MODELING TRAFFIC CRASH-FLOW RELATIONSHIPS FOR INTERSECTIONS: DISPERSION PARAMETER, FUNCTIONAL FORM, AND BAYES VERSUS EMPIRICAL BAYES

Shaw-Pin Miaou  
Research Scientist  
Texas Transportation Institute  
Texas A&M University System  
3135 TAMU  
College Station, Texas 77843-3135  
(979) 862-8753  
s-miaou@tamu.edu

Dominique Lord  
Associate Research Scientist  
Texas Transportation Institute  
Texas A&M University System  
3135 TAMU  
College Station, Texas 77843-3135  
(979) 458-1218  
d-lord@tamu.edu

ABSTRACT (218 WORDS)

Statistical relationships between traffic crash and traffic flows at roadway intersections have been extensively modeled and evaluated in recent years. This paper challenges the underlying assumptions adopted in the popular models for intersections. First, we challenge the assumption that the dispersion parameter is a fixed parameter across sites and time periods. Second, we examine mathematical limitations of some functional forms used in these models, particularly their properties at the boundaries. We also demonstrate 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, and introduce an alternative class of logical formulations that may enable a richer interpretation of the data. A comparison of site estimates from the empirical Bayes and full-Bayes methods is also presented. All discussion and comparison are illustrated with an urban 4-legged signalized intersection data set collected in Toronto, Canada, for years 1990 to 1995. In discussing functional forms, we emphasize and demonstrate the need for some “goodness-of-logic” (GOL) measures in addition to the GOF measure. Finally, we advise analysts to be mindful of the underlying assumptions adopted in the popular models, especially the assumption that dispersion parameter is a fixed parameter and limitations of the functional forms used. We conclude by discussing promising directions in which this study may be extended.

Prepared for Publication in Transportation Research Record  
Words: 5,449+1*250 (table) + 6*250 (figures) = 7,199  
March 2003
INTRODUCTION

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 $\mu_{it}$, is assumed to be Poisson distributed and independent over all intersections and time periods as:

$$Y_{it} \mid \mu_{it} \sim Po(\mu_{it})$$

where $i = 1,2,...,I$ and $t = 1,2,...,T$  \hspace{1cm} (1)

The mean of the Poisson is structured as

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

where $f(\cdot)$ is a function of traffic flows from the major and minor approaches of the intersection, represented by $F_{1it}$ and $F_{2it}$, respectively, and other covariates indicated by vector $x_{it}$; $\beta$ is a vector of unknown “fixed-effect” parameters; and $e_{it}$ is an unstructured random effect independent of all covariates which has a typical assumption that $\exp(e_{it})$ is independent and gamma distributed with mean equal to 1 and variance $1/\phi$ for all $i$ and $t$ (with $\phi > 0$). This particular formulation provides flexible and attractive statistical properties: For example, conditional on $f(\cdot)$ and $\phi$, $Y_{it}$ 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 un-modeled (or un-measured) heterogeneities due to omitted exogenous variables and intrinsic randomness. The key assumptions here are that (1) $\exp(e_{it})$ are independent (or, more strictly and statistically speaking, exchangeable) across all $i$ and $t$ and have a fixed variance, and (2) $e_{it}$ are independent of all covariates, including flows and $x_{it}$. In this paper, we will call $\phi$ “inverse dispersion parameter” in that the Poisson model can be regarded as a limiting model of the NB as $\phi$ approaches infinity. Without much loss of generality, we assume one time period is one year in the discussion. Note that for a Bayesian interpretation of “fixed” and “random” effects, the readers are referred to the book “Bayesian Data Analysis” (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 using an empirical Bayes (EB) method as:

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

where $\hat{\mu}_{it}$ is the EB estimate of the expected number of crashes, $\hat{\mu}_{it}$ is the maximum likelihood estimate (MLE) of $f(\cdot)$, and $y_{it}$ is the observed number of crashes for site $i$ and time $t$. In addition, the functional $g(\hat{\mu}_{it}, \hat{\phi})$ 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 $\infty$, the EB estimate would increase the weight on
many papers cited earlier that this EB estimate outperforms other estimates, such as the MLEs, and generally enjoys good precision.

Many functional forms for \( f(.) \) have been postulated. They differ by type of covariates included, forms of variable transformation applied, and/or parameterization schemes utilized. Even with a small number of covariates 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 (4) summarized commonly used functional forms into three types. Type 1 relates 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 which analyze crashes by impact type (rear-end, right-angle, etc.), severity (fatal, injury, etc.), and time-of-day, have also been developed (see, e.g., 5). In this paper, we will focus on the second type of the functional form, which is quite favorably considered and evaluated in all references cited earlier. Note, however, the discussion and findings presented in this paper apply to other types as well. The first 4 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, we first challenged the assumption that the dispersion parameter \( \phi \) is a fixed parameter across \( i \) and \( t \). Given the complexity of the traffic interaction around and within intersections and that driver behaviors are influenced by a multitude of factors pertaining to roadside development near intersections, especially “business environment” (e.g., shopping centers, schools, office compounds, etc.), it seems natural to suspect that the un-modeled 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 where 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, where \( \phi \) is assumed to be a fixed parameter. Second, we examined mathematical limitations of some popular functional forms of \( f(.) \), particularly their properties at the boundaries. We also demonstrate 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. We further introduced an alternative class of logical formulations, which could potentially provide richer interpretation of the data.

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, we had the opportunity to compare site estimates between full-Bayes and EB. 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 maximum
likelihood and the EB methods, on the other hand, tend to overestimate precision because they ignore this
uncertainty (11). For example, the EB estimate in Equation (3) was developed with the assumption
that \( f(\cdot) \) and \( \phi \) are both estimated precisely without errors, which are questionable assumptions in
practice. This advantage of the full Bayes approach is especially important when the sample size is
relatively small.

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
statistical field (12-16). At the same time, the complexity of statistical models in practice has increased
dramatically due mainly to the improved computer processing speed and lower data collection and storage
cost. 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 multi-
modal, 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 today’s desktop computing power, sampling of posterior distributions using MCMC methods as
needed in the full Bayes methods can now be relatively easily achieved. For some problems, existing
statistical programming software packages, such as WinBUGS (17) and MLwiN (18) can be employed
quite nicely, which 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 using programs coded in the WinBUGS language.

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

DATA
To motivate this study, we employed an urban 4-legged signalized intersection data set collected in
Toronto, Canada, for years 1990 to 1995. The data have been quality-checked and analyzed as part of a
network data set used in Lord (6), and further studied in Lord and Persaud and Persaud et al. (7,8). It
includes 868 intersections, which have a total of 54,989 reported crashes for the 6-year period, with
approximately a 30%-70% split of injury and non-injury crashes, respectively. Individual intersections
experienced crashes from 0 to 63 crashes per year. Traffic volumes vary widely from intersections to
intersections: from about 5,300 to 72,300 vehicles/day for main approaches (summing over both
directions) and from 52 to about 42,600 vehicles/day for minor approaches. “Traffic densities” (or entry
flows per lane per day to be precise) vary from 1,300 to 14,000 vehicles/lane/day for major approaches and from 52 to 13,000 vehicles/lane/day for minor approaches. Flow ratios, calculated as minor approach volume over major approach volume or $F_2/F_1$, range from 0.2% to almost 100%, i.e., from “link-like” (or “segment-like”) 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%. Number of lanes for the major approach varies from 2 to 10 lanes (including exclusive left-turning and right-turn lanes), while for the minor approach it varies from 1 to 8 lanes. The data set contains 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 is comprised of intersections that have permissive, semi-protected, and protected left turns. It also includes divided and undivided approaches. The speed limits on main approaches vary from 50 km/h (30 mph) to 70 km/h (43 mph).

Overall, the Toronto intersection data set covers a wide variation of 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 1000 vehicles/day), and densities ($D_1$ and $D_2$ in 1000 vehicles/lane/day) and their relationships, Figure 1 shows a set of scatter plots for the 1995 data. Note that the intent of this figure is to give the readers some sense of the data considered by this study, not meant to be very precise. For a detailed description of the data set, the readers are referred to Lord (6). We note here that the definition of intersection crashes has not been consistent in previous studies. For example, in Vogt and Bared (3), crashes considered “were ones labeled intersection or intersection-related …, and occurring within $\pm 76.2$ 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. Setting some standards to define, select, and report intersection crashes for modeling purposes within the roadway safety community is certainly needed. In this Toronto data set, crashes include both intersection and intersection-related crashes as reported by the police that are 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, and cyclists.

**DISPERSION PARAMETER**

As stated earlier, given the complexity of the traffic interaction around and within intersections and with the roadside environment, it is natural to suspect that the un-modeled heterogeneity, thus the inverse dispersion parameter, might be structured spatially in some way. The richness of the Toronto urban intersection data set is a good candidate data set 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 as presented in Equations (1) and (2) using the functional form listed as No. 5 in Table 1, which was identified as the best among others considered by Lord (6). Key estimated parameters are shown in Table 1. Note that the estimated parameters are similar to the MLE estimates reported in Lord (6). Second, we subject the same model to a different assumption
that the inverse dispersion parameter $\phi$ varies by individual sites and time period and are functions of traffic flows and flow ratios:

$$\phi_{i,t} = \exp(\eta_0 + \eta_1 F_{1,i,t} + \eta_2 F_{2,i,t} + \eta_3 F_{2,i,t} / F_{1,i,t})$$

(4)

In a sense, traffic flows are used here as surrogate variables for explaining the possible structure of unmodeled space-time heterogeneities. Modeling results are shown in Table 1 under form No. 5*. The estimated model suggests that $\phi$’s are indeed functions of flows. To see how $\phi$ varies by intersection and flows, the relative frequency of its mean (or posterior mean, to be precise) is presented in Figure 2(a), response surface by major flow and flow ratio in 2(b), and geographical distribution in 2(c). It varies widely from 3 to about 18, with a mean of 6.62 and a standard error of 1.61. Since all the models used in the papers cited earlier do not have spatial components, we suspect that these models may have structured $\phi$’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 $\phi$, which we will demonstrate in a later section. As will also be shown later that, for this particular data set, the estimated parameters in the mean function $f(\cdot)$, i.e., $\hat{\beta}$, only change slightly when altering the assumption from a fixed $\phi$ to flow-varying $\phi$’s.

**FUNCTIONAL FORM**

It has been suggested by most studies cited earlier that functional form No. 2, and thus No. 5, in Table 1 has two main advantages: (1) the form follows the logic of “no traffic flows, no accidents,” and (2) 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 1950s (21). The essence of the first cited advantage is based on the logic of having proper “boundary values,” as in the construct of differential equation systems. The second advantage is not a logical one, but rather one that based on previous experiences working with different data sets using a combination of visual inspections and statistical tests (e.g., 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. Results from the Toronto data set, including the posterior means of 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 1000 vehicles/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 (1) when the flow at main approach is close to zero (so is the flow for minor approach in this case), and (2) when the flow at minor approach is close to zero, or equivalently, the flow ratio $F_2/F_1$ approaches zero, which occurs to those “link-like” intersections
The deviance statistics, as a measure of overall goodness-of-fit, and their standard errors for the first 5 model forms in Table 1 indicate that there is no clear difference in performance among these forms from the statistical standpoint. Even though there are indications of differences in estimated $\phi$ for some of these models, they should be used with caution because, as in functional form No. 5, all models were found by this study to have $\phi$ potentially structured as functions of flows.

With the rapid development in local regression and smoothing techniques (including the spline-function based semi-parametric regressions) and the associated software in the last decade or so (e.g., 22), producing a good fit to data per se is no longer a challenge. Over-stretching and over-interpreting 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 non-statistical in nature to help balance the development and evaluation of functional forms. It is envisioned that these criteria and approaches should base 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, we have plenty of GOF measures from statistics, but no “goodness-of-logic” (GOL) measures from engineering and other perspectives.

With this limitation on the state-of-the-research in mind, what follows are our attempts to (1) offer some observations on the flexibility and boundary properties of the popular functional forms presented in Table 1, (2) illustrate alternative logic and functional forms, and (3) introduce potential logical constraints into the functional form. It should be noted that these attempts are meant to be demonstrative and suggestive in nature.

**Flexibility and Boundary Values**

- Functional form No. 5 is more flexible than No. 2 and No. 2 is more flexible than 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 zero when $F_2$ is zero, which is a limitation for modeling the “link-like” intersections mentioned earlier. This can be seen from Figure 3 by comparing the response surfaces of these functional forms with that of form No. 1 on which this restriction is not imposed. Note that even when $F_2$ is zero vehicles on the major approach (going through, making a left, or making a right) 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 forms in logics.

- At the first glance, it seems that form No. 1 is a special case of form No. 4. But it is clear from the response surfaces in Figure 3 that this is not the case. Form No. 4 suffers from the same boundary limitation as other forms discussed above. For a “link-like” intersection, form No. 1 has estimated an increase in the expected number of crashes per year from 0 to about 30 as the flow of major approach increases from 0 to 80,000 vehicles/day. Intuitively, this estimate seems high. It also appears to be
so when we compare its response surface with that of an alternative functional form, which will be discussed shortly.

- Compared to the first 4 functional forms, No. 5 is estimating a much higher expected number of crashes at high flows and high flow ratios. The difference of the response surface between No. 5 and No. 2 is included in Figure 3 as well. The additional term in form No. 5 for \( F_2 \) is playing a significant role here in producing higher estimates than those from No. 2.

**Interpretability and Alternative Logics**

To demonstrate that many other functional forms can be contemplated based on logics and can perform equally well statistically, we considered form No. 6 as presented in Table 1. That is,

\[
f(\cdot) = F_{1,i,t} \hat{\lambda}_{1,i,t} + F_{2,i,t} \hat{\lambda}_{2,i,t}
\]

where \( \hat{\lambda}_{1,i,t} = \exp(\beta_{0,1} + \beta_{1} F_{2,i,t}) \), \( \hat{\lambda}_{2,i,t} = \exp(\beta_{0,2} + \beta_{2} F_{2,i,t}) \), and \( \beta \)'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 \( \lambda_{1,t} \) and \( \lambda_{2,t} \), 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 risks involving the vehicle itself and vehicles in the same approach, which is captured by parameters \( \beta_{0,1} \). Second, this vehicle is exposed to risks due to vehicles entering from the minor approach, which is captured by the term \( \beta_{2} F_{2,i,t} \). 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 other forms discussed earlier by introducing the logic of differential risks. It is also clear that this form does not suffer the “boundary value” limitation of other forms discussed earlier. The Toronto data modeling results 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 the functional form between No. 6 and No. 5, their response surfaces matches quite well (when compared to 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 “link-like” intersections when major flows increase from 0 to 80,000 vehicles/day, which is much lower than what form No. 1 is suggesting (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 Lord (6), the expected number of crashes per year for a 6-lane segment at “mid-block” for 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 kilometer and flows in vehicles/day, respectively. By taking \( L=30 \) m = 0.03 km (as discussed earlier where intersection radius was defined as 15 m from the center) and \( F=80,000 \), we have \( \hat{f}(\cdot) = \exp(-13.828)(0.03)^{0.665}(80,000)^{1.616} = 8 \), which is very close to what the alternative functional form is estimating. This could be a pure coincidence, but it
certainly gives some comfort as to the consistency of the performance of the alternative form at the boundary.

**Extensibility and Logical Constraints**

Imposing 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. Using the system engineering language, logical constraints increase “observability” and enable us to better estimate the true state of the system under study.

To the best of our knowledge, logical constraints have not been attempted in developing the intersection crash estimation models as discussed in this paper. Here we only provide a limited demonstration of this concept using the alternative functional form in Equation (5).

So far, we have discussed the boundary value associated with “link-like” intersections where flow ratios, \( r_{it} = F_{it,2} / F_{it,1} \), are close to zero. Now, we look into the other end of the extreme where flow ratios \( r_{it} \) are close to 1, i.e., equal flows from the two approaches. At this end, under the logic behind Equation (5), we could expect a vehicle entering from the minor approach to have about the same risk level as a vehicle entering from the major approach, provided that the traffic density, lane configuration, and signal phasing plan are about the same between the two approaches. To demonstrate how this logical constraint may be introduced into Equation (5), we modify the crash rate as follows:

\[
\hat{\lambda}_{1,it} = \exp(\beta_{0,1} + \beta_{1} F_{it,2}), \\
\hat{\lambda}_{2,it} = \exp(\beta_{0,2} + \beta_{2} F_{it,1}), \\
\hat{\lambda}_{4,it} = \hat{\lambda}_{1,it} \exp(h(D_{it,1}; \omega)), \text{ and} \\
\hat{\lambda}_{2,it} = \{r_{it} \hat{\lambda}_{1,it} + (1-r_{it}) \hat{\lambda}_{2,it} \} \exp(h(D_{it,2}; \omega)) \\
\]

(6)

where \( h(., \omega) \) is a functional of the traffic density \( D_{it,1} \) and \( D_{it,2} \) with an unknown parameter vector \( \omega \) that is common to both approaches; \( h(., \omega) = 0 \), when the density is zero; and the parameter \( \delta \) is 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 the flow ratio increases from 0 to 1. Note that, for \( \delta = 1 \), the logical constraint is linear. It can be checked that \( \hat{\lambda}_{2,it} \) is a nominal crash rate for the minor approach when the flow ratio \( r_{it} \) is zero. That is, when \( r_{it} = 0, D_{it,2} = 0 \) and \( \hat{\lambda}_{2,it} = \hat{\lambda}_{2,it} \). Also, when \( r_{it} = 1 \) and \( D_{it,1} = D_{it,2} \), then \( \hat{\lambda}_{4,it} = \hat{\lambda}_{2,it} = \hat{\lambda}_{1,it} \exp(h(D_{it,1}; \omega)) \). Of course, 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 using $\delta = 0.5, 1.0, 2.0,$ and 3.0. The implementation also allowed the inverse dispersion parameter to be different for the two approaches and each dependent on the flows as in Equation (4). For the density effect function $h(\cdot; \omega)$, we used the second order grafted polynomial regression as described in Fuller (23) with breakpoints (or knots) set at traffic densities of 2, 5, 8, and 11 thousand vehicles/lane/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 $\exp(h(\cdot; \omega))$, at different $\delta$ values from the grafted polynomials are shown in Figure 4. A piece-wise linear function with $\delta = 1.0$ and with equal incremental density intervals of 1,000 vehicles/lane/day are also shown in Figure 4 for comparison.

In general, we found the estimated parameter values for $\beta$’s at different $\delta$’s are close. The estimated density effects are similar for $\delta = 1.0, 2.0, \text{and} \ 3.0$, and more pronounced density effects are indicated with $\delta = 0.5$. The piece-wise linear density effects indicate a wild fluctuation beyond 8 or 9 thousand vehicles/lane/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**

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 of individual site as presented in Equation (3) is developed with the assumption that $f(\cdot)$ and $\phi$ are both estimated precisely without errors, which is of course not true in practice. The extent to which this EB estimate for individual site deviates from the full Bayes estimate depends on the uncertainty of the estimates for $f(\cdot)$ and $\phi$.

For the Toronto data set used in this study, the uncertainties associated with the estimated $f(\cdot)$ and $\phi$ are relatively small. For the purpose of demonstration, we choose functional form No. 5 with a fixed inverse dispersion parameter. We then compare the site estimates from a model estimated using an MLE reported in Lord (6) and the full-Bayes model in Table 1. The result of this comparison for year 1995 is shown in Figure 5, where the percent difference for each site is computed as $(\text{EB} - \text{Full Bayes}) \times 100/\text{Full Bayes}$. The differences are less than 3 percent in 1995 (and less than 10 percent for all time periods). Of course, these differences can become quite large for other data sets if $f(\cdot)$ and $\phi$ are poorly estimated.

**FIXED VS. VARYING DISPERSION PARAMETER**

In contrast, ignoring that $\phi$ 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 estimates, whether it is EB or full-Bayes. To get a feel of the size of this error with the Toronto data, we compare
the same EB estimate obtained above with a full-Bayes that allows the inverse dispersion parameter to be varying with flows (which is presented under form No. 5* in Table 1). The differences for year 1995 are shown in Figure 6. It is observed that treating $\phi$ as a fixed parameter can seriously undermine the goodness of estimate of individual sites 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 $\phi$ is a fixed parameter.

DISCUSSION

There are many directions in which this study can be extended. Here are some promising extensions:

- Extend the analysis to multiple data sets for both urban and rural intersections. This will give us 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 estimation procedure for parameter $\delta$ in the logical constraint as expressed in Equation (6) require further research.
- There is a need to develop systematic criteria and perhaps new approaches for modeling vehicle crashes on roadway networks that are non-statistical and based on the logic (e.g., reason, consistency, and coherency), flexibility, extensibility, and interpretability of the functional form.
- While we demonstrated the methodologies for roadway traffic crashes at intersections, we recognize that fundamentally traffic crashes are “network-based” data 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 “stitch” 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 from 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 still recognizes the distinct differences on the physical, traffic, and driver behavior characteristics among these roadway entities (24).

REFERENCES


LIST OF TABLES (1 table in total)

TABLE 1. Commonly Used Functional Forms, Alternative Forms, and Estimated Parameters for 1995 ($t=6$) and Associated Statistics.
Table 1. Commonly Used Functional Forms, Alternative Forms, and Estimated Parameters for 1995 (t=6) and Associated Statistics.

<table>
<thead>
<tr>
<th>No.</th>
<th>Functional Form $f(\cdot)$</th>
<th>$\beta_{0, t}$</th>
<th>$\beta^*{0, t}$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\phi$</th>
<th>Deviance</th>
<th>$\eta_0$</th>
<th>$\eta_1$</th>
<th>$\eta_2$</th>
<th>$\eta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\beta_{0,t} (F_{1,t} + F_{2,t})^\beta_1$</td>
<td>0.05884 (±0.005)</td>
<td>1.4220 (±0.024)</td>
<td>4.053 (±0.170)</td>
<td>25410 (±102.9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$\beta_{0,t} F_{1,2,t}^{\beta_1} F_{2,2,t}^{\beta_2}$</td>
<td>0.41111 (±0.027)</td>
<td>0.5441 (±0.019)</td>
<td>0.6504 (±0.009)</td>
<td>6.643 (±0.238)</td>
<td>25480 (±101.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$\beta_{0,t} (F_{1,t} F_{2,t})^{\beta_1}$</td>
<td>0.32726 (±0.014)</td>
<td>0.6283 (±0.007)</td>
<td>6.592 (±0.229)</td>
<td>25470 (±102.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$\beta_{0,t} (F_{1,t} + F_{2,t})^\beta_1 \left( \frac{F_{2,t}}{F_{1,t}} \right)^{\beta_2}$</td>
<td>0.2001 (±0.013)</td>
<td>1.1950 (±0.016)</td>
<td>0.3770 (±0.009)</td>
<td>6.586 (±0.235)</td>
<td>25450 (±102.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$\beta_{0,t} F_{1,2,t}^{\beta_1} F_{2,2,t}^{\beta_2} \exp(\beta_3 F_{2,t})$</td>
<td>0.4711 (±0.031)</td>
<td>0.5259 (±0.018)</td>
<td>0.5663 (±0.021)</td>
<td>6.651 (±0.235)</td>
<td>25460 (±101.3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5*</td>
<td>$\beta_{0,t} F_{1,2,t}^{\beta_1} F_{2,2,t}^{\beta_2} \exp(\beta_3 F_{2,t})$ ($\phi$ is a function of flows and flow ratios)</td>
<td>0.4615 (±0.032)</td>
<td>0.5332 (±0.018)</td>
<td>0.5757 (±0.022)</td>
<td>0.00746 (±0.002)</td>
<td>See Eq. (4)</td>
<td>25380 (±103.0)</td>
<td>2.2200 (±0.180)</td>
<td>-0.02126 (±0.005)</td>
<td>0.05452 (±0.009)</td>
<td>-0.8783 (±0.265)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>$F_{1,t} \lambda_{1,t} + F_{2,t} \lambda_{2,t}$ where $\lambda_{1,t} = \exp(\beta_{0,1,t} + \beta_1 F_{2,t})$ and $\lambda_{2,t} = \exp(\beta_{0,2,t} + \beta_2 F_{1,t})$</td>
<td>-2.4280 (±0.083)</td>
<td>-0.54050 (±0.047)</td>
<td>0.00474 (±0.008)</td>
<td>0.01034 (±0.001)</td>
<td>6.578 (±0.232)</td>
<td>25440 (±99.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: (1) Flows are in 1000 vehicles/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 (2000) for this particular data set, (5) Functional form No. 5* is the same as No. 5 but with an assumption that the inverse dispersion parameter $\phi$ is a function of flows and flow ratios, (6) Functional forms 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 using the full-Bayes framework with non-informative priors (or hyper-priors), (8) Parameters ($\beta^*, \phi$, and Deviance) were estimated using MCMC techniques and the values shown in the table are their posterior means, (9) Values in parentheses are the estimated 1 standard error of parameters above, (10) Each year has a separate intercept term, and (11) “Deviance” is computed as $-2\log(\text{Prob}(y|\text{estimated parameters}))$ where “log” is the natural logarithm.
LIST OF FIGURES (6 figures in total)

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

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 year 1995.

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

FIGURE 4. “Density “ effects at various levels of logical constraints.

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

FIGURE 6. Differences in estimates of expected number of crashes for individual sites: EB with a fixed dispersion version Bayes with flow-dependent dispersions.
FIGURE 1. Ranges of and relationships among crash frequencies, traffic flows, and traffic densities.
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 year 1995.
FIGURE 3. Response surfaces of functional forms listed in Table 1 and their differences by major flow and flow ratio.
FIGURE 4. “Density” effects at various levels of logical constraints.
FIGURE 5. Differences in estimates of expected number of crashes for individual sites: EB versus Bayes, both with a fixed dispersion assumption.
FIGURE 6. Differences in estimates of expected number of crashes for individual sites: EB with a fixed dispersion version and Bayes with flow-dependent dispersions.