying assumptions adopted in the popular models, especially the assumption that the dispersion parameter is a fixed parameter, and the limitations of the functional forms used. Promising directions in which this study may be extended are also discussed.

Statistical relationships between traffic crash and traffic flows at roadway intersections have been extensively modeled and evaluated in recent years (*1–8*). The most popular and often recommended models have the following functional and probabilistic structures: the number of crashes at the *i*th intersection and *t*th time period (Y_{it}), when conditional on its mean (μ_{it}), is assumed to have a Poisson (Po) distribution and is independent (ind) over all intersections and time periods, as follows:

$$Y_{it}|\mu_{it} \overset{\text{ind}}{\sim} \text{Po}(\mu_{it}) \tag{1}$$

where i is equal to 1, 2, . . . , I and t is equal to 1, 2, . . . , T

The mean of the Poisson distribution is structured as

$$\mu_{it} = f(F_{1,it}, F_{2,it}, x_{it}; \beta)\exp(e_{it}) \tag{2}$$

Texas Transportation Institute, Texas A&M University System, 3135 TAMU, College Station, TX 77843-3135.

where

$f(\cdot)$ = function of traffic flows from the major and minor approaches of the intersection, represented by $F_{1,it}$ and $F_{2,it}$, respectively, and other covariates indicated by vector x_{it};
β = vector of unknown fixed-effect parameters; and
e_{it} = unstructured random effect independent of all covariates, which has a typical assumption that $\exp(e_{it})$ is independent and gamma distributed with a mean equal to 1 and variance $1/\phi$ for all i and t (where ϕ is the inverse dispersion parameter and ϕ is >0).

This particular formulation provides flexible and attractive statistical properties. For example, conditional on $f(\cdot)$ and ϕ, Y_{it} can be shown to be distributed as a negative binomial (NB) random variable with mean and variance of $f(\cdot)$ and $f(\cdot)\,[1+f(\cdot)/\phi]$, respectively. Also, $\exp(e_{it})$ can be viewed as unmodeled (or unmeasured) heterogeneities because of omitted exogenous variables and intrinsic randomness. The key assumptions here are that (*a*) $\exp(e_{it})$ values are independent (or, more strictly and statistically speaking, exchangeable) across all i and t and have a fixed variance, and (*b*) e_{it} is independent of all covariates, including flows and x_{it}. In this paper, ϕ is called the "inverse dispersion parameter," in that the Poisson model can be regarded as a limiting model of NB as ϕ approaches infinity. Without much loss of generality, it is assumed in the discussion that one time period is 1 year. See the book *Bayesian Data Analysis* for a Bayesian interpretation of "fixed" and "random" effects (*9*).

Under this formulation, it has been widely suggested that estimates of the expected number of crashes at individual sites and time periods can be relatively easily derived by using an empirical Bayes (EB) method, as follows:

$$\hat{\hat{\mu}}_{it} = g(\hat{\mu}_{it}, \hat{\phi})\hat{\mu}_{it} + [1 - g(\hat{\mu}_{it}, \hat{\phi})]y_{it} \tag{3}$$

where

$\hat{\hat{\mu}}_{it}$ = EB estimate of the expected number of crashes,
$\hat{\mu}_{it}$ = maximum-likelihood estimate (MLE) of $f(\cdot)$, and
y_{it} = observed number of crashes for site i and time t.

In addition, the function $g(\hat{\mu}_{it}, \hat{\phi})$ is equal to $1/(1 + \hat{\mu}_{it}/\hat{\phi})$, a function of the estimated inverse dispersion parameter $\hat{\phi}$ and $\hat{\mu}_{it}$. For a fixed $\hat{\mu}_{it}$, as $\hat{\phi}$ increases from 0 to infinity, the EB estimate would increase the weight on $\hat{\mu}_{it}$ from 0 to 1 and, conversely, decrease the weight on y_{it} from 1 to 0. For references, see Hauer et al. (*1*) and Persaud et al. (*8*). It has also been implied in many papers cited earlier that this

EB estimate outperforms other estimates, such as the MLEs, and generally enjoys good precision.

Many functional forms for $f(\cdot)$ have been postulated. They differ by the types of covariates included, the forms of variable transformation applied, or the parameterization schemes used (*10*). Even when a small number of covariates are considered, a large number of plausible functional forms are possible. Among others, instruments exercised by the analyst to determine the potential functional form include engineering logics; exploratory data and graphical analysis; and the resolution and availability of flows, other covariates, and crash data. With regard to the crash frequency–flow relationships, Turner and Nicholson (*3*) summarized commonly used functional forms into three types. Type 1 relates the total number of crashes to the total traffic volumes entering the intersection. Instead of total traffic volumes, Type 2 separates flows into two approaches entering the intersection (major and minor roads, respectively). Type 3 models the number of crashes involving individual pairs of conflicting vehicle movements at the intersection, which requires detailed turning flows and crash movement types. Other disaggregate models that analyze crashes by impact type (rear end, right angle, etc.), severity (fatal, injury, etc.), and time of day have also been developed (*5*). This paper focuses on the second type of functional form, which is quite favorably considered and evaluated in all references cited earlier. Note, however, that the discussion and findings presented in this paper apply to other types as well. The first four equations shown in Table 1 (column 2) are considered the most popular ones.

The objective of this study was to challenge some of the underlying assumptions adopted in the traffic–crash estimation model for intersections, as presented above. More specifically, first, the assumption that the dispersion parameter ϕ is a fixed parameter across i and t was challenged. Given the complexity of the traffic interaction around and within intersections and the fact that driver behaviors are influenced by a multitude of factors pertaining to roadside development near intersections, especially the business environment (e.g., shopping centers, schools, and office compounds), it seems natural to suspect that the unmodeled heterogeneity across sites might be structured spatially in some way. This is especially the case when a limited number of covariates are considered in the model and for urban intersections for which a wide range of traffic flow conditions are analyzed. As seen from Equation 3, if the inverse dispersion parameter is indeed structured, it would have a serious implication on the EB estimates, in which ϕ is assumed to be a fixed parameter. Second, the mathematical limitations of some popular functional forms of $f(\cdot)$, particularly their properties at the boundaries, were examined. It is also demonstrated that, for a given data set, a large number of plausible functional forms with almost the same overall statistical goodness of fit (GOF) are possible. Furthermore, an alternative class of logical formulations that could potentially provide richer interpretation of the data is introduced.

In this study, a full Bayes approach was taken for model specification and estimation. Thus, even though it was not the focus of this study, the opportunity to compare site estimates between the full Bayes and the EB approaches was available. The advantage of full Bayes treatment is that it takes into account the uncertainty associated with the estimates of the model parameters and can provide exact measures of uncertainty. The maximum-likelihood and the EB methods, on the other hand, tend to overestimate precision because they ignore this uncertainty (*11*). For example, the estimate derived by the EB method in Equation 3 was developed with the assumption that $f(\cdot)$ and ϕ are both estimated precisely, without errors, which is a questionable assumption in practice. This advantage of the full Bayes approach is especially important when the sample size is relatively small.

TABLE 1 Commonly Used Functional Forms, Alternative Forms, and Estimated Parameters for 1995 (Year 6) and Associated Statistics

No.	Functional Form $f(\cdot)$	$\beta_{0,6}$	$\beta^*_{0,6}$	β_1	β_2	β_3	ϕ	Deviance	η_0	η_1	η_2	η_3
1	$\beta_{0,t}(F_{1,it}+F_{2,it})^{\beta_1}$	0.05884 (±0.005)		1.4220 (±0.024)			4.053 (±0.170)	25410 (±102.9)				
2	$\beta_{0,t}F_{1,it}^{\beta_1}F_{2,it}^{\beta_2}$	0.41111 (±0.027)		0.5441 (±0.019)	0.6504 (±0.009)		6.643 (±0.238)	25480 (±101.2)				
3	$\beta_{0,t}(F_{1,it}F_{2,it})^{\beta_1}$	0.32726 (±0.014)		0.6283 (±0.007)			6.592 (±0.229)	25470 (±102.0)				
4	$\beta_{0,t}(F_{1,it}+F_{2,it})^{\beta_1}\left(\frac{F_{2,it}}{F_{1,it}}\right)^{\beta_2}$	0.2001 (±0.013)		1.1950 (±0.016)	0.3770 (±0.009)		6.586 (±0.235)	25450 (±102.2)				
5	$\beta_{0,t}F_{1,it}^{\beta_1}F_{2,it}^{\beta_2}\exp(\beta_3F_{2,it})$	0.4711 (±0.031)		0.5259 (±0.018)	0.5663 (±0.021)	0.00878 (±0.002)	6.651 (±0.235)	25460 (±101.3)				
5*	$\beta_{0,t}F_{1,it}^{\beta_1}F_{2,it}^{\beta_2}\exp(\beta_3F_{2,it})$ (ϕ is a function of flows and flow ratios)	0.4615 (±0.032)		0.5332 (±0.018)	0.5757 (±0.022)	0.00746 (±0.002)	See Eq. (4)	25380 (±103.0)	2.2200 (±0.180)	-0.02126 (±0.005)	0.05452 (±0.009)	-0.8783 (±0.265)
6	$F_{1,it}\lambda_{1,it}+F_{2,it}\lambda_{2,it}$ where $\lambda_{1,it}=\exp(\beta_{0,t}+\beta_1F_{2,it})$ $\lambda_{2,it}=\exp(\beta^*_{0,t}+\beta_2F_{1,it})$	-2.4280 (±0.083)	-0.54050 (±0.047)	0.00474 (±0.008)	0.01034 (±0.001)		6.578 (±0.232)	25440 (±99.4)				

NOTES: (1) Flows are in 1,000 vehicles per day. (2) Functional forms No. 1 to No. 4 are commonly used in previous studies. (3) Functional form No. 2 is the most popular form. (4) Functional form No. 5 was used by Lord (6) for this particular data set. (5) Functional form No. 5* is the same as functional form No. 5, but with an assumption that the inverse dispersion parameter ϕ is a function of flows and flow ratios. (6) Functional form No. 6 was an alternative form postulated by this study to represent two different risk levels for vehicles entering the two approaches. (7) All models were structured by using the full Bayes framework with noninformative priors (or hyperpriors). (8) Parameters (β's, ϕ, and deviance) were estimated by MCMC techniques, and the values shown in the table are their posterior means. (9) Values in parentheses are the estimated 1 standard error of the value given above. (10) Each year has a separate intercept term. (11) Deviance is computed as $-2\log[\text{Prob}(y|\text{estimated parameters})]$, where log is the natural logarithm.

The rediscovery of the Markov Chain Monte Carlo (MCMC) methods and new developments, including convergence diagnostic statistics, by statisticians in the last 15 years or so are revolutionizing the entire field of statistics (*12–16*). At the same time, the complexity of statistical models in practice has increased dramatically, mainly because of improved computer processing speeds and lower data collection and storage costs. These models are often hierarchical and high dimensional in both their probabilistic and functional structures. Furthermore, many models need to include dynamics of unobserved and unobservable (or latent) variables; handle data with distributions that are heavily tailed, highly overdispersed, or multimodal; and work with data sets with missing data points. MCMC provides a unified framework within which model identification and specification, parameter estimation, performance evaluation, inference, and communication of complex models can be conducted in a consistent and coherent manner.

With the computing power of today's desktop computers, sampling of posterior distributions by MCMC methods, which is needed in the full Bayes methods, can now be achieved relatively easily. For some problems, existing statistical programming software packages, such as WinBUGS (*17*) and MLwiN (*18*), can be used quite nicely. These provide Gibbs and other MCMC sampling for a variety of so-called hierarchical Bayes models. For most of the models presented in this study, the parameters and inferences were obtained by using programs coded in the WinBUGS language.

For the problems to be presented next, because the number of parameters (relative to the sample size) is very small and, in addition, MLEs are used as the initial estimates, the MCMC converges in just a few iterations. Note that the Gelman-Rubin statistics in WinBUGS were used to check for convergence (*17*). As in other iterative parameter estimation approaches, good initial estimates are always the key to a quick convergence.

DATA

To motivate this study, a set of data collected for an urban four-legged signalized intersection in Toronto, Ontario, Canada, from 1990 to 1995 was used. The quality of the data has been checked and analyzed as part of a network data set used by Lord (*6*) and further studied by Lord and Persaud (*7*) and Persaud et al. (*8*). It includes data for 868 intersections, for which a total of 54,989 crashes were reported over the 6-year period, with an approximately 30% and 70% split of injury and noninjury crashes, respectively. Individual intersections experienced from 0 to 63 crashes per year. Traffic volumes varied widely from intersection to intersection: from about 5,300 to 72,300 vehicles per day for main approaches (summing over both directions) and from 52 to about 42,600 vehicles per day for minor approaches. Traffic densities (or entry flows per lane per day, to be precise) varied from 1,300 to 14,000 vehicles per lane per day for major approaches and from 52 to 13,000 vehicles per lane per day for minor approaches. Flow ratios, calculated as the minor approach volume over the major approach volume (F_2/F_1), ranged from 0.2% to almost 100%, that is, from linklike (or segmentlike) intersections to intersections with equal flows from the two approaches. Note, however, for major flows with 51,000 vehicles per day or higher, the flow ratios are all less than 50%. The number of lanes for the major approach varied from 2 to 10 lanes (including exclusive left-turn and right-turn lanes), while for the minor approach it varied from 1 to 8 lanes. The data set contains data for intersections with a mix of fixed and actuated traffic signals, including intersections that are part of an adaptive traffic control system. In terms of signal phasing, the data set comprised data for intersections that have permissive, semiprotected, and protected left turns. It also includes divided and undivided approaches. The speed limits on the main approaches varied from 50 km/h (30 mph) to 70 km/h (43 mph).

Overall, the data in the Toronto intersection data set have wide variations in traffic flows, densities, flow ratios, and number of lanes. To give a general sense of the ranges of crash frequencies (Y), flows (F_1 and F_2, in 1,000 vehicles per day), and densities (D_1 and D_2, in 1,000 vehicles per lane per day) and their relationships, Figure 1 shows a set of scatter plots for the 1995 data. Note that the intent of Figure 1 is to give the readers some sense of the data considered in this study and is not meant to be very precise. For a detailed description of the data set, the readers are referred to Lord (*6*). It is noted here that the definition of intersection crashes has not been consistent in previous studies. For example, in Vogt and Bared (*4*), the crashes considered "were ones labeled intersection or intersection-related . . . and occurring within 76.2 m (250 ft) of the intersection milepost." In Kulmala (*19*), "Each junction was examined as a junction area: the road area within 200 metres from the centre of the junction," which was about 656 ft from the center. This definitional issue makes attempts to compare statistical relationships developed in different studies very difficult, if not impossible. The setting of some standards to define, select, and report intersection crashes for modeling purposes within the roadway safety community is certainly needed. In the Toronto data set used in the present study, crashes included both intersection and intersection-related crashes, as reported by the police, that were located within about 15 m (50 ft) from the center of the intersection. In addition, the data set does not include crashes involving pedestrians, animals, or cyclists.

DISPERSION PARAMETER

As stated earlier, given the complexity of the interaction of traffic around and within intersections and the interaction of traffic with the roadside environment, it is natural to suspect that the unmodeled heterogeneity, and thus the inverse dispersion parameter, might be structured spatially in some way. Because of the richness of the Toronto urban intersection data set, this is a good candidate data set with which to test this suspicion. A previous study by Heydecker and Wu (*20*) discussed how the dispersion parameter could be modeled as a function of covariates.

The test begins with a full Bayes estimate of the model presented in Equations 1 and 2 by using the functional form listed as No. 5 in Table 1, which was identified as the best among the others considered by Lord (*6*). Key estimated parameters are shown in Table 1. Note that the estimated parameters are similar to the MLEs reported by Lord (*6*). Then, the same model is subjected to a different assumption that the inverse dispersion parameter ϕ varies by individual sites and time period and is a function of traffic flows and flow ratios:

$$\phi_{it} = \exp(\eta_0 + \eta_1 F_{1,it} + \eta_2 F_{2,it} + \eta_3 F_{2,it}/F_{1,it}) \quad (4)$$

where η_1 to η_3 are unknown model parameters.

In a sense, traffic flows are used here as surrogate variables to explain the possible structure of unmodeled space–time heterogeneities. The modeling results are shown in Table 1 under functional form No. 5*. The estimated model suggests that ϕ's are indeed functions of flows. To see how ϕ varies by intersection and flows,

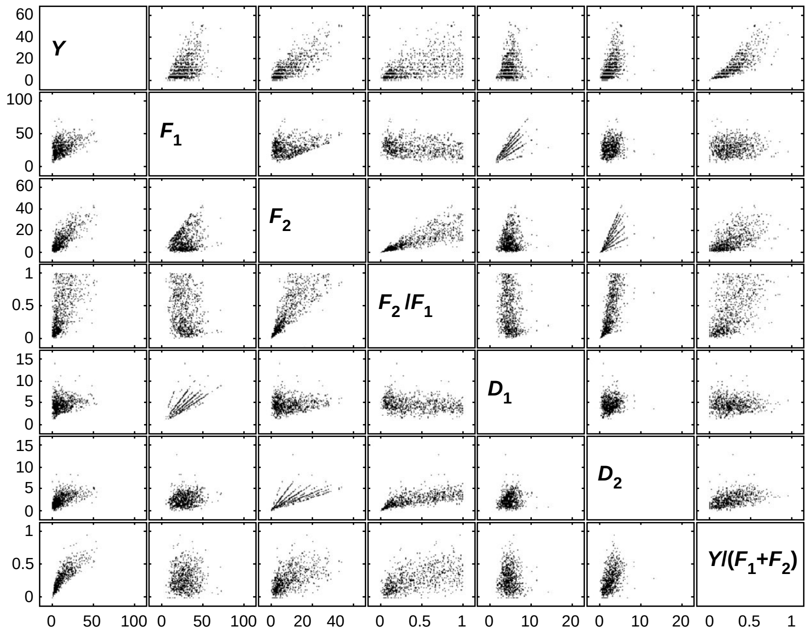


FIGURE 1 Ranges of and relationships among crash frequencies, traffic flows, and traffic densities for 1995.

the relative frequency of its mean (or posterior mean, to be precise) is presented in Figure 2*a*, the response surface by major flow and flow ratio is presented in Figure 2*b*, and the geographical distribution is presented in Figure 2*c*. It varies widely from 3 to about 18, with a mean of 6.62 and a standard error of 1.61. Because all the models used in the papers cited earlier do not have spatial components, the authors suspect that these models may have structured ϕ's as well. Ignoring the extent of this variation can seriously undermine the goodness of any estimate of individual sites, including both full Bayes and EB estimates with a fixed ϕ, which will be demonstrated in a later section. As will also be shown later, for this particular data set, the estimated parameters in the mean function $f(\cdot)$, that is, $\hat{\beta}$, change only slightly when the assumption is altered from a fixed ϕ to flow-varying ϕ's.

FUNCTIONAL FORM

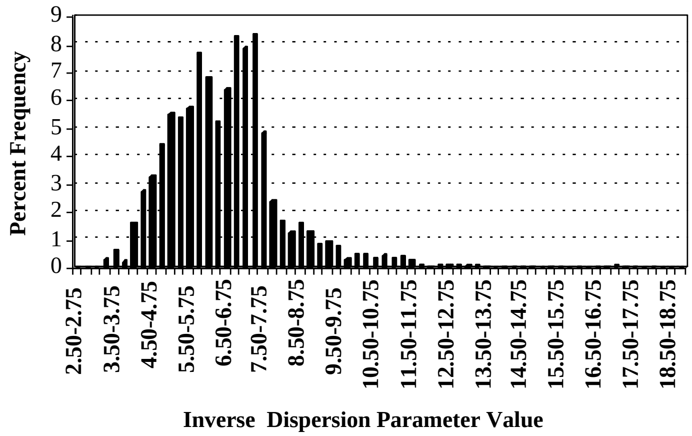
It has been suggested by most studies cited earlier that functional form No. 2—and, thus, functional form No. 5—in Table 1 has two main advantages: (*a*) the form follows the logic of "no traffic flows, no accidents," and (*b*) the form allows the relationship between accidents and traffic flows to be nonlinear, which has been advocated by most studies. This particular functional form has been investigated extensively and dates as far back as the 1950s (*21*). The essence of the first advantage cited is based on the logic of having proper boundary values, such as in the construct of differential equation systems. The second advantage is not a logical one but, rather, is one that is based on previous experience working with different data sets with a combination of visual inspections and statistical tests (*10*).

To take a closer look at the mathematical properties and performance of the four popular functional forms presented in Table 1, a full Bayes approach was again taken with Equations 1 and 2 as the basic hierarchy in the model. The results from the Toronto data set, including the posterior means of the estimated parameters and the associated statistics, are presented in Table 1. Figure 3 exhibits the estimated response surface for each of these functional forms. The coordinate on the right is for major flows from 0 to 80 in 1,000 vehicles per day, the one on the left is for flow ratios from 0 to 1, and the *z*-coordinate is the expected number of crashes. Note that the boundary conditions refer to, first, when the flow at the main approach is close to 0 (and so is the flow for the minor approach in this case) and, second, when the flow at the minor approach is close to 0 or, equivalently, the flow ratio F_2/F_1 approaches 0, which occurs in those linklike intersections.

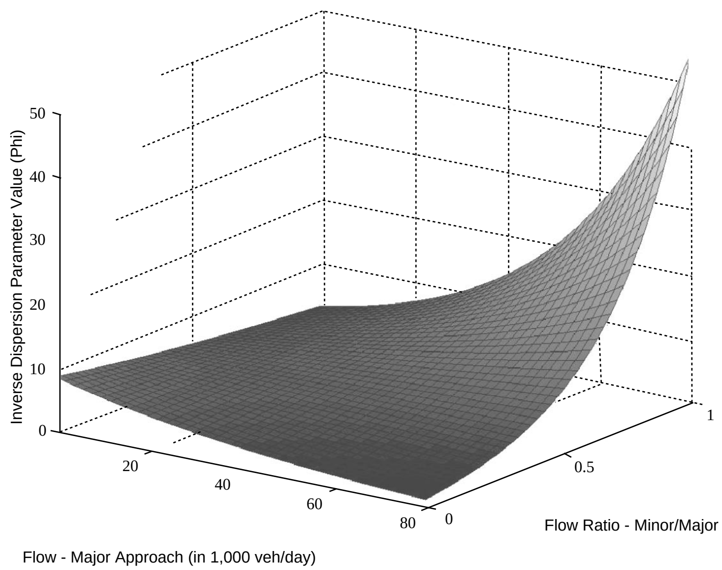
The deviance statistics, as a measure of overall GOF, and their standard errors for the first five model forms in Table 1 indicate that there is no clear difference in performance among these forms from a statistical standpoint. Even though there are indications of differences in estimated ϕ's for some of these models, they should be used with caution because, as in functional form No. 5, all models were found in this study to have ϕ potentially structured as a function of flows.

With the rapid development in local regression and smoothing techniques (including spline function-based semiparametric regressions) and the associated software in the last decade or so (*22*), production of a good fit to the data per se is no longer a challenge. Overstretching and overinterpretation of the data, on the other hand, can become serious concerns when such techniques are used. Therefore, there is a need to develop systematic criteria and perhaps new approaches that are nonstatistical in nature to help balance the development and evaluation of functional forms. It is envisioned that these criteria and approaches should be based on the logic (e.g., reason, consistency, and coherency), flexibility, extensibility, and interpretability of the functional form. At present, however, these criteria and approaches are almost nonexistent. In other words, plenty of GOF measures are available from statistics, but no "goodness-of-logic" measures are available from engineering and other perspectives.

With this limitation on the state of the research in mind, what follow are attempts to (*a*) offer some observations on the flexibility and boundary properties of the popular functional forms presented in Table 1, (*b*) illustrate alternative logic and functional forms, and (*c*) introduce potential logical constraints into the functional form. It should be noted that these attempts are meant to be demonstrative and suggestive in nature.



(a)



(b)

(c)

FIGURE 2 Estimated inverse dispersion parameter values: (*a*) relative frequency by site, (*b*) response surface by major flow and flow ratio, and (*c*) geographical distribution by site for 1995.

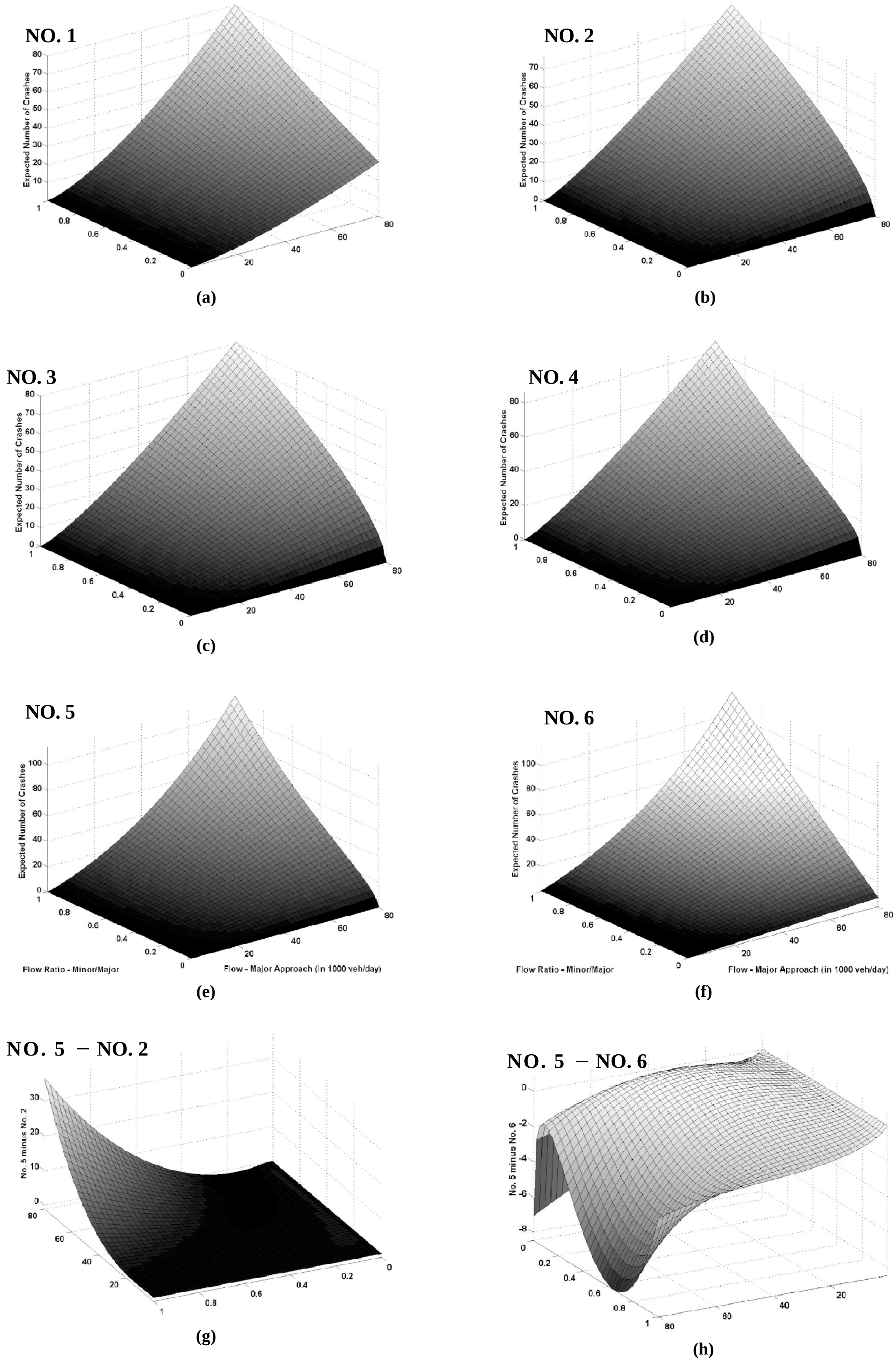


FIGURE 3 Response surfaces of functional forms listed in Table 1 and their differences by major flow and flow ratio.

Flexibility and Boundary Values

- Functional form No. 5 is more flexible than functional form No. 2, and functional form No. 2 is more flexible than functional form No. 3. This is simply because the less flexible one is a special case of the more flexible one, which includes more unknown parameters. All these forms suffer from a common limitation of restricting the response to 0 when F_2 is 0, which is a limitation for modeling the linklike intersections mentioned earlier. This can be seen in Figure 3 by comparison of the response surfaces of these functional forms with that of functional form No. 1, on which this restriction is not imposed. Note that even when F_2 is 0, vehicles on the major approach (going through, making a left turn, or making a right turn) can still be involved in crashes with vehicles on the same approach. Following exactly the same boundary value logic that made these forms advantageous, this limitation certainly goes against this type of functional form in logic.
- At first glance, it seems that functional form No. 1 is a special case of functional form No. 4, but it is clear from the response surfaces in Figure 3 that this is not the case. Functional form No. 4 suffers from the same boundary limitation as the other forms discussed above. For a linklike intersection, functional form No. 1 has estimated an increase in the expected number of crashes per year from 0 to about 30 as the flow of the major approach increases from 0 to 80,000 vehicles per day. Intuitively, this estimate seems high. It also appears to be so when its response surface is compared with that of an alternative functional form, which will be discussed shortly.
- Compared with the first four functional forms, functional form No. 5 estimated a much larger expected number of crashes at high flows and high flow ratios. The difference of the response surface between functional form No. 5 and functional form No. 2 is included in Figure 3 as well. The additional term in functional form No. 5 for F_2 plays a significant role in producing estimates higher than those from functional form No. 2.

Interpretability and Alternative Logics

To demonstrate that many other functional forms can be contemplated on the basis of logic and can perform equally well statistically, functional form No. 6 was considered as presented in Table 1. That is,

$$f(\cdot) = F_{1,it}\lambda_{1,it} + F_{2,it}\lambda_{2,it} \tag{5}$$

where $\lambda_{1,it} = \exp(\beta_{0,t} + \beta_1 F_{2,it})$, $\lambda_{2,it} = \exp(\beta^*_{0,t} + \beta_2 F_{1,it})$, and β's are unknown parameters. This functional form is based on a simple logic that vehicles entering from the major and minor approaches may have different risks, characterized by their crash rates, λ_{1t} and λ_{2t}, respectively. Take a vehicle entering the intersection from the major approach as an example. The functional form first postulates that this vehicle is exposed to a certain level of crash risk involving the vehicle itself and vehicles in the same approach, which is captured by parameter $\beta_{0,t}$. Second, this vehicle is exposed to a crash risk because of vehicles entering from the minor approach, which is captured by the term $\beta_1 F_{2,it}$. The same logic is applied to a vehicle entering from the minor approach.

Functional form No. 6 provides a richer interpretation of the data than the other forms discussed earlier because of introduction of the logic of differential risks. It is also clear that this form does not suffer the boundary value limitation of the other forms discussed earlier. The modeling results for the Toronto data set are included in Table 1 and Figure 3. The deviance statistics again suggest that this functional form is as good as other forms for this particular data set. Despite the apparent differences in functional form between functional form No. 6 and functional form No. 5, their response surfaces match quite well (compared with the other forms) for the ranges of flows and flow ratios, as shown in Figure 3.

The alternative form suggests an increase from 0 to about 7 crashes per year for linklike intersections when major flows increase from 0 to 80,000 vehicles per day, which is much lower than what functional form No. 1 suggests (from 0 to 30 crashes). One way to verify these results is to compare them with a crash estimation model for road segments in the same study area. In the work by Lord (*6*), the expected number of crashes per year for a six-lane segment at midblock for the year 1995 was estimated as $\hat{f}(\cdot) = \exp(-13.828)L^{0.665}\,F^{1.616}$, where L and F are segment length (in kilometers) and flows (in vehicles per day), respectively. By taking L equal to 30 m (0.03 km) (as discussed earlier, in which the intersection radius was defined as 15 m from the center) and F equal to 80,000, $\hat{f}(\cdot)$ is equal to $\exp(-13.828)(0.03)^{0.665}(80{,}000)^{1.616}$, which is equal to 8, which is very close to what the alternative functional form estimates. This could be a pure coincidence, but it certainly gives some comfort to the consistency of the performance of the alternative functional form at the boundary.

Extensibility and Logical Constraints

The imposition of logical constraints to the functional form reduces the solution space and thus decreases the achievable GOF of the resulting model. It can, however, enrich the logical interpretation of the functional form, complement the limitation of data in size and coverage, and potentially allow the estimated response surface to be more extensible beyond the data range. By using the system engineering language, logical constraints increase observability and enable a better estimate of the true state of the system under study.

To the best of the authors' knowledge, the use of logical constraints has not been attempted in the development of the intersection crash estimation models discussed in this paper. Only a limited demonstration of this concept with the alternative functional form in Equation 5 is provided here.

So far, the boundary value associated with linklike intersections has been discussed, in which flow ratios (r_{it}'s), which are equal to $F_{2,it}/F_{1,it}$, are close to 0. Now, the other end of the extreme is examined, in which r_{it}'s are close to 1, that is, equal flows from the two approaches. At this end, under the logic behind Equation 5, a vehicle entering from the minor approach could be expected to have about the same risk level as a vehicle entering from the major approach, provided that the traffic densities, lane configurations, and signal-phasing plans are about the same between the two approaches. To demonstrate how this logical constraint may be introduced into Equation 5, the crash rate is modified as follows:

$$\lambda^*_{1,it} = \exp(\beta_{0,t} + \beta_1 F_{2,it})$$

$$\lambda^*_{2,it} = \exp\left(\beta^*_{0,t} + \beta_2 F_{1,it}\right)$$

$$\lambda_{1,it} = \lambda^*_{1,it}\exp[h(D_{1,it};\omega)] \qquad \text{and}$$

$$\lambda_{2,it} = \left[r^{\delta}_{it}\lambda^*_{1,it} + \left(1 - r^{\delta}_{it}\right)\lambda^*_{2,it}\right]\exp[h(D_{2,it};\omega)] \tag{6}$$

where $h(\cdot\,; \omega)$ is a functional of traffic density $D_{1,it}$ and $D_{2,it}$ with an unknown parameter vector (ω) that is common to both approaches and that is equal to 0 when the density is 0, and δ is a parameter with a positive real value that governs the manner in which the risk level of a vehicle entering from the minor approach moves toward the risk level of a vehicle entering from the major approach as r_{it} increases from 0 to 1. Note that for δ equal to 1, the logical constraint is linear. It can be checked that $\lambda^*_{2,it}$ is a nominal crash rate for the minor approach when r_{it} is 0. That is, when r_{it} is equal to 0, $D_{2,it}$ is equal to 0 and $\lambda_{2,it}$ is equal to $\lambda^*_{2,it}$. Also, when r_{it} is equal to 1 and $D_{1,it}$ is equal to $D_{2,it}$, then $\lambda_{1,it} = \lambda_{2,it} = \lambda^*_{1,it} \exp[h(D_{1,it}; \omega)]$. Equation 6 is just one way of implementing the desired logical constraint, and there are other plausible alternatives.

Equation 6 was demonstrated with the Toronto data set by using δ equal to 0.5, 1.0, 2.0, and 3.0. The implementation also allowed the inverse dispersion parameter to be different for the two approaches and each inverse dispersion parameter to be dependent on the flows, as in Equation 4. For the density effect function $h(\cdot\,; \omega)$, the second-order grafted polynomial regression described by Fuller (*23*) was used with breakpoints (or knots) set at traffic densities of 2,000, 5,000, 8,000, and 11,000 vehicles per lane per day. Note that grafted polynomials passing through a set of given points and joined in such a way that they satisfy certain restrictions on the derivatives are also called "spline functions." The estimated density effects, represented by $\exp[h(\cdot\,; \omega)]$, from the grafted polynomials at different δ values are shown in Figure 4. A piecewise linear function with δ equal to 1.0 and with equal incremental density intervals of 1,000 vehicles per lane per day is also shown in Figure 4 for comparison.

In general, it was found that the estimated values for β at different δ's are close. The estimated density effects are similar for δ equal to 1.0, 2.0, and 3.0; and more pronounced density effects are indicated with δ equal to 0.5. The piecewise linear density effects indicate a wild fluctuation beyond 8,000 or 9,000 vehicles per lane per day, which suggests that the data may not be sufficient (in sample size) to allow a good estimate beyond that range.

BAYES VERSUS EMPIRICAL BAYES METHODS

As indicated in the introduction, the advantage of full Bayes treatment is that it takes account of the uncertainty associated with the estimates of the model parameters and can provide exact measures of uncertainty. The EB estimate for an individual site as presented in Equation 3 is developed with the assumption that $f(\cdot)$ and ϕ are both estimated precisely without errors, which is, of course, not true in practice. The extent to which this EB estimate for an individual site deviates from the full Bayes estimate depends on the uncertainty of the estimates for $f(\cdot)$ and ϕ.

For the Toronto data set used in this study, the uncertainties associated with the estimated values for $f(\cdot)$ and ϕ are relatively small. For the purpose of demonstration, functional form No. 5 with a fixed inverse dispersion parameter was chosen. Then, the site estimates from a model estimated by using an MLE reported by Lord (*6*) and the full Bayes model in Table 1 were compared. The result of this comparison for 1995 is shown in Figure 5, in which the percent difference for each site is computed as (EB estimate − full Bayes estimate) · 100/full Bayes estimate. The differences are less than 3% in 1995 (and less than 10% for all time periods). These differences can become quite large for other data sets if $f(\cdot)$ and ϕ are poorly estimated.

FIXED VERSUS VARYING DISPERSION PARAMETER

In contrast, ignoring the fact that ϕ is a function of flows and varies widely from 3 to about 18 by site, as demonstrated earlier under functional form No. 5*, posts a much more serious error in any estimate, whether it is from the EB or the full Bayes approach. To get a feel for the size of this error with the Toronto data, the same EB estimate obtained as described above was compared with that obtained by a full Bayes approach that allows the inverse dispersion parameter to vary with flows (which is presented under functional form No. 5* in Table 1). The differences for 1995 are shown in Figure 6. It is observed that treatment of ϕ as a fixed parameter can seri-

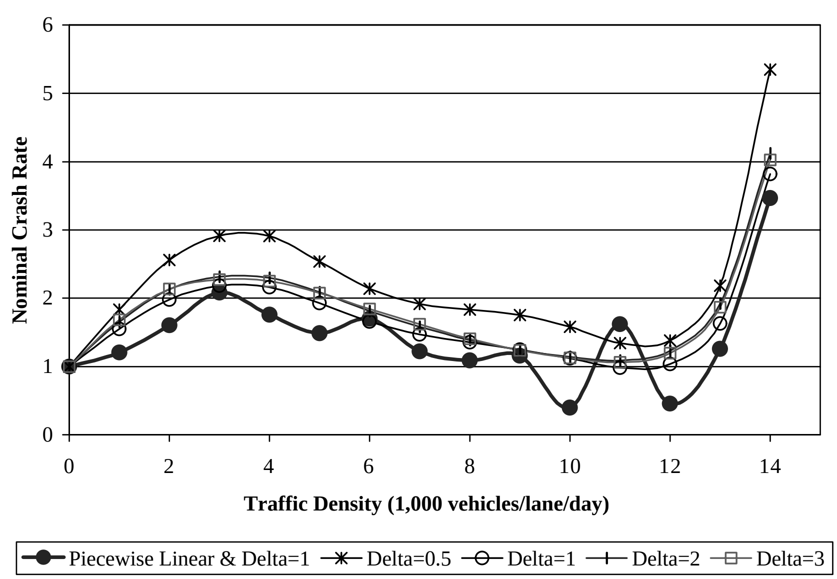


FIGURE 4 Density effects at various levels of logical constraints.

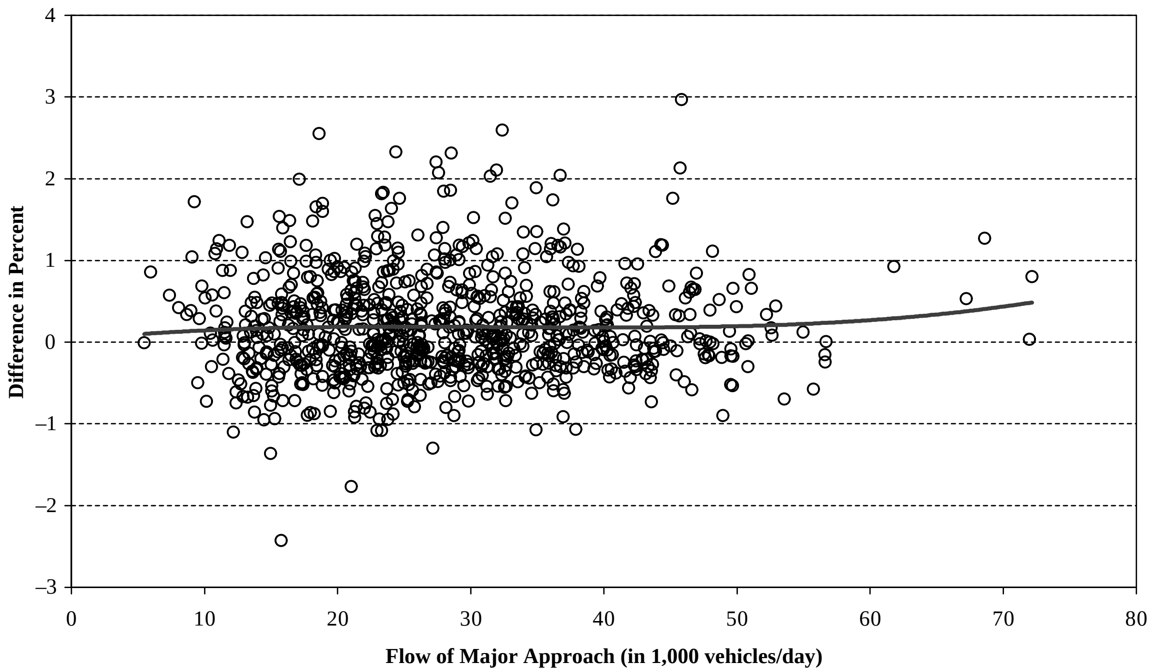


FIGURE 5 Differences in estimates of expected number of crashes for individual sites, 1995: EB method versus Bayes method, both with a fixed dispersion assumption.

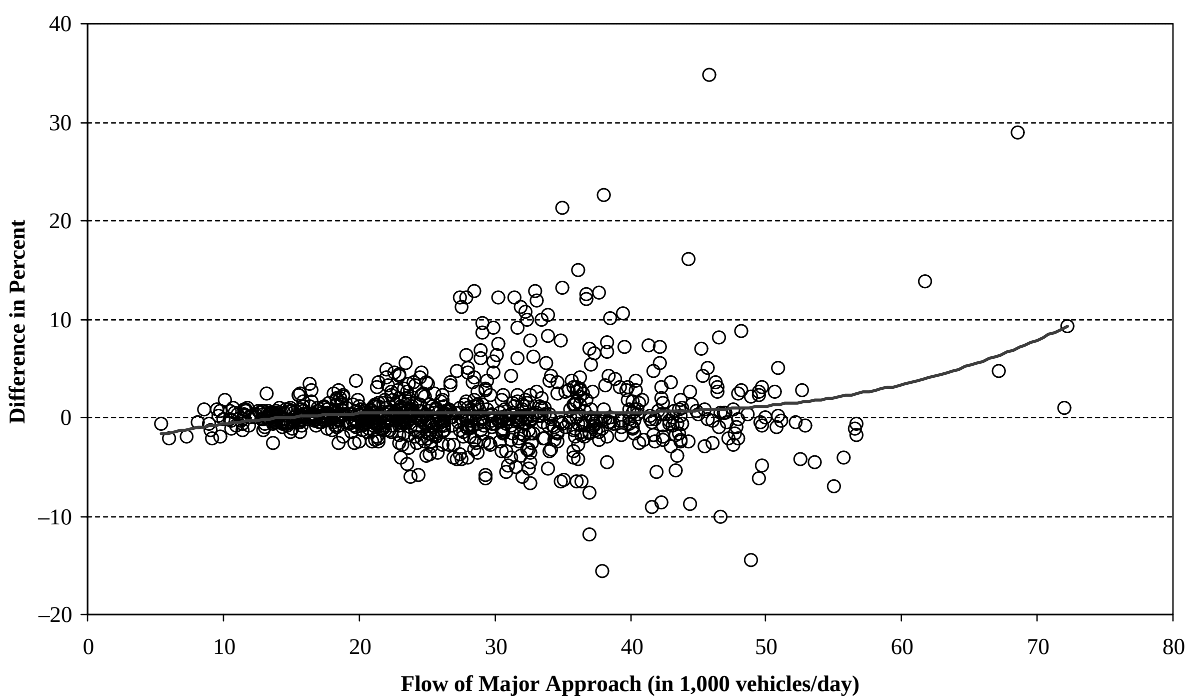


FIGURE 6 Differences in estimates of expected number of crashes for individual sites, 1995: EB method with a fixed dispersion versus Bayes method with flow-dependent dispersions.

ously undermine the goodness of the estimate for individual sites by up to about 35%. Thus, it is advisable that when modeling crash frequencies at intersections, the analysts should be mindful of the underlying assumptions adopted in the popular models, especially the assumption that ϕ is a fixed parameter.

DISCUSSION OF RESULTS

There are many directions in which this study can be extended. Here are some promising extensions of and needs identified from this research:

- The analysis can be extended to multiple data sets for both urban and rural intersections. This will give a sense of whether the varying dispersion parameter value is a common problem, an isolated problem, or specific to certain types of intersections.
- The appropriate value and the estimation procedure for parameter δ in the logical constraint expressed in Equation 6 require further research.
- There is a need to develop systematic criteria and perhaps new approaches for the modeling of vehicle crashes on roadway networks that are nonstatistical and that are based on the logic (e.g., reason, consistency, and coherency), flexibility, extensibility, and interpretability of the functional form.
- While methodologies for the modeling of roadway traffic crashes at intersections have been demonstrated, the authors recognize that, fundamentally, data for traffic crashes are network based and need to be modeled simultaneously for the probabilistic and functional structures to be statistically and logically consistent across the boundary points of various entities, such as segments, intersections, and ramps. The conventional modeling approach that treats these entities independently and then stitches the estimated models together in an ad hoc manner to represent the crash risk of the network basically assumes that vehicle crashes on these three types of roadway entities are independent of one another. Under this assumption, important spatial relationships, such as network connectivity and traffic interactions between these road entities, are ignored. This suggests the importance of having a network-based approach that analyzes the safety performance on the network as a whole and that still recognizes the distinct differences of the physical, traffic, and driver behavior characteristics among these roadway entities (*24*).

REFERENCES

1. Hauer, E., J. C. N. Ng, and J. Lovell. Estimation of Safety at Signalized Intersections. In *Transportation Research Record 1185,* TRB, National Research Council, Washington, D.C., 1988, pp. 48–61.
2. Mountain, L., and B. Fawaz. Estimating Accidents at Junctions Using Routinely Available Input Data. *Traffic Engineering and Control,* Nov. 1996, pp. 624–628.
3. Turner, S., and A. Nicholson. Intersection Accident Estimation: The Role of Intersection Location and Non-Collision Flows. *Accident Analysis and Prevention,* Vol. 30, No. 4, 1998, pp. 505–517.
4. Vogt, A., and J. Bared. Accident Models for Two-Lane Rural Segments and Intersections. In *Transportation Research Record 1635,* TRB, National Research Council, Washington, D.C., 1998, pp. 18–29.
5. Persaud, B., and T. Nguyen. Disaggregate Safety Performance Models for Signalized Intersections on Ontario Provincial Roads. In *Transportation Research Record 1635,* TRB, National Research Council, Washington, D.C., 1998, pp. 113–120.
6. Lord, D. *The Prediction of Accidents on Digital Networks: Characteristics and Issues Related to the Application of Accident Prediction Models.* Ph.D. dissertation. University of Toronto, Ontario, Canada, 2000.
7. Lord, D., and B. Persaud. Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations Procedure. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1717,* TRB, National Research Council, Washington, D.C., 2000, pp. 102–108.
8. Persaud, B., D. Lord, and J. Palmisano. Calibration and Transferability of Accident Prediction Models for Urban Intersections. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1784,* TRB, National Research Council, Washington, D.C., 2002, pp. 57–64.
9. Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman & Hall/CRC Press, London, 1995.
10. Hauer, E., and J. Bamfo. Two Tools for Finding What Function Links the Dependent Variable to the Explanatory Variables. *Proc., ICTCT 1997 Conference*, Lund, Sweden, 1997.
11. Goldstein, H. *Multilevel Statistical Models,* 1st Internet ed., 1999. www.ioe.ac.uk/multilevel/.
12. Besag, J., P. Green, D. Higdon, and K. Mengersen. Bayesian Computation and Stochastic Systems (with Discussion). *Statistical Science,* Vol. 10, 1995, pp. 3–66.
13. Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London, 1996.
14. Carlin, B. P., and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall/CRC Press, London, 1996.
15. Robert, C. P., and G. Casella. *Monte Carlo Statistical Methods.* Springer-Verlag, New York, 1999.
16. Roberts, G. O., and J. S. Rosenthal. Markov Chain Monte Carlo: Some Practical Implications of Theoretical Results. *Canadian Journal of Statistics,* Vol. 26, 1998, pp. 5–31.
17. Spiegelhalter, D. J., A. Thomas, and N. G. Best. *WinBUGS Version 1.3 User Manual.* MRC Biostatistics Unit, Cambridge, United Kingdom, 2000. www.mrc-cam.ac.uk/bugs.
18. Yang, M., J. Rasbash, H. Goldstein, and M. Barbosa. *MLwiN Macros for Advanced Multilevel Modelling,* Version 2.0. Multilevel Models Project, Institute of Education, University of London, Dec. 1999. www.ioe.ac.uk/multilevel/.
19. Kulmala, R. Measuring the Safety Effect of Road Measures at Junctions. *Accident Analysis and Prevention,* Vol. 26, No. 6, 1994, pp. 781–794.
20. Heydecker, B. G., and J. Wu. Identification of Sites for Road Accident Remedial Work by Bayesian Statistical Methods: An Example of Uncertain Inference. *Advances in Engineering Software,* Vol. 32, 2001, pp. 859–869.
21. Tanner, G. C. Accidents at Rural Three-Way Junctions. *Journal of Institution of Highway Engineers,* Vol. 2, No. 11, 1953, pp. 56–67.
22. Loader, C. *Local Regression and Likelihood.* Springer-Verlag, New York, 1999.
23. Fuller, W. A. *Introduction to Statistical Time Series,* 2nd ed. John Wiley and Sons, Inc., New York, 1996.
24. Miaou, S.-P., J. J. Song, and B. K. Mallick. Roadway Traffic-Crash Mapping: A Space-Time Modeling Approach. *Journal of Transportation and Statistics,* in press.

Publication of this paper sponsored by Committee on Statistical Methodology and Statistical Computer Software in Transportation Research.