

ITS IMPLEMENTATION RESEARCH CENTER

Speed/Headway Influence on Crashes

Byungkyu “Brian” Park

Hojun “Daniel” Son

Young-Jun Kweon

Dr. Byungkyu “Brain” Park
Email: bpark@virginia.edu

Ilsoo Yun
Email: hjs4x@virginia.edu

Dr. Young-Jun Kweon
Young-Jun.Kweon@VDOT.Virginia.gov



Research Report No. UVACTS-15-0-70
June 2008

Center for Transportation Studies at the University of Virginia produces outstanding transportation professionals, innovative research results and provides important public service. The Center for Transportation Studies is committed to academic excellence, multi-disciplinary research and to developing state-of-the-art facilities. Through a partnership with the Virginia Department of Transportation's (VDOT) Research Council (VTRC), CTS faculty hold joint appointments, VTRC research scientists teach specialized courses, and graduate student work is supported through a Graduate Research Assistantship Program. CTS receives substantial financial support from two federal University Transportation Center Grants: the Mid-Atlantic Universities Transportation Center (MAUTC), and through the National ITS Implementation Research Center (ITS Center). Other related research activities of the faculty include funding through FHWA, NSF, US Department of Transportation, VDOT, other governmental agencies and private companies.

Disclaimer: The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

CTS Website

<http://cts.virginia.edu>

Center for Transportation Studies

University of Virginia

351 McCormick Road, P.O. Box 400742

1. Report No. VACTS-15-0-70	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Development of Crash Prediction Models using Real Time Safety Surrogate Measures		5. Report Date June 2008
		6. Performing Organization Code
7. Author(s) Hojun "Daniel" Son, Young-Jun Kweon and Byungkyu "Brian" Park		8. Performing Organization Report No.
9. Performing Organization and Address Center for Transportation Studies University of Virginia PO Box 400742 Charlottesville, VA 22904-7472		10. Work Unit No. (TRAIS)
		11. Contract or Grant No.
12. Sponsoring Agencies' Name and Address Office of University Programs, Research and Special Programs Administration US Department of Transportation 400 Seventh Street, SW Washington DC 20590-0001		13. Type of Report and Period Covered Final Report
		14. Sponsoring Agency Code
15. Supplementary Notes		
16. Abstract <p>Typical engineering research on traffic safety focuses on identifying either dangerous locations or contributing factors through a post-crash analysis using aggregated traffic flow data and crash records. A recent development of transportation engineering technologies provides ample opportunities to enhance freeway traffic safety using individual vehicular information. However, methodologies on how to utilize and link such technologies to traffic safety analysis have not been thoroughly explored. Moreover, traffic safety research has not benefited from the use of hurdle-type models that treat excessive zeros in crash modeling.</p> <p>This study developed a new crash risk predictor, safe headway distance, to estimate traffic crash likelihood by using individual vehicular information and applying it to basic sections of interstates in Virginia. Individual vehicular data and crash data were used in the development of statistical crash prediction models including hurdle models. The results showed that safe highway distance measure was effective in predicting traffic crash occurrence and the hurdle negative binomial model outperformed other count data models including zero-inflated models.</p>		
17 Key Words Crash prediction; Safety surrogate measure, Safety headway distance; Count model; Hurdle model; Highway safety		18. Distribution Statement No restrictions. This document is available to the public.

ABSTRACT

Typical engineering research on traffic safety focuses on identifying either dangerous locations or contributing factors through a post-crash analysis using aggregated traffic flow data and crash records. A recent development of transportation engineering technologies provides ample opportunities to enhance freeway traffic safety using individual vehicular information. However, methodologies on how to utilize and link such technologies to traffic safety analysis have not been thoroughly explored. Moreover, traffic safety research has not benefited from the use of hurdle-type models that treat excessive zeros in crash modeling.

This study developed a new crash risk predictor, safe headway distance, to estimate traffic crash likelihood by using individual vehicular information and applying it to basic sections of interstates in Virginia. Individual vehicular data and crash data were used in the development of statistical crash prediction models including hurdle models. The results showed that safe highway distance measure was effective in predicting traffic crash occurrence and the hurdle negative binomial model outperformed other count data models including zero-inflated models.

INTRODUCTION

A fundamental goal of surface transportation agencies is to provide safe and reliable transportation services to road users. Being supported by legislation such as SAFETEA-LU (Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy Users) (1), U.S. transportation agencies have been making a variety of efforts in improving safety and mobility. Much effort in utilizing advanced transportation technologies such as Intelligent Transportation Systems (ITS) has been expended mainly for mobility improvement (2), whereas utilizing these technologies for safety improvement has not been thoroughly explored. Further, in recent years, safety seems to be receiving more attention than mobility as the former costs significantly more than the latter. One study showed that the total cost of traffic crashes is almost two and half times the cost of traffic congestion (3).

As an essential component of typical traffic safety studies, considerable effort was expended to identify traffic patterns that trigger crashes based on crash records and aggregated traffic flow data. For example, Oh et al. attempted to relate aggregated loop detector data to crash data using a Bayesian classification and non-parametric density function (4). They assumed that disruptive traffic flows trigger crashes and found that a 5-min standard deviation of speeds right before a crash was the best crash indicator. Golob et al. conducted an analysis with data collected on highways in southern California and concluded that median speed, speed variation, mean flow, and flow variation were the best crash indicators based on a statistical analysis of 30-sec-interval traffic flow data collected 30 min before the crashes (5). With the same dataset, Kockelman and Ma

employed binomial regression models and found that 30-sec average speed and 5-min speed variation were highly related to crash occurrence (6).

Abdel-Aty et al. considered spatial variation of traffic patterns for crash prediction by exploring data obtained from detectors located upstream and downstream of a crash location (7). They developed crash prediction models using case-control logistic regression and suggested 5-min average occupancy upstream and 5-min coefficient of variation in speed downstream of the crash location 5 to 10 min before crash occurrence as crash indicators. Lee et al. conducted a similar study using a log-linear model and concluded that coefficient of variation in speed, spatial variance of speed, and covariance of volume difference upstream and downstream of crash locations were important crash indicators (8).

As stated, the analysis methods mostly employed in the safety studies were based on site-specific crash records and aggregated traffic flow data. Given that a crash occurs through individual vehicular interactions, aggregated data often do not adequately capture crash characteristics. Individual vehicular data could significantly enhance crash modeling. For example, a large speed variance among vehicles likely indicates a high crash potential; it might not be if the time headways among the adjacent vehicles were large enough to lead to crash avoidance. Thus, the use of individual vehicular information could enhance crash prediction capability. However, few research efforts have attempted to use individual vehicular data in the development of crash prediction models mostly because of limitations in obtaining and using individual vehicular data.

As advanced communication technologies for vehicle to vehicle and vehicle to infrastructure through dedicated short range communication, IEEE 802.11p, etc., become

available for deployment, individual vehicular information is going to be available in the near future. It is expected that individual vehicular information can provide a unique opportunity to improve safety as well as mobility. Because of the potential benefits, an innovative concept of integrating information exchanges between vehicles and infrastructure, known as Vehicle Infrastructure Integration (VII), has been proposed by federal and state transportation agencies and vehicle manufacturers (9). In spite of a surge in recent attention with regard to the VII applications to traffic safety, few research efforts have been made to enhance the safety as well as the mobility of the surface transportation system. For example, using individual vehicular information such as vehicular speeds and headways can provide accurate information that can be used to understand crash risk in real time, but little is known how to link such information to traffic crash occurrence.

Park and Yadlapati studied simulation-based work-zone safety (10). They examined individual vehicular interactions within work zones in simulation and proposed a crash risk predictor. Oh et al. proposed a freeway safety index based on real-time individual vehicular data (11). However, their index was not empirically validated. Moreover, they simplified a continuous crash risk into a binary outcome (crash and non-crash), which caused loss of information regarding crash potential magnitude.

In an attempt to develop statistical models predicting crashes on highway, various count data models have been applied. Zero-inflated count models have been introduced to deal with an excess of zero counts in crash prediction modeling. They assume that zero crash counts are generated from two processes distinguishing sampling zeros from structural zero. Given that no roadway segments are absolutely safe from crashes, these

models with structural zero have an unrealistic assumption. Unlike zero-inflated models, zero-hurdle models are capable of addressing excessive zero counts without such an assumption (12). One popular example that describes the conceptual difference between zero-inflated and zero-hurdle models is the case with the following question: “How many fish did you get?” Zero responses would be either those who never try to get fish or those who did but did not get fish. In the case where those who never tried were included in the study, zero-inflated models are conceptually appropriate. However, if all zero responses are from those who tried but did not get fish, zero-hurdle models are appropriate. In spite of the conceptual relevance to crash data, hurdle count models have not been applied in traffic safety analysis.

To this end, this paper presents an approach to predict crash occurrence using individual vehicular information. Crash prediction models including zero-inflated hurdle models were developed on the basis of crash data and individual vehicular information collected from inductive loop detectors on urban interstate highways in Virginia.

METHODOLOGY

A new crash risk predictor, called safe headway distance (SHD), is proposed to calculate the crash potential between two consecutive vehicles. An aggregated SHD is computed by summing only negative SHD values, which indicate potential crash risks, over at a given time interval and a roadway segment is used as an input to a crash prediction model. Several count data models were employed

Safe Headway Distance Equation

A traffic conflict can be defined in several ways depending on the research purpose and design (10). This paper defines the conflict as a condition of two consecutively moving vehicles having inadequate SHD such that the following vehicle will crash into the leading vehicle when it makes an unexpected stop. It assumes that the leading vehicle abruptly reacts to a stimulus resulting in an emergency stopping maneuver while the following vehicle reacts to such a braking maneuver by the leading vehicle. Based on these assumptions, a safe (i.e., non-collision) condition of the two consecutive vehicles can be defined as a condition where a minimum stopping distance of the leading vehicle is greater than that of the following vehicle. Hence, using the minimum stopping distance formula (13), the safe condition can be mathematically expressed as:

Minimum Stopping Distance (leading vehicle) > Minimum Stopping Distance (following vehicle) \equiv

$$1.47 \times (V_{Leading} \times h) + \left[\frac{V_{Leading}^2}{30 \times \left(\frac{acc}{g} \pm Gr \right)} \right] > 1.47 \times (V_{Following} \times PRT) + \left[\frac{V_{Following}^2}{30 \times \left(\frac{acc}{g} \pm Gr \right)} \right]$$

(Eq. 1)

where $V_{Leading}$ = leading vehicle's speed in miles per hour (mph)

$V_{Following}$ = following vehicle's speed in mph

acc = deceleration rate in ft/sec²

g = gravity acceleration (32.2 ft/sec²)

Gr = grade in percentage

h = time headway in seconds

PRT = perception reaction time of the following vehicle in seconds.

Using Equation 1, an individual SHD for a pair of two consecutive vehicles can be defined as a difference of the two minimum stopping distances of vehicles as follows:

$SHD_i = \max[-Diff_i, 0]$ and

$$Diff_i = 1.47 \times (V_{Leading} \times h - V_{Following} \times PRT) + \left[\frac{V_{Leading}^2 - V_{Following}^2}{30 \times \left(\frac{acc}{g} \pm Gr \right)} \right] \quad (\text{Eq. 2})$$

where i = index of a pair of two consecutive vehicles

SHD_i = safe headway distance of the i th vehicle pair

$Diff_i$ = difference of two minimum stopping distances of the i th vehicle pair.

As noted, the individual SHD reflects a crash risk between the two consecutive vehicles. A larger SHD is translated into a greater deficiency in safe distance suggesting a crash-prone case.

When individual vehicular data become available on a real-time basis through advanced technologies such as the VII system, the SHD can be computed in real time and linked to actual crash occurrences. However, under the current data collection system, vehicular and crash data are not available in real time. Thus, to relate the SHD to crash

occurrences under the current system, the use of historical data, especially for crash data, is inevitable. Moreover, aggregation of such data over a certain time period is convenient to formulate crash prediction models empirically.

In addition to temporal aggregation, the aggregation needs to be done for a roadway segment. For developing a crash prediction model relating crashes to the proposed SHD, crash data were aggregated over a short stretch of highways to produce the number of crashes suitable for statistical crash modeling. To be consistent with the aggregated crash data, the SHD were also aggregated along the same road segment over the same time period.

The segment length for aggregation can vary from 100 ft to several miles, depending on the objective of an analysis. For example, if the analysis aims at a safety evaluation of a region, the individual SHD can be aggregated over an entire network within the region. If the analysis aims at a safety evaluation of a corridor, the SHD can be aggregated along the corridor. As such a 1-mile segment length, which is typical in traffic safety engineering studies for road segments, was used for this paper.

For crash prediction models to be estimated effectively using 1-mile-segment data, crash data need to be accumulated over a sufficient time period, 6 years in this paper. As the number of crashes over the 6 years can be divided by a time interval (e.g., 1 hour) within a day and potentially affect the crash predictions, this paper considers three time intervals: 15, 30, and 60 min. As a result, three datasets for crash prediction models were prepared. For example, the number of crashes that occurred on a particular 1-mile segment can be aggregated by each hour of the day, and the hourly crashes can be

accumulated over the entire crash records. This produces an aggregated dataset with 24 one-hour crash counts.

Obviously, the SHD needs to be aggregated by each hour of the day so that the crash prediction model can be developed on the basis of hourly crash counts and SHD values. The SHD was aggregated by the three same intervals, 15, 30, and 60 minutes over the 1-mile segment, and produced an aggregated SHD for each time interval. To predict the number of crashes for a 1-mile road segment during a certain time interval, the aggregated SHD was used as a crash risk predictor. These three time intervals were compared on the basis of the results of the crash prediction models.

Count Data Regression Models

To predict the number of crashes for a roadway segment during a specific time interval, the aggregated SHD was used as a crash predictor. It is noted that the count data models are commonly applied for crash prediction for the following reasons. The number of crashes is a non-negative integer and typically skewed in its distribution. Moreover, it typically shows an increasing variance in a relationship with predictors, which is heteroscedastic. These characteristics are apt to result in inefficient and biased parameter estimates when a classical linear regression is applied (12). Count data models such as the Poisson (P) and negative binomial (NB) models are designed specifically to account for such characteristics.

There exist several count models. The P model, a standard count data model, assumes that the mean and the variance of a dependent variable (e.g., crash counts) are equal, called equidispersion. However, crash counts often show overdispersion (i.e.,

variance is larger than mean). In such a case, an NB model is a usual alternative, and it explicitly relaxes the equidispersion assumption by adding a random term into the P model. In some cases, excessive zeros in crash counts contribute to an overdispersion. Zero-inflated and zero-hurdle count models are specifically designed to handle excessive zero crash counts.

The zero-inflated and zero-hurdle models look similar in that they integrate a binary response model and a count data model into a single modeling frame. However, a significant conceptual difference exists between these two in terms of how zeros were interpreted (14). The zero-inflated model assumes that there are two types of zeros: structural zeros and sampling zeros. In contrast, the zero-hurdle model assumes that all zeros are sampling zeros. A selection between these two seemingly similar models can be made to some extent based on a study design.

Rose et al. stated that a study design and purpose should be considered in selecting appropriate models handling excessive zeros (14). For an example of traffic safety, suppose a crash prediction model is being developed using crash data collected from 1-mile roadway segments over 1 year. Because of the short segment length and the relatively short time period, there likely exists a large number of segments with no crashes during the year, implying a preponderance of zero counts. By applying the zero-inflated models to these data, it is assumed that there exist two kinds of road segments, absolutely safe (i.e., crash-free) segments and crash potential (i.e., at-risk) segments. In contrast, by applying the zero-hurdle models, it is assumed that crashes can occur at any segment, implying all segments have crash potential. As such, it is determined that the zero-hurdle models assuming all road segments have crash potential are deemed to be

more appropriate than the zero-inflated models. Lord et al. concluded that zero crash data are not characterized by the two regimes noted in the zero-inflated models (15).

Six count data models including the standard P, NB, zero-inflated P (ZIP), zero-inflated NB (ZINB), zero-hurdle P (ZHP), and zero-hurdle NB (ZHNB) models were initially considered. The probability distribution functions of these models are presented in Table 1. A binary logit model is used to incorporate a probability of structured zeros in the ZIP and ZINB models and in the ZHP and ZHNB models.

Model Selection

Overdispersion of the crash counts and a percentage of zero counts were examined in selecting candidate models among the six models presented earlier. Three tests, the likelihood ratio test, the dispersion parameters based on deviance, and the regression-based test, were used for testing overdispersion. Since overdispersion was found by these tests, the P model was excluded from consideration because of a violation of its equidispersion assumption. Since there were many zeros in crash counts, zero-inflated and hurdle models were to be considered as candidate models. Once candidate models were determined, they were estimated using the maximum likelihood estimation (MLE) technique.

TABLE 1 Probability Functions of Count Models

Model	$P(y_{ij} X_{ij}, Z_{ij})$	
	For $y_{ij} = 0$	For $y_{ij} > 0$
P	$e^{-\mu_{ij}}$	$\frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!}$
NB	$\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_{ij}}\right)^{\alpha^{-1}}$	$\frac{\Gamma(y_{ij} + \alpha^{-1}) \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_{ij}}\right)^{\alpha^{-1}} \left(\frac{\mu_{ij}}{\alpha^{-1} + \mu_{ij}}\right)^{y_{ij}}}{y_{ij}! \Gamma(\alpha^{-1})}$
ZIP	$p_{ij} + (1 - p_{ij})e^{-\mu_{ij}}$	$(1 - p_{ij}) \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!}$
ZINB	$p_{ij} + (1 - p_{ij}) \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_{ij}}\right)^{\alpha^{-1}}$	$(1 - p_{ij}) \frac{\Gamma(y_{ij} + \alpha^{-1}) \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_{ij}}\right)^{\alpha^{-1}} \left(\frac{\mu_{ij}}{\alpha^{-1} + \mu_{ij}}\right)^{y_{ij}}}{y_{ij}! \Gamma(\alpha^{-1})}$
ZHP	p_{ij}	$(1 - p_{ij}) \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{(1 - e^{-\mu_{ij}}) y_{ij}!}$
ZHNB	p_{ij}	$(1 - p_{ij}) \frac{\Gamma(y_{ij} + \alpha^{-1}) \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_{ij}}\right)^{\alpha^{-1}} \left(\frac{\mu_{ij}}{\alpha^{-1} + \mu_{ij}}\right)^{y_{ij}}}{y_{ij}! \Gamma(\alpha^{-1}) \left(1 - (1 + \alpha \mu_{ij})^{-\alpha^{-1}}\right)}$

where y_{ij} = crash count at segment i in time j
 X_{ij} = vector of explanatory variables of the P and NB models
 Z_{ij} = vector of explanatory variables of the binary logit model
 $\mu_{ij} = \exp^{(X_{ij}\beta)}$
 $p_{ij} = \frac{\exp^{(Z_{ij}\delta)}}{1 + \exp^{(Z_{ij}\delta)}}$, logit probability of being a structured zero for the ZIP and ZINB models
and being a zero for the ZHP and ZHNB models
 α = dispersion parameter of the NB model in the ZINB and ZHNB models
 β = vector of coefficient parameters of the P and NB models
 δ = vector of coefficient parameters of the binary logit model.

To determine the best-fitted model among the candidate models, two MLE-based information criteria were used: the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) as shown in Eqs. 3 and 4, respectively (12).

$$AIC = -2 \ln L + k \quad (\text{Eq. 3})$$

$$BIC = -2 \ln L + (\ln n)k \quad (\text{Eq. 4})$$

where L = a log-likelihood value of a model

n = the number of observations

k = the number of parameters.

In practice, a model with the lowest AIC and BIC values is preferred. BIC rule-of-thumb criteria commonly used for model comparison were used for this study. For example, when the difference in BIC values of two models is greater than 10, a model with a lower BIC value is *very strongly* preferred (16).

In addition to the AIC and BIC values, comparisons between predicted and observed crash counts (or probabilities) were considered in the model selection. Plotting the difference between the mean predicted probabilities from estimated models and the observed proportions of each crash count can be helpful to compare the performance of the model prediction visually (see Figure 1). The mean probabilities were computed using Eq. 5 (17). Being close to zero in the probability difference indicates that the model fits the data well.

$$\bar{P}(y = m) = \frac{1}{N} \sum_{i=1}^N \hat{P}(y_i = m | X_i) \quad (\text{Eq. 5})$$

where m = non-negative integer number (e.g., crash count; 0, 1, 2, ...)

N = number of observations.

Data Collection

Data were collected from two urban interstates (I-64 and I-264) in Hampton Roads, Virginia. Two 1-mile-long eastbound segments, hereinafter called W64-05EB and E264-03EB, were selected from each highway; I-64 and I-264 segments have three and four lanes, respectively. Because the segments are straight basic highway sections (i.e., no horizontal/vertical curves and no merging/diverging points within or near the segments), impacts of geometric design factors were minimal. Individual vehicular data were collected from loop detector stations placed on the segments using a special traffic counter. Crash data for these segments were extracted from the Virginia Department of Transportation (VDOT) crash database and the Smart Travel Lab (STL) at the University of Virginia. These data are described here.

Individual Vehicular Data

Individual vehicular data (i.e., speed and time headway) were collected using PEEK's ADR-3000 automatic traffic counters connected to inductive loop detector stations under normal weather conditions on two weekdays at W64-05EB and three weekdays at E264-03EB between March and April 2005. The speed and time headway data of all vehicles passing over the loop detectors per each lane were collected for at least one 24-hour time period. To minimize the impact of suspicious vehicular data attributable to

malfunctioning of traffic counter and/or communication errors, the individual vehicular data were filtered. Collected individual vehicular speeds and headways were used to compute an individual SHD, and then each SHD was aggregated into 15-, 30-, and 60-min intervals. This resulted in an aggregated SHD by each time interval. An average value of SHD by each time interval was used as an aggregate SHD for crash prediction models.

Crash Data

Two sources of crash data were available for the study sites, the VDOT crash data based on the police crash report and the STL crash data based on CCTV cameras and the incident management system maintained by the Center for Transportation Studies at the University of Virginia. The two crash datasets were similar in nature. The VDOT crash data have reliable information on collision types and major reasons of crashes, but the crash occurrence time recorded by the police officer is often inaccurate. The STL crash data have a fairly accurate crash occurrence time observed by the traffic management center operators via CCTV cameras and the incident management system, yet they do not contain other crash information such as collision type. Since the proposed crash prediction models required collision types and accurate crash times, the two crash datasets were analyzed together to produce the final dataset. As discussed earlier, crash counts were aggregated by 15-, 30-, and 60-min intervals for each of the two sites.

The rear-end collision type, known as the most frequent, appears to be most suitable for an application of the proposed SHD, although the SHD can be applied to other collision types including sideswipe. To include a sufficient number of crash counts

for modeling, 6 years (2000-2005) of crash data were used. Rear-end crashes for these years accounted for 62.7 percent (161 of 256) crashes at E264-03EB and 37.1 percent (33 of 90) crashes at W64-05EB on weekdays under normal weather condition.

None of crash datasets contained crash lane information. In addition, at times, individual vehicular data were not available on certain lanes because of the malfunction of detectors or communication errors. As such, the number of crashes for each lane was estimated in proportion to a lane utilization factor (i.e., a percentage of traffic using a lane at each segment). It was assumed that crash occurrences were proportional to the traffic volume using each lane for the case of rear-end crashes. The lane utilization factor was computed from historical traffic flow data for these segments. For example, historical data revealed that 34.7 and 40.2 percent of traffic volume on the E264-03EB segment used Lane 1 (left-most lane) and Lane 2 (middle lane), respectively. Thus, 34.7 and 40.2 percent of all crashes were assumed to have occurred in Lane 1 and Lane 2, respectively.

RESULTS AND DISCUSSION

Three final datasets by 15-, 30-, and 60-min intervals were prepared for estimating crash prediction models, with 192, 96, and 48 observations, respectively. Since this paper aims to show empirically the usefulness of the SHD for predicting crash occurrences, only the aggregated SHD entered into the models as an explanatory variable. In the end, one final model was selected for each time interval. An application example is provided for a better understanding of how to interpret and utilize the final models.

Excessive Zeros and Overdispersion

Table 2 shows summary statistics for each of the three intervals. Overdispersion was found in the 30- and 60-min interval datasets, and considerable portions of zero counts were found in all three datasets. These findings led to the selection of four candidate models, ZIP, ZINB, ZHP, and ZHNB, which were estimated using each of the three interval datasets, resulting in a total of 12 estimated models (four models \times three datasets).

TABLE 2 Crash Counts by Time Intervals and Tests for Overdispersion

Time Interval	No. of Observations	No. of Zeros (% Total)	Min. (Max.)	Mean (Var.)	Overdispersion Test		
					LR test (p-value)*	Dispersion Parameter Based on Deviance**	Regression-based Test (p-value)***
15 min	192	104 (54.2%)	0 (4)	0.589 (0.641)	0.499	0.810	0.170
30 min	96	32 (33.3%)	0 (7)	1.167 (2.077)	0.037	1.118	0.048
60 min	48	5 (10.4%)	0 (13)	2.313 (7.113)	0	1.654	0.011

*Overdispersion concluded if a p-value of the likelihood ratio (LR) test is smaller than 0.05.

**Overdispersion concluded if a dispersion parameter is greater than 1.0.

***Overdispersion concluded if a p-value of the OLS regression coefficient is smaller than 0.05.

Model Selection

As noted, AIC, BIC, and log-likelihood values were computed and used to compare model performance for each time interval. Table 3 summarizes the goodness-of-fit measures for the four candidate models.

TABLE 3 AIC, BIC, and Log-Likelihood of Estimated Model

Time Interval	Model	AIC	BIC	Log-Likelihood
15 min	ZIP	338.7	351.7	-165.3
	ZINB	340.7	356.9	-165.3
	ZHP	333.5	350.5	-164.8
	ZHNB	326.2	348.4	-161.8
30 min	ZIP	249.8	260.0	-120.9
	ZINB	249.5	262.3	-119.7
	ZHP	248.6	262.9	-122.3
	ZHNB	227.4	246.2	-111.7
60 min	ZIP	193.1	200.6	-92.6
	ZINB	182.6	192.0	-86.3
	ZHP	178.6	190.1	-87.3
	ZHNB	149.2	164.6	-72.6

For the 15-min interval, the ZHNB model is the best based on AIC, BIC, and log-likelihood values because these values are smaller than those of the other three models.

The difference in BIC between the ZHNB and ZINB models is more than 6, which indicates that the ZHNB model is *strongly preferred* over the ZINB model according to the BIC rule-of-thumb criteria (16).

For the 30-min interval, the ZHNB model is also the best. The AIC and log-likelihood values are much smaller than those of the other three models. In addition, its BIC value is smaller than those of the other three models by more than 10. This concludes that the ZHNB model is *very strongly preferred* over the others.

For the 60-min interval, the ZHNB model is again the best. Its BIC value is much smaller than those of the ZIP and ZHP models by 37 and 25, respectively. Interestingly, the ZHNB model is the best for all three intervals based on AIC and BIC. Moreover, as the interval increases from 15 to 60 min, the degree of the performance toward the ZHNB model over the ZIP, ZINB, and ZHP models becomes greater in terms of BIC (i.e., the difference in BIC values between the ZHNB and the other three becomes larger).

Figure 1 depicts differences between observed and predicted probabilities of crash counts for the 60-min interval. All four models look similar at high crash counts, say, 6 or more. However, at low counts, the ZHNB model outperforms the other three and fits the data particularly well at very low counts (e.g., zero and 1 crash). This visual comparison supports the conclusion that the ZHNB model is the best based on AIC and BIC values. Findings from the visual comparison for the 15- and 30-min intervals, although the graphs are not presented here, are consistent with the previous findings.

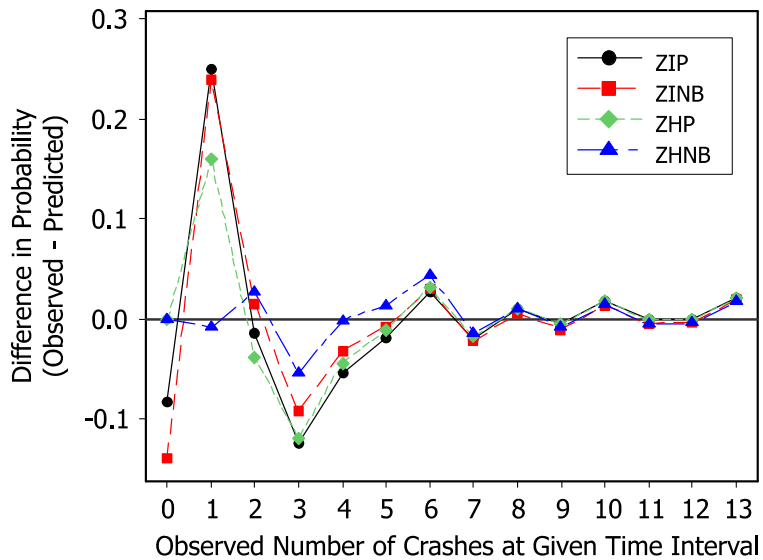


FIGURE 1 Comparison of observed and predicted mean probabilities of crash occurrence for 60-min interval. (ZIP = zero-inflated Poisson model; ZINB = zero-inflated negative binomial model; ZHP = zero-hurdle Poisson model; ZHNB = zero-hurdle negative binomial model).

In summary, the ZHNB model was determined to be final crash prediction model for the all three intervals, implying that it outperformed the ZIP, ZINB, and ZNH models. This means that excessive zeros contribute to an overdispersion, and the assumption regarding two types of zeros for zero-inflated models is not suitable for the data. Moreover, the longer time interval was found to offer stronger statistical preference on the ZHNB model in terms of BIC. However, choosing an appropriate time interval for aggregating data is also dependent on the study design and purpose.

Final Crash Prediction Model

Table 4 shows the final ZHNB models for the three time intervals. The aggregated SHD was found to be statistically significant in both the logit and NB models of the ZHNB

model, implying the SHD influences both whether a crash would occur and how many crashes would occur. A positive coefficient of the SHD in the logit model implies that an increase in the SHD is likely to increase the probability that a road segment will have at least one crash (i.e., crossing the zero hurdle). Meanwhile, a positive coefficient of the SHD in the NB model means that an increase in the SHD is likely to increase the expected number of crashes on that segment. This implies that as more drivers maintain an unsafe distance from their leading vehicle, producing a larger value of the aggregated SHD, their road segment will have a higher probability of crash occurrence and more crashes.

TABLE 4 Estimates of ZHNB Models

Time Interval		15 min		30 min		60 min	
		Coef.	p-value	Coef.	p-value	Coef.	p-value
Logit	Agg. SHD*	0.000187	0.000	0.000139	0.000	0.000568	0.149
	Constant	-1.545	0.000	-0.848	0.019	-0.235	0.755
NB	Agg. SHD*	0.000195	0.007	0.000102	0.003	0.0000527	0.006
	Constant	-15.49	0.591	-17.24	0.001	-2.576	0.234
	Ln(alpha)	12.60	0.662	15.55	0.003	1.624	0.466
No. of Observations		192		96		48	
Log-Likelihood		-161.1		-111.7		-72.6	

*Agg. SHD is the aggregated SHD, which was calculated by summing individual SHDs by the specified intervals, 15, 30, and 60 minutes.

Model Application

Since the SHD in the model was a summation of all negative SHD values of consecutive-vehicle pairs within the specified time interval, a straight interpretation of the estimated coefficients of parameters is difficult to make with respect to operational environments such as average driving speeds. To illustrate a potential application of the proposed crash prediction models, the following hypothetical example is given. Suppose a variable message sign is installed to warn drivers following too close and it affected drivers' behavior, resulting in a 10 percent reduction in the aggregated SHDs for 1 hour while no significant changes in other factors (e.g., volume and speeds) were observed.

For this example, the study data were used to provide basic inputs to the final models. The individual vehicular data collected on weekdays under normal weather condition from two lanes of four on a 1-mile segment with a 55 mph posted speed limit were used to calculate the aggregated hourly SHD during three times of day: peak (6 a.m.–9 a.m. and 4 p.m.–7 p.m.), mid-day (9 a.m.–4 p.m.), and night time (7 p.m.–6 a.m.). The calculated SHD values for the three times of day were 55,000 ft, 35,000 ft, and 3,000 ft, respectively. Inputting these SHD values into the ZHNB model developed in this paper for the 60-min interval resulted in predictions of rear-end crashes for the three times of day on that segment. Predicted values, probabilities, and expected numbers of such crashes were calculated with and without the impact of the variable message sign and are presented in Table 5.

TABLE 5 Safety Effects of Hypothetical Variable Message Sign (10% Decrease in Aggregated SHD for 60-min Interval)

Time of Day	Parameter	Before	After	% Change
Peak (6 a.m.–9 a.m. and 4 p.m.–7 p.m.)	Agg. SHD (ft)	55,000	49,500	-10.0
	$P(y = 0)$	3.50E-14	7.95E-13	2,168.7
	$P(y = 1 \text{ or } 2)$	0.518	0.572	10.4
	$P(y = 3 \text{ or more})$	0.482	0.428	-11.2
	$E(y)$	4.105	3.41	-16.9
Mid-Day (9 a.m.–4 p.m.)	Agg. SHD (ft)	35,000	31,500	-10.0
	$P(y = 0)$	2.98E-09	2.17E-08	629.1
	$P(y = 1 \text{ or } 2)$	0.722	0.758	5.0
	$P(y = 3 \text{ or more})$	0.278	0.242	-12.9
	$E(y)$	2.226	2.039	-8.4
Night Time (7 p.m.–6 a.m.)	Agg. SHD (ft)	3,000	2,700	-10.0
	$P(y = 0)$	0.187	0.215	14.6
	$P(y = 1 \text{ or } 2)$	0.776	0.75	-3.4
	$P(y = 3 \text{ or more})$	0.037	0.035	-5.4
	$E(y)$	1.022	0.984	-3.7

The aggregated SHD values were assumed to be uniformly decreased by 10 percent throughout a day as a result of the new variable message sign. In response to the installation of the variable message sign, during a 60-min period in 6 years, the

probabilities of observing no rear-end crash, $P(y = 0)$, were considerably increased, and the probabilities of observing more than three rear-end crashes, $P(y = 3 \text{ or more})$, were decreased regardless of time of day. Although percentage changes in the probabilities of zero crashes are enormous for peak and mid-day times, the probabilities after the sign installed are still practically zero.

For the peak time period, a 10 percent decrease in the aggregated SHD results in a 16.9 percent reduction in the expected number of rear-end crashes; for the mid-day and night time time periods, the reduction is 8.4 percent and 3.7 percent, respectively. This suggests that the variable message sign achieving a 10 percent reduction in the aggregated SHD would provide more safety benefits during peak times than during non-peak times.

CONCLUSIONS AND RECOMMENDATIONS

The major findings of the study were as follows:

- A statistically significant positive relationship was found between the aggregated SHD and the probability of observing rear-end crashes and the number of such crashes for all three intervals considered: 15, 30, and 60 min. This indicates that the aggregated SHD was proven to be useful in predicting the likelihood and number of rear-end crashes in an advanced transportation system.
- The ZHNB model outperformed the ZHP, ZIP, and ZINB models for all three intervals, and its superiority over the three models became stronger as the time interval increased. This suggests that the excessive zeros contribute to an overdispersion, and the

assumption of two types of zeros in zero-inflated models may not be suitable for the crash data analysis.

The study showed a promising opportunity in traffic safety analysis by applying SHD based individual vehicular car following behavior data in crash prediction. It also showed that the zero-inflated models with excessive zeros that have gained popularity recently for analyzing crash data may not be appropriate in that the ZHNB model was found to be superior to its zero-inflated counterpart: the ZINB model. This suggests that the presence of excessive zeros does not warrant the use of zero-inflated models, and characteristics of the zero counts (i.e., existence of structured zeros in the data) should be considered in selecting appropriate models.

The findings in this study can be used with future vehicle-infrastructure integration environment. The approach with a vehicle-based crash predictor would enable traffic engineers to have a reliable safety evaluation by location, time, and transportation management strategy under various traffic flow conditions. For example, a traffic engineer could evaluate the safety impact of typical transportation management strategies at particular freeway locations prior to their implementation by using well-calibrated and validated microscopic simulation models equipped with an individual-vehicle-based crash risk predictor such as the proposed SHD.

REFERENCES

1. Federal Highway Administration. *A Summary of Highway Provisions in SAFETEA-LU*. U.S. Department of Transportation, Washington, D.C.
www.fhwa.dot.gov/safetealu/summary.htm. Accessed June 30, 2008.

2. Institute of Transportation Engineers. *Intelligent Transportation Primer*. Washington, D.C., 2000.
3. Cambridge Systematics, and M. D. Meyer. *Crash vs. Congestion: What's the Cost to Society?* American Automobile Association, Heathrow, Fla., 2008.
4. Oh, C., J. Oh, S. G. Ritchie, and M. Chang. *Real-Time Estimation of Freeway Accident Likelihood*. Publication UCI-ITS-WP-00-21. Institute of Transportation Studies, University of California, Irvine, 2000.
5. Golob, T. F., W. W. Recker., and V. M. Alvarez. A Method for Relating Type of Crash to Traffic Flow Characteristics on Urban Freeways. *Transportation Research Part A*, Vol. 38, 2004, pp. 53-80.
6. Kockelman, K. M., and J. Ma. Freeway Speeds and Speed Variations Preceding Crashes, within and Across Lanes. *Journal of the Transportation Research Forum*, Vol. 46, No 1, 2007, pp. 43-61.
7. Abdel-Aty, M., N. Uddin, F. Abdalla, and A. Pande. Predicting Freeway Crashes From Loop Detector Data by Matched Case-Control Logistic Regression. In *Transportation Research Board: Journal of the Transportation Research Record*, No. 1897, Transportation Research Board of the National Academies, Washington D.C., 2004, pp. 88-95.
8. Lee, C., B. Hellinga, and F. Saccomanno. Evaluation of Variable Speed Limits to Improve Traffic Safety. *Transportation Research Part C*, Vol. 14, 2006, pp. 213-228.
9. Research and Innovative Technology Administration. *Vehicle Infrastructure Integration (VII)*. U.S. Department of Transportation, Washington, D.C. www.its.dot.gov/vii/index.htm. Accessed June 30, 2008.

10. Park, B. and S. Yadlapati. Development and Testing of Variable Speed Limit Logics at Work Zones using Simulation. In *Proceedings of the 82th Annual Meeting of the Transportation Research Board*. CD-ROM. Transportation Research Board of the National Academies, Washington, D.C., 2003.
11. Oh, C., S. Park, and S. G. Ritchie. A Method of Identifying Rear-End Collision Risks Using Inductive Loop Detectors. *Accident Analysis and Prevention*, Vol. 38, 2006, pp. 295-301.
12. Cameron, A. C., and P. L. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, U.K., 1998.
13. American Association of State and Highway Transportation Officials. *A Policy on Geometric Design of Highways and Streets, 5th Edition*. Washington, D.C., 2004.
14. Rose, C. E., S. Martin, K. Wanneumuehler, and B. Pikaytis. One the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data. *Journal of Biopharmaceutical Statistics*, Vol. 16, No. 4, 2006, pp. 463-481.
15. Lord, D., S. Washington, and J. Ivan. Poisson, Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and the Theory. *Accident Analysis and Prevention*, Vol. 37, No. 1, 2005, pp. 35-46.
16. Raftery, A. E. Bayesian Model Selection in Social Research. *Sociological Methodology*, Vol. 25, 1995, pp. 111-163.
17. Long, J. S. *Regression Models for Categorical and Limited Dependent Variables*. SAGE Publications, Thousand Oaks, Calif., 1997.