The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives

by

Dominique Lord
Assistant Professor
Zachry Department of Civil Engineering
Texas A&M University
College Station, TX 77843-3136
d-lord@tamu.edu

Fred Mannering
Charles Pankow Professor
School of Civil Engineering
550 Stadium Mall Drive
Purdue University
West Lafayette, IN 47907-2051
flm@purdue.edu

Forthcoming in Transportation Research, Part A

DOI 10.1016/j.tra.2010.02.001

March 22, 2010

ABSTRACT

Gaining a better understanding of the factors that affect the likelihood of a vehicle crash has been an area of research focus for many decades. However, in the absence of detailed driving data that would help improve the identification of cause and effect relationships with individual vehicle crashes, most researchers have addressed this problem by framing it in terms of understanding the factors that affect the frequency of crashes – the number of crashes occurring in some geographical space (usually a roadway segment or intersection) over some specified time period. This paper provides a detailed review of the key issues associated with crash-frequency data as well as the strengths and weaknesses of the various methodological approaches that researchers have used to address these problems. While the steady march of methodological innovation (including recent applications of random-parameter and finite mixture models) has substantially improved our understanding of the factors that affect crash frequencies, it is the prospect of combining evolving methodologies with far more detailed vehicle crash data that holds the greatest promise for the future.

Key words: highway safety, literature review, regression models, count data models

INTRODUCTION

With the enormous losses to society resulting from motor vehicle crashes, researchers have continually sought ways to gain a better understanding of the factors that affect the probability of crashes in the hopes that they will be able to better predict the likelihood of crashes and provide direction for policies and countermeasures aimed at reducing the number of crashes. Unfortunately, the detailed driving data (acceleration, braking and steering information, driver response to stimuli, etc.) and crash data (for example what might be available from vehicle black boxes) that would better enable identification of cause and effect relationships with regard to crash probabilities are typically not available.[1] As a result, researchers have framed their analytic approaches to study the factors that affect the number of crashes occurring in some geographical space (usually a roadway segment or intersection) over some specified time period (week, month, year, number of years). Such an approach handles the spatial and temporal elements associated with crashes, and ensures that adequate data are available for the estimation of statistical models (in terms of measurable explanatory variables). This results in crash-frequency data that are non-negative integers, and suggests the application of count-data regression methods or other approaches that can properly account for the integer nature of the data.

[1] In the U.S., the National Strategic Highway Research Program initiated a series of studies in recent years with the objective of addressing many of the fundamental questions relating to crash causation and involvement (see for example Dingus et al. 2006). These studies are based on naturalistic driving information, wherein a selected pool of drivers is observed over prolonged periods in terms of their crash, near-crash and incident involvements. Shankar et al. (2008) have attempted to construct one plausible statistical approach to extract insights from naturalistic driving data. However, for the most part, naturalistic driving data have not yet provided significant new insights with broad applicability. The use and statistical analysis of these data are also hampered by privacy issues relating to driver- identifying variables and other potential litigation issues.

The intent of this paper is to provide a review of contemporary thinking in the crash frequency-analysis field and to show how methodological approaches have evolved over the years to address this problem. To do this, we will first discuss the fundamental data and methodological issues associated with the analysis of crash frequencies. We then move on to a critical assessment of the strengths and weaknesses of the various methodological approaches that have been used to analyze crash-frequency data, and conclude with recommendations for future methodological directions.

DATA AND METHODOLOGICAL ISSUES

Important data and methodological issues have been identified in the crash-frequency literature over the years. These issues have been shown to be a potential source of error in terms of incorrectly specifying statistical models which may lead to erroneous crash-frequency predictions and incorrect inferences relating to the factors that determine the frequency of crashes. These issues are summarized in Table 1 and are discussed below.

Overdispersion

One notable characteristic of crash-frequency data that is that the variance exceeds the mean of the crash counts (see Equation 1). This is problematic because the properties of the most common count-data modeling approach (the Poisson regression model which is discussed below) restrict the mean and variance to be equal. When overdispersed data are present, estimating a common Poisson model can result in biased and inconsistent parameter estimates which in turn could lead to erroneous inferences regarding the factors that determine crash frequencies (Maycock and Hall, 1984; Miaou,

1994; Maher and Summersgill, 1996; Cameron and Trevedi, 1998; Park and Lord, 2007).[2]

Underdispersion

Although rare, crash data can sometimes be characterized by under-dispersion, where the mean of the crash counts on roadway entities is greater than the variance, especially when the sample mean value is very low. Previous work has shown that many traditional count-data models produce incorrect parameter estimates in the presence of underdispersed data (see Oh et al., 2006; Lord et al., 2009).

Time-Varying Explanatory Variables

Because crash-frequency data are considered over some time period, the fact that explanatory variables may change significantly over this time period is not usually considered due to the lack of detailed data within the time period. Ignoring the potential within-period variation in explanatory variables may result in the loss of potentially important explanatory information. For example, suppose we are modeling the number of crashes per month and precipitation is one of the explanatory variables. The distribution of precipitation over the month (by hour or even minute) is likely to be highly influential in generating crashes, but generally the analyst only has precipitation

[2] In count-data models, actual estimates of overdispersion can potentially be influenced by a variety of factors, such as the clustering of data (neighborhood, regions, etc.), unaccounted temporal correlation, and model miss-specification (Gourvieroux and Visser, 1997; Poormeta, 1999; Cameron and Trivedi, 1998). While model estimates of overdispersion can be attributed to these factors, Lord et al. (2005b) argued that there is a fundamental explanation for overdispersion that can be shown by viewing crash data as the product of Bernoulli trials with an unequal probability of events (this is also known as Poisson trials). Recently, some researchers have reported that model-estimated overdispersion can be greatly minimized by improving the model specification (Miaou and Song, 2005; Mitra and Washington, 2007).

data that is much more aggregated and thus important information is lost by using discrete time intervals – with larger intervals resulting in more information loss. This can introduce error in model estimation as a result of unobserved heterogeneity (see Washington et al., 2010).[3]

Temporal and Spatial Correlation

To avoid the information lost in time-varying explanatory variables, data are often considered in small time intervals. For example, one may have a years' worth of crash data and divide these data into 12 monthly observations and consider the number of crashes per month. However, this now means that the same roadway entity (roadway segment, intersection) will generate multiple observations, and these observations will be correlated over time because many of the unobserved effects associated with a specific roadway entity will remain the same over time. From a statistical perspective, this sets up a correlation in the disturbances used for model estimation, which is known to adversely affect the precision of parameter estimates. In a similar vein, there can be correlation over space, because roadway entities that are in close proximity may share unobserved effects. This again sets up a correlation of disturbances among observations and results in the associated parameter-estimation problems (Mountain et al., 1998; Sittikariya and Shankar 2009; Shankar et al., 1998; Ulfarsson and Shankar 2003; Lord and Persaud, 2000; Washington et al., 2003, 2010).[4]

---

[3] In addition, the aggregation of data over time periods can lead to a bias, especially in nonlinear models. This aggregation bias is a well known problem in the statistical analysis of discrete data (see for example Washington et al., 2003, 2010).

[4] The temporal aspect of these models raises the possibility of a modification continuous dependent variable time-series methodologies such as autoregressive conditionally heteroscedastic models (ARCH) and generalized autoregressive conditionally heteroscedastic models (GARCH) (see Engle, 1982; Bollerslev, 1986). However,

Low Sample Mean and Small Sample Size

Because of the large costs associated with the data collection process, crash data are often characterized by a small number of observations. In addition, crash data for some roadway entities may have few observed crashes which results in a preponderance of zeros. Data characterized by small sample size and low sample-mean can cause estimation problems in traditional count-frequency models. For example, with small sample sizes, the desirable large-sample properties of some parameter-estimation techniques (for example, maximum likelihood estimation) are not realized. With low sample means (and a preponderance of zeros), the distribution of crash counts will be skewed excessively toward zero which can result in incorrectly estimated parameters and erroneous inferences.[5]

Injury Severity and Crash Type Correlation

Crash data are often classified according to their injury severity or collision type. For example, classifying a crash by the most severely injured person could result in a crash being classified as fatal, incapacitating injury, non-incapacitating injury, possible injury and no injury. Classifying crashes by collision type could include outcomes such as rear-end, single-vehicle run-off-the-road, right-angle, and sideswipe among others. The most common modeling approach is to consider the frequency of all crashes

---

extensions of these continuous dependent variable methods to the count-data context have been limited.

[5] Please see Maycock and Hall (1984), Piegorsch (1990), Fridstrøm et al. (1995), Maher and Summersgill (1996), Wood (2002) and Lord and Bonneson (2005) for further discussion of this problem. Also, Lord (2006) showed that the dispersion parameter of the negative binomial model can be incorrectly estimated when data characterized by a small sample size and low sample mean values are used. The incorrect estimation of the dispersion parameter also negatively affects the inferences associated with the parameters of the model.

(including all severity and collision types together), and deal with the injury severities or crash types separately once the total number of crashes is determined.[6] However, some researchers have sought to develop separate crash-frequency models for various injury severities and collision types. If this is done, a potentially serious statistical problem results because there is a correlation among injury severities and collision types. For example, an increase in the number of crashes that are classified as incapacitating injury will also be associated with some change in the number of crashes that are classified by other injury types, which sets up a correlation among the various injury-outcome crash-frequency models. This necessitates the need for a more complex model structure to account for the cross-model correlation.[7]

Under Reporting

Because less severe crashes are less likely to appear in crash data bases, there is a potentially serious problem relating to under-reporting of crashes.[8] Although the magnitude of the under-reporting rate for each severity level is usually unknown, recent research has shown that count data models are likely to produce biased estimates when

---

[6] With this approach, separate models of injury severity, for example, are estimated conditioned on the fact that a crash has occurred. See for example Carson and Mannering (2001) and Lee and Mannering (2002) for applications of this method.

[7] The issues associated with this model formulation are discussed in Miaou and Song (2005), Bijleveld (2006), Song et al. (2006), Ma and Kockelman (2006), Park and Lord (2007), Ma et al. (2008), El-Basyouny and Sayed (2009a), Geedipally and Lord (2009).

[8] Incomplete reporting of crash data has been known to be a major problem in highway safety analysis for many years (Elvik and Mysen, 1999). The research on this topic has shown that fatal crashes are the most likely to be reported, while no-injury collisions are the ones most likely not to be reported (Aptel et al., 1999). Hauer and Hakkert (1988) and James (1991) have shown that the probability that a crash will be reported varies not only as a function of the crash severity but also as a function of reporting agency (city, regions, and so on).

under-reporting is not considered in the model-estimation process (Kumara and Chin, 2005; Ma, 2009).[9]

Omitted Variables Bias

It is often tempting to develop a simplified model with few explanatory variables (for example, using traffic flow as the only explanatory variable in the model). However, as with all traditional statistical estimation methods, leaving out important explanatory variables results in biased parameter estimates that can produce erroneous inferences and crash-frequency forecasts (see Washington et al., 2003, 2010). This would especially be the case if the omitted variable is correlated with variables included in the specification, which is often the case.

Endogenous Variables

There are times when the explanatory variables in models can be endogenous, in that their values may depend on the frequency of crashes. An example of this problem is the frequency of ice-related accidents and the effectiveness of ice-warning signs in reducing this frequency (this is the endogeneity problem studied in Carson and Mannering, 2001). When developing a crash-frequency model, an indicator variable for the presence an ice warning sign would be one way of understanding the impact of the warning signs. However, ice-warning signs are more likely to be placed at locations with high numbers of ice-related crashes, and are therefore endogenous (the explanatory variable will change as the dependent variable changes). If this endogeneity is ignored, the parameter estimates will be biased. In the case of the ice-warning sign indicator,

[9] Yamamoto et al. (2008) provide some insight into the extent of the injury-related under-reporting problem in their study of crash severities.

ignoring the endogeneity may lead to the erroneous conclusion that ice-warning signs actually increase the frequency of ice-related crashes because the signs are going to be associated with locations of high ice-crash frequencies (because these are the locations where the signs are most likely placed). Kim and Washington (2006) study a similar problem when studying the effectiveness of left-turn lanes at intersections. This is again endogenous because left-turn lanes are more likely to be placed at intersections with a high number of left-turn related crashes.

Accounting for endogenous variables in traditional least squares regression models is relatively straight forward (see Washington et al. 2003, 2010). However, for count-data models, the modeling processes typically applied (more on this below) do not lend themselves to traditional endogenous-variable correction techniques (such as instrumental variables). As a consequence, accounting for endogenous variables adds considerable complexity to the count-data modeling process (see Kim and Washington, 2006).

Functional Form

The functional form of the model establishes the relationship between the dependent variable and the explanatory variables and is a critical part of the modeling process.[10] Most count-data models assume that explanatory variables influence the

[10] The non-linear effect that explanatory variables have on crash-frequencies can be revealing. For example, some researchers have used traffic flow as a measure of exposure and found the relationship between crashes and traffic flow to be decreasing (Tanner, 1953; Mahalel, 1986; Hauer et al., 1988; Hauer, 1997; Persaud and Nguyen; 1998) implying a potentially contentious finding that the crash risk per unit of exposure becomes smaller as traffic flow increases (Maher et al., 1993; Lord 2002; Lord et al., 2005a). In a similar vein, segment length has also been used as a measure of exposure because researchers have noted that the probability of observing a crash tends to be smaller on shorter roadway segments and higher with longer segments. This makes

dependent variables in some linear manner. However, there is a body of work that suggests that non-linear functions better characterize the relationships between crash frequencies and explanatory variables. These non-linear functions can often be quite complex and may require involved estimation procedures (Miaou and Lord, 2003; Bonneson and Pratt, 2008).

Fixed Parameters

Traditional statistical modeling does not allow parameter estimates to vary across observations. This implies that the effect of the explanatory variable on the frequency of crashes is constrained to be the same for all observations (for example, the effect of an exposure variable such as the number of vehicle miles travelled over the time period being considered is the same across all roadway segments). However, because of unobserved variations from one roadway segment to the next (unobserved heterogeneity) one might expect the estimated parameters of some explanatory variables to differ across roadway segments. If some parameters do vary across observations and the model is estimated as if they were fixed, the resulting parameter estimates will be biased and possible erroneous inferences could be drawn. Estimation techniques do exist for allowing parameters to vary across observations, but the model estimation process becomes considerably more complex (Anastasopoulos and Mannering, 2009; El-Basyouny and Sayed, 2009b; Washington et al., 2010).

---

sense because multiplying traffic flow and segment length gives a traditional exposure measure (vehicle-miles traveled). However, similar to traffic flow, some researchers have found a non-linear relationship to exist between crashes and the length of a segment (e.g., a doubling of segment length more than doubles the crash frequency) while others have argued that exposure should be linear (Miaou et al., 2003; Lord et al., 2005a). Most likely, conflicting and counterintuitive findings with regard to exposure could be pointing to unobserved heterogeneity and possibly other specification problems.

MODELING METHODS FOR ANALYZING CRASH-FREQUENCY DATA

To deal with the data and methodological issues associated with crash-frequency data (many of which could compromise the statistical validity of an analysis if not properly addressed), a wide variety of methods have been applied over the years. Table 2 provides a listing of methods previously applied to crash-frequency analysis along with their strengths and weaknesses. Table 3 provides a listing of studies that have used specific methods. The details of these methods are discussed below.

Poisson Regression Model

Because crash-frequency data are non-negative integers, the application of standard ordinary least-squares regression (which assumes a continuous dependent variable) is not appropriate. Given that the dependent variable is a non-negative integer, most of the recent thinking in the field has used the Poisson regression model as a starting point. In a Poisson regression model, the probability of roadway entity (segment, intersection, etc.) i having y_i crashes per some time period (where y_i is a non-negative integer) is given by:

$$P(y_i) = \frac{EXP(-\lambda_i)\lambda_i^{y_i}}{y_i!} \quad (1)$$

where $P(y_i)$ is the probability of roadway entity i having y_i crashes per time period and λ_i is the Poisson parameter for roadway entity i, which is equal to roadway entity i's expected number of crashes per year, $E[y_i]$. Poisson regression models are estimated by specifying the Poisson parameter λ_i (the expected number of crashes per period) as a

function of explanatory variables, the most common functional form being $\lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i)$, where $\mathbf{X}_i$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of estimable parameters.

Although the Poisson model has served as a starting point for crash-frequency analysis for several decades, researchers have often found that crash data exhibit characteristics that make the application of the simple Poisson regression (as well as some extensions of the Poisson model) problematic. Specifically, Poisson models cannot handle over- and under-dispersion and they can be adversely affected by low sample means and can produce biased results in small samples.

Negative Binomial (Poisson-gamma) Regression Model

The negative binomial (or Poisson-gamma) model is an extension of the Poisson model to overcome possible overdispersion in the data. The negative binomial/Poisson-gamma model assumes that the Poisson parameter follows a gamma probability distribution. The model results in a closed-form equation and the mathematics to manipulate the relationship between the mean and the variance structures is relatively simple. The negative binomial model is derived by rewriting the Poisson parameter for each observation *i* as $\lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i + \varepsilon_i)$ where $EXP(\varepsilon_i)$ is a gamma-distributed error term with mean 1 and variance α. The addition of this term allows the variance to differ from the mean as $VAR[y_i] = E[y_i][1+ \alpha E[y_i]] = E[y_i]+ \alpha E[y_i]^2$.[11] The Poisson regression model is a limiting model of the negative binomial regression model as α approaches

[11] Other variance functions exist for negative binomial/Poisson-gamma models, but they are not covered here because they are seldom used in highway safety studies. The reader is referred to Cameron and Trevedi (1998) and Maher and Summersgill (1996) for a description of alternative variance functions.

zero, which means that the selection between these two models is dependent upon the value of α. The parameter α is often referred to as the overdispersion parameter.[12]

The Poisson-gamma/negative binomial model is the probably the most frequently used model in crash-frequency modeling. However, the model does have its limitations, most notably its inability to handle underdispersed data, and dispersion-parameter estimation problems when the data are characterized by the low sample mean values and small sample sizes (see Lord, 2006; Lord et al., 2009).

Poisson-Lognormal Model

Recently, some researchers have proposed using the Poisson-lognormal model as an alternative to the negative binomial/Poisson-gamma model for modeling crash data (Miaou et al., 2003; Lord and Miranda-Moreno, 2008; Aquero-Valverde and Jovanis, 2008). The Poisson-lognormal model is similar to the negative binomial/Poisson-gamma model, but the $EXP(\varepsilon_i)$ term used to compute the Poisson parameter (see above) is lognormal rather than gamma distributed.

Although the Poisson-lognormal potentially offers more flexibility than the negative binomial/Poisson-gamma, it does have its limitations. For example, model estimation is more complex because the Poisson-lognormal distribution does not have a closed form and the Poisson-lognormal can still be adversely affected by small sample sizes and low sample mean values (Miaou et al., 2003).[13]

---

[12] Usually the overdispersion parameter or its inverse is assumed to be fixed, but recent research in highway safety has shown that the variance structure can potentially be dependent on explanatory variables (Heydecker and Wu, 2001; Hauer, 2001; Miaou and Lord, 2003; Lord et al., 2005a; Cafiso et al., 2010b).

[13] Lord and Miranda-Moreno (2008) have shown that the conditions at which the small sample size and low sample mean problems occur are more extreme than those found

Zero-inflated Poisson and Negative Binomial

Zero-inflated models have been developed to handle data characterized by a significant amount of zeros or more zeros than the one would expect in a traditional Poisson or negative binomial/Poisson-gamma model. Zero-inflated models operate on the principle that the excess zero density that cannot be accommodated by a traditional count structure is accounted for by a splitting regime that models a crash-free versus a crash-prone propensity of a roadway segment. The probability of a roadway entity being in zero or non-zero states can be determined by a binary logit or probit model (see Lambert, 1992; Washington et al., 2003, 2010).

Since its inception, the zero-inflated model (both for the Poisson and negative binomial models) has been popular among transportation safety analysts (Shankar et al., 1997; Carson and Mannering, 2001; Lee and Mannering, 2002; Kumara and Chin, 2003; Shankar et al., 2003). Despite its broad applicability to a variety of situations where the observed data are characterized by large zero densities, others have criticized the application of this model in highway safety. For instance, Lord et al. (2005, 2007) argued that, because the zero or safe state has a long-term mean equal to zero, this model cannot properly reflect the crash-data generating process.[14]

---

for the negative binomial/Poisson-gamma. In other words, the problems start to be noticeable at smaller sample size and lower sample mean values than would be the case for the NB model.

[14] Recently, the problems associated with the long-term mean equal to zero have been discussed by other researchers (see, e.g., Malyshkina et al., 2009). As an alternative, Malyshkina and Mannering (2010a) have proposed a zero-sate Markov switching count-data model for circumventing the problem of having the long-term mean equal to zero.

Conway-Maxwell-Poisson Model

The Conway-Maxwell-Poisson distribution is a generalization of the Poisson distribution and was first introduced by Conway and Maxwell (1962) for modeling queues and service rates. Shmueli et al. (2005) further explored the statistical properties of the Conway-Maxwell-Poisson distribution, and Kadane et al. (2006) developed conjugate distributions for the parameters. The Conway-Maxwell-Poisson can handle both underdispersed and overdispersed data, and several common probability density functions are special cases of the Conway-Maxwell-Poisson (for example, the geometric distribution, the Bernoulli distribution, and the Poisson distribution). This flexibility greatly expands the types of problems for which the Conway-Maxwell-Poisson distribution can be used to model crash frequency data.

This model has been recently applied in highway-safety research, and has been found to be comparable to the Poisson-gamma model for data characterized by overdispersion (Lord et al., 2007). However, its main advantage is related to data characterized by underdispersion.[15] On the down side, this model can be negatively influenced by low sample mean, small-sample bias and, to date, there have not been any multivariate applications of the approach.

Gamma Model

The gamma model has been proposed by Oh et al. (2006) to analyze crash data exhibiting underdispersion (see also Cameron and Trivedi, 1998). This model can handle overdispersion and underdispersion and reduces to the Poisson model when the variance

[15] Please see Sellers and Shmueli (2010), Guikema and Coffelt (2008) and Lord et al. (2009) for further details on the estimation properties and applications of this model.

is roughly equal to the mean of the number of crashes. Although this model performs well statistically, it is still a dual-state model, with one of the states having a long-term mean equal to zero (see the previous discussion of zero-inflated models). The gamma model has seen limited use since it was first introduced by Oh et al. (2006).

Generalized Estimating Equation Model

The generalized estimating equation model has been applied to highway safety analysis by Lord and Persaud (2000) to model crash data with repeated measurements. As discussed previously, one often has data from roadway entities (roadway segments or intersections) over multiple time periods which set up a serial correlation problem (see Liang and Zeger, 1986). The generalized estimating equation is not actually a regression model per se, but a method used to estimate models with data characterized by serial correlation. The generalized estimating equation model offers different approaches to handle serial correlation including independence, exchangeable, dependence, and auto-regressive type 1 correlation structures. Usually, the correlation structure in the estimation process has a minimal influence on the modeling output when count-data models are used with a complete dataset (few if any omitted variables), however, the selection of the correlation type can be critical when the database has omitted variables (Lord et al., 2005a; Halekoh et al., 2006; Lord and Mahlawat, 2009).

Generalized Additive Models

The generalized additive model has more flexibility than the traditional count-data models (see Hastie and Tibshirani, 1990; Wood, 2006). As discussed by Xie and Zhang (2008), generalized additive models provide a more flexible functional form which involves smoothing functions for the explanatory variables of the model. The smoothing

function represents a more flexible relationship in how explanatory variables are taken into account and this not limited to linear or logarithm relationships as is frequently used in traditional count models.

Although the generalized additive model can be more flexible than traditional count models, there are still limitations. First, because these models include more parameters, the estimation process can become very complex, especially when the default values are not used. Second, because generalized additive models use spline functions, they are more difficult to interpret than traditional count models. Third, the modeling results between generalized additive model and traditional models are likely to be similar if the explanatory variables are exogenous and the dependent variable has a linear or exponential relationship with them. Thus far, applications of generalized additive models to crash-frequency analysis have been limited to a few papers including Xie and Zhang (2008) and Li et al. (2009).

Random-Effects Models

As discussed at length previously, there may be reason to expect correlation among observations. This correlation could arise from spatial considerations (data from the same geographic region may share unobserved effects), temporal considerations (such as in panel data – where data collected from the same observational unit over successive time periods could share unobserved effects), or a combination of the two. To account for such correlation, random-effects models (where the common unobserved effects are assumed to be distributed over the spatial/temporal units according to some distribution and shared unobserved effects are assumed to be uncorrelated with explanatory variables) and fixed effects models (where common unobserved effects are accounted for by

indicator variables and shared unobserved effects are assumed to be correlated with independent variables) models can be considered. In the context of count models, Hausman et al. (1984) first examined random-effects and fixed-effects negative binomial models for panel data (which has temporal considerations) in their study of research and development patents.

Random-effects models rework the Poisson parameter as $\lambda_{ij} = EXP\left(\boldsymbol{\beta}\mathbf{X}_{ij}\right)EXP\left(\eta_j\right)$ where λ_{ij} is the expected number of crashes for roadway entity i belonging to group j (for example, a spatial or temporal group expected to share unobserved effects), $\mathbf{X}_{ij}$ is a vector of explanatory variables, $\boldsymbol{\beta}$ is a vector of estimable parameters, and η_j is a random-effect for observation group j.

The most common model is derived by assuming η_j is randomly distributed across groups such that $EXP(\eta_j)$ is gamma-distributed with mean one and variance α (see Hausman et al., 1984).[16] As mentioned previously, the Poisson model restricts the mean and variance to be equal which in this case would be $E[y_{ij}] = VAR[y_{ij}]$. However, with random-effects the Poisson variance to mean ratio is $1 + \lambda_{ij}/(1/\alpha)$.[17]

Random-effects in the context of crash-frequencies have been studied by a number of researchers including Johansson (1996) (who studied the effect of a lowered speed limit on the number of crashes on roadways in Sweden), Shankar et al. (1998) (who compared standard negative binomial and random-effects negative binomial models in a study of crashes caused by median crossovers in Washington State), Miaou et al.

[16] It should be pointed out that other formulations exist for defining random effects models, see for example Gelman and Hill (2007), McCulloch et al. (2008) and Bivand et al. (2008).

[17] Using the same approach, the random effects negative binomial model can also be readily derived (see Hausman et al., 1984).

(2003) (who used random-effects in the development of crash-risk maps in Texas), and others (see Table 3).

Negative Multinomial Models

The problem of correlation among observations can also be addressed with a negative multinomial approach (Guo, 1996). This model is similar to the negative binomial in that it uses $\lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i + \varepsilon_i)$, except now $EXP(\varepsilon_i)$ is associated with a specific entity (roadway segment, intersection) as opposed to a specific observation. This is an important distinction because, for example, if one is considering annual crash frequencies and with 5 years of data, each roadway entity will produce 5 observations that would each have their own $EXP(\varepsilon_i)$ in a standard negative binomial, which would create a potential correlation problem. For the negative multinomial model, at the segment/intersection level, the $EXP(\varepsilon_i)$ is again assumed a gamma-distributed error term with mean 1 and variance α. Shankar and Ulfarsson (2003) presented an application of the negative multinomial model to crash frequency data and compare estimation results to standard negative binomial and random-effects negative binomial models (see also Hauer, 2004; Caliendo et al., 2007). Negative multinomial models cannot handle underdispersion and are susceptible to problems in the presence of low sample means and small sample sizes.

Random-Parameters Models

Random-parameter models can be viewed as an extension of random-effects models. However, rather than effectively only influencing the intercept of the model, random-parameter models allow each estimated parameter of the model to vary across

each individual observation in the dataset. These models attempt to account for the unobserved heterogeneity from one roadway site to another (Milton et al., 2008).

To allow for such random-parameters in count-data models, estimable parameters can be written as $\beta_i = \beta + \varphi_i$ where φ_i is a randomly distributed term (for example a normally distributed term with mean zero and variance σ^2). With this equation, the Poisson parameter becomes $\lambda_i|\boldsymbol{\varphi}_i=EXP(\boldsymbol{\beta}\mathbf{X}_i)$ in the Poisson model and $\lambda_i|\boldsymbol{\varphi}_i=EXP(\boldsymbol{\beta}\mathbf{X}_i+\varepsilon_i)$ in the negative binomial/Poisson-gamma with the corresponding probabilities for Poisson or negative binomial now $P(y_i|\boldsymbol{\varphi}_i)$. These models have been applied to crash-frequency data by Anastasopoulos and Mannering (2009) and El-Basyouny and Sayed (2009b). Because each observation has its own parameters, the final model will often provide a statistical fit that is significantly better than a model with traditional fixed parameters. However, random-parameter models are very complex to estimate, they may not necessarily improve predictive capability, and model results may not be transferable to other data sets because the results are observation specific (see Shugan, 2006; Washington et al., 2010).

Bivariate/Multivariate Models

Bivariate/Multivariate models become necessary in crash-frequency modeling when, instead of total crash counts, one wishes to model specific types of crash counts (for example, the number of crashes resulting in fatalities, injuries, etc.). Modeling the counts of specific types of crashes (as opposed to total crashes) cannot be done with independent count models because the counts of specific crash types are not independent (that is, the counts of crashes resulting in fatalities cannot increase or decrease without affecting the counts of crashes resulting in injuries and no injuries). To resolve this

problem, Bivariate/multivariate models are used because they explicitly consider the correlation among the severity levels (for example) for each roadway entity (Miaou and Song, 2005; Bijleveld, 2005, Song et al., 2006).

Bivariate models are used for jointly modeling two crash types (Subrahmainiam and Subrahmainiam, 1973; Maher, 1990; N'Guessan et al., 2006; Geedipally and Lord, 2009; N'Guessan, 2010).[18] Extensions to more than two crash types (multivariate-model formulations) have been proposed including the multivariate Poisson model (Ma and Kockelman, 2006), the multivariate negative binomial model (Winkerlman, 2003), and the multivariate Poisson-lognormal model (Park and Lord, 2007; Ma et al., 2008; El-Basyouny and Sayed, 2009a; Park et al., 2010).[19] On the downside, bivariate/multivariate models are complex to estimate in that they require a formulation of a correlation matrix.

Finite Mixture/Markov Switching Models

Finite mixture/Markov switching models are a new type of model that can be used to examine heterogeneous populations. Although this type of model has been around for some time, they have recently become more popular because of the advancement in computing power and technology (Frühwirth-Schnatter, 2006). For finite mixture models, the assumption is that the overall data are generated from several distributions that are mixed together implying that individual observations are generated from an unknown

[18] Other researchers have also employed a variant of bivariate models and established a non-linear function for modeling different crash types at the same time (Miaou and Lord, 2003; Bonneson and Pratt, 2008).

[19] Park and Lord (2007) have found the multivariate Poisson-lognormal model to be a preferred choice in most applications because it can handle overdispersion and has a fully general correlation structure.

number of subgroups. Markov Switching models also work on the assumption that a number of underlying distributions generate the data and that individual observations can switch among these distributions over time.

In recent years, a few researchers have examined the application of finite mixture models (Park and Lord, 2009; Park et al., 2009) and Markov switching models (Malyshkina et al., 2009; Malyshkina and Mannering, 2010a) to highway safety. Finite mixture and Markov switching models offer considerable potential for providing important new insights into the analysis of crash data, but these models are also quite complex to estimate.

Duration Models

Another way of framing the crash-frequency problem is to consider the time between crashes, as opposed to the frequency of crashes over some time period. The frequency of crashes and the time between crashes are obviously interrelated. In fact, count-data models (such as the standard Poisson model) imply an underlying distribution of time between crashes (for the standard Poisson the underlying time distribution is exponentially distributed), and a model of the duration of time between crashes can be aggregated to produce an expected frequency in any given period of time.

The most common duration-model approach is a hazard-based model that considers the conditional the probability of a crash happening at some time $t + dt$ given that it has been time t since the last crash occurred. Hazard-based models can be estimated under a wide variety of distributional assumptions and non-parametric forms

and allow important inferences to be made on how the probabilities of having a crash change over time (see Washington et al., 2003, 2010).[20]

Hazard-based duration models can be quite sophisticated in terms of their ability to handle data and common problems associated with crash data (unobserved heterogeneity, etc.) and can offer insights relating to duration effects. As an example, instead of considering the number of crashes individual drivers had over their lifetimes, Mannering (1993) used a hazard-based duration model to study the factors affecting the time between their crashes. The results of this study found interesting duration effects in that the longer males went without having a crash, the less likely they were to have a crash soon, but that the length of time females went without a crash was found to have no statistically significant effect on their crash probabilities.

There is considerable potential for the future application of duration models to crash frequency analysis, but the level of data required in terms of the timing of crashes and the values of explanatory variables and how they change over time can be prohibitive in many instances.

Hierarchical/Multilevel Models

Hierarchical models are used for analyzing data that are characterized by correlated responses within hierarchical clusters. In highway safety, crash data could be seen as exhibiting several levels of hierarchy. For instance, the lowest level of the hierarchy could be the crashes themselves. Then, the next level could be the type of vehicle (passenger cars, trucks, etc.). For the subsequent one, it could be the accident

[20] Jovanis and Chang (1989) applied hazard models to study the time until a crash on individual trips for commercial trucks. And, Chang and Jovanis (1990) conceptually explored some potential applications of these models to studying crash occurrences.

location on the transportation network, and so on. With this type of model, the primary assumption is that correlation may exists among crashes occurring for the same kind of vehicle and location, because they may share unobserved characteristics related to the vehicle type or location. Not considering the potential hierarchical structure of the data (the potential of a complex correlation structure) may lead to poorly estimated coefficients and associated standard errors, particularly when they are modeled using a traditional count-data modeling approaches (Skinner et al., 1989; Goldstein, 1995). On the other hand, depending upon the study objectives, these models may not be warranted, even if correlations considered are not large (de Leeuw and Kreft, 1995) and the modeling output may be difficult to interpret, especially by non-statisticians (Pietz, 2003).[21] There have been a number of applications of hierarchical models to crash data (Jones and Jørgenson, 2003; Kim et al., 2007).

Neural, Bayesian Neural Network, Support Vector Machine Models

Neural and Bayesian neural network models are functions that are defined using multilevel network structures (Liang, 2005). The network structure consists of a series of nodes and weight factors that link the various nodes together in hierarchical manner: input layer, hidden layer, and output layer. Although both Neural and Bayesian neural network models have similar modeling processes, they are actually different in the way they predict the outcome variables. For neural networks, the weights are assumed fixed, whereas for Bayesian neural networks, the weights follow a probability distribution and the prediction process needs to be integrated over all the probability weights. These

[21] It has been argued that hierarchical models are really a class of random effects models because they attempt to capture correlations among groups of data. Other researchers, however, consider these models on a class of their own (Gelman and Hill, 2007).

models have been used in highway safety (Abdelwahab and Abdel-Aty, 2002; Chang, 2005; Riviere et al., 2006; Xie et al., 2007) mainly as a predictive tool. Overall, these models tend to exhibit better linear/non-linear approximation properties than traditional count-model approaches. However, these models often cannot be generalized to other data sets (Xie et al., 2007).

Support vector machine models, which are based on statistical learning theory, are a new class of models that can be used for predicting count-frequencies (Kecman, 2005). These models are a set of related supervised learning methods used for classification and regression, and possess the well-known ability of being able to approximate any multivariate function to any desired degree of accuracy. Statistical learning theory and structural risk minimization are the theoretical foundations for the learning algorithms of support vector machine models. It has been found that these models show better or comparable results to the outcomes estimated by neural networks and other statistical models (Kecman, 2005).

Support vector machine models have recently been introduced for other transportation applications (see Zhang and Xie, 2007), including for predicting crashes (Li et al., 2008). However, they are complex to estimate and, like neural and Bayesian neural network models, these models often cannot be generalized to other data sets. Another general criticism of neural, Bayesian neural network, and support vector machine models is that they all tend to behave as black-boxes in that they do not provide the interpretable parameters one gets when using traditional crash-frequency models.[22]

[22] Other recently proposed methods include multivariate adaptive regression splines (Haleem and Abdel-Aty, 2010), the application of reliability processes (Haleem et al., 2010), and the use of genetic programming (Das et al., 2010).

PARAMETER ESTIMATION METHODS

Maximum likelihood estimation and Bayesian methods are the two most common methods used for crash-frequency models. The main advantage of the maximum likelihood estimation is that closed-form functions often exist for the most common distributions used. On the other hand, maximum likelihood estimation cannot be used when the likelihood function is difficult to characterize.

Bayesian estimating methods have been gaining in popularity due to advances in computing methods (Gilks et al., 1996). Bayesian models have the advantage of being able to handle very complex models, especially those that do not have easily calculable likelihood functions. Using Markov Chain Monte Carlo (MCMC) methods, a sampling-based approach to estimation that is well suited for Bayesian models, complex functional model forms can be handled. For instance, random-parameter and Markov switching models are more easily estimated using MCMC simulation. However, even with the great computational benefits provided by Bayesian models, the simulation time for the Markov Chain Monte Carlo simulation can still be a barrier to complex model forms. The simulation time, which is a function of the size of the sample and the complexity of the model structure, can take several days and this time-issue can still be a limiting factor in the complexity of the model.[23]

[23] For an example of this, Malyshkina and Mannering (2010a) estimated a Markov switching model of crash frequencies using Markov Chain Monte Carlo simulation. However, they had to aggregate the data over a year instead of the weekly data they used previously (Malyshkina et. al, 2009) to be able to estimate the model due to the computing time and computational capacity required.

SUMMARY AND CONCLUSIONS

As the preceding discussion indicates, crash-frequency data pose formidable problems in terms of data characteristics (overdispersion, underdispersion, time-varying explanatory variables, low sample means and size, crash-type correlation, underreporting of crashes, omitted variables bias, and issues related to functional form and fixed parameters). To deal with these data-related problems, innovative methodological approaches have been introduced in an attempt to improve the statistical validity of findings. In the past few years in particular, this stream of methodological innovation has introduced some very exciting statistical approaches. Random-parameter models, finite mixture models, Markov switching models and others all hold great promise in improving our understanding of the factors that affect the frequency of crashes. And, one would expect that in the coming years, variations and refinements of these more advanced models could help reveal new insights.[24]

[24] However, there is also a practical side of crash-frequency analysis which often presents researchers with the need to trade-off the level of methodological sophistication with the ability to forecast the number of crashes to help governmental agencies set safety policy and allocate safety resources. A classic example of this is the somewhat common practice of developing models to predict crash frequencies using only traffic volumes as explanatory variables. From a pragmatic point of view, such models are relatively simple and they use explanatory variables that can be readily forecast. Unfortunately, from a statistical point of view, such a simple model will have a clear omitted-variables bias which will result in incorrectly estimated parameters and thus errors in forecasting. Resolving this omitted-variables problem comes at a cost that includes the need for much more data for model estimation (in terms of additional explanatory variables needed for model estimation) and the ability to forecast the additional variables that are likely to be found significant in model estimation (levels of precipitation, friction coefficients of the pavement, and so on). Aside from this omitted variables trade-off example, there are countless other trade-offs one may have to make during the model development (for example, how one handles unobserved heterogeneity, time varying explanatory variables and so on).

This paper shows a steady advancement of research in crash-frequency data analysis over the years. For the most part, researchers have spent an extraordinary amount of effort in developing models with superior statistical fit and/or predictive capabilities. But it is important to keep in mind that this work has been inherently limited by the available data which has been overly restrictive in terms of the insights that could be made and the statistical methodologies that could be developed to generate these insights. The anticipated availability of new data that includes detailed driving data (acceleration, braking and steering information, driver response to stimuli, etc.) and crash data (from vehicle black boxes) holds considerable promise for the future development of the field. When these data become available to the full research community, an entirely new direction of research could potentially open up – one that would provide exciting new insights into fundamental cause and effect relationships as they relate to motor-vehicle crash frequencies.[25]

ACKNOWLEDGEMENTS

This paper benefited from the input of Dr. Mohamed Abdel-Aty, Dr. James A. Bonneson, Dr. Venky N. Shankar, Dr. Gudmundur F. Ulfarsson, and Dr. Simon P. Washington. Their comments and suggestions are gratefully acknowledged.

[25] There is a body of work from other fields that could be used as a starting point in this endeavor. See, for example, the work on causal-relationship models by Rubin (1978; 1991) and Pearl (2000). Recently, a Transportation Research Board sub-committee associated with the development of the forthcoming Highway Safety Manual has put forward a document detailing the importance of including causal-relationships in future model development (Transportation Research Board, 2009). Elvik (2003) and Davis (2004) have also discussed this important issue for highway safety analyses.

REFERENCES

Abbas, K.A., 2004. Traffic Safety Assessment and Development of Predictive Models for Accidents on Rural Roads in Egypt. Accident Analysis and Prevention 36(2), 149-163.

Amoros, E., Martin, J.L., Laumon, B., 2003. Comparison of road crashes incidence and severity between some French counties. Accident Analysis and Prevention 35 (4), 537-547.

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41(1), 153-159.

Abdelwahab, H.T., and Abdel-Aty, M.A. (2002) Artificial neural networks and logit models for traffic safety analysis of toll plazas. Transportation Research Record 1784, pp. 115-125.

Aptel, I., Salmi, L. R., Masson, F., Bourdet, A, Henrion, G., Erny, P., 1999. Road accident statistics: discrepancies Between police and hospital data on a French island. Accident Analysis and Prevention 31(1) 101–108.

Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of road Crash Frequency with Spatial Models. Transportation Research Record 2061, 55-63.

Aguero-Valverde, J., and P.P. Jovanis, 2009. Bayesian Multivariate Poisson Log-Normal Models for Crash Severity Modeling and Site Ranking. Paper presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C.

Bijleveld, F. D., 2005. The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. Accident Analysis and Prevention 37(4), 591-600.

Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V., 2008.Applied Spatial Data Analysis with R (Use R). Springer Science, New York, N.Y.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroscedasticity. Journal of Econometrics 31(3), 307-327.

Bonneson, J.A., McCoy, P., 1993. Estimation of Safety at Two-Way Stop-Controlled Intersections on Rural Roads. Transportation Research Record 1401, 83-89.

Bonneson, J.A., Pratt, M.P., 2008. Procedure for developing accident modification factors from cross-sectional data. Transportation Research Record 2083, 40-48.

Brüde, U., Larsson, J., 1993. Models for predicting accidents at junctions where pedestrians and cyclists are involved. How well do they fit? Accident Analysis and Prevention 25 (5), 499-509.

Brüde, U., Larsson, J., Hegman, K.-O., 1998. Design of Major Urban Junctions-Accident Prediction Models and Empirical Comparison, VTI, Linkööping, Sweden.

Cafiso, S., di Graziano, A., Di Silvestro, G., La Cava, G., Persaud, B., 2010a. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. Accident Analysis and Prevention, in press.

Cafiso, S., Di Silvestro, G., Persaud, B., Begum, M.A., 2010b. Revisiting the Variability of the Dispersion Parameter of Safety Performance Functions Using Data for Two-Lane Rural Roads. Paper 10-3572, 89th Annual Meeting of the Transportation Research Board, Washington, D.C.

Caliendo, C., Guida, M., Parisi, A., 2007. A crash-prediction model for multilane roads. Accident Analysis and Prevention 39(4), 657-670.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge, U.K.

Carson, J., Mannering, F., 2001. The effect of ice warning signs on accident frequencies and severities. Accident Analysis and Prevention 33(1), 99-109.

Chang, L.-Y., 2005. Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. Safety Science 43 (8), 2005, 541-557.

Chang, H.-L., Jovanis, P.P., 1990. Formulating accident occurrence as a survival process. Accident Analysis and Prevention 22(5) 407-419.

Chung, Y., 2010. Development of an accident duration prediction model on the Korean Freeway Systems. Accident Analysis & Prevention 42(1), 282-289.

Conway, R.W., Maxwell, W.L., 1962. A queuing model with state dependent service rates. Journal of Industrial Engineering, 12, 132-136.

Daniels, S., Brijs, T., Nuyts, E., Wets, G., 2010. Explaining variation in safety performance of roundabouts. Accident Analysis and Prevention, in press.

Das A., Abdel-Aty M., Pande A., 2010. Using genetic programming to investigate the design parameters contributing to crash occurrence on urban arterials, Preprint No. TRB 10-1409, 89th Annual Meeting of the Transportation Research Board, January 2010.

Davis, G.A., 2004. Possible aggregation biases in road safety research and a mechanism approach to accident modeling. Accident Analysis and Prevention, 36(6), 1119-1127.

de Leeuw, J., Kreft, I.G.G., 1995. Questioning Multilevel Models. Journal of Educational and Behavioral Statistics 20(2), 171-189.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accident Analysis and Prevention 40(4), 1257-1266.

Dingus, T. A., Klauer, S.G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., Perez, M. A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z. R., Jermeland, J., and Knipling, R.R., 2006. The 100-car naturalistic driving study: phase II – results of the 100-car field experiment, DOT HS 810 593, Washington, DC.

El-Basyouny, K., Sayed, T., 2006. Comparison of two negative binomial regression techniques in developing accident prediction models**.** Transportation Research Record 1950, 9-16.

El-Basyouny, K., Sayed, T., 2009a. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention, 41(4), 820-828.

El-Basyouny, K., Sayed, T., 2009b. Accident prediction models with random corridor parameters. Accident Analysis and Prevention, 41(5), 1118-1123.

Elvik, R., 2003. Assessing the validity of road safety evaluation studies by analysing causal chains. Accident Analysis and Prevention, 35(5), 741-748.

Elvik, R., Mysen, A.B., 1999. Incomplete accident reporting: meta-analysis of studies made in 13 countries. Transportation Research Record 1665, 133-140.

Engle, R., 1982. Autoregressive conditional heteroscedasticity with estimates of variance of UK inflation. Econometrica 50(4), 987-1008.

Flahaut, B., Mouchart, M., San Martin, E., Thomas, I., 2003. The local spatial autocorrelation and the kernel method for identifying black zones: A comparative approach. Accident Analysis and Prevention 35(6), 991-1004.

Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., Thomsen, L.K., 1995. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accident Analysis and Prevention 27(1), 1–20.

Frühwirth-Schnatter, S., 2006 Finite Mixture and Markov Switching Models. Springer Series in Statistics, Springer, New York.

Geedipally, S.R., Lord, D., 2009. Investigating the effect of modeling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson-gamma models. Submitted for publication in Accident Analysis and Prevention.

Gelman, A., Hill, J., 2007. Data Analysis using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York, NY.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. Chapman and Hall, London, UK.

Goldstein, H., 1995. Multilevel Statistical Models, 2nd Edition. Edward Arnold, London, UK.

Gourieroux, C.A., Visser, M., 1997. A count data model with unobserved heterogeneity. Journal of Econometrics 79(2), 247-268.

Guikema, S.D., Coffelt, J.P., 2007. A flexible count data regression model for risk analysis. Risk Analysis 28(1), 213-223.

Guo, F., Wang, X., Abdel-Aty, M., 2010. Modeling signalized intersection safety with corridor spatial correlations. Accident Analysis and Prevention 42(1), 84-92.

Guo, G., 1996. Negative multinomial regression models for clustered event counts. Sociological Methodology. 26(1), 113-132.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized additive models. Chapman and Hall, New York.

Haleem K., Abdel-Aty M., 2010. Multiple applications of the multivariate adaptive regression splines technique in Predicting rear-end crashes at unsignalized intersections, Preprint No. TRB 10-0292, 89th Annual Meeting of the Transportation Research Board, January 2010.

Haleem K., Abdel-Aty M., Mackie K., 2010. Using a reliability process to reduce uncertainty in predicting crashes at unsignalized intersections, Accident Analysis and Prevention, 42(2), 654-666.

Halekoh, U., Højsgaard, S., Yan, J., 2006. The R Package geepack for generalized estimating equations. Journal of Statistical Software, 15(2), 1-11.

Hauer, E., 1997. Observational Before–After Studies in Road Safety. Pergamon Press, Elsevier Science Ltd., Oxford, England.

Hauer, E., 2001. Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation. Accident Analysis and Prevention, 33(6), 799-808.

Hauer, E., 2004. Statistical Road Safety Modeling. Transportation Research Record 1897, 81–87.

Hauer, E., Hakkert, A. S., 1988. Extent and some implications of incomplete accident reporting. Transportation Research Record 1185, 1–10.

Hauer, E., Ng, J.C.N., Lovell, J., 1988. Estimation of Safety at Signalized Intersections. Transportation Research Record 1185, 48-61.

Hausman, J. A., Hall B. H., Griliches, Z. 1984. Econometric models for count data with an application to the patents-R&D relationship. Econometrica 52(4), 909–938.

Heydecker, B.G., Wu, J., 2001. Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference. Advances in Engineering Software 32(10), 859-869.

Hirst, W.M., Mountain, L.J., Maher, M.J., 2004. Sources of error in road safety scheme evaluation: a method to deal with outdated accident prediction models. Accident Analysis and Prevention 36 (5), 717-727.

James, H. F., 1991. Under-reporting of road traffic accidents. Traffic Engineering and Control 32(12), 574–583.

Johansson, P., 1996. Speed limitation and motorway casualties: A time series count data regression approach. Accident Analysis and Prevention 28(1), 73-87.

Jones, A.P., Jørgensen, S.H., 2003. The use of multilevel models for the prediction of road accident outcomes. Accident Analysis and Prevention 35(1), 59-69.

Jones, B., Janssen, L., Mannering, F., 1991. Analysis of the frequency and duration of freeway accidents in Seattle. Accident Analysis and Prevention 23(2), 239-255.

Joshua S.C., Garber, N.J., 1990. Estimating truck accident rate and involvements using linear and Poisson regression models. Transportation Planning and Technology 15(1), 41-58.

Jovanis, P.P., Chang, H.L., 1986. Modeling the Relationship of Accidents to Miles Traveled Transportation Research Record 1068, 42-51.

Jovanis, P.P., Chang, H.L., 1989. Disaggregate model of highway accident occurrence using survival theory. Accident Analysis and Prevention 21(5), 445-458.

Kadane, J.B., Shmueli, G., Minka, T.P., Borle, S., Boatwright, P., 2006. Conjugate analysis of the Conway-Maxwell-Poisson distribution. Bayesian Analysis, 1(2), 363-374.

Karlaftis, M., Tarko, A., 1998. Heterogeneity considerations in accident modeling. Accident Analysis and Prevention 30(4), 425-433.

Kecman, V., 2005. Support Vector Machines – An Introduction. In: Support Vector Machines: Theory and Applications, ed. By L. Wang, Springer-Verlag Berlin Heidelberg, New York, pp.1-48.

Kim, D., Washington, S., 2006. The significance of endogeneity problems in crash models: An examination of left-turn lanes in intersection crash models. Accident Analysis and Prevention 38(6), 1094-1100.

Kim, D.-G., Lee, Y., Washington, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: Application of hierarchical binomial logistic models. Accident Analysis & Prevention 39(1), 125-134.

Kumala, R., 1995. Safety at Rural Three- and Four-Arm Junctions: Development and Applications of Accident Prediction Models. VTT Publications 233, Technical Research Centre of Finland, Espoo, Finland.

Kumara, S.S.P., Chin, H.C., 2003. Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. Traffic Injury Prevention 3(4), 53-57.

Kumara, S.S.P., Chin, H.C., 2005. Application of Poisson underreporting model to examine crash frequencies at signalized three-legged intersections. Transportation Research Record 1908, 46-50.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34(1), 1-14.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accident Analysis and Prevention 34(2), 149-161.

Li, W., Carriquiry, A., Pawlovich, M., Welch, T., 2008. The choice of statistical models in road safety countermeasure effectiveness studies in Iowa. Accident Analysis and Prevention 40 (4), 1531-1542.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accident Analysis and Prevention 40(4), 1611-1618.

Li, X., Lord, D., Zhang, Y., 2009. Development of accident modification factors for rural frontage road segments in Texas using results from generalized additive models. Working Paper, Zachry Department of Civil Engineering, Texas A&M University, College Station, TX.

Liang, F., 2005. Bayesian neural networks for nonlinear time series forecasting. Statistics and Computing 15(1), 13-29.

Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73(1), 13-22.

Lord, D., Persaud, B.N., 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure. Transportation Research Record 1717, 102-108.

Lord, D., 2002. Issues related to the application of accident prediction models for the computation of accident risk on transportation networks. Transportation Research Record 1784, 17-26.

Lord, D., Bonneson, J.A., 2005. Calibration of predictive models for estimating the safety of ramp design configurations. Transportation Research Record 1908, 88-95.

Lord, D., Manar, A., Vizioli, A., 2005a. Modeling crash-flow-density and crash-flow-v/c ratio for rural and urban freeway segments. Accident Analysis and Prevention 37(1), 185-199.

Lord, D., Washington, S.P., Ivan, J.N., 2005b. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 37(1), 35-46.

Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the Estimation of the fixed dispersion parameter. Accident Analysis and Prevention 38(4), 751-766.

Lord, D., Bonneson, J.A., 2007. Development of Accident Modification Factors for Rural Frontage Road Segments in Texas. Transportation Research Record 2023, 20-27.

Lord, D., Washington, S.P., Ivan, J.N., 2007. Further notes on the application of zero inflated models in highway safety. Accident Analysis and Prevention 39(1), 53-57.

Lord, D., Guikema, S., Geedipally, S.R., 2008. Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis and Prevention 40(3), 1123-1134.

Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: A Bayesian perspective. Safety Science 46(5), 751-770.

Lord, D., Geedipally, S.R., Guikema, S., 2009. Extension of the application of Conway-Maxwell-Poisson models: analyzing traffic crash data exhibiting under-dispersion. Submitted to the 89th Annual Meeting of the Transportation Research Board, Washington, D.C.

Lord, D., Mahlawat, M., 2009. Examining the application of aggregated and disaggregated Poisson-gamma models subjected to low sample mean bias. Transportation Research Record 2136, 1-10.

Ma, J., Kockelman, K.M., 2006. Bayesian multivariate Poisson regression for models of injury count by severity. Transportation Research Record 1950, 24-34.

Ma, J., Kockelman, K.M., and Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis and Prevention 40 (3), 964–975.

Ma J., 2009. Bayesian analysis of underreporting Poisson regression model with an application to traffic crashes on two-lane highways. Paper #09-3192. Presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C.

MacNab, Y.C., 2004. Bayesian spatial and ecological models for small-area crash and injury analysis. Accident Analysis and Prevention 36(6), 1019-1028.

Mahalel, D., 1986. A note on accident risk. Transportation Research Record 1068, 85-89.

Maher, M.J., 1990. A bivariate negative binomial model to explain traffic accident migration. Accident Analysis and Prevention 22(5), 487-498.

Maher, M.J., Hughes, P.C., Smith, M.J., Ghali, M.O., 1993. Accident- and travel time-minimizing routing patterns in congested networks. Traffic Engineering and Control 34(9) 414-419.

Maher M.J., Summersgill, I., 1996. A comprehensive methodology for the fitting predictive accident models. Accident Analysis and Prevention 28(3), 281-296.

Malyshkina, N.V., Mannering, F.L., Tarko, A.P., 2009. Markov switching negative binomial models: an application to vehicle accident frequencies. Accident Analysis and Prevention 41(2), 217–226.

Malyshkina, N., Mannering, F., 2010a. Zero-state Markov switching count-data models: An empirical assessment. Accident Analysis and Prevention 42(1), 122-130.

Malyshkina, N., Mannering, F., 2010b. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. Accident Analysis and Prevention 42(1), 131-139.

Mannering, F., 1993. Male/female driver characteristics and accident risk: Some new evidence. Accident Analysis and Prevention 25(1), 77-84.

Maycock, G., Hall, R.D., 1984. Accidents at 4-Arm Roundabouts. TRRL Laboratory Report 1120, Transportation and Road Research Laboratory, Crowthorne, U.K.

McCulloch, C.E., Searle, S.R., Neuhaus, J.M., 2008. Generalized, Linear, and Mixed Models, 2nd Ed., John Wiley & Sons, Hoboken, N.J.

Miaou, S.-P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accident Analysis and Prevention 25(6), 689-709.

Miaou, S.-P. 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention 26(4), 471-482.

Miaou, S.-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus Empirical Bayes. Transportation Research Record 1840, 31-40.

Miaou, S.-P., Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping: A space-time modeling approach. Journal of Transportation and Statistics 6(1), 33-57.

Miaou, S.-P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion and spatial dependence. Accident Analysis and Prevention 37(4), 699-720.

Miaou, S.-P., Bligh, R.P., Lord, D., 2005. Developing median barrier installation guidelines: a benefit/cost analysis using Texas data. Transportation Research Record 1904, 3-19.

Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor vehicle accident frequencies. Transportation 25(4), 395-413.

Milton, J., Shankar, V., Mannering, F., 2008. Highway accident severities and the mixed logit model: An exploratory empirical analysis. Accident Analysis and Prevention 40(1), 260-266.

Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 39(3), 459-468.

Mountain, L., Fawaz, B., Jarrett, D., 1996. Accident prediction models for roads with minor junctions. Accident Analysis and Prevention 28(6), 695-707.

Mountain, L., Maher, M.J., Fawaz, B., 1998. The influence of trend on estimates of accidents at junctions. Accident Analysis and Prevention 30(5), 641-49.

N'Guessan, A., 2010. Analytical Existence of solutions to a system of nonlinear equations with application. Journal of Computational and Applied Mathematics, 234, 297-304.

N'Guessan, A., Essai, A., N'Zi, M., 2006. An Estimation method of the average effect and the different accident risks when modeling a road safety measure: a simulation study. Computational Statistics and Data Analysis, 51, 1260-1277.

N'Guessan A., Langrand C. 2005a. A covariance components estimation procedure when modelling a road safety measure in terms of linear constraints. Statistics, 39(4), 303-314.

N'Guessan, A., Langrand, C., 2005b. A Schur complement approach for computing subcovariance matrices arising in a road safety measure modeling. Journal of Computational and Applied Mathematics, 177, 331-345.

Noland, R. B., Quddus, M A., 2004. A spatially disaggregated analysis of road casualties in England. Accident Analysis and Prevention 36(6), 973-984.

Oh, J., Washington, S.P., Nam, D., 2006. Accident prediction model for railway-highway interfaces. Accident Analysis and Prevention 38(2), 346-56.

Park, E.-S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record 2019, 1-6.

Park, E-S., Park, J., Lomax, T.J., 2010. A fully Bayesian multivariate approach to before-after safety evaluation. Accident Analysis and Prevention, in press.

Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis and Prevention, 41(4), 683-691.

Park, B.J., Lord, D., Hart, J.D., 2010. Bias properties of Bayesian statistics in finite mixture of negative regression models for crash data analysis. Accident Analysis and Prevention, 42(2), 741-749.

Payne, R.W., 2000. The Guide to Genstat. Lawes Agricultural Trust, Rothamsted Experimental Station, Oxford, U.K.

Pearl, J., 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, UK.

Persaud, B.P., 1994. Accident prediction models for rural roads. Canadian journal of civil engineering 21 (4), 547-554.

Persaud, B.P., Nguyen, T., 1998. Disaggregate safety performance models for signalized intersections on Ontario provincial roads. Transportation Research Record 1635, 113-120.

Piegorsch W.W., 1990. Maximum likelihood estimation for the negative binomial dispersion parameter, Biometrics, 46(3), 863–867.

Pietz, K., 2003. An Introduction to Hierarchical Modeling. Seminar, U.S. Department of Veteran Affairs, Houston, TX (accessed Nov 5, 2009: www.hsrd.houston.med.va.gov/Documents/Linear%20Models.ppt)

Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection-accident frequencies. Journal of Transportation Engineering 122(2), 105-113.

Poormeta, K., 1999. On the modelling of overdispersion of counts. Statistica Neerlandica 53 (1), 5-20.

Qin, X., Ivan, J.N., and Ravishankar, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis and Prevention 36 (2), 183–191.

Quddus, M.A., 2008. Time series count data models: An empirical application to traffic accidents. Accident Analysis and Prevention, 40 (5), 1732-1741.

Riviere, C., Lauret, P., Ramsamy, J.F.M., and Page, Y., 2006. A Bayesian neural network approach to estimating the energy equivalent speed. Accident Analysis and Prevention 38(2), 248-259.

Rubin, D., 1978. Bayesian inference for causal effects: the role of randomization. Annals of Statistics 6(1), 1-26.

Rubin, D., 1991. Practical implications of modes of inference for causal effects and the critical role of the assignment mechanism. Biometrics 47, 1213-1234.

Sellers, K. F., Shmueli, G. 2010. A Flexible Regression Model for Count Data. The Annals of Applied Statistics, in press.

Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural accident frequencies. Accident Analysis and Prevention 27(3), 371-389.

Shankar, V., Milton, J., Mannering, F.L., 1997. Modeling accident frequency as zero-altered probability processes: an empirical inquiry. Accident Analysis and Prevention 29(6) 829-837.

Shankar, V.N., Albin, R.B., Milton, J.C., Mannering, F.L., 1998. Evaluating median cross-over likelihoods with clustered accident counts: an empirical inquiry using random effects negative binomial model. Transportation Research Record 1635, 44-48.

Shankar, V.N., Ulfarsson, G.F., Pendyala, R.M., and Nebergall, M.B., 2003. Modeling crashes involving pedestrians and motorized traffic. Safety Science, 41(7), 627-640.

Shankar V., Jovanis P., Aguerde J. and Gross F., 2008. Analysis of Naturalistic Driving Data: Prospective View on Methodological Paradigms. Transportation Research Record 2061, 1-9.

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., Boatwright, P., 2005. A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. Journal of the Royal Statistical Society: Series C 54(1), 127-142.

Shugan, S.M., 2006. Editorial: errors in the variables, unobserved heterogeneity, and other ways of hiding statistical error. Marketing Science 25(3), 203-216.

Sittikariya S., Shankar V., 2009. Modeling Heterogeneity: Traffic Accidents. VDM-Verlag, 80 pp, 2009.

Skinner, C.J., Holt, D., Smith, T.M.F., 1989. Analysis of Complex Surveys. Wiley, Chichester, UK.

Song, J. J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of Multivariate Analysis 97(1), 246-273.

Subrahmaniam, K and K. Subrahmaniam (1973). On the estimation of the parameters in the bivariate negative binomial distribution, Journal of the Royal Statistical Society Series B 35, 131–146.

Tanner, J.C., 1953. Accidents at rural three-way junctions. Journal of the Institution of Highway Engineers 2(11), 56-67.

Transportation Research Board, 2009. Theory, Explanation and Prediction in Road Safety: Identification of Promising Directions and a Plan for Advancement. Workshop Circular. Task Force for the Development of a Highway Safety Manual, Transportation Research Board, Washington D.C., (Access on January 20, 2010: http://tcd.tamu.edu/FDsub/Safety_Workshop_Circular_3_.pdf)

Turner, S., Nicholson, A., 1998. Using accident prediction models in area wide crash reduction studies. In Proceedings of the 9th Road Engineering Association of Asia and Australasia Conference, Wellington, NZ.

Ulfarsson, G.F., and Shankar, V.N., 2003. An accident count model based on multi-year cross-sectional roadway data with serial correlation. Transportation Research Record 1840, 193-197.

Wang, C., Quddus, M.A., Ison, S., 2009. The effects of area-wide road speed and curvature on traffic casualties in England. Journal of Transport Geography 17(5), 385-395.

Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. Accident Analysis and Prevention 38(6), 1137-1150.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2003. Statistical and Econometric Methods for Transportation Data Analysis. Chapman Hall/CRC, Boca Raton, FL.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2010. Statistical and Econometric Methods for Transportation Data Analysis. Second Edition, Chapman Hall/CRC, Boca Raton, FL.

Winkelman, R., 2003. Econometric Analysis of Count Data (4th ed.). New York: Springer.

Wood G.R., 2002. Generalised linear accident models and goodness of fit testing. Accident Analysis and Prevention 34 (4), 417-427.

Wood, S.N., 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC, Boca Raton, Florida.

Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural networks: an empirical analysis. Accident Analysis and Prevention 39(5), 922-933.

Xie, Y., Zhang, Y., 2008. Crash frequency analysis with generalized additive models. Transportation Research Record 2061, 39-45.

Yamamoto T. Hashiji J. and Shankar V., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models, Accident Analysis and Prevention 40(4), 1320-1329.

Ye, X., Pendyala, R.M., Washington, S.P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. Safety Science 47(3), 443-452.

Zhang, Y. Xie, Y., 2007. Forecasting of short-term freeway volume with v-Support vector machines. Transportation Research Board 2024, 92-99.

Table 1. Data and Methodological Issues Associated with Crash-Frequency Data.

Data/Methodological Issue	Associated Problems
Overdispersion	Can violate some the basic count-data modeling assumptions of some modeling approaches
Underdispersion	As with overdispersion, can violate some the basic count-data modeling assumptions of some modeling approaches
Time-varying explanatory variables	Averaging of variables over studied time intervals ignores potentially important variations within time intervals – which can result in erroneous parameter estimates
Temporal and spatial correlation	Correlation over time and space causes losses in estimation efficiency
Low sample mean and small sample size	Causes an excess number of observations where zero crashes are observed which can cause errors in parameter estimates
Injury severity and crash type correlation	Correlation between severities and crash types causes losses in estimation efficiency when separate severity-count models are estimated
Under reporting	Under reporting can distort model predictions and lead to erroneous inferences with regard to the influence of explanatory variables
Omitted variables bias	If significant variables are omitted from the model, parameter estimates will be biased and possibly erroneous inferences with regard to the influence of explanatory variables will result
Endogenous variables	If endogenous variables are included without appropriate statistical corrections parameter estimates will be biased and erroneous inferences with regard to the influence of explanatory variables may be drawn
Functional form	If incorrect functional for is used, the result will be biased parameter estimates and possibly erroneous inferences with regard to the influence of explanatory variables

Fixed parameters	If parameters are estimated as fixed when they actually vary across observations, the result will be biased parameter estimates and possibly erroneous inferences with regard to the influence of explanatory variables

Table 2. Summary of Existing Models for Analyzing Crash-Frequency Data

Model Type	Advantages	Disadvantages
Poisson	Most basic model; easy to estimate	Cannot handle over- and under-dispersion; negatively influenced by the low sample mean and small sample size bias
Negative binomial/Poisson-gamma	Easy to estimate can account for overdispersion	Cannot handle under-dispersion; can be adversely influenced by the low sample mean and small sample size bias
Poisson-lognormal	More flexible than the Poisson-gamma to handle over-dispersion	Cannot handle under-dispersion; can be adversely influenced by the low sample mean and small sample size bias (less than the Poisson-gamma); cannot estimate a varying dispersion parameter
Zero-inflated Poisson and negative binomial	Handles datasets that have a large number of zero-crash observations	Can create theoretical inconsistencies; zero-inflated negative binomial can be adversely influenced by the low sample mean and small sample size bias
Conway-Maxwell-Poisson	Can handle under- and over-dispersion or combination of both using a variable dispersion (scaling) parameter	Could be negatively influenced by the low sample mean and small sample size bias; no multivariate extensions available to date
Gamma	Can handle under-dispersed data	Dual state model with one state having a long term mean equal to zero
Generalized estimating equation models	Can handle temporal correlation	May need to determine or evaluate the type of temporal correlation a priori; results sensitive to missing values

Generalized additive models	More flexible than the traditional generalized estimating equation models; allows non-linear variable interactions	Relatively complex to implement; may not be easily transferable to other datasets
Random-effects models	Handles temporal and spatial correlation	May not be easily transferable to other datasets
Negative multinomial	Can account for overdispersion and serial correlation; panel count data.	Cannot handle under-dispersion; can be adversely influenced by the low sample mean and small sample size bias
Random-parameters models	More flexible than the traditional fixed parameter models in accounting for unobserved heterogeneity	Complex estimation process; may not be easily transferable to other datasets
Bivariate/multivariate models	Can model different crash types simultaneously; more flexible functional form than the generalized estimating equation models (can use non-linear functions)	Complex estimation process; requires formulation of correlation matrix
Finite mixture/Markov Switching	Can be used for analyzing sources of dispersion in the data	Complex estimation process; may not be easily transferable to other datasets
Duration models	By considering the time between crashes (as opposed to crash frequency directly), allows for a very in-depth analysis of data and duration effects	Requires more detailed data than traditional crash frequency models; time-varying explanatory variables are difficult to handle
Hierarchical/Multilevel Models	Can handle temporal, spatial and other correlations among groups of observations	May not be easily transferable to other datasets; correlation results can be difficult to interpret

Neural Network, Bayesian Neural Network, and support vector machine	Non parametric approach does not require an assumption about distribution of data; flexible functional form; usually provides better statistical fit than traditional parametric models	Complex estimation process; may not be transferable to other datasets; work as black-boxes; may not have interpretable parameters

Table 3. Summary of Previous Research Analyzing Crash-Frequency Data[a]

Model Type	Previous Research
Poisson	Jovanis and Chang (1986); Joshua and Garber (1990); Jones et. al. (1991); Miaou and Lum (1993); Miaou (1994)
Negative binomial/Poisson-gamma	Maycock and Hall (1984); Hauer et al. (1988); Brüde and Larsson (1993); Bonneson and McCoy (1993); Miaou (1994); Persaud (1994); Kumala (1995); Shankar et al. (1995); Poch and Mannering (1996); Maher and Summersgill (1996); Mountain et al. (1996); Milton and Mannering (1998); Brüde et al. (1998); Mountain et al. (1998); Karlaftis and Tarko (1998); Persaud and Nguyen, 1998; Turner and Nicholson (1998); Heydecker and Wu (2001); Carson and Mannering (2001); Miaou and Lord (2003); Amoros et al. (2003); Hirst et al. (2004); Abbas (2004); Lord et al. (2005a); El-Basyouny and Sayed (2006); Lord (2006); Kim and Washington (2006); Lord and Bonneson (2007); Lord et al. (2009); Malyshkina and Mannering (2010b); Daniels et al. (2010); Cafiso et al. (2010a)
Poisson-lognormal	Miaou et al. (2005); Lord and Miranda-Moreno (2008); Aquero-Valverde and Jovanis (2008)
Zero-inflated Poisson and negative binomial	Miaou (1994); Shankar et al. (1997); Carson and Mannering (2001); Lee and Mannering (2002); Kumara and Chin (2003); Shankar et al. (2003); Qin et al., 2004; Lord et al. (2005b); Lord et al. (2007); Malyshkina and Mannering (2010a)
Conway-Maxwell-Poisson	Lord et al. (2008); Sellers and Shmueli (2010)
Gamma	Oh et al. (2006); Daniels et al. (2010)
Generalized estimating equation models	Lord and Persaud (2000); Lord et al. (2005a); Halekoh et al. (2006); Wang and Abdel-Aty (2006); Lord and Mahlawat (2009)
Generalized additive models	Xie and Zhang (2008); Li et al. (2009)

Random-effects models[b]	Johansson (1996); Shankar et al. (1998); Miaou and Lord (2003); Flahaut et al. (2003); MacNab (2004); Noland and Quddus (2004); Miaou et al., (2003); Miaou et al., (2005); Aquero-Valverde and Jovanis (2009); Li et al. (2008); Quddus (2008); Sittikariya and Shankar (2009); Wang et al. (2009); Guo et al. (2010)
Negative multinomial	Ulfarsson and Shankar (2003); Hauer (2004); Caliendo et al. (2007)
Random-parameters models	Anastasopoulos and Mannering (2009); El-Basyouny and Sayed (2009b)
Bivariate/multivariate models	Miaou and Lord (2003); Miaou and Song (2005); N'Guessan and Langrand (2005a); N'Guessan and Langrand (2005b); Bijleveld (2005); Song et al. (2006); Ma and Kockelman (2006); Park and Lord (2007); N'Guessan et al. (2006); Bonneson and Pratt (2008); Geedipally and Lord (2009); Ma et al. (2008); Depaire et al. (2008); Ye et al. (2009); Aguero-Valverde and Jovanis (2009); El-Basyouny and Sayed (2009a); N'Guessan (2010); Park et al. (2010)
Finite mixture/Markov switching	Malyshkina et al. (2009); Park and Lord (2009); Malyshkina and Mannering (2010a); Park et al. (2010)
Duration models	Jovanis and Chang (1989); Chang and Jovanis (1990); Mannering (1993); Chung (2010)
Hierarchical/Multilevel Models	Jones and Jørgensen (2003); Kim et al. (2007)
Neural Network, Bayesian Neural Network, and support vector machine	Abdelwahab and Abdel-Aty (2002); Chang (2005); Riviere et al. (2006); Xie et al. (2007); Li et al. (2008)

[a] This is a representative, but not a comprehensive list of references.
[b] Includes spatial statistical models.