Crash Frequency Analysis with Generalized Additive Models

Yuanchang Xie and Yunlong Zhang

Recent crash frequency studies have been based primarily on generalized linear models, in which a linear relationship is usually assumed between the logarithm of expected crash frequency and other explanatory variables. For some explanatory variables, such a linear assumption may be invalid. It is therefore worthwhile to investigate other forms of relationships. This paper introduces generalized additive models to model crash frequency. Generalized additive models use smooth functions of each explanatory variable and are very flexible in modeling nonlinear relationships. On the basis of an intersection crash frequency data set collected in Toronto, Canada, a negative binomial generalized additive model is compared with two negative binomial generalized linear models. The comparison results show that the negative binomial generalized additive model performs best for both the Akaike information criterion and the fitting and predicting performance.

Statistical models are extensively used in traffic safety studies for identifying major contributing factors to crashes and injuries, establishing proper relationships between crashes and explanatory variables, and predicting crash frequency and injury severity (*1*). The most popular statistical models used in crash frequency studies so far are generalized linear models (GLMs), including Poisson (*2*), Poisson–gamma or negative binomial (NB) (*3–5*), gamma (*6*), and zero-inflated models (*3, 7, 8*). One aspect common in all these models is that they usually assume a linear relationship between the logarithm of the expected crash frequency and the explanatory variables. One example of this linear relationship, often referred to as functional form, is

$$\eta_i = \ln(\mu_i) = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} \quad (1)$$

where

μ_i = expected number of accidents at study location i,
η_i = logarithm of μ_i,
β_0 = intercept,
β_j = coefficient for jth explanatory variable ($j = 1, \ldots, n$),
x_{ij} = value of jth explanatory variable for ith study location, and
n = number of explanatory variables.

The assumption of a linear relationship between the logarithm of expected crash frequency and explanatory variables in a GLM is usually out of convenience and common practice. It may not be the best assumption, though. Conceivably, there could be functional forms other than the linear one that better describe the relation between the logarithm of the expected crash frequency and other explanatory variables. For example, it is possible that as the value of an explanatory variable such as traffic volume increases, the rate of increase for expected crash frequency may slow down or even change sign because of the reduced level of speed variability. However, a linear functional form is not flexible enough to describe this type of relationship accurately.

In response to this limitation of GLMs, some researchers proposed using neural networks and support vector machines for crash injury (*9*) and crash frequency (*10, 11*) studies. Neural networks and support vector machines have strong nonlinear approximation ability and do not require specification of any functional forms. It has been shown that neural networks and support vector machines can outperform conventional GLMs for predicting crash frequencies (*10, 11*). However, neural networks and support vector machines are criticized for being black boxes because they cannot generate explicit functional relationships and statistically interpretable results. Although sensitivity analysis used by Xie et al. and Li et al. (*10, 11*) can partially address the functional form issue, neural networks and support vector machines still cannot produce statistically interpretable results, which are of great interest to traffic safety researchers.

With the aim of finding a method that has strong nonlinear approximation ability and can also produce statistically interpretable results, generalized additive models (GAMs) (*12*) are introduced and investigated in this study. GAMs can be considered as the extension of GLMs using nonparametric methods. In the functional forms of GAMs, nonparametric smooth functions are used to replace or in conjunction with the parametric terms shown in Equation 1. Since GAMs are based on statistical theory and still retain the basic framework of GLMs, they can generate statistically interpretable results. Also, since smooth functions are able to better fit nonlinear functions, it is expected that GAMs will have better nonlinear approximation ability than GLMs. To validate this conjecture, both GLMs and GAMs are applied to an intersection crash frequency data set. Since this data set exhibits overdispersion, an NB generalized linear model (NBGLM) and an NB generalized additive model (NBGAM) are fitted. The fitting and predicting results, along with the Akaike information criterion (AIC) (*13–15*) values of the NBGAM, are compared with those of the NBGLM to demonstrate the advantages and potential of GAMs in traffic safety studies.

THEORETICAL BACKGROUND

Negative Binomial Generalized Linear Models

An intersection crash data set is assumed that consists of n records $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_i, y_i), \ldots, (\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x}_i$ is a vector representing

Y. Xie, Department of Civil and Mechanical Engineering Technology, South Carolina State University, Orangeburg, SC 29117. Y. Zhang, Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136. Corresponding author: Y. Xie, yxie@scsu.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2061, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 39–45.
DOI: 10.3141/2061-05

the accident-related characteristics of observation i, and y_i is the corresponding number of crashes reported. A typical NBGLM is given by the following equations (*5*):

$$\text{Prob}(Y_i = y_i) = \frac{\Gamma(y_i + \phi)}{\Gamma(y_i + 1)\Gamma(\phi)}\left(\frac{\mu_i}{\mu_i + \phi}\right)^{y_i}\left(\frac{\phi}{\phi + \mu_i}\right)^{\phi} \quad (2)$$

Expectation of Y_i is

$$E(Y_i) = \mu_i = g(\mathbf{x}_i) \quad (3)$$

Variance of Y_i is

$$\text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\phi} \quad (4)$$

where

Y_i = independently and identically NB-distributed variable;
y_i = reported number of crashes for observation i, i = 1, 2, . . . , n;
ϕ = inverse dispersion parameter of NB distribution;
μ_i = expected number of accidents for observation i; and
$g(\mathbf{x}_i)$ = functional form of NBGLM.

For successful application of NBGLMs, one of the most important parts is to find an appropriate functional form. The following two functional forms are commonly used in NBGLMs (*4*, *5*):

$$\mu_i = g(x_i) = \exp\left\{\beta_0 + \sum_{j=1}^{n}\beta_j x_{ij}\right\} \quad (5)$$

$$\mu_i = g(x_i) = \text{off}_i * \exp\left\{\beta_0 + \sum_{j=1}^{n}\beta_j x_{ij}\right\} \quad (6)$$

where off_i is the offset. It could be a constant, an explanatory variable, or the result of combining several explanatory variables.

Taking the logarithm on both sides of Equation 5 results in Equation 1. Thus for the functional form in Equation 5, the relationship between the logarithms of the expected crash frequency (η_i) and other explanatory variables (x_{ij}) is linear. If the same operations are applied to Equation 6, the result is

$$\eta_i = \ln(\mu_i) = \ln(\text{off}_i) + \beta_0 + \sum_{j=1}^{n}\beta_j x_{ij} \quad (7)$$

The functional form in Equations 6 and 7 also considers a simple nonlinear logarithmic relationship between the logarithms of the expected crash frequency (η_i) and the offset (off_i). Nevertheless, NBGLMs can only model linear and simple nonlinear relationships. In addition, the nonlinear relationship has to be prespecified explicitly. To account for more complicated nonlinear relationships, NBGAMs will be introduced next.

Negative Binomial Generalized Additive Models

NBGAMs extend NBGLMs by introducing smooth functions into the functional forms. By so doing, Equations 5 and 6 become

$$\mu_i = g(x_i) = \exp\left\{\beta_0 + \sum_{j=1}^{n} f_j(x_{ij})\right\} \quad (8)$$

$$\mu_i = g(x_i) = \text{off}_i * \exp\left\{\beta_0 + \sum_{j=1}^{n} f_j(x_{ij})\right\} \quad (9)$$

where $f_j(x_{ij})$ is a smooth function, which is a nonparametric function without a rigid form (*12*). It should be noted that no coefficients are associated with these smooth functions. The functional forms for NBGAMs can be very flexible. They can include both parametric and nonparametric terms. In addition, smooth functions can have more than one input variable. In this way, the joint effects of multiple variables on expected crash frequency can be modeled. The following equation shows an example of more complicated functional forms that can be used in NBGAMs:

$$\mu_i = g(x_i) = \exp\left\{\beta_0 + \sum_{j=1}^{k}\beta_j x_{ij} + f_{k12}\left(x_{i(k+1)}, x_{i(k+2)}\right) + \sum_{j=k+3}^{n} f_j(x_{ij})\right\} \quad (10)$$

where

$\sum_{j=1}^{k}\beta_j x_{ij}$ = parametric term that is commonly used in GLMs,
$f_{k12}(x_{i(k+1)}, x_{i(k+2)})$ = smooth function taking two input variables, and
$\sum_{j=k+3}^{n} f_j(x_{ij})$ = summation of nonparametric smooth functions.

As can be seen from Equations 8 through 10, the main difference between GLMs and GAMs is the smooth function. There are many different types of smooth functions, including kernel smoothers, cubic regression splines, thin-plate regression splines, and *P*-splines (*12*, *13*). In this study, a commonly used cubic regression spline is used as the smooth function.

A simple example of cubic regression splines is presented here to explain the NBGAM used in this study. Interested readers can refer to the work by Wood (*13*) for more details on smooth functions and cubic regression splines. A univariate smooth function is considered with x and y the independent and dependent variables, respectively. A cubic regression spline uses several cubic polynomial curves joined together to represent the relationship between x and y. As shown in Figure 1, there are five curves that are joined at four selected data points. These points as well as the two data points at each end are called knots. A special requirement for these curves is that their values and second derivatives be continuous at the knots. The model for this univariate smooth function can be expressed as follows (*13*):

$$y_i = s(x_i) + \epsilon_i = \sum_{j=1}^{k} b_j(x_i)\alpha_j + \epsilon_i \quad (11)$$

where

$s(\cdot)$ = smooth function, which consists of linear combination of several basis functions;
$b_j(\cdot)$ = jth basis function;
α_j = parameter for jth basis function;
ϵ_i = independently and identically distributed random term with normal distribution; and
k = total number of basis functions.

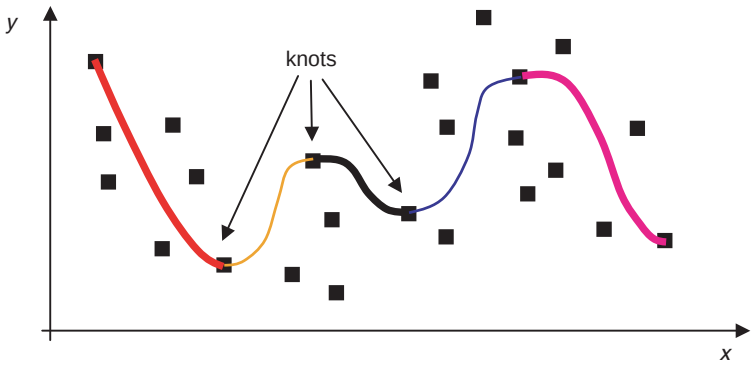

FIGURE 1 Cubic regression spline (*20*).

The choice of basis functions for cubic regression splines is not unique. For instance, the following basis functions can be used (*13*):

$$\begin{cases} b_1(x_i) = 1 \\ b_2(x_i) = x_i \\ b_{j+2}(x_i) = R(x_i, x_j^*), j = 1, \ldots, k-2 \end{cases} \quad (12)$$

where

$$R(x_i, x_j^*) = \frac{\left[\left(x_j^* - \frac{1}{2}\right)^2 - \frac{1}{12}\right]\left[\left(x_i - \frac{1}{2}\right)^2 - \frac{1}{12}\right]}{4 - \left[\left(\left(\left|x_i - x_j^*\right| - \frac{1}{2}\right)^2 - \frac{1}{4}\right)^2 - \frac{1}{30}\right]} \Bigg/ 24$$

and x_j^* = x-value of jth knot. Values of the knots at the two ends (Figure 1) are represented by x_0^* and x_{k-1}^* and are not used in the basis functions.

Once the knots are decided on, the basis functions can be determined. The parameters α_j can thus be estimated by using methods such as least squares, which is to minimize the following objective (*13*):

$$\text{obj} = \sum_{i=1}^{n} (y_i - s(x_i))^2 \quad (13)$$

Choosing the best number of knots can be tricky. Too many knots can make the model overly smoothed and result in overfitting the data. However, too few knots can significantly reduce the model's data-fitting ability. To solve this problem, Wood (*13*) introduced a smoothing parameter gamma (γ) into Equation 13 to form a modified minimization objective function as in the following:

$$\text{obj} = \sum_{i=1}^{n} (y_i - s(x_i))^2 + \gamma \int_{\min(x_i)}^{\max(x_i)} s''(x)dx \quad (14)$$

In general, the larger the gamma value is, the smoother the smooth function will be. Wood also suggested that when $\gamma \approx 1.4$, the overfitting problem can be properly accounted for without significantly affecting the model's goodness of fit (*13*).

The least-squares method introduced in the previous paragraph is for ordinary additive models with a single explanatory variable. For NBGAMs with more than one explanatory variable, a penalized iteratively reweighted least squares (P-IRLS) developed by Wood (*13*) is used. Since the P-IRLS method is fairly complicated, it is not discussed here. Interested readers can find details elsewhere (*13*).

DATA DESCRIPTION

For comparison of the NBGLMs and NBGAMs, a data set collected from 59 three-leg signalized intersections located in Toronto, Canada, over a 6-year period is used. The crash count for each year at each intersection is taken as one observation. There is a total of 354 records in the data set. For each observation i, values of three variables are collected: number of accidents (y_i), entering traffic flow rates from major approaches (x_{i1}), and entering flow rates from minor approaches (x_{i2}). The mean and variance of y_i are 3.7 and 9.9, respectively. Other descriptive statistics of this data set are summarized in Table 1.

The variance of y_i is significantly larger than its mean, suggesting that this data set might be overdispersed. However, a large ratio of variance to mean of y_i does not necessarily mean that there is overdispersion in the data, since this large ratio can be caused by large variations in x_{i1} and x_{i2}. Thus, later in this paper, statistical tests are performed to further confirm the existence of overdispersion. The test results show that there is overdispersion and application of NBGLMs and NBGAMs is appropriate. To give a better idea of this data set, its crash frequency distribution is plotted in Figure 2, which shows that for 43 observations, no crashes occurred. It should be noted that since the main purpose of this study is to compare GAMs with GLMs, different years' data for the same intersection are treated independently in this study, even though there might be some space–time correlations among different observations.

TABLE 1 Descriptive Statistics of Toronto Crash Data

	y_i	x_{i1} (vehicles per day)	x_{i2} (vehicles per day)
Mean	3.7	26,055	1,879
Minimum	0.0	5,669	378
Maximum	14.0	53,531	8,638
Variance	9.9	117,682,022	2,619,775
Std. Dev.	3.1	10,848	1,619

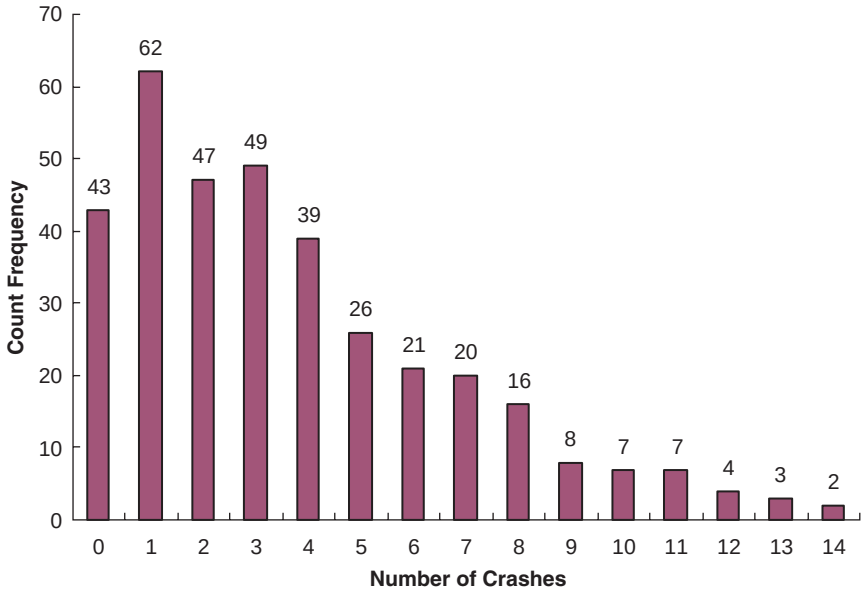

FIGURE 2 Crash frequency distribution.

TEST DESIGN AND MODEL IMPLEMENTATION

Test Design

To compare the two types of regression models, the data set is randomly separated into two parts. The first part consists of 200 records and is used for model fitting. The second part has 154 records and is used for model testing. To make the comparison results more convincing, four scenarios are considered. The random data separation process is repeated for each scenario, which generates a set of fitting and testing data sets while keeping the number of data records for training and testing unchanged. Thus, in total there are four fitting data sets and four testing data sets. The NBGLM and NBGAM are fitted and tested on these four sets of data.

Functional Forms

Both NBGLMs and NBGAMs can have many different functional forms. The following two functional forms are often used for NBGLMs and are adopted in this study:

$$\mu_i = g(x_i) = \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\} \quad (15)$$

$$\mu_i = g(x_i) = \beta_0 x_{i1}^{\beta_1} x_{i2}^{\beta_2} = \exp\{\ln(\beta_0) + \beta_1 \ln(x_{i1}) + \beta_2 \ln(x_{i2})\} = \exp\{\theta_0 + \theta_1 x'_{i1} + \theta_2 x'_{i2}\} \quad (16)$$

where

$\theta_0 = \ln(\beta_0)$,
$\theta_1 = \beta_1$,
$\theta_2 = \beta_2$,
$x'_{i1} = \ln(x_{i1})$, and
$x'_{i2} = \ln(x_{i2})$.

It should be noted that Equation 16 is essentially a nonlinear function. However, by rearranging its right-hand side and replacing the logarithmic terms with two new variables and a new coefficient, it can effectively be modeled as a GLM. Equation 16 was recommended for intersection crash frequency analysis by Miaou and Lord (*16*) to fit a simple nonlinear logarithmic relationship between $\eta_i = \ln(\mu)$ and x_{ij}.

By taking the logarithm on both sides of Equation 15, an equation similar to Equation 1 can be obtained. Applying the same operations to Equation 16, the following equation is obtained:

$$\eta_i = \ln(\mu_i) = \ln(\beta_0) + \beta_1 \ln(x_{i1}) + \beta_2 \ln(x_{i2}) \quad (17)$$

It is obvious that the first functional form (Equation 15) assumes a linear relationship between η_i and x_{ij}, whereas the second functional form (Equations 16 and 17) considers a simple logarithmic relationship on the right-hand side. For ease of description, the two NBGLMs based on Equations 15 and 16 are hereinafter referred to as NBGLM I and NBGLM II, respectively.

For the NBGAM in this study, the following functional form is used:

$$\mu_i = g(x_i) = \exp\{\beta_0 + f_1(x_{i1}) + f_2(x_{i2})\} \quad (18)$$

This functional form is the result of replacing the two parametric terms in Equations 15 and 16 with two corresponding smooth functions. Although more complicated functional forms (as shown in Equation 10) can be formulated, this simple one was chosen to investigate how a simple substitution of smooth functions for explanatory variables can affect crash frequency models' fitting and predicting performance.

Performance Indexes

The NBGLMs and the NBGAM are compared in terms of fitting and testing performance and a goodness-of-fit test statistic. For fitting and testing performance comparison, mean absolute error (MAE) and mean squared error (MSE) are used. For each scenario, these two performance indexes are calculated as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{19}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \tag{20}$$

where n is the fitting or testing data size, and y_i and $\hat{y}_i$ are the observed and predicted numbers of accidents for observation i, respectively.

In addition to the MAE and MSE, the AIC, a goodness-of-fit test statistic (*17*), is also used to compare the performance between NBGLMs and the NBGAM. In general, a lower AIC value means better model fit (*17*).

Model Implementation

Both the NBGLMs and the NBGAM are fitted using R software (*18*). More specifically, the NBGLMs are fitted by using the MASS package (*19*), and the NBGAM is fitted using the MGCV package (*20*). For all scenarios, the smoothing parameter γ is set to be 1.4 for the NBGAM.

RESULTS ANALYSIS

Test of Overdispersion

Although the values of the mean and variance of the original data set suggest that there might be overdispersion in the data, a statistical confirmation of overdispersion is still meaningful to substantiate the validity of using NB instead of Poisson models for the same functional form. In this research, a Lagrange multiplier (LM) (*21*) test is performed to test the existence of overdispersion in the four generated data sets. The LM test statistic is defined as follows:

$$\text{LM} = \left[\frac{\sum_{i=1}^{n}\left[\left(y_i - \mu_i\right)^2 - y_i\right]}{2\sum_{i=1}^{n}\mu_i^2}\right]^2 \tag{21}$$

TABLE 2 Results of Overdispersion Test Using LM Statistics

Scenario	Functional Form 1 [Equation (15)]	Functional Form 2 [Equation (16)]
1	134.20	76.60
2	142.80	101.64
3	168.07	89.71
4	152.70	120.77

where

y_i = reported number of crashes for observation i,
μ_i = expected number of accidents for observation i from Poisson GLMs, and
n = number of observations.

The null hypothesis for this test statistic is that the Poisson GLM is appropriate. Under this null hypothesis, the LM test statistic has an asymptotically chi-square distribution with one degree of freedom.

To calculate the LM test statistic, two Poisson GLMs using the functional forms in Equations 15 and 16 are estimated for each set of data. In total eight Poisson GLMs are estimated and eight LMs are calculated. The calculated LM test statistic values are given in Table 2. All LM values in Table 2 are much greater than $\chi^2_{1,0.05} \approx 3.84$, meaning that the null hypotheses for all cases should be rejected at the 0.05 level of significance. The LM test result suggests that there is overdispersion in all four generated data sets, and it would be better to use an NB model instead of a Poisson model.

Comparison of AICs

Table 3 shows the comparison results of the NBGLMs and the NBGAM. It is obvious that for all four scenarios, the NBGAM gives the lowest AIC values. Since lower AIC values suggest better model fit (*17*), this result indicates that the NBGAM fits the four training data sets consistently better than do the two NBGLMs. Another interesting finding is that the NBGLM II produces lower AIC values than the NBGLM I does, suggesting that the relationships between η_i and x_{ij}'s are more likely to be nonlinear than linear.

TABLE 3 AIC and Performance Index Values

Scenario	Performance Index	NBGLM I		NBGLM II		NBGAM	
		Fitting	Predicting	Fitting	Predicting	Fitting	Predicting
1	MAE	2.28	2.49	2.10	2.45	1.91	2.30
	MSE	7.90	10.04	6.94	9.48	5.75	8.55
	AIC	920.15		892.19		878.78	
2	MAE	2.34	2.42	2.26	2.27	1.91	2.23
	MSE	8.48	9.16	7.84	8.20	5.81	8.06
	AIC	940.93		921.05		891.68	
3	MAE	2.39	2.41	2.20	2.37	1.96	2.21
	MSE	9.03	8.63	7.76	8.66	6.45	7.53
	AIC	944.38		911.26		889.89	
4	MAE	2.36	2.38	2.28	2.25	2.08	2.07
	MSE	8.30	9.57	7.85	8.46	6.66	7.42
	AIC	930.53		916.83		909.48	

Comparison of Performance Indexes

A further comparison of MAE and MSE values, also in Table 3, shows that for all scenarios the NBGAM outperforms the two NBGLMs as well. Although the NBGLM II underperforms the NBGAM, it is again considered to be better than the NBGLM I in terms of almost all MAE and MSE values. This finding also indicates that nonlinear relationships between η_i and x_{ij}'s might be more appropriate, and the NBGAM has better nonlinear approximation ability than the NBGLM II. Because the four sets of fitting and testing data are randomly divided from the original data set, the consistently better performance from the NBGAM is a strong indication of the NBGAM's superiority.

Fitted Curves

One output of the NBGAM is the fitted smooth function curve. Figure 3 shows the curves of the two smooth functions in Equation 18 using Scenario 1 data as an example. The fitted relationships between $f_j(x_{ij})$ and x_{ij}'s are plotted as the solid lines, and the dashed lines are 95% confidence limits. Since

$$\eta_i = \ln(\mu_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) \quad (22)$$

the two fitted smooth function curves also reflect the effects of each flow variable on the logarithm of the expected crash frequency $\ln(\mu_i)$.

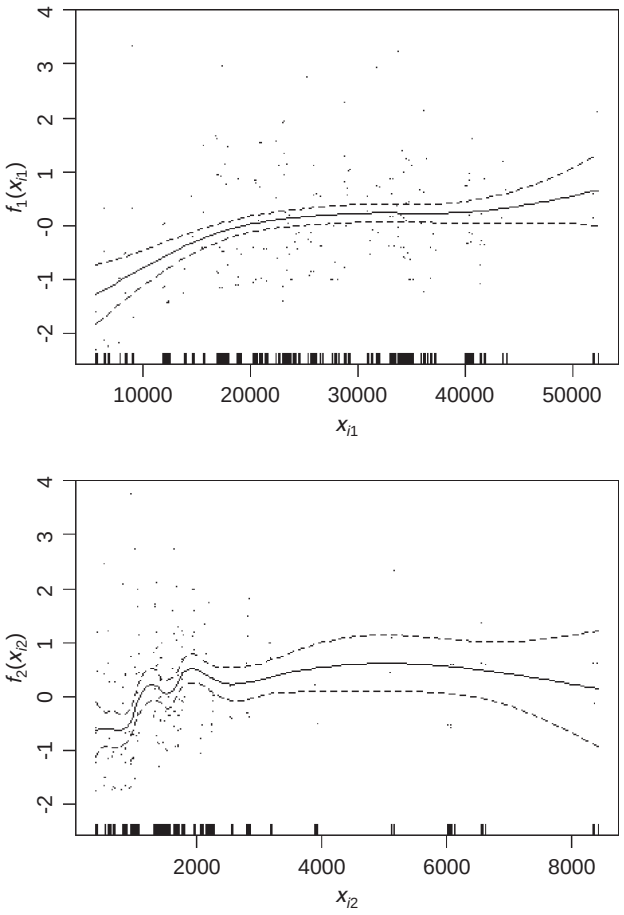

FIGURE 3 Fitted curves based on NBGAM.

Because of the additive nature of Equation 22, the effect of x_{i1} on $\ln(\mu_i)$ is independent of the effect of x_{i2} on $\ln(\mu_i)$.

The fitted curves in Figure 3 suggest that the relationships between $\ln(\mu_i)$ and x_{ij}'s are nonlinear and have logarithmic trends in general. This conclusion is supported to a certain degree by the fact that the NBGLM II performs better than the NBGLM I, since the NBGLM II assumes logarithmic relationships between $\ln(\mu_i)$ and x_{ij}'s whereas the NBGLM I assumes linear ones. Although the NBGLM II can adequately capture the logarithmic trend, it is not flexible enough since it is unable to capture other types of nonlinear relationships or any deviations from a strictly logarithmic trend. This shortcoming may be the reason that NBGAM II's fitting and testing performance and AIC are not as good as those of the NBGAM.

CONCLUSIONS

In this study, a new type of generalized regression model, the GAM, is applied to crash frequency analysis for the first time. Compared with GLMs, GAMs introduce smooth functions into the functional forms. The introduction of nonparametric smooth functions theoretically provides GAMs with better nonlinear approximation ability than GLMs.

To confirm the advantages of using GAMs over GLMs for crash frequency modeling, both model categories are tested and compared on the basis of an intersection crash frequency data set collected in Toronto, Canada. Since the data appear to be overdispersed, two NBGLMs and an NBGAM are fitted. Statistical tests are also performed to show that NB models are more appropriate than Poisson models for this data set.

AIC values are used to compare the goodness of fit of these three models. In addition, MAE and MSE are used to compare their fitting and predicting performance. The comparison results consistently show that the NBGAM performs the best in terms of all three performance indexes. Also, judging from all three indexes, the NBGLM II with a nonlinear functional form noticeably outperforms the NBGLM I with a linear functional form. The curves of the fitted smooth functions in the NBGAM are plotted and analyzed, and the curves clearly suggest nonlinear trends. The analysis shows that GAMs can be very useful in modeling complicated nonlinear relationships.

Overall, GAMs are excellent methods that bridge the gap between neural networks or support vector machines and GLMs. Similar to GLMs, GAMs can generate statistically interpretable results. Like neural networks and support vector machines, GAMs are very flexible and have powerful nonlinear modeling ability.

This study is limited to the investigation of the application of GAMs to crash frequency studies. In the future, it would be interesting to further explore their potential in other traffic safety research areas such as crash injury severity studies.

ACKNOWLEDGMENTS

Part of this work was done when the first author was at Texas A&M University. The authors thank Dominique Lord for providing the data used in this study.

REFERENCES

1. Xie, Y., Y. Zhang, and F. Liang. Crash Injury Severity Analysis Using a Bayesian Ordered Probit Model. Presented at 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2007.

2. Miaou, S. P., and H. Lum. Modeling Vehicle Accidents and Highway Geometric Design Relationships. *Accident Analysis & Prevention*, Vol. 25, No. 6, 1993, pp. 689–709.
3. Lord, D., S. P. Washington, and J. N. Ivan. Poisson, Poisson-Gamma, and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. *Accident Analysis & Prevention*, Vol. 37, No. 1, 2005, pp. 35–46.
4. Milton, J., and F. Mannering. The Relationship Among Highway Geometrics, Traffic-Related Elements and Motor-Vehicle Accident Frequencies. *Transportation*, Vol. 25, No. 4, 1998, pp. 395–413.
5. Miaou, S. P. The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson Versus Negative Binomial Regressions. *Accident Analysis & Prevention*, Vol. 26, No. 4, 1994, pp. 471–482.
6. Oh, J., S. P. Washington, and D. Nam. Accident Prediction Model for Railway-Highway Interfaces. *Accident Analysis & Prevention*, Vol. 38, No, 2, 2006, pp. 346–356.
7. Shankar, V., J. Milton, and F. Mannering. Modeling Accident Frequencies as Zero-Altered Probability Processes: An Empirical Inquiry. *Accident Analysis & Prevention*, Vol. 29, No. 6, 1997, pp. 829–837.
8. Qin, X., J. N. Ivan, N. Ravishanker, and J. Liu. Hierarchical Bayesian Estimation of Safety Performance Functions for Two-Lane Highways Using Markov Chain Monte Carlo Modeling. *Journal of Transportation Engineering*, ASCE, Vol. 131, No. 5, 2005, pp. 345–351.
9. Abdelwahab, H. T., and M. A. Abdel-Aty. Artificial Neural Networks and Logit Models for Traffic Safety Analysis of Toll Plazas. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1784*, Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 115–125.
10. Xie, Y., D. Lord, and Y. Zhang. Predicting Motor Vehicle Collisions Using Bayesian Neural Network Models: An Empirical Analysis. *Accident Analysis & Prevention*, Vol. 39, No. 5, 2007, pp. 922–933.
11. Li, X., D. Lord, Y. Zhang, and Y. Xie. Predicting Motor Vehicle Crashes Using Support Vector Machine Models. Presented at 87th Annual Meeting of the Transportation Research Board, Washington, D.C., 2008.
12. Hastie, T. J., and R. J. Tibshirani. *Generalized Additive Models.* Chapman and Hall, New York, 1990.
13. Wood, S. N. *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC, Boca Raton, Fla., 2006.
14. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In *Proceedings of the Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki, eds.), Akademiai Kiado, Budapest, Hungary, 1973, pp. 267–281.
15. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, 1974, pp. 716–723.
16. Miaou, S.-P., and D. Lord. Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes Methods. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840*, Transportation Research Board of the National Academies, Washington D.C., 2003, pp. 31–40.
17. Liu, W., and J. Cela. Improving Credit Scoring by Generalized Additive Model. Presented at SAS Global Forum, Orlando, Fla., 2007. http://www2.sas.com/proceedings/forum2007/078-2007.pdf. Accessed July 3, 2007.
18. *The R Project for Statistical Computing (Version 2.5.1).* Vienna University of Economics and Business Administration. http://www.r-project.org/. Accessed July 3, 2007.
19. Ripley, B. *The VR Package (Version 7.2-34).* http://cran.r-project.org/doc/packages/VR.pdf. Accessed July 3, 2007.
20. Wood, S. *The mgcv Package (Version 1.3-25).* http://cran.r-project.org/doc/packages/mgcv.pdf. Accessed July 3, 2007.
21. Greene, W. H. *Econometric Analysis*, 4th ed. Prentice Hall, Upper Saddle River, N.J., 2000.

The Statistical Methodology and Statistical Computer Software in Transportation Research Committee sponsored publication of this paper.