Volume 10 Number 1, 2014
ISSN 1094-8848

SPECIAL ISSUE ON SAFETY

JOURNAL OF TRANSPORTATION AND STATISTICS

U.S. Department of Transportation
Bureau of Transportation Statistics

JOURNAL OF TRANSPORTATION AND STATISTICS

A PEER-REVIEWED JOURNAL

JOURNAL OF TRANSPORTATION AND STATISTICS

Volume 10 Number 1, 2014
ISSN 1094-8848

U.S. Department of Transportation
Bureau of Transportation Statistics

The *Journal of Transportation and Statistics* is published by the

Bureau of Transportation Statistics
Office of the Assistant Secretary for Research and Technology
U.S. Department of Transportation
1200 New Jersey Avenue, SE
Washington, DC 20590
USA

Subscription information

Mail Product Orders
Bureau of Transportation Statistics
Office of the Assistant Secretary for Research and Technology
U.S. Department of Transportation
1200 New Jersey Avenue, SE
Washington, DC 20590
USA

Internet www.bts.dot.gov/publications

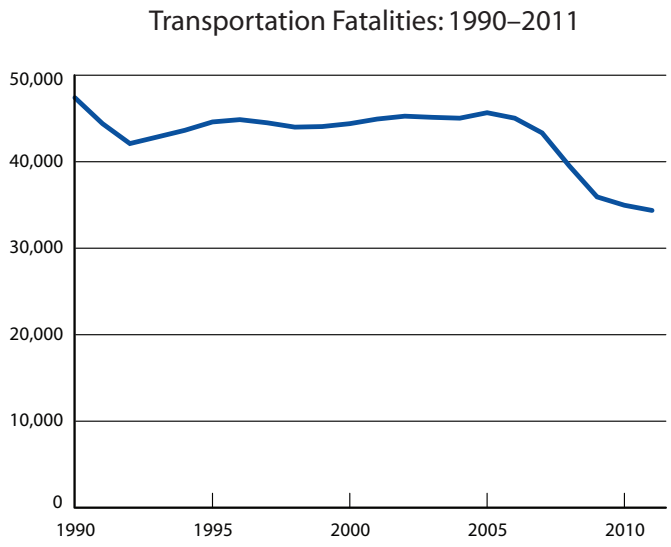
Information Service
Email answers@dot.gov
Phone 800.853.1351

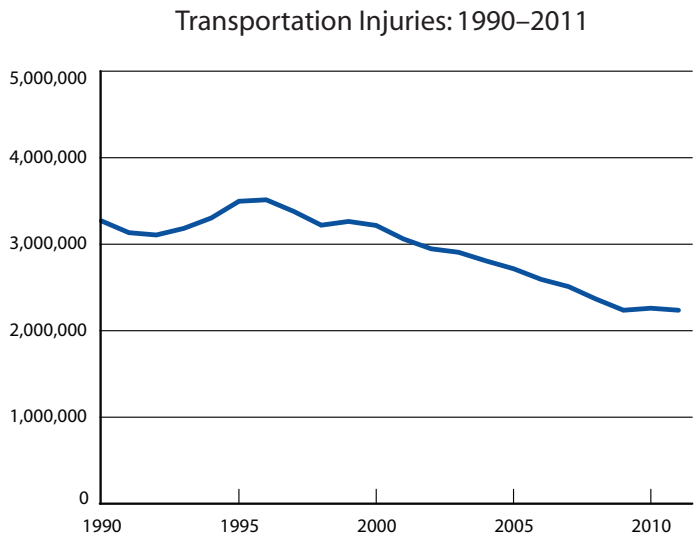
Introduction

Improving transportation safety remains the U.S. Department of Transportation's (USDOT's) top priority. The overarching objective is to reduce transportation-related fatalities and injuries by addressing driving behavior and vehicle-related and infrastructure safety issues. USDOT uses a data-driven approach to identify risk factors and develop countermeasures and assess their effectiveness.

The United States and much of the world have made considerable progress in improving safety across all modes of transport— strides made possible through technological advances such as more effective safety belts, regulatory actions such as vehicle safety standards, effective law enforcement, and public outreach. Despite the progress, transportation, including highways and the other modes, accounts for about one-third of the accidental deaths in the United States and is the leading cause of death for people between the ages of 5 and 24 [USHHS CDC 2012].

Transportation fatalities in 2011 were 34,388, a decline of 22.5% over 2000, while transportation injuries were 2,237,029 after a decline of 30.5%. This is in contrast to the preceding 1990-2000 period when fatalities declined 6.3% and injuries declined 1.6%.

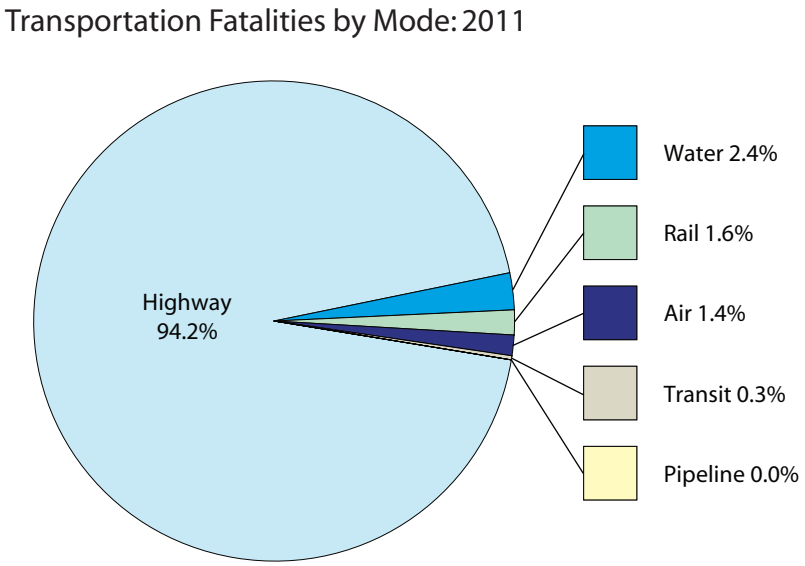




These decreases in the number of fatalities and injuries were observed despite U.S. Census data that show a 24.9% increase in the U.S. population—from 249 million in 1990 to nearly 312 million in 2011 [USDOT BTS 2013].

The majority of transportation fatalities and injuries occurred on the Nation's highways, which carry most of the passenger and freight traffic in the United States. Even though 2011 was the safest year on the highways since 1949 in terms of the number of traffic fatalities [USDOT NHTSA 2011], on average 89 people died and over 6,074 per day were injured on the Nation's highways.

While new and emerging technologies like vehicle-to-vehicle communications and next-generation air traffic control systems offer great promise, no solution would be complete without high-quality data and robust statistical analysis. Recognizing the importance of data and statistical analysis in improving transportation safety, Volume 10 of the *Journal of Transportation and Statistics* is focused on safety research. This Special Issue features six compelling studies that explore the frontier of applied statistical analysis and modeling to offer potentially life-saving insights for both researchers and policymakers. The authors are from some of the leading transportation research institutes in the world, and their exemplary work is indicative of the scope and depth of their expertise. Furthermore, the papers featured in this issue are the product of a multidisciplinary approach to transportation research—in a world where the lines between academic fields, industries, and business sectors are becoming less defined, such a perspective is crucial.



Research published in this Special Issue ranges from assessing crash-risk in roadway corridors where the absence of crash-related incidents has skewed the perception of danger, but not the tangible threat to a comprehensive analysis of the National Highway Traffic Safety Administration's Fatality Analysis Reporting System (FARS) data that sought to uncover previously unseen correlations between the different types of crash-related deaths and the factors that led to those deaths. In all of these studies, the underlying current driving their research is the idea that somewhere in the ever-expanding sea of data are answers capable of saving lives.

This special issue of the *Journal* includes six papers:

Modeling School Bus Crashes Using Zero-Inflated Model

When a school bus crashes, it is almost always breaking news. While motor-vehicle crashes during the morning or evening commute are a relatively common occurrence across much of the country, school bus crashes are rare events, and when children are injured or worse, it can be devastating for a community. This study explores the potential of the zero-inflated negative binomial (ZINB) model to shed light on previously unknown risk factors that could threaten the safe transport of children on specific segments of the roadway.

Crash Injuries in Four Midwestern States: Comparison to Regional Estimates

This study looks into factors that contribute to the most deadly motor-vehicle crashes in Iowa, Kansas, Missouri, and Nebraska, and why the magnitude of outcomes associated with factors such as adverse weather or seatbelt use in these four states varies so greatly from previous regional estimates. The findings raise questions about current methodologies used to guide new safety measures as well as the absence of a standard framework for crash reporting.

Investigation of the Impact of Corner Clearance on Urban Intersection Crash Occurrence

Signalized intersections contain numerous crash risk factors that have been subject to extensive study. However, there has been little research on corner clearance—the distance between a corner of two intersecting roads and the first driveway—which poses a unique safety risk to drivers exiting from such driveways. This study analyzed crash count data collected from all major, signalized intersections in Las Vegas and North Las Vegas, Nevada, to determine how corner clearance impacts roadway safety. The results provide several key findings that could support future measures to reduce risks associated with corner clearance.

Application of the Bayesian Model Averaging in Predicting Motor Vehicle Crashes

Reliable statistical models underpin the validity of roadway safety research. Typically, analysts will apply multiple models during a study and then apply the one that provides the single "best fit" for the relevant data. This methodology is inherently limited because it does not incorporate the uncertainties presented by the disparate models. In this study, the authors explore the efficacy of applying Bayesian Model Averaging to account for this problem.

Lane Width Crash Modification Factors for Curb-and-Gutter Asymmetric Multilane Roadways: Statistical Modeling

This study is the result of an analysis of crash frequency on multilane, urban roadways and the possible correlation of asymmetrical lanes to both frequency and severity. Asymmetric lanes occur when then the outside lane is wider than the inside lane. The authors' conclusions point to simple changes in roadway design which could reduce the number and severity of crashes along corridors identified as being at-risk.

A Multidimensional Clustering Algorithm for Studying Fatal Road Crashes

Building on existing research on correlative relationships linking fatal crash factors, this study applies a specialized theoretical method, called "graph-cuts," to analyze all fatal car crashes occurring in the prior 2-, 5-, and 10-year spans. This approach searches for clusters that indicate subtle correlations that emerge in a comparative analysis of the historical crashes to the 84 enumerated parameters that can describe a fatal crash event. Using this method, the authors found strong correlations between certain parameters that had not been reported in prior studies.

REFERENCES

U.S. Department of Health and Human Services. Center for Disease Control. Deaths: Preliminary Data for 2010. *National Vital Statistics Reports* Vol. 60, No. 4, Jan. 11, 2012. Available at www.cdc.gov.

U.S. Department of Transportation. Bureau of Transportation Statistics. 2013. The American Landscape. *Pocket Guide to Transportation*. Available at www.rita.dot.gov/bts/publications.

U.S. Department of Transportation. National Highway Traffic Safety Administration. 2011. DOT Estimates Three Percent Drop Beneath 2009 Record Low. *Press Release*. April. Available at www.nhtsa.gov.

Volume 10 Number 1, 2014

Contents

Modeling School Bus Crashes Using Zero-Inflated Model

DEO CHIMBA [1,*]
THOBIAS SANDO [2]
VALERIAN KWIGIZILE [3]
BONIPHACE KUTELA [1]

[1]Department of Civil Engineering
Tennessee State University
Nashville, TN 37209

[2]University of North Florida

[3]Western Michigan University

ABSTRACT

School bus crashes are rare, but their occurrence can have devastating effects on the school children involved. Such crashes are infrequent and random, and some roadway segments may not experience any school bus related crashes for a number of years (zero crashes). Despite the fact that no crashes may have occurred along particular stretches of road, these zero-crash road segments cannot be termed as safe sites, and they cause a dual state of crash experience (no crashes, but still at risk for crashes) compared to a single state of non-zero crash prone sections where risk is confirmed. Literature indicates that for extremely rare and random count data, such as school bus crashes, Poisson and Negative Binomial (NB) distributions become more applicable for modeling. Apart from Poisson and NB, there exists an alternative discrete distributional model that is used to model extra-zero discrete data, such as school bus crashes,that allows exploration of the impact of zero segments. This alternative modeling approach called zero-inflated negative binomial (ZINB) model is introduced in this study for evaluation of variables influencing school bus crashes. Although crash data rarely reveal variability, the ZINB model provides a more flexible modeling framework for school bus crashes. The study found that, ZINB yields better prediction (tight standard errors and higher z-statistics), compared to NB model though same variable coefficient signs.

[*] Corresponding author.
Tel.: +1 615 963 5430
Email address: dchimba@tnstate.edu

KEYWORDS: school bus crashes; zero-inflated model, Poisson, Negative Binomial

Presence of median and outside shoulders was found to have tendency of reducing school bus crashes. On the other end, wider medians, outside shoulders, inside shoulders, and lane widths were found to reduce the probability of these crashes. Presence of curb and gutter and two-way left turn lane (TWLTLL, high posted speed limits, multilane segments, and congested segments were found to increase the probability of school bus crashes.

INTRODUCTION

School bus crashes are rare, but their occurrence can have devastating effects on the school children involved. Limited studies have been published on the statistical modeling of school bus related crashes. Yang et al. (2009) indicated that findings on the few available published studies on school bus crashes and injuries vary widely depending on the source of information and study population. In the same study, which was conducted in Iowa, they found that the school bus fatality rate was 0.4 per 100 million miles traveled. The study concluded by recommending that safety of school bus transportation over other vehicles should be a factor in making school transportation policies. McGeehan et al. (2006) reported that there were an estimated 51,100 school bu–-related injuries treated in U.S. emergency departments from 2001 to 2003 and that head injuries accounted for more than half (52.1%) of all injuries among children<under 10 years of age, whereas lower extremity injuries predominated among children 10 to 19 years of age (25.5%). In their study, Lapner et al. (2003) found that head, neck, and spine are the most common injuries when children are involved in rollover school bus collisions and that additional safet, changes to the current school bus design are needed.

Based on their dramatic effects, analysis of roadway, traffic, human, and environmental factors impacting school bus crashes is needed. One of the known methodologies in studying and analyzing crash data similar to those related to school buses is through statistical analysis. Though one of the popular method, statistical evaluation of factors impacting school bus crashes becomes more challenging due to the rarity of these types of crashes. In connection to crash statistical analysis, various modeling approaches have been proposed to identify the genuine relationship between crash (in general) occurrences and roadway geometrics, traffic characteristics, environmental conditions, and human factors. In contrast, not much effort in terms of modeling has been invested in school bus related crashes. This might be contributed by many factors, one of them being data availability as many school buses use local roads whose crash data may not easily be available. Other factor may be unavailability of school bus counts as an exposure variable in calculating crash rates. Furthermore, adequate modeling methodology that takes into consideration extreme rarity of these types of crashes may have hindered study progress.

However, in the past two decades, the Poisson, negative binomial (NB, and various model extensions, such as zero-inflated, have been extensively studied and applied to modeling general crash data, Chimba et al. 2010. Fairly comprehensive reviews were given by Maher and Summergill (1996) and Lord et al. (2005). To make necessary connections with the scope of work proposed in this paper, though, several important milestones in model developments and significant progresses made in recent years are relevant to the discussion. Jovanis and Chang (1986), Joshua and Garber (1990), Chimba et al. (2010) and Miaou and Lump (1993) compared model fitness using the Poisson and multivariate linear regression mod-

els, concluding that the Poisson outperforms the multivariate linear regression model due largely to its more appropriate statistical properties for describing non-negative discrete data like crashes. However, they noted that if the crash data reveal significant over-dispersion around the estimated mean, the Poisson model becomes inadequate and more general distributional models, such as the negative binomial (NB), are needed.

Over-dispersion is a phenomenon which occurs when the model is fitted using Poisson or negative binomial. Hardin and Hilbe (2001) listed the following as the source or over-dispersion in the data or model:

- when some important independent variables are omitted from the model;
- when the data contains a lot of outliers resulting either from unreliable data collection or mistake and errors during data recording;
- when the model fails to include sufficient number of interaction terms;
- when the variable by itself is not appropriate and it needs transformation;
- if the distribution assumed is quite different from the real distribution which relates the data, e.g., using linear model instead of quadratic.

The earliest recognition of this limitation dates back to Maycock and Hall (*1984*) who used the NB model for analyzing crashes at junctions. The NB distribution, also known as the Poisson-Gamma distribution, is derived by conjugating the Poisson distribution with the Gamma distribution in which the true mean is assumed to account for extra data variability (over-dispersion). The over-dispersion parameter, in an earlier stage, is assumed identical across all roadway segments, Hauer et al. (1989), Bonneson and McCoy (1993), Miaou (1994) and Shankar et al. (1997). Inspired by Cameron and Trivedi (1986), Maher and Summersgill (1996) adopted the site-dependent (indexed by i) over-dispersion factor (ϕ_i) to account for extra variability around the estimated mean (μ_i). The over-dispersion parameter is linked with the estimated mean in the form of $\phi_i = \phi\mu_i^d$, where d is an additional parameter that can be estimated along with all other parameters. Hauer (*15*) demonstrated that, when roadway segments differ in length, parameters estimated by the maximum likelihood estimation (MLE) method will be unduly influenced by very short segments if the over-dispersion parameter is assumed to be equal across all roadway segments. The most plausible remedy suggested by Hauer is to set the over-dispersion parameter in proportion to the segment length (L), i.e., for segment i, $\phi_i = \phi L_i^d$, where d=1. When $d\neq1$, the sum of crash estimates from su-segments based on the empirical Bayes (EB) method will not be consistent with the estimates for the roadway as a whole. As an alternative to the estimated mean function, one could also model the over-dispersion parameter as a separate function of explanatory variables different from the mean prediction function (2003). They found that the dispersion parameter varies by site (intersection) as a function of the approach flows, the ratio of the flows from minor and major approaches, and geographical locations. Ignoring this variation, namely, treating the over-dispersion factor as a fixed parameter, can significantly undermine the fitness of the estimation. However, they found the parameter estimates only change slightly with the specific crash data used.

As a plausible extension from the Poisson family, the zero-inflated has been applied to model the over-dispersion phenomenon due to

an excess of zero observations, Miaou (1994), Shankar et al. (1997) and Lee and Mannering (2002). The concept originated from a mixture of distributions, where the split parameter can be modeled as a constant (Johnson and Kotz (1969) and Washington et al (2002)) or, because it is bounded between 0 and 1, a logit function of the estimated mean (Miaou (1994) and Mullahy (2001)). The relative effectiveness of these two treatments remains to be investigated although it has been demonstrated that both of them could make great contributions toward enhancing model performance.

ZERO-INFLATED (ZI) DISTRIBUTION

Recently, zero inflated has surfaced as a plausible approach for use in crash analysis. The use of zero-inflated has been justified from the fact that Poisson and Negative Binomial (NB) models with or without their extensions as well as several variations seem to model non-negative discrete response variables, with over-dispersion and the underlying assumption that the occurrence of crashes observed at a given time and space scale follows a Poisson process. Lord et al. (2005) challenges this assumption by arguing that the occurrence of crashes is in fact a binomial process, which can be approximated by a Poisson process when the number of trials (e.g., traffic exposure) is large with a small likelihood (risk) of crashes. This argument roots from modeling crashes as a dual-state data generation process or a series of Bernoulli trials, i.e., the outcome of a school bus entering into a roadway section is either perfectly safe, involving no crashes, or unsafe, assuming crashes do take place (Lord et al. (2005), Shankar et al. (1997) and Qin et al. (2004)). If the probability of independent events is the same for all trials, the dual-state data generation process will naturally give rise to a binomial distribution for describing school bus crash frequencies. However, the equal probability assumption is questionable in reality since the crash-involved probability varies across roadway segments as a function of drivers, traffic and roadway geometric characteristics among others; for this reason, further research into modeling the unequal probability of independent events is needed.

As discussed in previous sections, zero-inflated models are mainly used for modeling excessive zero count data. Zero count may refer to the situations where the likelihood of an event occurring is extremely rare se.g., school bus crashes) in comparison to normal expectatio, (Cameron and Trivedi (1986), Lee and Mannering (2002), and Mullay (1986)). Zero-inflated (ZI) can be modeled as Zero Inflated Poisson (ZIP) or as Zero Inflated Negative Binomial (ZINB) models. Poisson regression model probability function is given in the following form (*14*)

$$P(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \quad y_i = 0, 1, 2\ldots \text{ and } \mu > 0 \qquad (1)$$

The mean parameter

$$E[y_i / x_i] = \mu = \exp(x_i\beta), \text{ Variance} = \mu,$$

Where y_i = a random variable representing number of school bus crashes,

x_i = Parameter related to the occurrence of school bus crash (Vector of explanatory variable)

ß = the coefficient of the corresponding factor (vector of estimable parameter).

For the ZIP model, it assumes that the events y_i=(y_1, y_2…..y_N) are independent and the model is

$$\Pr[y_i = 0] = \varphi_i + (1-\varphi_i)e^{-\mu_i}$$

$$\Pr[y_i = r] = (1-\varphi)\frac{e^{-\mu_i}\mu_i^r}{r!}, \; r = 1, 2\ldots n \qquad (2)$$

Where φ=proportion of zeros.

Maximum likelihood estimates are used to estimate the parameters of the ZIP regression model and confidence intervals are constructed by likelihood ratio tests.

The negative binomial (NB) model can be expressed as:

$$p(y)=\frac{\Gamma\left(y+\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma(y+1)}\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \quad (3)$$

where the mean $\mu=E(y)=v\exp(X\beta)$.

The corresponding variance is $Var(y)=\mu+\alpha\mu^2$. Similar extensions to the NB model are considered, including the zero-inflated model (ZINB) with constant and mean-dependent split parameters, and the mean-dependent over-dispersion factor.

The ZINB model with constant split parameter (Shankar et al. (1997) and Washington et al. (2002)) can be expressed as:

$$\begin{cases}\gamma+(1-\varphi)p_{Y=0}=\varphi+(1-\varphi)\left(\frac{1}{1+\alpha\mu}\right)^{1\alpha} & Y=0\\ (1-\varphi)p_{Y=y} & Y=1,2,3,\ldots\end{cases} \quad (4)$$

Furthermore, appropriateness of using the zero inflated model rather than the traditional model, Poisson or Negative Binomial can be tested. The common known test statistic is through Vuong's Value, estimated as shown below, Washington et al. (2002);

$$m_i=\ln\left\{\frac{f_1(y_i/X_i)}{f_2(y_i/X_i)}\right\} \quad (5)$$

Where f_1(yi/*Xi*) is the probability density function for one model, say Zero-Inflated Negative Binomial, ZINB) and $f_2(y_i/Xi)$ is the probability density for comparison model, say Standard Negative Binomial, NB)

$$V=\frac{\sqrt{n}\left(\overline{m}\right)}{S_m}, \text{ where}$$

$$\overline{m}=Mean=\left[(1/n)\sum_{i=1}^{n}m_i\right], \quad (6)$$

S_m=Standard Deviation, n=Sample Size, and V=Vuong's Value.

If Absolute(V)<$V_{critical}$(1.96 for 95% Confidence Interval), the test does not support the selection of one model over the other. Large Positive values of V greater than $V_{critical,}$ e.g. V> $V_{critical}$ favor first model over second model whereas large negative values support second model.

DESCRIPTIONS OF DATA AND VARIABLES

Data for this study originated from Tennessee Department of Transportation (TDOT) through Tennessee Roadway Information Management System (TRIMS) database system. This database includes crash data comprising attributes such as harmful event, contributing causes, injury severities, traffic characteristics and geometric characteristics among others. The data have the exact log mile where the crash occurred. Furthermore, crashes are listed if they are school bus related or not. Downloaded crashes were therefore screened to identify only those that were school bus related. School bus only crashes occurring on state roadways (SR) in Davidson and Shelby counties were considered in this study (Davidson and Shelby are the most populous counties in the state of Tennessee, hosting the largest two cities of Nashville and Memphis, respectively). Local streets ann non-state roads were not included in the analysis due to unavailability of the traffic counts, an essential element in crash modeling. A total of 493 school bus crashes for the span of eight years from 2002 to 2009 with an

average of 61.6 crashes per year were gathered for the study as shown below. It should be noted that the data for year 2002 were not complete; hence, if taken out of consideration, the average crash frequency per year becomes 67.7 crashes per year.

Apart from crash data, roadway geometrics of the analyzed roadway segments were also downloaded from the TRIMS database. Roadway geometry is defined by the beginning and end log miles. Utilizing the county and route number which are present in both crash and geometric data, total number of crashes for each segment were tallied utilizing a small written computer program in Stata software. The written program had the capability of searching the route ID and segment boundaries (beginning and end log miles) in the geometric data, then matching and tallying the corresponding number of crashes by counting log miles within that particular segment. It then merges the two data into a single dataset. Overall, 1903 roadway segments ranging from 53 ft (0.01 mile) to 6.508 miles were identified for the study. As expected, most of the segments had zero school bus crashes (83.8%), followed by one crash (11.8%) and 2.89% for two crashes. Only 1.54% of the segments had more than two school bus crashes for the analyzed eight years span. The presence of excessive zero crash segments supports the use of zero inflated model as a modeling distribution for this study.

Table 1 summarizes the statistics on crash, roadway features, and categorical variables created as a derivative of other variables. For estimation purposes, some variables were modeled as indicator (categorical) variables as listed in table 1—such as posted speed limit (35 mph or below, 40 mph to 45 mph, and 50 mph to 55 mph) and so forth. In addition, a new variable called directional peak hour volume (DPHV) per lane was created to represent the traffic intensity during rush hours, which is computed as the product of AADT, directional split, and K-factor divided by number of lanes. Note that DPHV might be correlated with AADT to some extent but the causal effects are different from AADT. The former characterizes the effect of congestion on crashes, while the latter represents an exposure measure for crashes. Another variable created was vehicle miles of travel (VMT) as the product of AADT and length of the segment.

EMPIRICAL DISTRIBUTIONS OF CRASHES

The crash frequency distributions (line and histogram) were first analyzed and fitted with the over-dispersed distributional models, including Poisson, NB, and their zero-inflated versions, to determine if the data was over-dispersed. As can be seen in Figure 1, NB and ZINB models closely follow the observed crash distribution compared to Poisson and ZIP. The over-dispersion factor tested highly significant, indicating the school bus crash data was over-dispersed. This was furthermore confirmed by alpha log-likelihood ratio test (in table 2).

In comparison, data were fitted with the non-over-dispersed models, including, Poisson, and zero-inflated Poisson (ZIP). In contrast, the Poisson gave the least desirable results. The ZIP model yielded good fit to the zero

Year	2002	2003	2004	2005	2006	2007	2008	2009	Total	Average/Year (without 2002 data)
Crashes	19	72	67	69	85	69	59	53	493	67.7

TABLE 1 Study Data Summary

Numerical variables

	Mean	Min	Max
Number of crashes	0.2	0	7
Segment length	0.2	0.01	6.508
Daily volume (AADT)	23,062	1,190	61,230
Percent of peak hour volume	9.0	4	14
Directional split	61.6	50	100
VMT	4,693	24	70,265
DPHV	370	46	1,234
Number of through lanes	4	2	8
Median width	5.8	0	60
Lane width	11.6	9	15
TWLTL width	4.3	0	24
Outside shoulder width	4.3	0	32
Inside shoulder width	0.3	0	24
Posted speed limit	40	15	55
School speed limit	16.9	15	30

Categorical variables

Street lights	0-No lights, 1-Lights present
Posted speed limit categories	0-No posted speed limit up 35 mph, 1-(40-45 mph Posted Speed and 2-(50-55 mph posted speed)
Presence/absence of school speed limit	0-No school speed limit, 1- School speed limit
Terrain	0-Flat, 1-Rolling or Mountain
Land use	0-Commercial, 1-Mixed residential & commercial, and 2-Residential
Presence or absence of median	0-No median, 0-There is a median
Median composition	0-No median, 1-Concrete, 2-Grass plot, 3-Painted
Presence or absence of TWLTL	0-No TWLTL, 1-There is TWLTL
Presence or absence of outside shoulder	0--No outside shoulder, 1-There is outside shoulder
Outside shoulder composition	0-Ditch, 1-Asphalt, 2-Gravel and dirt, and 3-Concrete
Presence or absence of inside shoulder	0-No inside shoulder, 1-There is inside shoulder
Presence or absence of curb & gutter	0-No curb & gutter, 1-There is curve and gutter
Presence or absence of sidewalk	0-No sidewalk, 1-There is a sidewalk

observation, but fitted the rest of observations rather poorly. This indicates that mixing the nonover-dispersed model with a simple spike mass function at zero-crash observations is not sufficient to produce a satisfactory fit when the distribution is highly skewed. Apart from the probability plots, decision of whether to use a Poisson or Negative Binomial can also be based on the dispersion parameter, σ_d by Poisson error structure (*23*) given as;

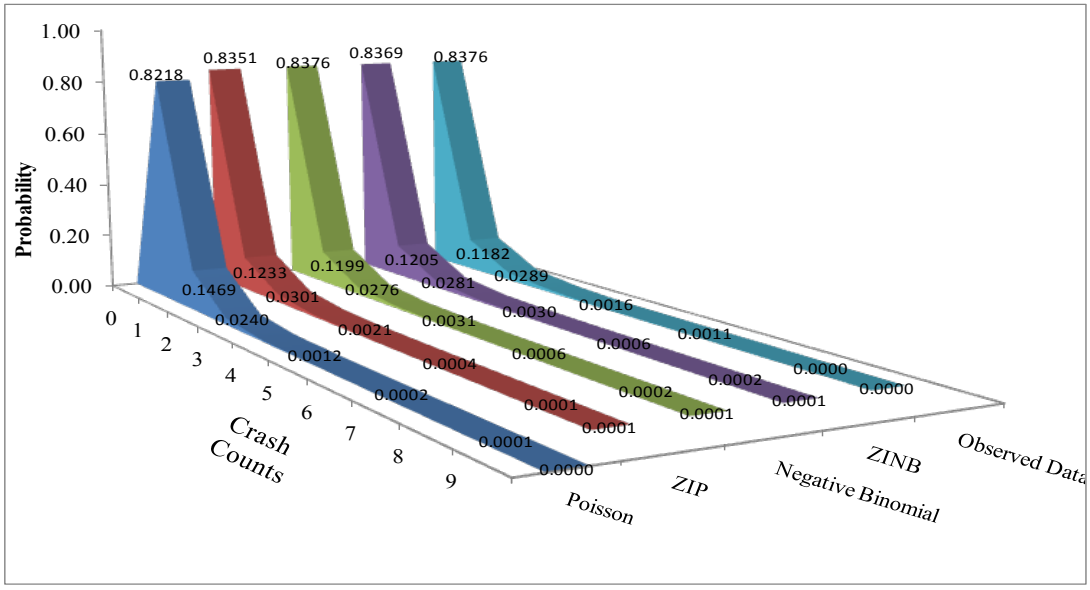
$$\sigma_d = \frac{Pearson\chi^2}{n-p}$$

, where n is the number of observations, p is the number of model parameters, and $Pearson\chi^2$ is defined as

$$Pearson\chi^2 = \sum_{i=1}^{n} \frac{\left[y_i - \hat{E}(y_i)\right]^2}{Var(y_i)} \quad (7)$$

Where y_i is the observed number of accidents on section i, $E(y_i)$ is the predicted accident fre-

FIGURE 1 Probability Histogram Plots of the Poisson, NB and Zero-Inflated Models



quency for section i, and Var(y_i) is the variance of accident frequency for section i. If σ_d turns out to be significantly greater than 1.0, then the data has greater dispersion than is explained by Poisson distribution, and Negative Binomial regression model is fitted to the data, Hardin and Hilbe (2001). The crash over-dispersion parameter was found to be 3.06 favoring NB over Poisson regression. The following distributions were therefore retained for final model estimation:

- Standard NB
- Zero Inflated Negative Binomial

RESULTS ON MODEL ESTIMATION

Based on the findings from fitting the empirical distributions, the NB and its zero-inflated version were therefore utilized for model estimation. Estimations were performed using Stata (Stata Corp LP (2008)) software. The *z*-statistic was used to assess the statistical significance of the variables. As expected, not all variables were statistically significant. The adequacy of individual models and effectiveness of the model extensions were evaluated based on the log-likelihood test. Beyond that, the corrected R^2, reasonableness of the fitted values, such as their mean and maximum, were also monitored. Finally, the normalized Bayesian Information Criterion (BIC) was used to assess the most appropriate model. The BIC can be defined as:

$$\text{Normalized BIC} = \frac{\text{Log-likelihood}}{N} - \frac{P}{2}\frac{\ln(N)}{N} \quad (8)$$

Where P and N are the number of parameters and samples, respectively. Although the distributional models under consideration may have different scales for the log-likelihood function value, this criterion provides a preliminary model assessment within the same distributional model family.

ZINB AND NB MODEL FITNESS PERFORMANCES

Model performance results are summarized in table 2. As shown in the results

1. Vuong test of ZINB vs. standard NB was found to be 2.31 favoring ZINB over NB

2. Sign and magnitude of the variable coefficients were almost identical in both the NB and ZINB models, but differed in standard error and z-statistics.

3. The ZINB standard errors were observed to be slightly lower (tighter) compared to those in NB indicating ZINB was well fitted compared to the later.

The results clearly suggested that the zero-inflated model with mean-dependent over-dispersion factor performed better for school bus crash data modeling compared to other models tested. On the other end, comparing the two models, one can argue that it requires more computational time for ZINB than the NB model primarily due to the necessary intensive computation of the proportion of zeros. In addition, the Hessian matrix for determining the *z*-test values is much more complicated in the ZINB model.

From the **"**inflate**"** part of the ZINB model in table 2, the probability of being in zero school bus crash segments "φ" is determined as shown in equation 9. The inflate model is used to calculate the probability of zero crashes

$$\varphi_i = \frac{e^{(-45.54 - 0.207 * MW + 3.45 * PPHV - 16.9 * PCC - 31.13 * TWLTL}}{1 + e^{(-45.54 - 0.207 * MW + 3.45 * PPHV - 16.9 * PCC - 31.13 * TWLTL}} \quad (9)$$

which inflates the number of zeroes then used in equation 10.

The ZINB first estimates the effects of the independent variables on the crash frequency; these coefficients are interpreted just like standard NB coefficients. Secondly it estimates the equivalent of a binary logit model where the outcome variable is the log odds of being in the zero-school bus crash segment compared to being in the non-zero crash segments. The coefficients in the inflate part; labeled "INFLATE" (table 2) correspond to the binary model predicting group which considers probability of zero crashes. It specifies the model that determines whether the observed count is zero. These coefficients are interpreted just as the coefficients for a binary logit model. A positive coefficient means the independent variable has the effect of increasing the odds that the dependent variable equals a given value, usually 1 for binary dependents. A negative coefficient means that the independent variable has the effect of decreasing the odds that the dependent variable equals the given value. Utilizing the probability of zero crash model "inflate" in equation 9, school bus crash frequency are then predicted as follows;

ANALYSIS OF THE MODEL RESULTS WITH RESPECT TO CRASH ATTRIBUTES

As stated earlier, the primary objective of this modeling effort was to evaluate school bus crash frequency as a function of explanatory variables. The school bus crash frequency here is defined as the number of crashes per year. The model performs reasonably well in term of the reasonableness of the model coefficients with respect to school bus crashes. As expected, roadway segments with posted school zone speed limit (PSSL) has negative coefficient with strong z-value showing its significance in reducing school bus crashes. Some literatures have pointed out traffic volume as surrogate for congestion and characteristics of increasing crash. As a common wisdom, AADT showed positive coefficient which supports previous findings (though not necessarily for school bus crashes) by Miaou and Lump (1993), Miaou (1994), Garber and Ehrhart

$$\text{School Bus Cashes} = (1 - \varphi_i) * e^{\left[\begin{array}{l} -0.66 + 0.0000026AADT + 0.026NL - 0.146LW - 0.008LW - 0.004OSW - 0.059ISW \\ +0.434TWLTL + 0.207SL_1 + 0.285SL_2 + 0.249PSSL - 0.367PM + 0.135PCC - 0.556POS \end{array}\right]} \quad (10)$$

(2000), Chimba et al. (2010) and others. As the number of vehicles per lane increases on the highway, fewer gaps allow lane changing, turning movements, or merging, which eventually increase likelihood of crashes. The analyzed segments ranged between a minimum of 1,190 vpd to 61,230 vpd AADT. Though not directly found from this study, a slight change in traffic characteristics can have noticeable impact to school bus operations which eventually can lead to risk of accident. Most often the school buses operate during peak hours (especially morning rush hours) where interaction with congested passenger cars becomes very common and may lead to scenario or likelihood of crashes.

The coefficient of the percentage of peak hour volumes (PPHV) is positive in the inflated portion, which is consistent with some previous research findings, Shankar et al. (1997) and Ivan et al. (1999). This suggests that as the percentage of peak hour volumes increases, the likelihood of school bus crashes increases too. Number of lanes (NL) appears in "crash" portion only of the ZINB with a significant positive coefficient indicating increasing likelihood of school bus crashes if the segments have more lanes. Most of the previous studies also concluded with the same findings that the higher the number of lanes, the higher the crash rate, Noland and Oh (2004), Aty and Radwan (2000), Chimba et al. (2010) and Garber and Ehrhart (2000). As a general rule more lanes roadway sections are associated with more flow per lane which can be unsuitable for school bus safety.

Posted speed limit appeared to be significant only on the crash model portion and non-significant in the "inflate" portion. Both 40–45 mph (SL_1) and 50–55 mph (SL_2) posted speed limit categories have positive coefficients indicating the tendency to increase the probability of school bus crashes compared to lower 15–35 mph speed segments or no speed limit sections. However, the coefficient of 50–55 mph speed limit is slightly larger than that of 40–45 mph, denoting the higher the posted speed limit, the riskier are the school buses to be involved in a crash. The conclusion which can be drawn from the speed limit to school bus safety is school buses should avoid high speed routes.

Lane width (LW) was found to be significant only to the crash part of the developed ZINB model and not to the "inflate" portion. The variable was incorporated with the intention of evaluating how lane width affects school bus crash frequency. The coefficient of lane width is negative in the model with *z*-value of -1.84 which approximately 93% significant level. This means wider lanes are likely to reduce school bus crashes compared to narrow lanes. One can suggest that wider the lane provide extra separation between the vehicles which can give school bus buffer deviation in case of crash leading incidents. In addition, the buffer between vehicles can provide a room for the driver to correct before the crash. Wider lanes can as well give the driver more driving comfortability then reduce delays and improve capacity. On the other part, the study used median under two scenarios; median width (MW) and presence or absence of the median (PM). Median width was found to be a significant variable in both crash and inflate portions of the ZINB. Presence of median has a negative coefficient, the depiction that the school bus passing on the segments with medians will have low probability of being involved in crash compared to "no-median" (undivided) segments. Furthermore, roadway segments with wider medians according to model outputs are shown to be safer (negative coefficient) compared to narrow median segments.

TABLE 2 Modeling Results Using the ZINB and NB Models

	ZINB			Negative Binomial (NB)		
CRASHES	**Coef.**	**Std error**	**Z-Value**	**Coef.**	**Std error**	**Z-Value**
AADT	2.6E-05	6.29E-06	4.2	2.9E-05	6.38E-06	4.5
Number of lanes (NL)	0.026	0.0139	1.89	0.028	0.0175	1.59
Lane width (LW)	-0.146	0.0794	-1.84	-0.144	0.0809	-1.78
Median width (MW)	-0.008	0.0041	-1.89	-0.008	0.0043	-1.75
Outside shoulder width (OSW)	-0.004	0.0021	-1.99	-0.002	0.0021	-1.13
Inside shoulder width (ISW)	-0.059	0.0309	-1.92	-0.055	0.0314	-1.75
Presence of TWLTL	0.434	0.1497	2.9	0.484	0.1509	3.21
Speed limit 40-45 mph (SL1)	0.207	0.1100	1.88	0.206	0.1115	1.85
Speed limit 50-55 mph (SL2)	0.285	0.0993	2.87	0.262	0.1323	1.98
Presence of school speed limit (PSSL)	-0.249	0.064	-3.92	-0.265	0.07	-3.77
Presence of median (PM)	-0.367	0.1954	-1.88	-0.344	0.2058	-1.67
Presence of curb & gutter (PCG)	0.135	0.0708	1.9	0.191	0.1028	1.86
Presence of outside shoulder (POS)	-0.556	0.1890	-2.94	-0.579	0.1937	-2.99
Constant	-0.660	0.8916	-0.74	-0.832	0.9138	-0.91
Length		(offset)			(offset)	
INFLATE						
Median width (MW)	-0.207	0.136	-1.52			
Percent of peak hour volume (PPHV)	3.454	1.229	2.81			
Presence of curb & gutter (PCG)	-16.901	8.326	-2.03		NA	
Presence or absence of TWLTL	-31.127	15.486	-2.01			
Constant	-45.535	15.488	-2.94			
Length		(offset)				
Alpha	1.376			1.524		
Likelihood-ratio test of alpha=0			18.52			121.78
Vuong test of ZINB vs. standard NB			2.31		NA	

As stated, median (MW and PM), outside shoulder widths (OSW) and presence of outside shoulder (POS) both have negative coefficients. For the school bus safety point of view, these cross-sectional elements can be used for emergency stops, hence the wider they are the better for the driver correction and for bus safety in general. Inside shoulder width (ISW) is significant with a negative coefficient, indicating likelihood of bus crashes decreases as shoulder width increases. These findings are consistent with some previous studies on general crashes, Chimba et al. (2010), Milton and Mannering (1998), Aty and Radwan (2000), and Lee and Mannering (2002). From a highway safety standpoint, shoulder can be used for vehicle stopping in an emergency or during an incident, and drivers can take advantage of wider shoulders to avoid hitting roadside obstacles. Besides that, wider shoulders can also be used for deceleration to avoid a crash.

Median composition (no median, concrete barrier, planted grass or painted) does not appear in the final model due to insignificancy. Though not appearing in final model, they are still considered very crucial factors with respect to school bus crash. Painted and grass/

lawn median (without concrete barrier) can allow the school bus to accelerate further from the main travel way which can help avoid a crash. As shown in table 2, two-way left turn lane (TWLTL) median generated mixed results. In the crash portion of the ZINB, this variable has positive coefficient, the sign that presence of TWLTL creates safety concerns for school bus crashes. As raised median play a very vital part in separating opposing traffic streams and reduces access from the mainline, TWLTL does the opposite. The environment in which vehicles use the TWLTL for making U-turns or for left turns access is seen as a generator of conflict points especially for school buses. However TWLTL is negative in the "inflate" part of the model.

Apart from drainage purposes, presence of curb and gutter (PCG) can be considered to prevent vehicles from accelerating beyond the travel lane. Table 2 shows PCG to be significant in both "crash" and "inflate" parts of the ZINB. The variable is positive in the "crash" portion but negative in the "inflate" portion. The common wisdom could have considered the presence of curb and gutter to reduce probability of school bus crashes, but in most cases segments with curb and gutter also have sidewalks which may skew the expectations. As sidewalks accommodate pedestrians, one could expect segments with curb and gutter to be associated with minor crashes resulting from hitting the curb walls. The positive coefficient might have been caused by the effect of population which also do influence crash frequency. Roadway segments in densely populated areas and which have high density of school buses often have curb and gutter compared to low population areas. Effect of curb and gutter as well as sidewalks can furthermore be linked with adjacent land uses.

Finally, it should be noted that school buses usually run in a designed route and mostly during peak hours to pick up and drop off students. Most of the bus drivers are familiar with the hazards along the route with the exception of new drivers. That means, some of bus crashes may therefore have nothing to do with road characteristics as found in modeling results discussed.

CONCLUSIONS

Based on the postulation that crashes follow a Bernoulli process with an unequal probability of independent events, the authors applied the zero-inflated negative binomial (ZINB) distributional model to analyze school bus crash data. The ZINB model is derived in Bayesian statistics by conjugating the negative binomial model. Statistically, the ZINB model could explain extra zero variability beyond the standard negative binomial model and is therefore capable of modeling a wider range of data variability especially for school bus crashes which are very rare than the Poisson and NB models. Comparison of ZINB and NB variable coefficients yielded almost similar sign (negative or positive coefficients) but ZINB resulted with slightly tighter standard errors hence slightly stronger z-values compared to NB. This recaps that for crash data such as those related to school buses, the use of zero inflated models is more applicable in lieu of standard count models. The results show that the ZINB model with a mean-dependent over-dispersion factor yields better performance and is recommended for use in school bus crash data modeling, with the caveat that the model may slightly over estimate the mean of crash frequency. This implies that, in high-risk roadway segments with low traffic exposure, the occurrence of school bus crashes will possibly deviate from the Poisson or NB process. In this case, the ZINB becomes more appropriate and provide more flexible modeling framework.

During the modeling process, some of the explanatory variables showed inconsistent signs and significance levels. The study found that high traffic volume (AADT), more number of through lanes (multilane section), presence of two-way left turn lanes, high posted speed limit sections, presence of curb and gutter have positive effect in increasing the probability of school bus related crashes. On the other hand, wider lanes, median, outside shoulder, inside shoulder, presence of posted school speed limits, presence of median (divided roadways), and presence of outside shoulders both have negative coefficients describing their effect in reducing the chances of school bus related crashes.

The study analyzed factors to consider when planning for school bus routes. Minimization of school bus crashes will be achieved by avoiding routes with roadway cross-section features found to influence probability of crashes. For instance, as found in the study, a school bus route planned through a multilane roadway with congested traffic volume, higher posted speed limits, TWLTL will have higher probability of being involved in a crash compared to section with opposite characteristics.

As stated in previous sections, local streets and non-state roads were not included in the analysis due to unavailability of the traffic counts, an essential element in crash modeling. Future studies should evaluate school bus crashes on these roadway classes, the findings might vary due to low traffic volumes, low operating speed as high volume of pedestrians.

REFERENCES

Aty, M. A. and A. E. Radwan, 2000. Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention* Vol. 32, No. 5, 633–642.

Bonneson, J. A. and P. T. McCoy, 1993. Estimation of safety at two-way stop-controlled intersections on rural highways. *Transportation Research Record* Vol. 1401, 83–89.

Cameron, A.C. and P.K. Trivedi, 1986. Econometric models based on count data: comparisons and applications of some estimators and tests. J. *Appl. Econometrics* Vol. 1, 29–53.

Chimba, D., T. Sando, and V. Kwigizile, 2010. Effect of bus size and operation to crash occurrences. *Accident Analysis & Prevention* Vol. 42, (6), 2063–2067.

Garber, N. J. and A. A. Ehrhart, 2000. The Effect of Speed, Flow, and Geometric Characteristics on Crash Rates for Different Types of Virginia Highways. Virginia Transportation Council.

Hardin, J. and J. Hilbe, 2000. *Generalized Linear Models and Extension*. Stata Corp.

Hauer, E. 2001. Overdispersion in modeling accidents on road sections and in empirical Bayes estimation. *Accident Analysis and Prevention* Vol. 33, 799–808.

Hauer, E., J.C.N. Ng, and J. Lovell, 1989. Estimation of safety at signalized intersections. *Transportation Research Record* Vol. 1185, 48–61.

Ivan, J. N., R. K. Pasupathy, and P. J. Ossenbruggen, 1999. Differences in causality factors for single and multi-vehicle crashes on two-lane roads. *Accident Analysis and Prevention* Vol. 31, No. 6, 695–704.

Johnson, N. L. and S. Kotz, 1969. *Discrete Distributions: Distributions in Statistics*. Wiley, New York, N.Y.

Joshua, S. C. and N. J. Garber, 1990. Estimating truck accident rate and involvements using linear and Poisson regression models. *Transportation Planning and Technology* Vol. 15, 41–58.

Jovanis P. P., and H. L. Chang, 1986. Modeling the relationship of accidents to miles traveled. *Transportation Research Record* Vol. 1068, 42–51.

Lapner, P. C., D. Nguyen, and M. Letts, 2003. Analysis of a school bus collision: mechanism of injury in the unrestrained child, *Can. J. Surg*. Vol. 46, No. 4, 269–272.

Lee, J. and F. Mannering, 2002. Impact of Roadside Features on the Frequency and Severity of Run-off-Roadway Accidents: Empirical Analysis. *Accident Analysis and Prevention* Vol. 34, 149–161.

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

Maher M. J., and I. Summersgill, 1996. A Comprehen-

sive Methodology for the Fitting of Predictive Accident Models. *Accident Analysis and Prevention* Vol. 28, 281–296.

Maycock, G. and R. D. Hall, 1984. Accidents at 4-arm roundabouts. Laboratory Report LR 1120, Crowthorne, Berks, U.K., Transport Research Laboratory.

McGeehan, J., J.L. Annest, M. Vajani, M.J. Bull, P.E. Agran, and G. A. Smith, 2006. School bus-related injuries among children and teenagers in the United States 2001–2003. *Pediatrics* Vol. 118, No. 5, 1978–1984.

Miaou, S., 1994. The relationship Between Truck accidents and Geometric Design of Road Sections: Poisson versus negative Binomial Regressions. *Accident Analysis and Prevention* Vol. 26, 471–482.

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

Miaou, S. and H. Lump, 1993. Modeling Vehicle Accidents and Highway Geometric Design Relationships. *Accident Analysis and Prevention* Vol. 25, 689–709.

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

Mullahy, J., 1986. Specification and testing of some modified count data models. *J. Econometrics* Vol. 33, 341–365.

Noland, R.B. and L. Oh, 2004. The Effect of Infrastructure and Demographic Change on Traffic-Related Fatalities and Crashes: A Case Study of Illinois County-Level Data. *Accident Analysis and Prevention* Vol. 36, 525–532.

Qin, X., J. N. Ivan, and N. Ravishanker, 2004. Selecting Exposure Measures in Crash Rate Prediction for Two-Lane Highway Segments. *Accident Analysis and Prevention* Vol. 36, 183–191.

Sawalha, Z., 2003. *Statistical Issues in Traffic Accident Modeling*. In Proceedings of the 82th Annual Meeting of the Transportation Research Board, January 12–16, Washington, D.C.

Shankar, V., J. Milton, and F. L. Mannering, 1996. Modeling accident frequency as zero-altered probability processes: an empirical inquiry. *Accident Analysis and Prevention* Vol. 29, No. 6, 829–837.

StataCorp LP, 2008. *Data Analysis and Statistical Software*. College Station, Texas.

Washington, S.P., M.G. Karlaftis, and F. L. Mannering, 2002. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman & Hall/CRC.

Yang, J. C., Peek-Asa, G. Cheng, E. Heiden, S. Falb, and M. Ramirez, 2009. Incidence and characteristics of school bus crashes and injuries. *Accident Analysis and Prevention* Vol. 41, 336–341.

Crash Injuries in Four Midwestern States: Comparison to Regional Estimates

MAHTAB GHAZIZADEH[1]
LINDA NG BOYLE[2,*]

[1]Dept. of Industrial and Systems Engineering
University of Wisconsin-Madison
3217 Mechanical Engineering Building
1513 University Ave
Madison, WI 53706
Phone: (319) 430-7583
Email: ghazizadeh@wisc.edu

[2]Associate Professor, Dept. of Industrial & Systems Engineering
Dept. of Civil & Environmental Engineering
University of Washington
G5 Mechanical Engineering Building
Seattle, WA , USA 98195-2650
Phone: (206) 616-0245
Fax: (206) 685-3072
Email: linda@uw.edu

[*] Corresponding author

ABSTRACT

This study used binary logit models to examine the crash factors that contribute to severe injuries to the drivers across four Midwestern states: Iowa, Kansas, Missouri, and Nebraska. The findings regarding the association between many crash factors (e.g., driver's age, gender, seat belt use, and alcohol use) and severe injuries are consistent with previous studies. However, the magnitude of the associations varies greatly with some outcomes not even significant in some states (e.g., adverse weather). Findings were then compared to those from regional crash estimates using the General Estimates System (GES) and differences were obtained for rural road crashes. The outcomes bring up issues on the appropriateness of implementing safety countermeasures based on geographical regions and underscore the need for standard crash reporting procedures.

INTRODUCTION

Many traffic regulations and countermeasures are aimed at reducing the risk of driver fatalities and injuries. However, traffic safety is still a major concern in the United States. U.S. crash data from the year 2008 show that over 37,000 people were killed and about 2.5 million were injured from motor vehicle crashes (NHTSA 2009). In one study, highway crashes were estimated to be about 3.2% of the total medical cost in the United States, and over 14% for those in the 15–24 years range (Miller, Lestina, and Spicer 1998). Although several studies have provided some insights on

KEYWORDS: injury severity, crash data, General Estimates System (GES), Midwestern crashes, rural areas, crash type

the driver, vehicle, and road and environmental factors associated with these motor vehicle crash injuries and fatalities (e.g., Bedard et al. 2002; Connor et al. 2004; Evans and Frick 1994; Huelke and Compton 1995; Kim et al. 1995; O'Donnell and Connor 1996), there are differences that exist across states and many of these differences correspond to the data used as well as the analytic techniques employed.

Driver characteristics related to elevated crash risks include age and experience (Zhang et al. 1998; Kweon and Kockelman 2003), weather conditions (Khattak, Kantor, and Council 1998; Khattak and Knapp 2001), alcohol impairment (Zador, Krawchuk, and Voas 2000; Keall, Frith, and Patterson 2004), and driver distraction (Klauer et al. 2006; Neyens and Boyle 2008; Violanti 1998). However, the patterns of injury risk do differ across regions. For example, a model of injury severity based on data from Hawaii showed no significance differences for age and gender (Kim et al. 1995), while studies on Wisconsin (Tavris, Kuhn, and Layde 2001) and Iowa do reveal differences in age and gender. Estimates from the Iowa crash data also differed from the national estimates (Hill and Boyle 2005). These findings demonstrate the impact of aggregating data to too high a level. That is, a model based on national data may not be able to capture patterns specific to a state or region.

The Midwestern states in the United States do have common characteristics, including many rural roads and sparsely populated areas. These rural areas also contribute to a large proportion of crash fatalities in the United States (NHTSA 2008b). A study on four Midwestern states, Kansas, Nebraska, South Dakota, and North Dakota, showed that there is an inverse relationship between motor vehicle crash fatality rates (per 100,000 persons) and population density (Muelleman and Mueller 1996). That is, the more sparse the population in these rural areas, the higher the fatality rates. A 5-mph increase in roadway speed limit increases the odds of fatalities and injuries (Baum, Lund, and Wells 1989; Renski, Khattak, and Council 1999). Although many studies tend to group this region into one cluster, there may be differences between these states with respect to traffic patterns.

The present study examines different factors surrounding traffic crashes and the severity of driver injuries within four Midwestern states: Iowa, Kansas, Missouri, and Nebraska. The analyses attempt to estimate the likelihood of severe injuries, given a crash has occurred. In other words, exposure data is not incorporated into the analysis, as the goal is not to estimate the general crash likelihood of driver groups or driving conditions, but instead to estimate the odds of severe outcomes when a crash occurs. Each state is examined individually and then compared to estimates from the regional model. It is hypothesized that the injury trends will be similar to those previously observed in other studies using Midwestern states, but the magnitude of such associations may differ. Comparisons will then be made to the representative sample at the region level of all 12 Midwestern states, extracted from the General Estimates System (GES) (NHTSA 2008a). Conclusions and the impacts on policies are then considered relative to the results of the analyses and comparisons.

METHODS

Data

Data for this study was obtained from the Departments of Transportation and Roads for Iowa, Kansas, Missouri, and Nebraska. The four databases contained information on crashes for the years 2001 to 2006. The scope of this

study was limited to passenger vehicles, so data from other vehicle types (e.g., buses, trucks, and motorcycles) were eliminated prior to analyses. Moreover, only data related to drivers (not passengers or pedestrians) were included in the analysis in order to achieve consistency for model comparisons. That is, passengers could be situated in different locations within the vehicle, but the driver is always in the driver seat. Each state also had different formats for their crash data. Hence, the databases were standardized and reformatted to facilitate comparisons. The usable crash records available for analysis (i.e., records with sufficient crash information) encompassed 78.33% of Iowa, 84.49% of Kansas, 85.68% of Missouri, and 70.53% of Nebraska's reported crashes. The differences noted are based on the amount of available information extracted from each state's database.

A model of the Midwestern region of the United States based on data from the National Automotive Sampling System (NASS)-GES for the same years (2001 to 2006) was then used as a comparison to the individual findings (NHTSA 2008a). The GES data is a stratified sample of crashes weighted to represent national crash patterns. The GES obtains its data from a national representative probability sample that is extracted from police accident reports (PARs). The sampling from PARs is accomplished in three stages: 1) sampling of geographic areas, which provides the Primary Sampling Units (PSUs), 2) sampling of police jurisdiction within each PSU, and 3) selection of crashes within the sampled police jurisdictions (NHTSA 2005).

Injury Level Classification in Crash Data

The classification of injury level in crash reports is based on the KABCO scale, which was introduced by the National Safety Council in the late 1960s (Compton 2005). This rating system, also used in GES (NHTSA 2005), categorizes occupant injuries into five groups: fatal (K), incapacitating (A), non-incapacitating (B), possible injuries or complaint of pain (C), and not injured (O). In addition, categories such as "unknown" and "not reported" are included for some states because discerning the level of injury may not always be possible. All four states examined in this study employ the KABCO scale; although there were some differences in definitions (table 1).

Statistical Analysis

Four separate models were developed to examine the factors that may increase the likelihood of a severe injury for each state. Although an ordering to the severity level may initially seem obvious, there are two general problems with employing ordered models in the context of crash injury severity as noted by Savolainen and Mannering (2007): 1) non-injury crashes may be underreported in crash data and this can lead to biased coefficient estimation by the model, and 2) ordered models restrict variable influences. In other words, the hypothesis that the parameters from an ordered logit model are equal across all the levels of the dependent variable was rejected. Rather, many researchers have used multinomial logit models to examine the severity of occupant injuries (Awadzi et al. 2008; Bedard et al. 2002; Khorashadi et al. 2005; Watt et al. 2006).

The primary goal of this study was to compare outcomes across multiple states and the use of the "KABCO" scale was different in each one (see table 1). Hence, a more simplistic binary logistic regression model (or logit model), as used by Al-Ghamdi (2002), is employed to provide insights on injuries while also allowing a clear comparison across the four states, which can then be generalized to the regional level. The injury severity levels were therefore grouped into two general categories of "severe"

TABLE 1 Detailed KABCO Injury Categories in Crash Databases

Crash database		Iowa	Kansas	Missouri	Nebraska	GES national sampled data
KABCO injury levels	K	Fatal	Fatal injury	Fatal	Fatal	Fatal injury
	A	Incapacitating	Disabled (incapacitating)	Disabling injury	Disabling	Incapacitating injury
	B	Non-incapacitating	Injury, not incapacitating	Evident injury (not disabling)	Visible	Non-incapacitating injury
	C	Possible	Possible injury	Probable injury (not apparent)	Possible	Possible injury
	O	Uninjured	Not injured	Not apparent	No injury	No injury
Augmenting categories		Unknown	Unknown	Unknown	na	Died prior to crash
		Not reported	na	na	na	Unknown if injured
		na	na	na	na	na

KEY: na = not available

(including codes K and A) and "non-severe" injuries (including codes B, C, and O) and examined using simultaneous binary logit models developed with SAS (Statistical Analysis System) version 9.1 and the CATMOD procedure (Allison 1999). The CATMOD procedure is used to estimate the likelihood of a driver sustaining severe injuries when compared to non-severe injuries. The model is represented in equation 1.

$$\log\left(\frac{\Pr(Y_i = m)}{\Pr(Y_i = k)}\right) = \sum_{r=1}^{R} (\beta_{mr} - \beta_{kr}) X_{ir} \qquad (1)$$

Where X_{ir} is the value of the explanatory variable r for driver i, and β_{jr} is the coefficient associated with the rth variable ($r = 1, \ldots, R$) for the jth injury severity level. Y_i is a random variable whose value ($j = 1$ or 2) indicates the severity level of the injuries sustained by driver i. The CATMOD procedure uses maximum likelihood estimation (MLE) and outputs logarithmic ratio estimates of the likelihood of severe (versus non-severe) injuries, based on the levels of each explanatory variable. By exponentiating the logarithmic ratio estimates, odds ratios for sustaining severe (versus non-severe) injuries were obtained. The adjusted odds ratios [AORs] are odds ratios that have been adjusted for the other explanatory variables in the model because they are calculated based on a multivariate model that controls for other factors. Wherever the logarithmic ratio estimate is positive, exponentiating this estimate will give a value greater than 1, and thus the odds of sustaining severe injuries are higher than non-severe injuries. Conversely, when this estimate is negative, exponentiating will give a value less than 1, and the odds of having a severe injury are less than a non-severe injury. The likelihood ratio test was used to compare the goodness of fit (Cochran 1952) of the fitted model to a saturated model (i.e., backward elimination) (Ananth and Kleinbaum 1997). A high p-value would suggest that the fitted model was a good fit and that no significant terms were omitted.

Explanatory Variables

The statistical models included explanatory variables shown to have an impact on the

likelihood of a crash or a severe injury in a crash. Drivers were categorized into three age groups: 24 years old and younger (younger drivers), aged between 25 and 65 (reference group), and drivers older than 65 (older drivers) as similarly done in other studies (Zhang et al. 1998; Khattak, Kantor, and Council 1998; Farmer, Braver, and Mitter 1997). Weather conditions were divided into two categories, normal and adverse. The adverse weather category encompassed situations where rain, snow, freezing rain, fog/smoke, mist, sleet, severe winds, blowing sand/soil/dirt, or combinations of these conditions were present. If none of the above conditions were present, then weather was labeled as normal. Lighting was considered in two categories, daylight and non-daylight situations with the latter consisting of night, dawn, and dusk. Roadway speed limit was set up into three groups: less than 35 mph, between 35 and 55 mph, and higher than 55 mph. The categories used for weather and lighting conditions are consistent with those used in similar studies (Khattak, Kantor, and Council 1998; Zhang et al. 1998; Abdel-Aty 2003). The point of impact variable (the first point that produced damage or injury) was examined using five categories; front, driver side, passenger side, top/under, and rear of the car.

Five crash types were considered: rear-end, head-on, angular, sideswipe, and single-vehicle crashes. Angular, rear-end, sideswipe, and head-on crashes are the four categories of "collision with motor vehicle in transport" used by US DOT, while single-vehicle crashes correspond to "collisions with fixed object" and "collision with object not fixed" (NHTSA 2009). In addition to crash type, the (initial) crash point of impact was included for states whose crash database supported this variable (i.e., Iowa and Nebraska).

Two driver-related factors were also of interest given the abundance of literature demonstrating increase crash risk, driver distraction and blood-alcohol content (BAC). However, the crash databases did not include sufficient information regarding these two factors for the years examined. More specifically, the proportion of crashes that included any details about the distraction-related factors encompass only 1.27% in Iowa, 1.33% in Kansas, 1.18% in Missouri, and 0.81% in Nebraska. Surprisingly, driver BAC information was not available in any of the states' databases. Those states that did include this variable had a large proportion of non-reporting (e.g., about 51% of Iowa crashes with drivers under the influence of alcohol lacked BAC level). Considering these limitations, only the more general factor of "being under the influence of alcohol or drugs" (yes or no) was used in the analyses. It should be noted that several other factors could contribute to the severity of injuries sustained by driver. For example, vehicle size and mass are known to influence driver fatality (Evans and Frick 1992, 1994). However, this information was not available in the datasets that were used for this study.

RESULTS

State Level

A separate model was developed for each state, with each state's databases including the majority of variables of interest. The Iowa crash database included all variables of interest. Kansas did not have sufficient information regarding point of impact, while Nebraska lacked data on air bag deployment. Missouri did not have information on drug use, air bag deployment, and point of impact. All models fitted well based on the likelihood ratio test. The significance level was set to 0.0001.

There were similar demographic patterns across the four states (table 2). Drivers' mean age ranged from 36.64 (in Nebraska) to 37.90 (in Missouri). The proportion of female drivers ranged from 43.23% (Kansas) to 45.09% (Nebraska). Among crash types, angular and rear-end crashes were the most common in each of the five databases, comprising 65–85% of crashes (table 3).

Driver and Vehicle Characteristics

The four binary logit models are shown in table 4, with a general finding that female drivers were more susceptible to serious injuries in the four Midwestern states examined. There were similar estimates between Iowa and Kansas (adjusted odds ratios [AORs] = 1.07 and 1.08, respectively) and between Missouri and Nebraska (AORs = 1.21 and 1.24, respectively). With respect to driver age, younger drivers (younger than 25) were less likely to sustain serious injuries when compared to the middle-aged group (aged 25-65). Older drivers were more likely to be severely injured. There was also an age and gender interaction in Missouri only, with young females being less likely to sustain severe injuries compared to middle-aged male drivers (AOR = 0.96).

Passengers were shown to have a protective effect in Iowa and Kansas, with drivers being less severely injured driving with passengers when compared to driving alone. In contrast, drivers with passengers in Missouri were more likely to sustain severe injuries. No significant association was observed between injury severity and passengers in Nebraska.

TABLE 2 Descriptive Statistics of State Crash Data

			Gender (%)			
State	**Number of crashes**	**Mean age (SD)**	**Male**	**Female**	**Seat belt use (%)**	**Drug/alcohol use (%)**
Iowa	370,428	37.85 (18.46)	55.82	44.18	57.82	4.01
Kansas	598,070	36.66 (17.34)	56.77	43.23	83.84	5.68
Missouri	1,465,219	37.90 (17.91)	56.43	43.57	82.25	2.91
Nebraska	271,445	36.64 (17.23)	54.91	45.09	79.1	1.52

TABLE 3 Frequencies of Crash Types in Crash Databases

	Iowa		**Kansas**		**Missouri**		**Nebraska**		**GES (Midwest)**	
Crash type	**Count**	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**
Angular	137,687	37.17	196,864	32.9	487,583	33.28	104,099	38.35	4,424,740	36.11
Rear-end	114,150	30.82	190,377	31.8	573,904	39.17	125,349	46.18	4,446,780	36.29
Sideswipe	48,569	13.11	42,100	7.04	145,999	9.96	39,316	14.48	1,062,281	8.67
Head-on	8,951	2.42	11,005	1.84	41,184	2.81	2,053	0.76	239,021	1.95
Single-vehicle	48,569	16.49	157,724	26.4	216,549	14.78	628	0.23	2,079,440	16.97
All crashes	370,428		598,070		1,465,219		271,445		12,252,262*	

*Based on weighted observations

As expected, there was a higher likelihood of a severe injury when the driver did not use a seat belt, and this was consistently observed in all four states with Nebraska and Missouri being more similar in odds (AORs = 2.70 and 2.74, respectively) and Iowa and Kansas having higher odds ratios (3.59 and 4.24, respectively). Air bag deployment data was available in Iowa and Kansas only, and drivers in these two states were more likely to be severely injured with an airbag deployment. In each state, drivers under the influence of alcohol or drug were significantly more likely to sustain severe injuries compared to sober drivers. The magnitude of this effect varied slightly from 1.32 (Kansas) to 1.74 (Nebraska).

Crash Types and Points of Impact

The odds of sustaining severe injuries were higher for head-on crashes when compared to rear-end crashes, and ranged from 3.18 in Iowa to 4.92 in Kansas. Drivers in sideswipes were less likely to sustain severe injuries in all four states, with quite similar odds ratios (from 0.38 to 0.50). Observations for single-vehicle crashes were consistent for Iowa and Missouri, indicating higher likelihoods of serious injuries (AORs = 1.29 and 1.93, in Iowa and Missouri, respectively). However, the odds of having severe injuries were not significantly different between single-vehicle and rear-end crashes in Kansas and Nebraska. No significant difference was observed between angular and rear-end crashes in any of the states in terms of severe injury odds. In Iowa and Nebraska, drivers in crashes with impacts on the driver side were 1.16 and 1.82 times, respectively, more likely to sustain severe injuries compared to those whose vehicles were impacted on the rear side. No other significant differences were observed with respect to crash types and points of impact.

Environmental Conditions

In all four states, drivers involved in crashes in rural settings were more likely to sustain severe injuries when compared to those having crashes in urban areas (AORs ranged from 1.71 (Iowa) to 2.55 (Missouri)). Non-dry surfaces were associated with lower likelihoods of severe crashes in all four states, i.e., the odds of sustaining severe injuries on non-dry surfaces were between 0.81 (for Nebraska) and 0.92 (for Missouri) compared to dry surfaces. There was an interaction effect between crash location and crash type, with drivers more likely to be severely injured if they were involved in head-on crashes in rural settings (AORs = 1.26, 1.48, and 1.20 in Iowa, Kansas, and Missouri, respectively), compared to those involved in rear-end crashes in urban settings. By contrast, drivers in single-vehicle crashes in rural settings were less likely to sustain severe injuries (AORs = 0.72, 0.48, and 0.83 in Iowa, Kansas, and Missouri, respectively). In Missouri, two additional contrasts were significant as well; drivers in angular crashes in rural settings were 0.88 times less likely and those in sideswipes were 1.22 times more likely to have severe injuries.

Findings regarding lighting conditions were not consistent across the states. In Iowa and Kansas, drivers were slightly less likely to sustain severe injuries in crashes occurring in non-daylight situations, i.e., during night, dawn, or dusk (AORs = 0.94 and 0.92, respectively). In Nebraska, contrary to Iowa and Kansas, the odds of sustaining severe injuries were higher in daylight hours (AOR = 1.11). The Missouri model showed no significant association between lighting and injury severity. Weather condition at the time of crash was a significant factor only in Missouri, where drivers were slightly less likely to be severely injured in crashes occurring in adverse weather conditions (AOR = 0.94).

TABLE 4 State Models for the Likelihood of Severe Injuries

	Iowa				Kansas			
Parameter	Estimate	SE	χ^2	Adjusted OR	Estimate	SE	χ^2	Adjusted OR
Intercept	-2.21	0.03	4012.3	0.11	-3.33	0.04	5816.6	0.04
Head-on crashes	1.16	0.04	1049.6	3.18	1.59	0.04	1517.2	4.92
Angular crashes	0.07	0.02	ns	1.08	0.10	0.03	ns	1.10
Sideswipes	-0.70	0.04	346.8	0.5	-0.89	0.05	271.1	0.41
Single-vehicle crashes	0.26	0.02	114.9	1.29	0.09	0.03	ns.	1.10
Rural settings	0.54	0.02	1097.4	1.71	0.65	0.03	575.1	1.92
Female drivers	0.07	0.01	22.9	1.07	0.08	0.01	43.2	1.08
Age < 25	-0.36	0.02	346.3	0.7	-0.26	0.02	180.5	0.77
Age > 65	0.40	0.02	279.7	1.49	0.33	0.03	164.7	1.39
Passenger(s) present in the car	-0.10	0.01	64.0	0.9	-0.07	0.01	30.3	0.93
Adverse weather	0.02	0.02	ns	1.02	0.05	0.03	ns	1.05
No daylight	-0.07	0.01	23.8	0.94	-0.09	0.01	42.4	0.92
Non-dry surface	-0.17	0.02	87.8	0.85	-0.18	0.02	52.9	0.84
Under influence of alcohol/drug	0.50	0.02	753.7	1.64	0.28	0.02	250.9	1.32
No seat belt in use	1.28	0.02	2813.8	3.59	1.44	0.03	2105.1	4.24
Air bag deployed	0.76	0.02	1254.0	2.13	0.27	0.03	115.7	1.32
Speed limit < 35 mph	-0.48	0.03	361.5	0.62	-0.93	0.03	1348.6	0.39
Speed limit > 55 mph	0.29	0.03	110.5	1.34	0.79	0.03	837.6	2.20
Point of impact: front	-0.03	0.02	ns	0.97	na	na	na	na
Point of impact: driver side	0.15	0.03	30.0	1.16	na	na	na	na
Point of impact: passenger side	-0.11	0.03	ns	0.9	na	na	na	na
Point of impact: top/under	0.10	0.06	ns	1.1	na	na	na	na
Head-on crashes in rural settings	0.23	0.03	44.6	1.26	0.39	0.05	57.1	1.48
Angular crashes in rural settings	0.04	0.02	ns	1.04	0.01	0.04	ns	1.01
Sideswipes in rural settings	0.07	0.04	ns	1.07	0.14	0.07	ns	1.15
Single-vehicle crashes in rural settings	-0.33	0.02	217.9	0.72	-0.73	0.03	546.5	0.48
Female drivers younger than 25	na	na	na	na	na	na	na	na
Female drivers older than 65	na	na	na	na	na	na	na	na
Likelihood ratio	19,894.74				9,361.05			
Number of observations	370,428				598,070			

NOTE: All parameters are significant at p ≤ 0.0001 unless otherwise noted (ns). For variables not found statistically significant, no contrast estimate is reported.

KEY: na = not applicable; ns = not significant

Continued, next page

TABLE 4 State Models for the Likelihood of Severe Injuries *(continued)*

	Missouri				Nebraska			
				Adjusted				Adjusted
Parameter	**Estimate**	**SE**	**χ^2**	**OR**	**Estimate**	**SE**	**χ^2**	**OR**
Intercept	-3.03	0.02	18146.8	0.05	-2.69	0.10	738.2	0.07
Head-on crashes	1.24	0.02	2759.4	3.44	1.45	0.09	277.6	4.26
Angular crashes	-0.06	0.02	ns	0.95	0.18	0.06	ns	1.20
Sideswipes	-0.96	0.04	641.7	0.38	-0.68	0.07	90.4	0.50
Single-vehicle crashes	0.66	0.02	1529.0	1.93	-0.22	0.23	ns	0.80
Rural settings	0.94	0.01	4211.4	2.55	0.76	0.02	1325.1	2.15
Female drivers	0.19	0.01	618.1	1.21	0.21	0.02	143.4	1.24
Age < 25	-0.21	0.01	420.0	0.81	-0.31	0.03	120.1	0.73
Age > 65	0.25	0.01	336.2	1.28	0.35	0.04	101.0	1.42
Passenger(s) present in the car	0.49	0.01	4606.2	1.64	-0.01	0.02	ns	0.99
Adverse weather	-0.06	0.01	20.1	0.94	0.07	0.04	ns	1.07
No daylight	0.00	0.01	ns	1.00	0.11	0.02	26.9	1.11
Non-dry surface	-0.09	0.01	78.6	0.92	-0.21	0.03	51.6	0.81
Under influence of alcohol	0.31	0.01	927.4	1.36	0.56	0.04	226.3	1.74
No seat belt in use	1.01	0.02	4020.4	2.74	0.99	0.04	550.3	2.70
Air bag deployed	na	na	na	na	na	na	na	na
Speed limit < 35 mph	-0.59	0.01	1833.5	0.56	-0.54	0.04	191.6	0.58
Speed limit > 55 mph	0.47	0.01	1969.0	1.60	0.74	0.04	351.8	2.09
Point of impact: front	na	na	na	na	-0.08	0.07	ns	0.92
Point of impact: driver side	na	na	na	na	0.60	0.07	68.8	1.82
Point of impact: passenger side	na	na	na	na	0.15	0.08	ns	1.16
Point of impact: top/under	na	na	na	na	0.01	0.24	ns	1.01
Head-on crashes in rural settings	0.19	0.03	48.8	1.20	na	na	na	na
Angular crashes in rural settings	-0.12	0.02	39.5	0.88	na	na	na	na
Sideswipes in rural settings	0.20	0.04	22.1	1.22	na	na	na	na
Single-vehicle crashes in rural settings	-0.19	0.02	109.7	0.83	na	na	na	na
Female drivers younger than 25	-0.05	0.01	21.0	0.96	na	na	na	na
Female drivers older than 65	0.04	0.01	ns	1.04	na	na	na	na
Likelihood ratio	14,845.97				7,880.31			
Number of observations	1,465,219				271,445			

The likelihood of driver's sustaining severe injuries also increased on roads with higher posted speed limits. The odds ratios for severe injuries on roads with lower speed limits (less than 35 mph) compared to the reference speed limit (35–55 mph) were very similar for Iowa, Missouri, and Nebraska (0.62, 0.56, and 0.58, respectively), while Kansas revealed a slightly lower odds ratio (0.39). For roads with higher speed limits, the odds ratios ranged from 1.34 in Iowa to 2.20 in Kansas.

Comparisons Across States

There were some common and consistent findings across all four Midwestern states for various driver characteristics (gender, age, alcohol and drug use, and seat belt use), as well as environmental conditions including surface condition, posted speed limit, and rural/urban settings. However, differences were observed for crash type. Single-vehicle crashes significantly impacted the likelihood of a severe injury in Iowa and Missouri, but in Kansas and Nebraska, there was no difference between single-vehicle and rear-end crashes.

Similarly, the interaction between crash type and location (rural/urban) was significant for all states but Nebraska. The interaction between age and gender, on the other hand, was only significant in Missouri. Results pertaining to weather condition showed significant differences only in Missouri. Driving in nondaylight conditions was associated with a decrease in injuries in both Iowa and Kansas, but increased injuries in Nebraska. No significant difference was observed in Missouri. Passengers were found to be similarly associated with a protective effect for drivers in Iowa and Kansas, but an increase in severe injuries in Missouri. No significance was observed in Nebraska.

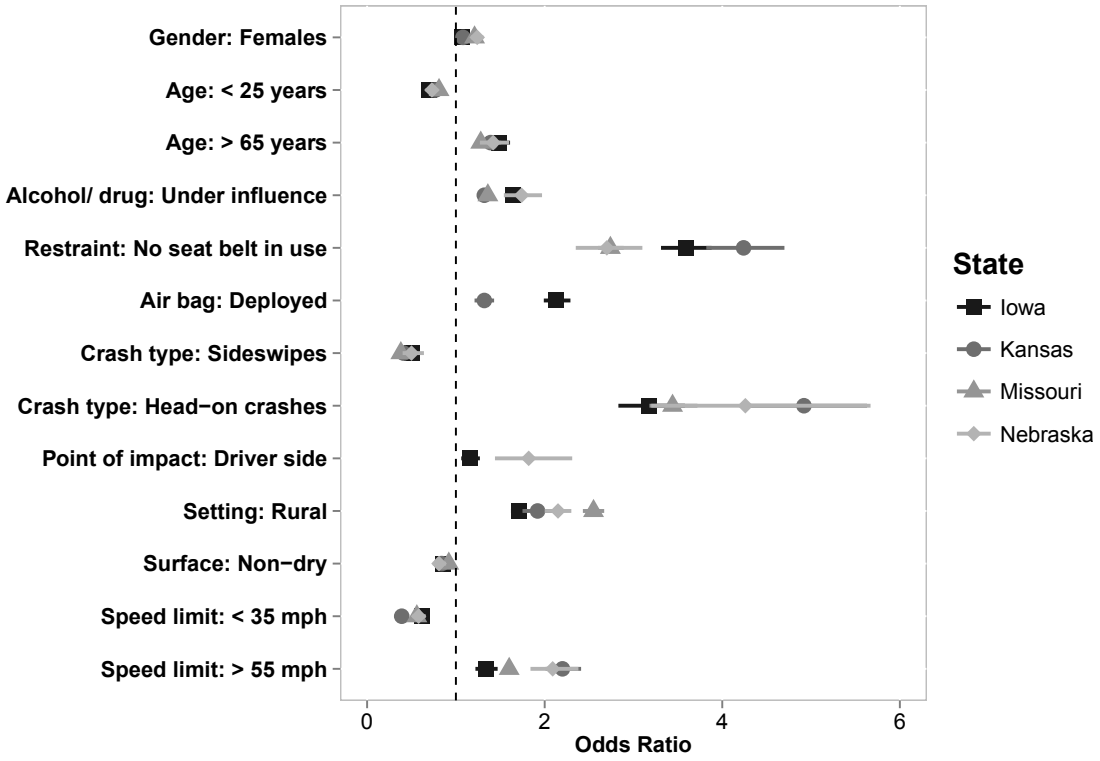
Point of impact information was only available for Iowa and Nebraska, where it produced patterns in the same direction (although the magnitude was different). Air bag deployment information was also available in Iowa and Nebraska only. Here again, results indicated associations in the same direction but different magnitude.

Figure 1 summarizes the results (point estimates and confidence intervals) for variables that showed a significant association with severe injuries across the four states examined. Note that in many cases, the confidence intervals overlap heavily (e.g., age, under the influence of alcohol/drug) which indicates great similarities across the four states examined. In other cases, however, the estimates are more diverse (e.g., seat belt use, head-on crashes) which indicates considerable differences among the states. These findings motivate developing a Midwestern crash injury severity model, extract driver injury patterns from it, and compare them to the four state models to assess the extent to which sampled crash databases can describe injury patterns across the region.

Regional Level

The Midwestern region of the United States was examined using the sampling region established by the National Automotive Sampling System (NASS) and collected as part of the GES data (NHTSA 2005). This region consisted of 12 states (i.e., Ohio, Indiana, Illinois, Michigan, Wisconsin, Minnesota, North Dakota, South Dakota, Nebraska, Iowa, Missouri, and Kansas). The goal of this analysis was to assess the level of agreement between the outcomes observed from the states' crash databases and the outcomes from a sample of crashes in the same region (GES).

FIGURE 1 State Estimates and Confidence Intervals for the Association Between Crash Variables and Severe Injuries



Midwestern crash data used in the injury severity model included 97,070 weighted records, representing 12,252,262 crashes. The binary logit model demonstrated a good fit (based on the likelihood ratio test*)*. Results obtained from this regional model are summarized in table 5 (only estimates pertaining to significant factors are included in the table) and discussed in the forthcoming section.

Driver and Vehicle Characteristics

Driver gender was found significant in the regional model with female drivers being 8% more susceptible to severe injuries than males (AOR = 1.08). Driver age was a significant factor as well. Younger drivers were less likely to be severely injured while older drivers were more likely to sustain severe injuries when compared to those aged between 25 and 65 (AORs = 0.71 and 1.47, respectively). The interaction between age and gender was also significant; younger female drivers were less and female drivers aged more than 65 were more likely to be severely injured in car crashes (AORs = 0.95 and 1.03).

Restraint use was also significant. The likelihood of having severe injuries for drivers with no restraint was 5.29 times more than drivers wearing seatbelts. As expected, air bag deployment was associated with severe injuries (AOR = 3.42). Drivers under the influence of alcohol or drugs were found to be 2.47 times more likely to have severe injuries compared to sober drivers. Conversely, drivers with passengers in their cars were slightly less likely to be seriously injured compared to drivers who traveled alone (AOR = 0.97).

Crash Types and Points of Impact

Drivers involved in head-on crashes were 2.82 times more likely to have severe injuries com-

pared to those in rear-end crashes. Drivers in single-vehicle and angular crashes were also more likely to sustain severe injuries (AORs = 1.86 and 1.10, respectively). As expected, sideswipes were mainly associated with minor or no injuries (AOR = 0.34).

Vehicles impacted on the driver side were more likely to have a severely injured occupant than those vehicles impacted on the back (AOR = 1.82). Other areas of the vehicle (i.e., front, passenger side, top, or undercarriage) were associated with lower likelihoods of severe injuries, compared to rear of the vehicle (AORs = 0.92, 0.90, and 0.75, respectively).

Environmental Conditions

The regional model revealed that drivers were more likely to sustain severe injuries in crashes occurring in non-daylight (i.e., dark, dawn, or dusk) conditions compared to crashes during daylight hours (AOR = 1.08). By contrast, crashes on non-dry surfaces (e.g., snow covered, icy, wet, dirty) were associated with less likelihood of severe injuries (AOR = 0.82). Weather conditions (i.e., adverse versus normal weather) were found insignificant.

Crashes in rural settings were significantly less injurious for drivers (AOR = 0.92). The interaction between crash type and rural or urban setting was also significant; drivers who had been in angular and single-vehicle crashes in rural settings were more likely to be seriously injured (AORs = 1.08 and 1.09), and those in sideswipes were less likely to have severe injuries (AOR = 0.71). This interaction was insignificant for head-on crashes.

Roadway speed limit was found significant: drivers involved in crashes on roadways with speed limits lower than 35 mph were 0.55 times less likely to have severe injuries compared to those in crashes on roads with a 35 to 55 mph speed limit. Crashes on roadways with posted speed limits higher than 55 mph were 1.47 times more likely to result in severe injuries than those on roadways with speed limits between 35 and 55 mph.

State and Regional Level Comparison

The goal of comparing the state outcomes with the sampled data collected as part of GES is to assess the capability of gaining insights on the Midwestern states when aggregated to the regional level. It should be noted that the GES data for the Midwest does cover 12 states within the region. The additional eight Midwestern states are Ohio, Indiana, Illinois, Michigan, Wisconsin, Minnesota, North Dakota, and South Dakota. GES does not provide data at the state level and as such, it was not possible to isolate the four states for which the individual analyses had been done.

The odds ratios (and corresponding confidence intervals) for the parameter estimates common across the four states and at the regional level are listed in table 6 and are graphically depicted in figure 2. The greatest similarities are for driver age and roadway surface condition, where the odds ratios estimated by the four state models are close and the odds ratios calculated by the GES-based model fall in their range. The same pattern is evident for the contrast between lower (less than 35 mph) and reference (35–55 mph) speed limits. For driver gender, the odds ratio calculated for the contrast between female and male drivers (1.08) is equal to the odds ratio for the same contrast in Kansas and very close to that of Iowa (1.07); however, the value of the odds ratio for this contrast is higher for Missouri and Nebraska (1.21 and 1.24, respectively). For higher speed

TABLE 5 Regional Model for the Likelihood of Severe Injuries

Parameter	Estimate	SE	χ2	Adjusted OR
Intercept	-3.36	0.017	37,015.7	0.03
Head-on crashes	1.04	0.011	8,500.3	2.82
Angular crashes	0.09	0.007	197.7	1.1
Sideswipes	-1.09	0.017	4,261.2	0.34
Single-vehicle crashes	0.62	0.007	7,491.5	1.86
Rural settings	-0.09	0.005	286.3	0.92
Female drivers	0.07	0.004	310.8	1.08
Age < 25 years old	-0.34	0.006	3,820.6	0.71
Age > 65 years old	0.38	0.007	2,809.7	1.47
Passenger(s) present in the car	-0.03	0.004	62.5	0.97
Adverse weather	0.01	0.006	1.1	1.01
No daylight	0.08	0.003	480.6	1.08
Non-dry surface	-0.19	0.005	1,423.6	0.82
Under influence of alcohol/drug	0.91	0.023	1,557.2	2.47
No restraint in use	1.67	0.009	32,223.5	5.29
Air bag deployed	1.23	0.005	62,029.3	3.42
Speed limit < 35 mph	-0.59	0.006	10,441.2	0.55
Speed limit > 55 mph	0.39	0.007	2,872.1	1.47
Point of impact: front	-0.09	0.009	93.5	0.92
Point of impact: driver side	0.6	0.01	3,307.1	1.82
Point of impact: passenger side	-0.11	0.011	97.3	0.9
Point of impact: top/under	-0.29	0.032	82	0.75
Head-on crashes in rural settings	-0.03	0.011	9.1	0.97
Angular crashes in rural settings	0.08	0.006	151.1	1.08
Sideswipes in rural settings	-0.35	0.016	465.6	0.71
Single-vehicle crashes in rural settings	0.08	0.007	155.5	1.09
Female drivers aged less than 25	-0.05	0.005	89.9	0.95
Female drivers aged more than 65	0.03	0.007	17.8	1.03
Likelihood ratio			438,835.9	
Number of unweighted observations				97,070
Number of weighted observations				12,252,262

limits (above 55 mph), the odds of sustaining severe injuries is 1.47 based on the GES model, which is between the odds ratios calculated for the states of Iowa and Missouri (1.34 and 1.60, respectively). However, the odds ratios estimated for the same contrast in Kansas and Nebraska are considerably higher (2.20 and 2.09, respectively).

In many cases, there was general agreement among the state models and the GES-based model on the association between certain levels of a factor (e.g., no restraint in use by driver) and severe injuries, but as expected, the strength of such association was not always similar. For crash type, the likelihood of a head-on crash sustaining greater injuries

TABLE 6 Parameters Related to the Findings Across the Four Midwestern States and the Region

Parameter		Logit models (AOR and confidence intervals)				
		Iowa	**Kansas**	**Missouri**	**Nebraska**	**Midwest**
Crash type	Head-on crashes	3.18	4.92	3.44	4.26	2.82
		(2.83, 3.58)	(4.30, 5.63)	(3.18, 3.72)	(3.20, 5.67)	(2.72, 2.92)
	Sideswipes	0.50	0.41	0.38	0.50	0.34
		(0.44, 0.56)	(0.34, 0.49)	(0.34, 0.43)	(0.40, .64)	(0.32, 0.36)
(compared to rear-end)						
Rural setting		1.71	1.92	2.55	2.15	0.92
(compared to urban)		(1.62, 1.80)	(1.75, 2.10)	(2.43, 2.67)	(2.00, 2.30)	(0.90, 0.93)
Females		1.07	1.08	1.21	1.24	1.08
(compared to males)		(1.02, 1.12)	(1.04, 1.13)	(1.18, 1.25)	(1.17, 1.31)	(1.06, 1.09)
Driver age	< 25 years	0.70	0.77	0.81	0.73	0.71
		(0.65, 0.74)	(0.73, 0.82)	(0.79, 0.84)	(0.66, 0.80)	(0.70, 0.72)
	> 65 years	1.49	1.39	1.28	1.42	1.47
		(1.38, 1.61)	(1.27, 1.51)	(1.23, 1.34)	(1.27, 1.60)	(1.43, 1.50)
(compared to 25-65)						
Non-dry surface		0.85	0.84	0.92	0.81	0.82
(compared to dry surface)		(080, 0.90)	(0.77, 0.91)	(0.89, 0.95)	(0.74, 0.89)	(0.81, 0.84)
Alcohol/ drug impairment		1.64	1.32	1.36	1.74	2.47
(compared to sober driving)		(1.55, 1.74)	(1.25, 1.40)	(1.31, 1.40)	(1.54, 1.97)	(2.29, 2.67)
No restraint in use		3.59	4.24	2.74	2.70	5.29
(compared to seat belt in use)		(3.31, 3.88)	(3.82, 4.70)	(2.60, 2.89)	(2.35, 3.10)	(5.13, 5.45)
Speed limit	< 35 mph	0.62	0.39	0.56	0.58	0.55
		(0.57, 0.67)	(0.36, 0.43)	(0.53, 0.58)	(0.51, 0.66)	(0.54, 0.56)
	> 55 mph	1.34	2.20	1.60	2.09	1.47
		(1.22, 1.47)	(2.01, 2.41)	(1.55, 1.66)	(1.84, 2.38)	(1.44, 1.51)
(compared to 35-55 mph)						
Air bag deployed		2.13	1.32	NA	NA	3.42
(compared to no air bag deployment)		(1.99, 2.29)	(1.21, 1.43)			(3.36, 3.47)
Point of impact: Driver side		1.16	NA	NA	1.82	1.82
(compared to rear side)		(1.06, 1.27)			(1.44, 2.31)	(1.76, 1.89)

when compared injuries associated with rear-end crashes (OR = 2.82) was lower at the regional level than at the state level. The same is observed for sideswipes where the odds ratio of sustaining serious injuries (0.34) is lower at the regional level than at the individual states' models (range of 0.38 to 0.50). On the contrary, the GES-based odds ratios for alcohol and drug use and restraint use are higher than the highest odds ratios found in the individual states' models, indicating stronger associations between being under the influence of alcohol or drugs and having no restraint in use, and sustaining severe injuries by drivers. No confidence interval overlap is evident between the GES-based Midwestern model and the individual states models.

The air bag deployment factor could only be incorporated in the models of Iowa and Kansas due to the unavailability of precise data for the other two states. The odds of having serious injuries for cases in which air bags had been deployed were 3.42 times the cases without air bag deployment, based on the GES Midwestern model. Iowa and Kansas models showed weaker incompatible contrasts; i.e., odds ratios of 2.13 and 1.32, respectively. The comparison of confidence intervals revealed no overlap between the results of the three models. Therefore, the observations for air bag deployment yield no consensus for the Midwestern states considered in this study.

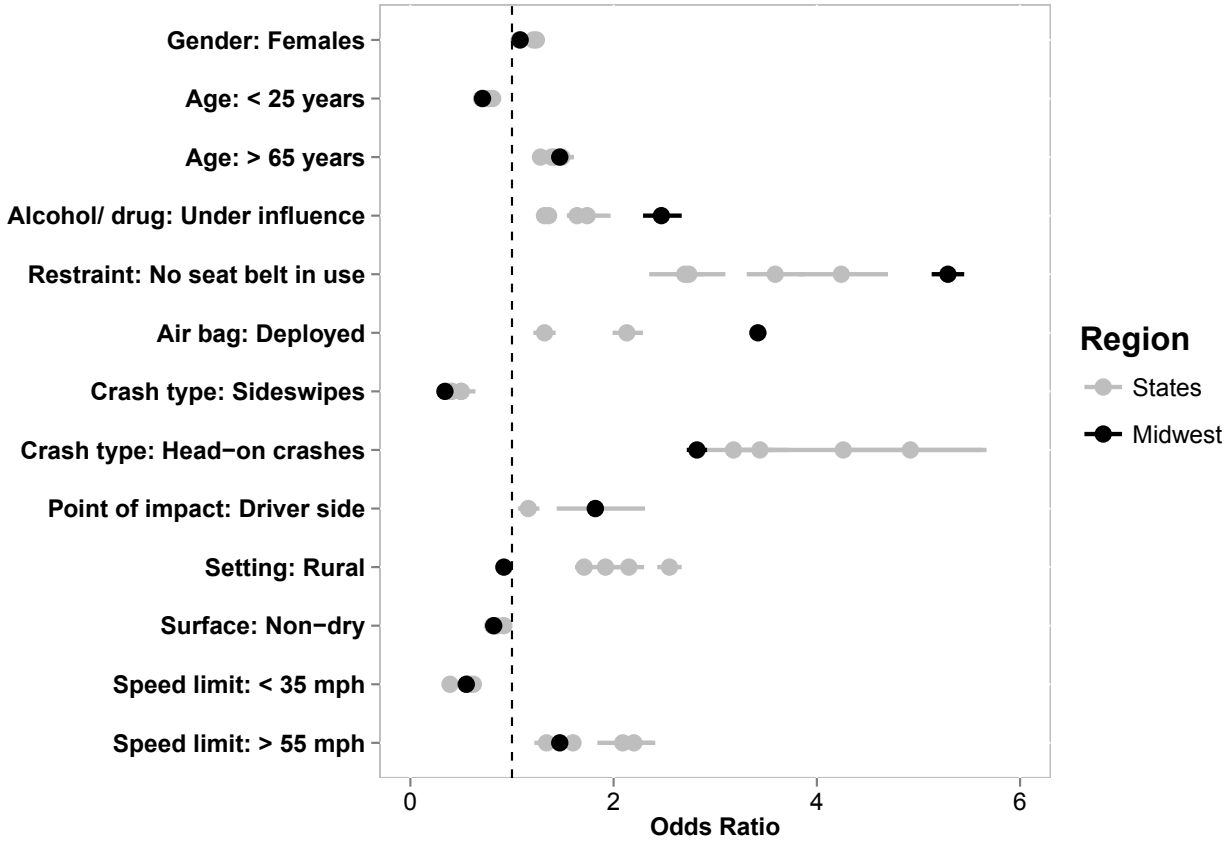
As noted earlier, point of impact was only available in for Iowa and Nebraska. The contrast between driver side and rear side of the vehicle was significant in predicting driver injury severity for both states, indicating higher likelihoods of serious injuries for drivers whose cars were impacted on driver side versus those involved in crashes in which the rear of the car was affected (odds ratios of 1.16 for Iowa and 1.82 for Nebraska). While the pattern observed in Nebraska was exactly the same as that calculated based on the GES data (with a wider confidence intervals for Nebraska), the contrast was smaller for Iowa, depicting a weaker difference between the levels of injury sustained by drivers for the two points of impact.

For rural versus urban settings, the directions of findings were completely opposite. The GES Midwestern model depicted slightly lower likelihoods of severe injuries for drivers in crashes occurring in rural settings (odds ratio of 0.92), whereas all the individual state models predicted higher likelihoods of such injuries in rural regions compared to urban settings, with odds ratios in the range of 1.71 to 2.55. This is the only contradiction between the states data and GES data-based models.

DISCUSSION

The goal of this paper was to investigate the factors associated with severe (fatal or incapacitating) injuries sustained by drivers in crashes, with a focus on the Midwestern states in the central part of the United States. The majority of the findings from each state were consistent with the literature. For example, our findings showed that females and older drivers were more susceptible to severe injuries in car crashes, and this has been observed in previous studies (O'Donnell and Connor 1996; Bedard et al. 2002). Seat belt use had an even greater effectiveness at the state level when compared to estimates from other studies (Martin, Crandall, and Pilkey 2000; Malliaris, Digges, and DeBlois 1995; Evans 1993; Bedard et al. 2002). Alcohol and drug use is another explanatory variable in our model that has consistently shown to increase the likelihood of severe injuries (Evans and Frick 1993; Evans 1990; Keall, Frith, and Patterson 2004; Mayhew et al. 1986; Sjogren et al. 1997; Zador, Kraw-

FIGURE 2 Comparison Between State and Regional Estimates
(Lines Indicate Confidence Intervals of the Estimates)



chuk, and Voas 2000). Head-on crashes were associated with the highest odds of sustaining severe injuries, and this is consistent with the findings of O'Donnell and Connor (1996).

There were differences that are worth noting. The four Midwestern states were not consistent with respect to crash type, with severe injuries more likely in single-vehicle crashes compared to rear-end crashes in Iowa and Missouri, and equally likely in the two crash types in Kansas and Nebraska. Crashes in rural settings were more likely to cause severe injuries than those occurring in urban crashes at the state level. However, the opposite was observed at the regional level, underscoring the impact of potential information loss when aggregating to the general region. Although this study had crash data for only four states, it clearly demonstrates that differences do exist from the state to regional level. Population distribution differences and geographical properties of different regions of the Midwest are influential in the disparity observed, even though the same modeling technique was used in all the models developed.

The four states examined may also have more rural characteristics when compared to other Midwestern states such as Illinois, Michigan, and even Indiana, with much larger metropolitan areas (e.g., Chicago, Detroit, and Indianapolis). Research has shown that differences in injury patterns in rural and urban settings may be due to the variations in availability of trauma care systems and distance from these

facilities (Brodsky and Hakkert 1988; Bentham 1986). Therefore, with more crashes occurring in areas with access to advanced medical facilities, these differences may lessen with other factors influencing urban crashes, e.g., roadway geometry, type of vehicle, distractions, etc., playing a larger role in the severity of injuries.

There are many data quality issues with using crash data at the state and national level related to underreporting, misclassification, and omitted data. At the national level, crashes are sampled and reported as four separate regions: Northwest, Midwest, South, and West. There is no information to the analyst to connect the data back to a specific state. Given that GES data is a weighted sample, it is not actually possible to have a direct comparison of the models developed at the state level to the sampled data state using GES data.

The state crash databases used in this study had several shortcomings that resulted in the need to exclude many crash records from the data used in statistical analyses. There were also missing crash, vehicle, and driver attributes for some state models (i.e., point of impact in Kansas, air bag deployment, point of impact, and drug use in Missouri, and air bag deployment in Nebraska). The same problem was identified by Ghazizadeh and Boyle (2009) in their study of driver distraction. Crash factors were not as comprehensive at the state level as initially expected. Missouri had the highest percentage (over 85%) of reported crashes that included the explanatory variables needed for the statistical model.

The crash report forms for each state for the years studied provide some insights on the relatively low numbers for some explanatory factors. In all four states, there was no specific callout for the various types of distraction, but instead all four states had "contributing circumstances" as a variable with distraction or inattention as a category. The Missouri and Nebraska forms did include distraction as a check box, whereas in Iowa and Kansas, categories of distraction were to be entered under a generic contributing circumstances area. The form used in Missouri included check-boxes for factors with several potential categories, which eliminated the need to refer to code sheets (as was the case in Iowa). Standardization of information could provide researchers better insights on safety issues and also allow better comparisons across states, which can have implications at the regional and national level. It is recognized that some improvements may actually lengthen the already cumbersome task of data entry, but could actually decrease the chance of non-reporting and even misclassification. Prioritizing information based on the findings in the literature of injury severity may also help officers in collecting the most critical information surrounding a crash.

Non-reporting and misclassification of conditions surrounding a crash is another potential issue that can impact the reliability of the estimates. However, past studies have shown that even though the crash reporting systems might not be ideal, estimates driven based solely on crash databases still offer valuable insights. For example, Cummings (2002) compared estimates of fatalities based on seat belt use for police-reported data and data based on trained crash investigators' reports and found no substantial difference, and Guo, Eskridge, Christensen, Qu, and Safranek (2007) showed that misclassifications of seat belt and alcohol use in Nebraska biased the odds ratio estimates of injury only slightly.

Recent studies have explored more rigorous statistical methods to predict crash injury severity outcomes (for a recent review of the methodol-

ogies, see Savolainen et al. 2011). For example, Anastasopoulos and Mannering (2002) examined the utility of the random-parameter logit model that used a less detailed crash profile (including injury outcomes, roadway geometrics, pavement condition, general weather, and traffic characteristics only) relative to the fixed-parameter model. Although the models based on individual crash-data provide better fit, their findings suggest that these models would be difficult to use for assessing the changes in injury severities caused by safety countermeasures because of the large number of variables that need to be determined for each crash. The random-parameter model, on the other hand, provides reasonable accuracy while also being easier to build. In other studies, Chang and colleagues used a non-parametric classification tree model in analyzing traffic injury severity (Chang and Chien 2013; Chang and Wang 2006). These methods and others can aid researchers in appropriately connecting the right model to the data being examined.

Crash data clearly has limitations in terms of exposure and standardization of information, but they do provide useful information on traffic, vehicle, and environmental factors that can be examined further in other test settings (e.g., simulator, test track, naturalistic studies). However, it is important to recognize the differences that exist, even within states in one geographic region. Future studies should examine the differences in rural/urban areas and crash type over a larger portion of the Midwest, and over a longer time period. It would also be of great interest to examine the underlying reasons for the disparity observed in injury trends across various states. Research in this direction can help provide insights for more effective crash countermeasures that can guide safer driver behaviors and driving environments.

ACKNOWLEDGEMENTS

A version of this paper was presented at the 3rd International Conference on Road Safety and Simulation (in Indianapolis, IN), September 2011. We would like to acknowledge the Midwest Transportation Consortium for sponsoring this project, and Iowa Department of Transportation (DOT), Kansas DOT, Missouri DOT, and Nebraska Department of Roads for providing us with crash data. We would also like to thank the editor and two anonymous reviewers for their helpful comments on previous versions of this article.

REFERENCES

Abdel-Aty, M. 2003. "Analysis of driver injury severity levels at multiple locations using ordered probit models." *Journal of Safety Research* no. 34 (5):597-603.

Al-Ghamdi, A.S. 2002. "Using logistic regression to estimate the influence of accident factors on accident severity." *Accident Analysis & Prevention* no. 34 (6):729-741.

Allison, P. D. 1999. *Logistic regression using the SAS system: Theory and application*. Cary, NC: SAS Institute Inc.

Ananth, C.V., and D.G. Kleinbaum. 1997. "Regression models for ordinal responses: a review of methods and applications." *International Journal of Epidemiology* no. 26 (6):1323-1333.

Awadzi, K. D., S. Classen, A. Hall, R. P. Duncan, and C. W. Garvan. 2008. "Predictors of injury among younger and older adults in fatal motor vehicle crashes." *Accident Analysis & Prevention* no. 40 (6):1804-1810.

Baum, H. M., A. K. Lund, and J. K. Wells. 1989. "The mortality consequences of raising the speed limit to 65 mph on rural interstates." *American Journal of Public Health* no. 79 (10):1392-1395.

Bedard, M., G. H. Guyatt, M. J. Stones, and J. P. Hirdes. 2002. "The independent contribution of driver, crash, and vehicle characteristics to driver fatalities." *Accident Analysis & Prevention* no. 34 (6):717-727.

Bentham, G. 1986. "Proximity to hospital and mortality from motor vehicle traffic accidents." *Social Science & Medicine* no. 23 (10):1021.

Brodsky, H., and A. S. Hakkert. 1988. "Risk of a road

accident in rainy weather." *Accident Analysis & Prevention* no. 20 (3):161-176.

Chang, L.Y., and J.T. Chien. 2013. "Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model." *Safety Science* no. 51 (1):17-22.

Chang, L.Y., and H.W. Wang. 2006. "Analysis of traffic injury severity: An application of non-parametric classification tree techniques." *Accident Analysis & Prevention* no. 38 (5):1019-1027.

Cochran, W.G. 1952. "The $\chi 2$ test of goodness of fit." *The Annals of Mathematical Statistics*:315-345.

Compton, C. P. 2005. "Injury severity codes: A comparison of police injury codes and medical outcomes as determined by NASS CDS Investigators." *Journal of Safety Research* no. 36 (5):483-484.

Connor, J., R. Norton, S. Ameratunga, and R. Jackson. 2004. "The Contribution of Alcohol to Serious Car Crash Injuries." *Epidemiology* no. 15 (3):337.

Cummings, P. 2002. "Association of seat belt use with death: a comparison of estimates based on data from police and estimates based on data from trained crash investigators." *Injury Prevention* no. 8 (4):338-341.

Evans, L. 1990. "The fraction of traffic fatalities attributable to alcohol." *Accident Analysis & Prevention* no. 22 (6):587-602.

Evans, L. 1993. "Restraint effectiveness, occupant ejection from cars, and fatality reductions." *Accident Analysis & Prevention* no. 22:167-175.

Evans, L., and M. C. Frick. 1993. "Alcohol's effect on fatality risk from a physical insult." *Journal of Studies on Alcohol* no. 54 (4):441-449.

Evans, L., and M. C. Frick. 1994. "Car mass and fatality risk: has the relationship changed?" *American Journal of Public Health* no. 84 (1):33-36.

Farmer, C. M., E. R. Braver, and E. L. Mitter. 1997. "Two-vehicle side impact crashes: The relationship of vehicle and crash characteristics to injury severity." *Accident Analysis & Prevention* no. 29 (3):399-406.

Ghazizadeh, M., and L. N. Boyle. 2009. "Influence of driver distractions on the likelihood of rear-end, angular, and single-vehicle crashes in Missouri." *Transportation Research Record* no. 2138:1-5.

Guo, H., K. M. Eskridge, D. Christensen, M. Qu, and T. Safranek. 2007. "Statistical adjustment for misclassification of seat belt and alcohol use in the analysis of motor vehicle accident data." *Accident Analysis & Prevention* no. 39 (1):117-124.

Hill, J. D., and L. N. Boyle. 2005. Analyzing Severe Injury Risk for Crashes Nationally and Within Iowa. Paper read at Mid-Continent Transportation Research Symposium, at Ames, Iowa.

Huelke, D. F., and C. P. Compton. 1995. "The effects of seat belts on injury severity of front and rear seat occupants in the same frontal crash." *Accident Analysis & Prevention* no. 27 (6):835-838.

Keall, M. D., W. J. Frith, and T. L. Patterson. 2004. "The influence of alcohol, age and number of passengers on the night-time risk of driver fatal injury in New Zealand." *Accident Analysis & Prevention* no. 36 (1):49-61.

Khattak, A. J., P. Kantor, and F. M. Council. 1998. "Role of Adverse Weather in Key Crash Types on Limited-Access: Roadways Implications for Advanced Weather Systems." *Transportation Research Record* no. 1621 (-1):10-19.

Khattak, A. J., and K. K. Knapp. 2001. "Interstate Highway Crash Injuries During Winter Snow and Nonsnow Events." *Transportation Research Record* no. 1746 (-1):30-36.

Khorashadi, A., D. Niemeier, V. Shankar, and F. Mannering. 2005. "Differences in rural and urban driver-injury severities in accidents involving large-trucks: An exploratory analysis." *Accident Analysis & Prevention* no. 37 (5):910-921.

Kim, K., L. Nitz, J. Richardson, and L. Li. 1995. "Personal and behavioral predictors of automobile crash and injury severity." *Accident Analysis & Prevention* no. 27 (4):469-481.

Klauer, S. G., T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey. 2006. The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data. Washington, DC: NHTSA.

Kweon, Y. J., and K. M. Kockelman. 2003. "Overall injury risk to different drivers: combining exposure, frequency, and severity models." *Accident Analysis & Prevention* no. 35 (4):441-450.

Malliaris, A. C., K. H. Digges, and J. H. DeBlois. 1995. "Evaluation of Airbag Field Performance." *SAE transactions* no. 104 (6):1513-1534.

Martin, P. G., J. R. Crandall, and W. D. Pilkey. 2000. "Injury trends of passenger car drivers in frontal crashes in the USA." *Accident Analysis & Prevention* no. 32 (4):541-557.

Mayhew, D. R., A. C. Donelson, D. J. Beirness, and H. M. Simpson. 1986. "Youth, alcohol and relative risk of crash involvement." *Accident Analysis & Prevention* no. 18 (4):273-287.

Miller, T. R., D. C. Lestina, and R. S. Spicer. 1998. "Highway crash costs in the United States by driver age,

blood alcohol level, victim age, and restraint use." *Accident Analysis & Prevention* no. 30 (2):137-50.

Muelleman, R. L., and K. Mueller. 1996. "Fatal Motor Vehicle Crashes: Variations of Crash Characteristics within Rural Regions of Different Population Densities." *The Journal of Trauma: Injury, Infection, and Critical Care* no. 41 (2):315.

Neyens, D. M., and L. N. Boyle. 2008. "The influence of driver distraction on the severity of injuries sustained by teenage drivers and their passengers." *Accident Analysis & Prevention* no. 40 (1):254-259.

NHTSA. 2005. National Automotive Sampling System (NASS) General Estimates System (GES): Analytical user's manual 1988-2005. Washington, DC: U.S Department of Transportation.

———. 2008. *General Estimates System (GES) 2001-2006* 2008a [cited June 2008]. Available from ftp://ftp.nhtsa.dot.gov/GES.

———. 2008b. Traffic Safety Facts 2007 Data: Rural/Urban Comparison. Washington, DC: U.S. Department of Tranportation.

———. 2009. Traffic Safety Facts 2008: A compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system. Washington, DC: U.S Department of Transportation.

O'Donnell, C. J., and D. H. Connor. 1996. "Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice." *Accident Analysis & Prevention* no. 28 (6):739-753.

Renski, H., A. J. Khattak, and F. M. Council. 1999. "Effect of speed limit increases on crash injury severity: Analysis of single-vehicle crashes on North Carolina Interstate highways." *Transportation Research Record* no. 1665 (1):100-108.

Savolainen, P., and F. Mannering. 2007. "Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes." *Accident Analysis & Prevention* no. 39 (5):955-963.

Savolainen, P.T., F.L. Mannering, D. Lord, and M.A. Quddus. 2011. "The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives." *Accident Analysis & Prevention* no. 43 (5):1666-1676.

Sjogren, H., U. Bjornstig, A. Eriksson, U. Ohman, and A. Solarz. 1997. "Drug and Alcohol Use Among Injured Motor Vehicle Drivers in Sweden: Prevalence, Driver, Crash, and Injury Characteristics." *Alcoholism: Clinical and Experimental Research* no. 21 (6):968-973.

Tavris, D. R., E. M. Kuhn, and P. M. Layde. 2001. "Age and gender patterns in motor vehicle crash injuries: importance of type of crash and occupant role." *Accident Analysis & Prevention* no. 33 (2):167-172.

Violanti, J. M. 1998. "Cellular phones and fatal traffic collisions." *Accident Analysis & Prevention* no. 30 (4):519-524.

Watt, K., D. M. Purdie, A. M. Roche, and R. McClure. 2006. "Injury severity: role of alcohol, substance use and risk-taking." *Emergency Medicine Australasia* no. 18 (2):108-117.

Zador, P. L., S. A. Krawchuk, and R. B. Voas. 2000. "Alcohol-related relative risk of driver fatalities and driver involvement in fatal crashes in relation to driver age and gender: An update using 1996 data." *Journal of Studies on Alcohol* no. 61 (3):387-395.

Zhang, J., S. Fraser, J. Lindsay, K. Clarke, and Y. Mao. 1998. "Age-specific patterns of factors related to fatal motor vehicle traffic crashes focus on young and elderly drivers." *Public Health* no. 112 (5):289-295.

Investigation of the Impact of Corner Clearance on Urban Intersection Crash Occurrence

VALERIAN KWIGIZILE, PH. D., P.E.
ENELIKO MULOKOZI
XUECAI XU, PH. D., F.E.
HUALIANG (HARRY) TENG, PH. D.
CAIWEN MA, PH.D.

Western Michigan University
1903 W. Michigan Avenue
Kalamazoo, MI 49008-5316
Phone: (269) 276-3218
Fax: (269) 276-3211
Email: Valerian.Kwigizile@wmich.edu

University of Nevada Las Vegas
Las Vegas, NV
Huazhong

University of Science and Technology
Wuhan, China

University of Nevada Las Vegas
Las Vegas, NV

Department of Transportation Engineering
Daliang Jiaotong University
Dalian, China

ABSTRACT

Corner clearance is defined as the distance between the corner of an intersection of two roadways and the first driveway. Vehicles turning into a driveway adjacent to an intersection or vehicles merging into the mainline from such a driveway may pose a safety hazard to other traffic. Adequate corner clearance is important to effectively separate conflict points and allow drivers enough time to make safe maneuvers. Although previous studies have investigated and identified factors influencing crash frequency at intersections, corner clearance has not been well studied. In this study, we used crash count data collected from all signalized intersections of major roadways in the cities of Las Vegas and North Las Vegas, Nevada, to investigate the impact of corner clearance on crash frequency. We estimated and compared results from four models: Poisson, Negative Binomial, and Zero-Inflated (Poisson and Negative Binomial). Model comparison test results indicated that the Zero-Inflated Negative Binomial was the best fitted model for the data at hand. As expected, it was revealed that longer corner clearance tends to reduce the number of crashes occurring at an urban intersection. In addition to corner clearance, the results indicated that landuse type, entering volume, number of left-turn lanes, as well as number of through lanes, have significant impact on the number of crashes occurring at an intersection. Sensitivity results revealed that adequate corner clearances have greater potential of improving safety at signalized intersections when compared to other factors considered in this study.

KEYWORDS: corner clearance, urban intersection safety

BACKGROUND AND MOTIVATION

Corner clearance is defined as the distance between the corner of an intersection of two roadways and the first driveway. Vehicles turning into a driveway adjacent to an intersection or vehicles merging into the mainline from a driveway may pose a safety hazard to other traffic. Adequate corner clearance is important to effectively separate conflict points and allow drivers enough time to make safe maneuvers. Although many studies (e.g., Oh et al. 2004; Guo et al. 2009; Wang and Abdel-Aty 2007; Poch and Mannering 2007; and Kumara and Chin 2010) have investigated the impact of roadway and traffic characteristics on intersection safety, corner clearance has not been fully investigated. Most studies have primarily evaluated the impact of corner clearance on other intersection performance aspects, while treating the impact of corner clearance on intersection safety as secondary. For example, McCoy and Heimann (1990) evaluated the impact of corner clearance on saturation flow rates at signalized intersections in Lincoln, Nebraska. Similarly, Long and Gan (2007) developed a model for minimum driveway corner clearances at signalized intersections by considering saturation flow rates. Long and Gan (1993) and Gluck et al. (1999) focused on how to specify the corner clearance criteria for practical implementation. Although the above studies did not explicitly address the impact of corner clearance on safety of signalized intersections, their findings suggested the importance of corner clearance as well as driveway density at signalized intersections. For example, a finding by Long and Gan (1997) that corner clearance has significant impact on saturation flow rates implies that this could lead to some types of crashes, such as rear-end crashes due to interruption of traffic flow resulting from vehicles entering and/or exiting driveways.

Very often, roadway designers ask themselves which design feature has greater potential to improve safety at a signalized intersection. Although it is clear that longer corner clearances may result in improved safety of a signalized intersection, quantification of their impacts is lacking in the literature. Also, the relative impact of corner clearances on safety of signalized intersections is not well documented. As a result, the main objective of this study is to investigate the impact of corner clearance and other variables on the number of crashes occurring at urban signalized intersections. Data from signalized intersections in the Las Vegas and North Las Vegas urban areas were used to conduct the analyses. Count models were developed to investigate the impact of corner clearance and other variables on the number of crashes occurring at such intersections.

LITERATURE ON MODELING INTERSECTION SAFETY

Proper design of roadway features around signalized intersection can result in increased intersection safety. To achieve such a good design, safety studies are required to identify high risk factors related to these features. Guo et al. (2009) utilized five full Bayesian models on 170 signalized intersections of Orange and Hillsborough counties in central Florida to show that intersections in close proximity along a corridor are correlated and proper signal coordination has significant impact on safety. In addition to spatial correlation between intersections, it was found that Average Daily Traffic (ADT) per through-lane and left-turn traffic, landuse, speed limits, intersection size, and exposure have a significant impact on the safety of signalized intersections. It was also shown that larger intersections are more dangerous than smaller intersections. However, this study did not investigate the impact

of corner clearance on signalized intersection safety. Different severity levels of crash incidents at signalized intersections have been studied. Wong et al. (2007) used Poisson and negative binomial regression models on 262 signalized intersections features from Hong Kong to quantify the influence of various factors on fatal and severe injury, and slightly injury crashes. The negative binomial regression model indicated that an increase in curvature and the presence of tram stops significantly increased the incidence of slightly injury crashes. Also, the marginal increase of the incidence of slightly injury crashes diminished under high traffic flow conditions. With the Poisson regression model it was found that the presence of tram stops, the increase in the proportional of commercial vehicles, the increase in the number of pedestrian streams, and the decrease in the average lane width significantly increased the incidence of killed and severe injury crashes. Similar to the study by Guo et al. (2009), this study did not incorporate corner clearance in the models.

Although Poisson and traditional negative binomial regression models are widely used in crash frequency analysis (Yaacob et al, 2011; Zlatoper, 1989; Lord, 2006; Chin and Quddus, 2003; Miaou and Lum, 1993; and Noland and Quddus, 2004), they may lead to biased estimators and invalid statistics when applied to longitudinal crash data. The correlation features in longitudinal data for signalized intersections require a different modeling approach leading to consistent estimates. Wang et al. (2006) applied generalized estimating equations (GEEs) to identify significant factors and their temporal correlation effect on crashes at signalized intersections. Using 208 signalized intersections in central Florida from Brevard and Seminole Counties in suburban areas, they modeled the relationship between crash frequencies and other variables at signalized intersections. Speed limits, traffic volume (ADT), intersection size (indicated by total number of lanes), and intersection within highly populated areas were found to be associated with high crash frequency. Using the same model of GEEs but with different link functions, Wang and Abdel-Aty (2007) investigated the relationship between different patterns of left-turn crash occurrence and intersection features using 197 four-legged signalized intersections from Orange and Hillsborough counties in the central Florida area. Selection of the particular link function in the GEEs for modeling different functions was a function of the number of crashes and the proportion of zero crashes and one crashes recorded. GEEs with binomial logit link function were applied to crash patterns with a higher proportion of zeros and one crashes. Negative binomial link function was used to model crash frequency for patterns with fewer crashes. The modeling results showed that the amount of conflicting flows (traffic volumes), the type of left-turn phasing, crossing distance (indicated by the number of through lanes), and speed limit are significant in influencing the crash occurrence frequency.

Other researchers have used negative binomial and zero-inflated negative binomial regression in modeling crashes at signalized intersections. Using 104 three-legged signalized intersections from Singapore, Kumara, and Chin (2010) indicated that right-turn channelization, acceleration section on the left-turning lane, median railing, and existence of more than a 5% gradient may reduce accident occurrence. Although surprising, the finding that existence of more than a 5% gradient may reduce accident occurrences may be attributed to possible proper signage and extra carefulness by drivers resulting from existence of such steep grade. The same

research team indicated that traffic volumes (total and left-turn), an uncontrolled left-turn slip road, signal phases per cycle, existence of horizontal curve, and permissive right-turn phase may increase accident occurrence. Poch and Mannering (2007) used negative binomial regression modeling on 64 intersections from Bellevue in Washington to show that traffic volumes (separated according to traffic movements), number of lanes, sight distance restriction, and speed limit have a negative effect on intersection safety while signal controlled intersections and protected left turn movements have a positive effect on intersection safety. A study by Oh et al. (2004) was aimed at developing macrolevel crash prediction models that can be used to understand and identify effective countermeasures for improving signalized highway intersections and multilane stop-controlled highway intersections in rural areas. The results indicated that traffic flow variables significantly affected the overall safety performance of the intersections regardless of intersection type and that the geometric features of intersections varied across intersection type and also influenced crash type.

There are many issues related with modeling crash counts. Lord and Mannering (2010) 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. Among the issues discussed include dispersion (over- or under-), temporal and spatial correlations, endogeneity, low sample mean, underreporting, etc. Different model types designed to handle these issues were also discussed. The authors identified Zero-Inflated models (Poisson or Negative Binomial) as models that can handle datasets that have a large number of zero-crash observations. However, the authors cautioned that the zero-inflated negative binomial can be adversely influenced by the low sample-mean and small sample size bias.

Despite all the efforts to investigate different factors contributing to intersection crash frequencies using different modeling approaches, other important factors have not been included in the previous researches. In this study, investigation of the impact of corner clearance on urban intersection crash occurrence using data from the cities of Las Vegas and North Las Vegas, Nevada, is performed. Crashes considered in the model were those that happened within a 250 ft radius measured from the center of a signalized intersection. Crashes occurring within 250 ft of the intersection have traditionally been considered to be influenced by intersection performance (e.g., Oh et al. 2004; and Ye et al 2009)

METHODOLOGY

When modeling crash counts, Poisson regression analysis or Negative Binomial (NB) regression analysis can be used (Yaacob et al, 2011; Zlatoper, 1989; Lord, 2006; Chin and Quddus, 2003; Miaou and Lum, 1993; and Noland and Quddus, 2004). The choice between the two model types depends on the relationship between the mean and the variance of the data. If the mean is equal to the variance, the data is assumed to follow a Poisson distribution, and hence the Poisson regression analysis can be performed. However, as a result of possible positive correlation between observed accident frequencies, overdispersion may occur (Hilbe, 2011). Accident frequency observations are said to be overdispersed if their variance is greater than their mean. If overdispersion is detected in the data, NB regression analysis should be used. Another issue arising with modeling accident frequencies is presence of sites with zero counts. Hurdle and

zero-inflated Poisson or NB regression models are the two foremost methods used to deal with count data (e.g., accident frequencies) having zero counts (Hilbe, 2011). This study explored the suitability of Poisson, NB, and zero-inflated (Poisson and NB) models.

Standard textbooks (e.g., Hilbe 2011; Greene 2012; and Washington et al 2011) present clear derivation of the Poisson, Negative Binomial (NB), and zero-inflated models (Poisson (ZIP) or Negative Binomial (ZINB)). According to Poisson distribution, the probability $P(y_i)$ of intersection i having y_i crashes in a given time period (usually one year) can be written as:

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

where λ_i denotes the Poisson parameter for intersection i. By definition, λ_i is equal to the expected number of crashes in a given time period for intersection i, $E[y_i]$. According to Washington et al. (2011), the expected number of crash occurrences λ_i, can be related to a vector of explanatory variables, $\boldsymbol{X}\,i$ as follows:

$$\lambda_i=EXP(\boldsymbol{\beta X}_i) \quad (2)$$

where $\boldsymbol{\beta}$ represents a vector of estimable parameters. Under Poisson assumption, the mean and variance of crashes occurring at an intersection in a year are equal (i.e., $E[y_i]=Var[y_i]$,). With N observations, the parameters of the Poisson model can be estimated by maximum likelihood method with a function that can be shown to be as follows:

$$LL(\beta)=\sum_{i=1}^{N}[-EXP(\boldsymbol{\beta X}_i)+y_i\boldsymbol{\beta X}_i-\ln(y_i!)] \quad (3)$$

The Poisson assumption of equal mean and variance of the observed crash occurrences is not always true. To handle the cases where the mean and variance of crashes are not equal, the Poisson model is generalized by introducing an individual, unobserved effect, ε_i, in the function relating crash occurrences and explanatory variables (equation 2) as follows:

$$\lambda_i=EXP(\boldsymbol{\beta X}_i+\varepsilon_i) \quad (4)$$

in which $EXP(\varepsilon_i)$ is a gamma-distributed error term with mean one and variance α^2. With such a modification, the mean λ_i becomes a variable that follows binomial distribution. The mean-variance relationship becomes:

$$Var[y_i]=E(y_i)\cdot[1+\alpha E(y_i)]=E[y_i]+\alpha E(y_i)^2 \quad (5)$$

If α is equal to zero, the negative binomial distribution reduces to Poisson distribution. If α is significantly different from zero, the crash data are said to be overdispersed (positive value) or underdispersed (negative value). As stated earlier, overdispersion is a result of possible positive correlation between observed accident frequencies. When α is significantly different from zero, the resulting negative binomial probability distribution is:

$$P(y_i)=\frac{\Gamma\left(\left(1/\alpha\right)+y_i\right)}{\Gamma\left(1/\alpha\right)y_i!}\left(\frac{1/\alpha}{\left(1/\alpha\right)+\lambda_i}\right)^{1/\alpha}\left(\frac{\lambda_i}{\left(1/\alpha\right)+\lambda_i}\right)^{y_i} \quad (6)$$

where Γ(x) is a value of the gamma function, y_i is the number of crashes for intersection i and α is an overdispersion parameter. Because crash counts involve intersections with zero observations, possible remedies include the estimation of models such as zero-inflated (ZIP or ZINB) and hurdle models. The zero-inflated models have two parts: a binary part for distinguishing the intersections that will always have zero counts from those that, although they now have zero counts, will not always have zero counts, and a Poisson regression model (for ZIP) or Negative Binomial

model (for ZINB), which models the intersections with zero or positive counts. Washington et al, 2011, shows that with the ZINB regression model the probability of an intersection having zero crash counts can be estimated as:

$$\Pr(y_i = 0) = p_i + (1 - p_i)\left[\frac{1/\alpha}{1/\alpha + \lambda_i}\right]^{1/\alpha} \quad (7)$$

and the probability of having positive crash counts (y = 1, 2, 3,…) can be estimated as:

$$\Pr(y_i = y) = (1 - p_i)\left[\frac{\Gamma\left((1/\alpha) + y\right)\psi_i^{1/\alpha}(1-\psi_i)^y}{\Gamma(1/\alpha)y!}\right] \quad (8)$$

in which. $\psi_i = \frac{1/\alpha}{1/\alpha + \lambda_i}$.

It is imperative to test whether using the zero-inflated model is necessary. This can be achieved by conducting the Vuong test (Vuong 1989). Although alternative tests have been developed to improve model selection reliability, the Vuong test is commonly used. The Vuong test is far more conservative than alternative tests such as the distribution-free test and therefore does a better job of protecting against an incorrect decision (Clarke 2007). Given two models, model 1 with $P_1(y_i/x)$ as the probability of observing y crashes on the basis of variable x, and model 2 with the probability denoted as $P_2(y_i/x)$, the log ratio of the sum of probabilities for each observation can be computed as:

$$\phi_i = \ln\left(\frac{\sum_i P_1(y_i/x_i)}{\sum_i P_2(y_i/x_i)}\right) \quad (9)$$

and the Vuong test statistic can be computed as:

$$V = \frac{\sqrt{N}(\bar{\phi})}{SD(\phi_i)} \quad (10)$$

in which $\bar{\phi}$ is the average of the log ratios and $SD(\phi_i)$ is the standard deviation of the log ratios. The Vuong test statistic has been proved to follow a normal distribution. Greene (2012) states that if $|V|$ is less than 1.96, then the test does not favor one model over the other. If V is greater than 1.96, model one is favored while if V is less than -1.96, model two is favored. In addition to comparing the models using the Vuong test statistic, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were calculated for each model and compared. Standard statistical software such as Stata (2008) can be used to estimate the models and statistics described above.

To better interpret the results of a count data models, elasticities can be computed. Elasticity of a continuous variable is used to quantify the effect of a small change (1%) in the mean of the variable on the outcome (expected crash occurrences λ_i). Elasticity of a k^{th} continuous variable x for observation i (x_{ik}), can be estimated as:

$$E_{x_{ik}}^{\lambda_i} = \frac{\partial \lambda_i}{\lambda_i} \times \frac{x_{ik}}{\delta x_{ik}} = \beta_k x_{ik} \quad (11)$$

For indicator variables (which strictly take on values of 0 or 1), such a small change is meaningless. As a result, the "pseudo-elasticity" can be used to accommodate such variables. It shows the difference in the outcome (crash occurrence) with a specific variable taking the value of 1 versus 0. It can be computed as:

$$E_{x_{ik}}^{\lambda_i} = \frac{EXP(\beta_k) - 1}{EXP(\beta_k)} \quad (12)$$

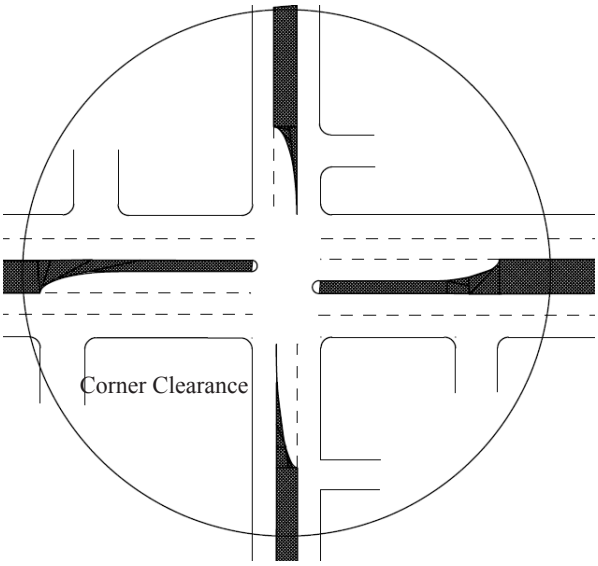
DATA DESCRIPTION

Data used in this study were collected from all signalized intersections in the cities of Las Vegas and North Las Vegas, Nevada, in

the USA. With the Geographic Information System (GIS) and aerial images from Google Earth, we located the identified signalized intersections. The roadway network map for Las Vegas and North Las Vegas area was used as the base layer in the ArcGIS. The geographic coordinates of each intersection obtained from Google Maps were used to geo-code the intersections to create a GIS layer. The Safety and Traffic Engineering Division of the Nevada Department of Transportation (NDOT) maintains a database of all crashes occurring in the state. From this database, North Las Vegas and Las Vegas crashes were selected. The crash data were mapped with ArcGIS to identify crashes that occurred at signalized intersections. This was accomplished by creating a buffer with a radius of 250 ft around each signalized intersection in Las Vegas and North Las Vegas and counting the number of crashes within that buffer. These crashes were deemed to be influenced by the intersection. Although the 250-ft buffer around an intersection may omit some intersection crashes and/or include some nonintersection crashes, it is commonly used in the United States as it is a nonarbitrary criterion that is easily repeatable and generalizable across jurisdictions (Ye et al. 2009). The number of corner clearances was obtained by counting the number of driveways within the intersection area (a circle with a radius of 250 ft) and the corner clearances were measured as shown in figure 1. For example, the intersection depicted in figure 1 has five corner clearances.

Traffic volume is a very important determinant of crash occurrence at intersections. This data item was obtained from a database maintained by the Southern Nevada Regional Transportation Commission (RTC). For the intersections selected, the RTC database contained complete traffic volume data for year 2004 only.

FIGURE 1 Corner Clearance Measurements and Counting



Therefore, this study used crash data for year 2004 only. In addition, landuse type for each intersection was collected.

For each approach, functional classification was obtained from the GIS database provided by the Regional Transportation Commission (RTC) of Southern Nevada. From the same database, the number of lanes on both directions and the posted speed limit were also extracted. The number of lanes was confirmed using the Google map. To designate major and minor approaches, the AADT volumes were used. The opposing approaches with the highest sum of AADT were designated major while the ones with the lowest were labeled minor. In addition to AADT data, other roadway attributes were extracted. After cleaning data to remove intersections with incomplete information, only 170 intersections remained. Table 1 presents the descriptive statistics for the selected variables.

Table 1 indicates that on average, over 22 crashes occurred at signalized intersections in Las Vegas and North Las Vegas during 2004.

TABLE 1 Descriptive Statistics of all Extracted Variables

Variable	Average	Std. dev.	Minimum	Maximum
2004 crash count	22.71	19.89	0	96
Commercial landuse	0.53		0	1
No. of lanes on major approach	4.77	1.12	1	7
No. of left turn lanes on major approach	1.29	0.51	0	4
No. of right turn lanes on major approach	1.01	0.10	1	2
No. of lanes on minor approach	3.15	1.56	1	8
No. of left turn lanes on minor approach	1.20	0.44	1	3
No. of right turn lanes on minor approach	1.03	0.17	1	2
No. of corner clearances	6.24	1.79	2	8
Average corner clearance (ft)	134.85	69.28	43.63	250.00
AADT on major approach	13241.19	29936.59	554	153896
AADT on minor approach	6315.17	14929.43	469	99890
Average speed on major approach	41.22	4.93	25	50
Average speed on minor approach	33.84	7.28	15	45

The results also show that the abutting land on about 53% of the signalized intersections was commercial landuse. On average, major approaches had about five lanes (in both directions) while minor approaches had about three lanes (in both directions). Both major and minor approaches had an average of one left-turn lane and one right-turn lane. Corner clearance for each driveway was measured all averaged to determine the average corner clearance. On average, the observed average corner clearance was 134.85 ft. The average AADT on major approaches was 13,242 vehicles/day while for minor approaches it was 6,316 vehicles/day. There was an average of six corner clearances (equivalent to driveways) at the intersections observed.

The extracted variables were to derive explanatory variables used in the models. Only significant variables were retained and are presented in the next section. In modeling, number of left turn lanes and right turn lanes were combined into one category of "turning lanes" for each approach. The "turning lanes" were used distinctively from "through lanes". Commercial landuse was another significant variable used in modeling. The AADT on minor approach was divided by the AADT on major approach to generate "flow ratio." Average corner clearance was transformed by taking natural logarithm before using it in modeling. Also, the number of corner clearances was used in modeling. Table 2 presents summary statistics for variables used in the model.

MODELING RESULTS AND DISCUSSION

Using commercially available software, Stata (2008), we estimated and compared results from four models: Poisson, Negative Binomial, Zero-Inflated Poisson, and Zero-Inflated Negative Binomial. Table 3 presents the coefficient estimates from these models. As it can be seen, all models (Poisson, Negative Binomial, Zero-Inflated Poisson ZIP) and Zero-Inflated Negative Binomial (ZINB)) produced results consistent with intuition in terms of the impact of the variables on crash count. The models were compared to identify the "best" fitted

TABLE 2 Descriptive Statistics of Modeling Variables

Variable	Average	Std. dev.	Minimum	Maximum
Commercial landuse	0.53	0.50	0	1
Flow ratio	1.30	1.55	0.05	11.00
Natural log. of average corner clearance	4.97	0.35	3.78	5.75
Left turning lanes	2.49	0.74	1	5
Through lanes	7.95	2.18	2	13
Number of corner clearance	6.25	1.79	2	8

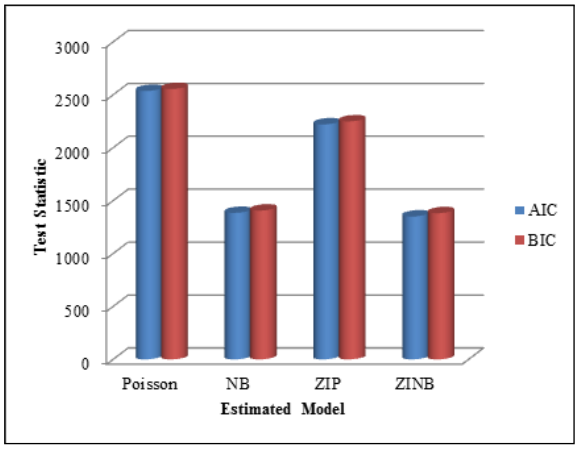
TABLE 3 Model Estimation Results

	Poisson		Negative binomial		ZIP		ZINB	
Explanatory variables	**Coef.**	**Statistic**	**Coef.**	**Statistic**	**Coef.**	**Statistic**	**Coef.**	**Statistic**
Regression part								
Commercial landuse	0.511	14.31	0.463	3.72	0.394	11.05	0.377	3.59
Flow (AADT) ratio (minor/major)	0.065	8.33	0.074	1.9	0.062	7.84	0.063	2.00
Natural log. of avg. corner clearance	-0.210	-6.46	-0.193	-1.14	-0.404	-9.00	-0.509	-3.26
Number of left turning lanes	0.096	4.24	0.094	1.02	0.185	7.32	0.208	2.60
Number of through Lanes	0.147	15.49	0.147	4.28	0.110	11.35	0.112	3.80
Constant	2.304	13.36	2.242	2.49	3.479	12.78	3.929	4.74
Inflation part								
No. of corner clearance					-0.551	-2.67	-0.564	-2.57
Natural log. of avg. corner clearance					-0.845	-1.59	-0.873	-1.57
Constant					4.237	1.49	4.375	1.48
Auxiliary statistics								
Alpha (α)			0.528	8.07			0.338	7.61
Number of observations		170		170		170		170
Final log-likelihood		-1262.20		-685.38		-1101.62		-665.18
Likelihood-ratio test of α=0 (p-value)		-		1153.64 (0.000)		-		872.89 (0.000)
Vuong statistic (p-value)		-		-		2.90 (0.002)		2.58 (0.004)

model. The resulting test statistic of 1,153.64 with a p-value of 0.0000 for the likelihood-ratio test of zero overdispersion (α=0) indicates that the Negative Binomial model is preferred to the Poisson model. Even the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) support this assertion. As shown in figure 2, both the AIC and the BIC for the Negative Binomial model are less than those of the Poisson model, signifying superiority of the Negative Binomial model over the Poisson model. Furthermore, the Vuong test statistics of 2.90 (with p-value of 0.002) and 2.58 (with p-value of 0.004) from the Zero-Inflated Poisson (ZIP) and the Zero-Inflated Negative Binomial (ZINB) models, respectively, indicate that the zero-inflated models (ZIP and ZINB) are preferred to their respective standard models (Poisson and Negative Binomial).

Comparing the AIC and the BIC for the ZIP and the ZINB indicates that the ZINB has slightly lower values (figure 2), which signi-

FIGURE 2 Selection of the "Best" Fitted Model



fies slight preference of the ZINB over the ZIP. However, the Vuong test statistic for comparing ZINB to ZIP was 0.84, signifying that neither of the two models is preferable over the other for the data at hand. Table 4 presents the model selection results.

It should be recalled that the test on alpha (testing equality of mean and variance) indicated that alpha is significantly different from zero, which indicates that Poisson is not the closest estimation of the distribution for the data at hand. Therefore, the Zero-Inflated Negative Binomial (ZINB) model was identified as the "best" fitted model for the data at hand.

The results from the ZINB model estimates (table 3) indicate that increased length of corner clearance leads to decreased crash frequency. This is because a driveway that is far from the intersection allows sufficient distance for drivers exiting the businesses (whether they are familiar or unfamiliar with the area) to perform the desired maneuver. Also, with longer corner clearance, the drivers of through traffic could perceive and respond more quickly and safely to the maneuvers by traffic leaving or entering the adjacent lands because their attention is not already preoccupied by the maneuver they desire to perform at the intersection. In other words, with longer corner clearances, the maneuver needed at the intersection has less influence on decisions by through-traffic drivers and those leaving or entering businesses. Also, shorter corner clearances imply more driveways at intersections, which increase the chance of conflicts to occur between turning and through traffic.

The results also show that an intersection surrounded by commercial landuse is more likely to experience more crashes compared to an intersection surrounded by residential landuse. Reasons for this finding might include the fact that drivers entering or exiting businesses around an intersection may include those who are unfamiliar with the roadway (noncommuters). Such drivers who are unfamiliar with the roadway are more likely to perform unpredictable maneuvers that increase the chance of crash occurrence.

The results also show that signalized intersections with traffic volume on minor street close to the traffic on major street (high ratio) tend to have higher crashes. In previous studies such as Chin and Quddus (2003), traffic flow was also found to be an important predictor

TABLE 4 Model selection results

	Test Statistic				Vuong Statistic	
Model	**BIC**	**AIC**	**Vuong**	**"Best" Model**	**for ZINB vs. ZIP**	**Conclusion**
NB	1407	1385				
ZINB	1382	1350	2.58	ZINB	0.84	Neither ZINB nor ZIP is superior over the other
Poisson	2555	2536				
ZIP	2249	2221	2.90	ZIP		

of crashes at intersections. Some studies have used total entering traffic volume (e.g., Greibe (2003) and Chin and Quddus (2003)) while others have separated traffic volume on minor street from that on major street (e.g., Lord and Persaud 2000). Other studies have used traffic volume per lane (e.g., Wang et al 2006). The finding in this study indicates that with higher traffic on the minor approach, there is an increased probability of higher conflicts and therefore higher crashes. Although not examined in this study, this could be associated with permitted right-turn and left-turn movements. This study examined the separate impact of turning lanes and through lanes at signalized intersections. Left-turning lanes and right-turning lanes as well as through lanes were counted for each intersection. The modeling results also indicated that crash increases with increase in number of both left-turning lanes and through lanes. However, the number of through lanes has the highest impact on the number of crashes (higher elasticity presented in table 5). With left-turning lanes, vehicles exiting adjacent businesses around the intersection and desiring to make left turn may experience relatively higher difficulty in performing their maneuvers.

Extra care is needed when interpreting the results of the "inflation" part of the model. This is a binary process with a prediction of success being a prediction that the response will certainly be a zero. The negative coefficients associated with both the number of corner clearances and the natural logarithm of corner clearance signify that an increase in these variables reduces the chance of having zero crashes at signalized intersections. In other words, increase in these variables could potentially lead to crash occurrence at intersections.

To better investigate the impact of corner clearance and other variables on the number of crashes occurring at signalized intersections we calculated their elasticities presented in table 4. Elasticity of a continuous variable is used to quantify the effect of a small change (1%) in the mean of the variable on the outcome (expected crash occurrences λ_i). Because a small change is meaningless for indicator variables (which strictly take on values of 0 or 1), the "pseudo-elasticity" was calculated. Again, the results indicate that corner clearance is very sensitive to the number of crashes occurring at signalized intersections. There could be an 83% reduction in the number of crashes by increasing the natural logarithm of average corner clearance by 1%. However, an increase of 1% on the number of left-turn lanes would increase the number of crashes by about 16%. Compared to the number of left turn lanes, an increase of 1% in the number of through lanes would result into a 29% increase on the number of crashes. Flow ratio and landuse have the lowest sensitivity to the number of crashes. An intersection being surrounded by commercial landuse would result into a 7% increase on number of crashes while a 1% increase on the flow ratio would result into just 3% increase

Table 5 Estimated Elasticities

Variable	Elasticity	Std. Err.	z-statistic	p-value
Commercial landuse	0.07	0.02	3.75	0.000
Flow (AADT) ratio (minor/major)	0.03	0.01	2.06	0.039
Natural log. of avg. corner clearance	-0.83	0.26	-3.22	0.001
Number of left turning lanes	0.16	0.06	2.62	0.009
Number of through Lanes	0.29	0.08	3.81	0.000

on the number of crashes. This indicates that intersections with higher traffic volumes on minor approaches (i.e., ratio close to 1) may experience relatively more crashes. Large number of left-turn lanes may be an indication of higher left-turn traffic, making maneuverability more difficult and dangerous.

CONCLUSIONS AND RECOMMENDATIONS

Adequate corner clearance is important to effectively separate conflict points and allow drivers enough time to make safe maneuvers. Although most studies have investigated the impact of roadway and traffic characteristics on intersection safety, corner clearance has not been fully investigated. The main objective of this study was to investigate the impact of corner clearance and other variables on the number of crashes occurring at urban signalized intersections. Data from all signalized intersections in the Las Vegas and North Las Vegas urban areas were used to conduct the analyses. This study explored the suitability of Poisson, Negative Binomial (NB), Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models. Statistical tests such as the Vuong test, Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC) were calculated to identify the best model. Also, the accuracy of the Negative Binomial (NB) and the Zero-Inflated Negative Binomial (ZINB) models in predicting crash occurrence was compared by computing the difference between predicted probability and observed probability for each model. With all comparison tests, the ZINB outperformed other models and was selected as the best fitted model for the data at hand. It was revealed that the ZINB is very accurate in predicting zero crash occurrences when compared with the regular NB but their difference fades away for high crash occurrences. To better interpret the results of a count data models, elasticities can be computed.

The results from the ZINB model estimates indicated that increased length of corner clearance leads to decreased crash frequency. This finding is consistent with intuition because shorter corner clearances imply more driveways at intersections, which increase the chance of conflicts to occur between turning and through traffic. The sensitivity results (table 4) indicated that corner clearance is very sensitive to the number of crashes occurring at signalized intersections. The results also showed that an intersection surrounded by commercial landuse is more likely to experience more crashes compared to an intersection surrounded by residential landuse. Reasons for this finding might include the fact that drivers entering or exiting businesses around an intersection may include those who are unfamiliar with the roadway (noncommuters). The results indicated that 65% of the driveways with corner clearance less than 150 ft are from intersections surrounded by commercial landuse and 78% of the driveways with corner clearance less than 100 ft are from intersections surrounded by commercial landuse. The results also showed that signalized intersections with traffic volume on minor street close to the traffic on major street (high flow ratio) tend to have higher crashes. With higher traffic on the minor approach, there is an increased probability of higher conflicts and therefore higher crashes. Although not examined in this study, this could be associated with permitted right-turn and left-turn movements. The modeling results also indicated that, generally, crash increases with increase in number of both left-turning lanes and through lanes. However, the number of through lanes has the highest impact on the number of crashes (higher elastic-

ity presented in table 4). With turning lanes, vehicles leaving or entering businesses around the intersection make their maneuvers from low speed lanes (turning lanes). In addition to confirming the impact of corner clearance on safety of signalized intersections, the findings of this study also reveal the importance of properly designed corner clearances at signalized intersections when compared with other geometrics. For example, compared to the number of left turn lanes at signalized intersections, adequate corner clearances may produce higher safety gains by reducing number of crashes. Such an understanding is important to access managers and roadway designers as they consider potential geometric design parameters with potential to improving safety.

Intersections located on a given urban arterials may share common but unobserved attributes such as similar traffic volume patterns. Such unobserved common attributes may influence statistical inferences. One way to addressing this issue is to develop panel count models. Such modeling consideration is relatively complex especially with zero-inflated models. It is recommended that future research consider incorporate panel structure to address the possible problem of unobserved common attributes.

REFERENCES

Chin, H. C., and M. A. Quddus. 2003. Applying the Random Effect Negative Binomial Model to Examine Traffic Accident Occurrence at Signalized Intersections. *Accident Analysis & Prevention* 35: 253–259.

Clarke, K. A. 2007. A Simple Distribution-Free Test for Non-nested Hypotheses. Published by Oxford University Press on behalf of the Society for Political Methodology. Doi:10.1093/pan/mpm004.

Gluck, J., H. S. Levinson, and V. Stover. 1999. *Impact of Access Management Techniques*, NCHRP Report 420, Transportation Research Record, National Research Council, Washington, D. C.

Greene, W. H. 2012. *Econometric Analysis*, 7th Edition, Prentice Hall, Upper Saddle River, New Jersey, United States of America.

Greibe, P. 2003. Accident Prediction Models for Urban Roads. Accident *Analysis & Prevention* 35: 273–285.

Guo, F., X. Wang, and M. Abdel-Aty. 2010. Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis and Prevention* 42: 84–92.

Hilbe, J.M. 2011. *Negative Binomial Regression*, Second Edition, Cambridge University Press, United Kingdom.

Kumara, S. P., and H. C. Chin. 2010. Modeling Accident Occurrence at Signalized Tee Intersections with Special Emphasis on Excess Zeros. *Traffic Injury Prevention* 4(1): 53-57.

Long, G., and C. Gan. 1993. *Driveway Impact on Saturated Traffic Flow*. Transportation Research Center, University of Florida, Gainesville, FL.

Long, G., and C. Gan. 1997. Model for Minimum Driveway Corner Clearances at Signalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board* 1579: 53-62.

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: 751-766.

Lord, D., and B. N. Persaud. 2000. Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations Procedure. *Transportation Research Record: Journal of the Transportation Research Board* 1717: 102–108.

Lord, D. and F. L. Mannering. 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research* Part A 44: 291–305.

McCoy, P.T., and J. E. Heimann. 1994. Effect of Driveway Traffic on Saturation Flow Rates at Signalized Intersections. *ITE Journal* February 1990: 12-15.

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

Noland, R.B., and M. A. Quddus. 2004. Analysis of pedestrian and bicycle casualties with regional panel data. *Transportation Research Record: Journal of the Transportation Research Board* 1897: 28-33.

Oh, J., S. Washington, and K. Choi. 2004. Development of accident prediction models for rural highway intersections. *Transportation Research Record: Journal of the Transportation Research Board* 1897: 18–27.

Poch, M., and F. Mannering. 2007. Negative Binomial Analysis of intersection accident frequencies. *Journal of Transportation Engineering* 122(2): 105-113.

StataCorp LP. *Data Analysis and Statistical Software*. College Station, Texas, 2011.

Vuong, Q.H. 1989. Likelihood ratio tests for model selection and non-nested hypothesis. *Econometrica* 57(2): 307–333.

Wang, X., and M. Abdel-Aty. 2007. Right-angle crash occurrence at signalized-intersections." *Transportation Research Record: Journal of the Transportation Research Board* 2019: 156-168.

Wang, X., and M. Abdel-Aty. 2008. Modeling left-turn crash occurrence at signalized intersections by conflicting patterns. *Accident Analysis and Prevention* 40: 76-88.

Wang, X., M. Abdel-Aty, and P. Brady. 2006. Crash Estimation at Signalized Intersections: Significant Factors and Temporal Effect. *Transportation Research Record: Journal of the Transportation Research Board* 1953: 10–20.

Washington, P. S., G. M. Karlaftis, and F. L. Mannering. 2011. *Statistical and Econometric Methods for Transportation Data Analysis*. 2nd Edition. Chapman & Hall/CRC, United States of America.

Wong, S. C., N. N. Sze, and Y. C. Li. 2007. Contributory factors to traffic crashes at signalized intersections in Hong Kong. *Accident Analysis and Prevention* 39: 1107–1113.

Yaacob, W. F. W., M. A. Lazim, and Y. B. Wah. 2011. Applying Fixed Effects Panel Count Model to Examine Road Accident Occurrence. *Journal of Applied Sciences* 11: 1185-1191.

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

Zlatoper, T.J. 1989. Models explaining motor vehicle death rates in the United States. *Accident Analysis and Prevention* 21: 125-154.

Application of the Bayesian Model Averaging in Predicting Motor Vehicle Crashes

YAJIE ZOU
DOMINIQUE LORD, PH.D.
YUNLONG ZHANG, PH.D.
YICHUAN PENG

Zachry Department of Civil Engineering
Texas A&M University, 3136 TAMU
College Station, TX 77843-3136
Phone: 979/595-5985,
Fax: 979/845-6481
Email: yajiezou@tamu.edu

ABSTRACT

Developing reliable statistical models is critical for predicting motor vehicle crashes in highway safety studies. However, the conventional statistical method ignores model uncertainty. Transportation safety analysts typically select a single "best" model from a series of candidate models (called model space) and proceed as if the selected model is the true model. This paper proposes a new approach for deriving more reliable and robust crash prediction models than the conventional statistical modeling method. This approach uses the Bayesian model averaging (BMA) to account for model uncertainty. The derived BMA crash model is an average of the candidate models included in the model space weighted by their posterior model probabilities. To examine the applicability of BMA to the Poisson and negative binomial (NB) regression models, the approach is applied to the crash data collected on 338 rural interstate road sections in Indiana over a five-year period (1995 to 1999). The results show that BMA was successfully applied to Poisson and NB regression models. More importantly, in the presence of model uncertainty, the proposed approach can provide better prediction performance than single models selected by conventional statistical techniques. Thus, this paper provides transportation safety analysts with an alternative methodology to predict motor vehicle crashes when model uncertainty is suspected to exist.

KEYWORDS: crash model, Poisson, negative binomial, Bayesian model averaging, prediction

INTRODUCTION

In highway safety analysis, regression models play a significant role in identifying relationships between motor vehicle crashes and different explanatory variables, predicting accident frequency and screening variables. Up to now, a large number of analysis tools and models for analyzing crash data have been proposed by transportation safety analysts (Lord and Mannering 2010). Among these models, the negative binomial (NB) model remains the most frequently used tool for crash-frequency modeling (e.g., Lord and Mannering 2010; Miaou 1994; Miaou and Lord 2003; Malyshkina et al. 2009). Recently, some new methodologies and models have been proposed for the purpose of modeling and predicting motor vehicle crashes. For example, artificial neural networks (ANN) have been suggested as an alternative method for analyzing and predicting accident frequency (e.g., Abdelwahab and Abdel-Aty 2002; Chang 2005). However, these models can sometimes overfit the data. To overcome this problem, a few researchers (Xie et al. 2007) have examined the Bayesian neural networks (BNN) and concluded that BNN are more efficient than NB models for predicting crashes. The support vector machine (SVM) model (Li et al. 2008) was recently applied to crash data collected in Texas and was found to predict crashes more accurately than both NB and BNN models. Haleem et al. (2010) used the multivariate adaptive regression splines (MARS) technique to predict motor vehicle crashes and showed that the MARS predicts crashes almost as effectively as the traditional NB models, and its goodness-of-fit performance seems to show promise for adequately predicting crashes.

Despite extensive efforts on modeling and predicting crash data, the conventional statistical approach faces a few important challenges. The selection of subsets of explanatory variables is a basic part for building a crash prediction model. Given the dependent variable accident frequency y_i and the candidate explanatory variables $X_1,...,X_k$, the general routine is to find the "best" regression model based on a selected number of variables to describe the crash frequency. In highway safety research, one typical approach is often to select a single "best" model based on some model selection criteria, such as log-likelihood, Akaike information criterion (AIC), Bayesian information criterion (BIC), Deviance information criterion (DIC), etc. (e.g., Park and Lord 2009; Pei et al. 2011). After the model is selected, further inferences are made with the assumption that the selected model is the true model. However, this approach neglects the uncertainty associated with the choice of models, especially those from the same category (e.g., Poisson model, NB or Poisson-lognormal model) but with different combinations of explanatory variables. The uncertainty between models may be important in making inference particularly in the cases where more than one models are considered plausible but differ in predictions (Li and Shi 2010). If the uncertainty about the model is ignored, the quantities of interest (accident frequency) may be underestimated. BMA combines and averages all possible models (models with different combinations of explanatory variables) when making inferences about the quantities of interest (crash frequency) (Raftery et al. 1997). By computing the average over many different competing models, BMA incorporates model uncertainty into modeling output related to the parameter estimation and prediction. BMA has been applied successfully in various fields including engineering (Li and Shi 2010), meteorology (Raftery et al. 2005), epidemiology (Viallefont et al. 2001), water resources (Duan et al. 2007), etc., and in most cases, BMA can

improve the prediction performance. In this study, we have two objectives: the first objective is to examine the applicability of BMA to the Poisson and NB regression models for traffic accident analysis (the most basic models for count data); the second objective is to compare the model prediction performance between BMA and the conventional statistical approach used in transportation safety analysis. To accomplish these two objectives, BMA is examined using accident data collected on 338 rural interstate road sections in Indiana.

The next section outlines the methodology used in this study.

METHODOLOGY

This section describes the characteristics of the NB regression and BMA, as well as the Occam's Window Method. This latter method is used for discarding models that predict much more poorly than their competitors in the model space.

Negative Binomial Regression

Because the crash-frequency data on a highway section are non-negative and discrete integers, the most basic model for modeling crash data is the Poisson regression model. The advantage of Poisson regression model is that it is easy to estimate the parameters. However, past studies (Lord and Mannering 2010) have indicated that the Poisson regression model cannot accommodate observed over-dispersion in crash data. Moreover, this model (and its sister the NB model described below) can be adversely influenced by the low sample-mean and small sample size bias (Lord, 2006). The NB regression model is an extension of the Poisson regression model and is used for handling the over-dispersion often observed in crash data. The derivation of the NB regression model is as follows: the number of crashes yi at roadway entity i during some time period is assumed to be Poisson distributed and independent over all entities, which is defined by:

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

where $P(y_i)$ is the probability of roadway entity i having y_i crashes for a given time period and λ_i is the expected accident frequency $E[y_i]$ for roadway entity i. The expected accident frequency λ_i is structured as a function of explanatory variables,

$$\lambda_i = \exp(\beta X_i) \quad (2)$$

where X_i is a vector of explanatory variables and β is a vector of estimable coefficients.

The NB regression model arises if we assume that the parameter λ_i follows a gamma distribution. A gamma-distributed error term is added to the parameter λ_i and equation (2) is rewritten as follows:

$$\lambda_i = \exp(\beta X_i + \varepsilon_i) \quad (3)$$

where exp( ) is the added error term with mean 1 and variance α, and α is the dispersion parameter. With this new structure, the mean is allowed to differ from the variance such that $VAR[y_i] = E[y_i][1+\alpha E[y_i]] = E[y_i]+\alpha E[y_i]^2$. Despite the documented limitations (Lord and Mannering 2010; Hilbe 2011; Zou et al., 2012), the NB model is popular for modeling crash data for several reasons. First, most statistical software programs have built-in functions that can handle such models. Second, two types of analysis commonly used in highway safety are available within the NB modeling framework. The first type of analysis is Empirical Bayesian method, and the second one

is related to the estimation of confidence and prediction intervals for NB models (see Lord 2006). Besides, Hauer (1997) also concluded that the NB model is the most common distribution used for modeling crash data because its marginal distribution has a closed form and this mixture results in a conjugate model.

Bayesian Model Averaging

When describing BMA, consider a model space M of K models M_k (k=1, 2,..., K) and let y denote the quantity of interest (a future observation of the accident frequency using new input data). The posterior distribution of y given the observed data D, is

$$p(y \mid D) = \sum_{k=1}^{K} p(y \mid M_k, D) p(M_k \mid D) \quad (4)$$

where $p(y \mid M_k, D)$ is the posterior distribution of y under model M_k given data D and $p(M_k \mid D)$ is the likelihood of M_k being the correct prediction model given the observational data D, which is also known as the posterior model probability (PMP). The output of BMA method is an average of the posterior distribution $p(y \mid M_k, D)$, weighted by the corresponding posterior probabilities, $w_k = p(M_k \mid D)$. For any model space, the sum of w_k equals 1. The posterior model probability is given by:

$$p(M_k \mid D) = \frac{p(M_k) p(D \mid M_k)}{\sum_{l=1}^{K} p(M_l) p(D \mid M_l)} \quad (5)$$

where $p(M_k)$ is the prior probability that M_k is the true model and $p(D \mid M_k)$ is the corresponding marginal model likelihood. In this study, the model space M is initially considered to be equal to the set of all possible combinations of explanatory variables. For a given set of N models, the results of the BMA approach depend on the specification of prior probability. When there is little prior information about the relative plausibility of the models considered, the assumption that all models are equally likely a priori is a reasonable "neutral" choice (Hoeting et al. 1999). The marginal model likelihood $p(D \mid M_k)$ is calculated by:

$$p(D \mid M_k) = \int p(D \mid \theta_k, M_k) p(\theta_k \mid M_k) d\theta_k \quad (6)$$

where θ_k is the vector of parameters in model M_k, $p(\theta_k \mid M_k)$ is the prior density of θ_k under model M_k, and $p(D \mid \theta_k, M_k)$ is the likelihood.

The posterior mean and variance of the BMA prediction can be defined as follows:

$$E[y \mid D] = \sum_{k=1}^{K} E(y \mid D, M_k) w_k \quad (7)$$

$$Var[y \mid D] = \sum_{k=1}^{K} (Var[y \mid D, M_k] + E[y \mid D, M_k]^2) w_k - E[y \mid D]^2 \quad (8)$$

Although BMA is theoretically attractive, two practical difficulties need to be solved before its implementation. First, the results of BMA heavily rely on the model space and it is necessary to select a proper set of candidate models. One obvious approach is to include all possible models. However, when the number of possible models is large, the process of the BMA method becomes very time-consuming. Currently, two approaches are available to solve this problem. One approach is called the Occam's window method, which will be introduced in the following section. The other approach, the Markov chain Monte Carlo model composition (MC3), uses a Markov chain Monte Carlo method to directly approximate model space in equation (4) (see Madigan and York 1995). The implementation of the MC3 is very complicated and the Occam's window

tends to be much faster computationally (Raftery et al. 1997). Thus, we adopted the Occam's window method.

The second difficulty associated with the BMA approach is that the marginal model likelihood may be analytically intractable especially in many cases where no closed form integral is available. Several alternative methods have been proposed in the literature to calculate or approximate the likelihood (Gibbons et al. 2008): (i) The most popular approximation of the marginal likelihood is the Laplace approximation which can be calculated at the posterior mode or at the maximum likelihood parameter estimates; (ii) Another approximation of the marginal likelihood is the harmonic mean estimator. This estimator is relatively simple, but it is quite unstable and sensitive to small likelihood values and hence is not recommended in this study; (iii) Kass and Wasserman (1995) derived the Bayesian Information Criterion as a rough but adequate approximation, and this BIC approximation was used in this paper.

Occam's Window Method

Because the number of terms in equation (4) can be very large, the Occam's window approach was used to discard models that predict much poorer than their competitors. The Occam's window algorithm was first developed by Madigan and Raftery (1994). Raftery et al. (1997) later applied this method to linear regression models. There are two basic principles under the Occam's window method. First, if a model predicts the data far poorer than the model which provides the best predictions, then this model should be excluded from the model space and no longer be considered. Those models not belonging to

$$A' = \left\{ M_k : \frac{\max_l \{p(M_l \mid D)\}}{p(M_k \mid D)} \le C \right\} \quad (9)$$

should be discarded in equation (4). The $\max_l \{p(M_l \mid D)\}$ is the model with the highest PMP and the value of C is determined by the data analyst. Usually, the value of C is equal to 20 and we also used $C=20$ in this study.

The second (optional) principle is called Occam's razor and this method is used to exclude complex models that receive less support from the data than any of their simpler submodels. Those models excluded from model space belong to

$$B = \left\{ M_k : \exists M_l \in M, M_l \subset M_k, \frac{p(M_l \mid D)}{p(M_k \mid D)} > 1 \right\} \quad (10)$$

This method can significantly reduce the number of models in the sum in equation (4). Typically the number of terms in equation (4) can be reduced to fewer than 20 models and often to as few as one to two models. The equation (4) can be rewritten as:

$$p(y \mid D) = \sum_{M_k \in A} p(y \mid M_k, D) p(M_k \mid D) \quad (11)$$

where $A = A' \setminus B \in M$.

To implement the proposed principles, this study adopted the leaps and bounds algorithm as the search strategy. For more details about the search strategy, interested readers should read Raftery's paper (1995).

DATA DESCRIPTION

The dataset used for this study contains crash data collected on 338 rural interstate road sections in Indiana over a five-year time period from 1995 to 1999. The data have been investigated in previous studies (e.g., Anastasopoulos et al. 2008; Geedipally et al. 2012). Explanatory variables in table 1 are considered

to construct a set of model space M for the Bayesian model averaging in the study. The available highway geometric design information includes length of section, minimum friction reading, pavement surface type, median width, presence of median barrier, presence of interior shoulder and interior shoulder width; while the available traffic information contains average daily traffic (ADT) of various vehicle types and truck percentage. During the five-year study period, there were 5,737 crashes. The summary statistics for the model variables are presented in table 1. As shown in this table, the observed crash frequency ranges from 0 to 329, and the mean frequency is 16.97. For a complete list of variables in this dataset, interested readers can consult (Washington et al. 2011).

RESULTS AND DISCUSSION

This section describes the modeling results for Poisson regression and NB regression models using the BMA approach. Despite the fact that Poisson regression model has significant disadvantages and is now rarely used for analyzing crash data (Lord and Mannering 2010), this study considers this regression model as an example to demonstrate the usefulness of BMA. When analyzing the crash data, we consider the segment length as an offset term which means that the number of crashes is linearly proportional to the segment length. Thus, we have 8 candidate explanatory variables, and these variables can potentially result in $2^8 = 256$ different models. For the model averaging strategies, all possible combinations of candidate explanatory variables are assumed to be equally likely a priori. The Occam's window method is implemented to exclude the models with poor prediction performance. The results show that the BMA approach can provide additional insight in interpreting the explanatory variables and averaging over the selected models provides better prediction performance than basing inference on a single model in the NB regression example. All statistical analyses were carried out in an R package.

Poisson Regression Model

The BMA approach was performed using the leaps and bounds algorithm and the results are provided in tables 2 and 3. Table 2 contains the selected models with the highest posterior probabilities using the Occam's window

TABLE 1 Summary Statistics of Characteristics for the Data

Variable	Minimum	Maximum	Mean(SD)	Sum
Number of crashes (5 years) X_1*	0	329	16.97 (36.30)	5,737
Average daily traffic over the 5 years (ADT) X_2	9,442	143,422	30,237.6 (28776.4)	
Minimum friction reading in the road section over the 5-year period (FRICTION) X_3	15.9	48.2	30.51 (6.67)	
Pavement surface type (1: asphalt, 0: concrete) (PAVEMENT) X_4	0	1	0.77 (0.42)	
Median width (in feet) (MW) X_5	16	194.7	66.98 (34.17)	
Presence of median barrier (1: present, 0: absent) (BARRIER) X_6	0	1	0.16 (0.37)	
Presence of interior shoulder (1: present, 0 absent) (SHOULDER) X_7	0	1	0.93 (0.26)	
Interior shoulder width (in feet) (SW) X_8	2.7	24.1	5.35 (2.80)	
Percentage of trucks (average daily) (TRUCKS) X_9	7.32%	44.87%	31.74%	
Segment length (in miles) (L) X_{10}	0.009	11.53	0.89 (1.48)	300.09

* X_1 is the serial number of variable number of crashes.

method. As shown in this table, only two models are selected based on the Occam's window method. The model with the higher posterior model probability accounts for 90% of the total posterior probability. Although the amount of model uncertainty is not significant for this case, there still exists model uncertainty to some degree. Compared to Model 1, Model 2 excludes the variable X_7, presence of interior shoulder. Table 3 lists the posterior means of $\beta \mid D$, standard deviations of $\beta \mid D$ and posterior effect probabilities $P(\beta \neq 0 \mid D)$ for the coefficient associated with each variable using the BMA approach. The posterior effect probability $P(\beta \neq 0 \mid D)$ for one explanatory variable is obtained by summing the posterior model probabilities of models that contain that explanatory variable. Using the conventional statistical technique and assuming the full model, the estimates, standard errors and p-values for the coefficients are also provided in table 3. Note that all standard deviations using the BMA approach are larger than their corresponding standard errors using the full model. This is because those parameter estimates and standard deviations directly incorporate model uncertainty (Hoeting et al. 1999). Another point to note is that the posterior effect probability of coefficient associated with variable X_7 is 90%, and this is because only Model 1 in the model space includes variable X_7 in the analysis.

As shown in table 3, we can see that all explanatory variables except variable X_7 are highly important when predicting the crash frequency. Both posterior effect probabilities and p-values indicate that there is a very strong evidence of an effect. The posterior effect probabilities are all 100% and p-values are less than 0.0001. The estimated coefficient values

TABLE 2 Models with Highest Posterior Model Probabilities for Poisson Regression

Model number	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	PMP***
1	T*	T	T	T	T	T	T	T	0.90
2	T	T	T	T	T	F**	T	T	0.10

* T denotes that the explanatory variable is considered in the corresponding model.
** F means that the explanatory variable is NOT considered in the corresponding model.
*** PMP is the posterior model probability.

TABLE 3 Comparison of BMA Results to Full Model for Poisson Regression

	Bayesian model averaging			Full model		
Variable	Mean $\beta \mid D$	SD* $\beta \mid D$	$P(\ \neq 0 \mid D)$	Estimate	SE**	p value
Ln(ADT) X_2	0.7069974	0.035122	100	0.706032	0.035035	< 2e-16
FRICTION X_3	-0.02241202	0.002124	100	-0.02246	0.00212	< 2e-16
PAVEMENT X_4	0.32118	0.044125	100	0.322125	0.044051	2.62e-13
MW X_5	-0.003429911	0.000758	100	-0.0034	0.000752	6.28e-06
BARRIER X_6	-3.498241	0.357748	100	-3.55938	0.315016	< 2e-16
SHOULDER X_7	-0.9032813	0.437834	90	-0.99939	0.340566	0.00334
SW X_8	-0.07734632	0.018532	100	-0.07867	0.018197	0.0000154
TRUCKS X_9	-1.497767	0.164029	100	-1.50363	0.163252	< 2e-16

* SD is the standard deviation.
** SE means the standard error.

of the variables from both approaches demonstrate that: first, an increase in ADT is found to be linked to an increase in the crash frequency (although non-linear). Road sections with asphalt surface tend to have more crashes than sections with concrete surface. Second, the increases of other variables are found to be associated with a decrease in the crash frequency. For the explanatory variable X_7, the p-value indicates that the effect is significant and posterior effect probability concludes that there is a strong effect.

Negative Binomial Regression Model

The BMA approach was also applied to the NB regression model and the results are presented in tables 4 and 5. As the BMA results in table 4 indicate, the model with the highest posterior model probability accounts for 89.7% of the total posterior probability. Thus, we can conclude that there is a certain amount of model uncertainty. Compared with other selected models, Model 1 is in a dominant position and this model considers only two explanatory variables X_6 and X_9. Table 5 gives the statistics of the coefficient associated with each variable using the BMA approach and the conventional statistical technique. For the two explanatory variables X_6 and X_9, since the corresponding posterior effect probabilities are equal to 100% and the p-values are less than 0.001, both posterior effect probabilities and p-values demonstrate that they have very strong effects on the crash frequency. For the other six explanatory variables, the results show that there is a qualitative difference between the two methods. If 0.01 is chosen as the significance level, then five variables (X_2, X_4, X_5, X_7 and X_8) are rejected based on the reported p-values. On the one hand, for the variables X_2, X_4, X_5, X_7 and X_8, the p-values indicate that the effect is insignificant and the posterior effect probabilities conclude that there is a weak or no effect. On the other hand, for the variable X_3, the posterior effect probability indicates that the variable minimum friction has no effect on the crash frequency, while the corresponding p-value shows that the effect of minimum friction on the crash frequency is significant. Overall, the posterior effect probabilities of the four variables (X_2, X_3, X_4 and X_5) imply weaker evidence for these effects given the corresponding p-values. This is because the p-values from the full model do not take account of model uncertainty, and the p-values thus overstate the evidence for the effects (Hoeting et al. 1999). For some variables, the posterior mean of coefficients are 0, which means the results shrink the estimates toward zero (Hoeting et al. 1999).

Prediction Performance Comparisons

In problems where model uncertainty is present, BMA can yield prediction performance improvements over single selected models. This conclusion has been verified in various fields, as discussed above. In order to measure the applicability of BMA in predicting the crash data, the mean absolute deviance (MAD), the mean squared predictive error (MSPE) and the logarithmic score (LS) were used to compare the model prediction performance between BMA and the conventional statistical approach. The first two performance indexes (MAD and MSPE) were calculated as follows: $MAD = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$ and $MSPE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$, where n is the testing data size, and y_i and $\hat{y}_i$ are the observed and predicted numbers of accidents for observation i, respectively (Oh et al. 2003). The LS was introduced by Good (1952) and previous studies (e.g., Hoeting et al. 1999; Madigan and Raftery 1994) have used the

TABLE 4 Models with Highest Posterior Model Probabilities for Negative Binomial Regression

Model number	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	PMP
1	F	F	F	F	T	F	F	T	0.897
2	F	F	F	F	T	T	F	T	0.052
3	F	F	F	T	T	F	F	T	0.051

TABLE 5 Comparison of BMA Results to Full Model for Negative Binomial Regression

	Bayesian model averaging			Full model		
Variable	**Mean** $\beta \mid D$	**SD** $\beta \mid D$	$P(\beta \neq 0 \mid D)$	**Estimate**	**SE**	**p value**
Ln(ADT) X_2	0	0	0	0.345986	0.161449	0.032*
FRICTION X_3	0	0	0	-0.02822	0.010116	0.005
PAVEMENT X_4	0	0	0	0.393769	1.72E-01	0.022*
MW X_5	-0.00026525	0.001242	5.1	-0.00362	0.002089	0.083*
BARRIER X_6	-2.830011	0.280228	100	-3.08466	0.406729	3.35e-14
SHOULDER X_7	-0.02144128	0.142661	5.2	-0.57188	0.502978	0.256*
SW X_8	0	0	0	-0.02819	0.038981	0.470*
TRUCKS X_9	-3.847502	0.627213	100	-2.66588	0.779984	0.000631

* Insignificant at 0.01 level of significance.

LS to measure the prediction performance of BMA. The observed data in this study are randomly split into two subsets. The first subset is referred to as the build data. We apply the BMA method and conventional statistical approach to this subset of data. Then, the second subset defined as the test data is used to measure the prediction performance. The number of sections used for building model is 238, and the number of sections used for testing is 100. The logarithmic score measures the prediction ability of an individual model, M_i, using the equation, $-\sum_{y \in D^T} \mathrm{Ln}\left(p(y \mid M_i, D^B)\right)$. D^B is the build data and D^T is the test data. Then the prediction performance of BMA is examined using the equation, $-\sum_{y \in D^T} \mathrm{Ln}\left(\sum_{M_i \in A} p(y \mid M_i, D^B) p(M_i \mid D^B\right)$. To make the comparison results more convincing, the random data separation process was repeated for four times and four scenarios were considered. Smaller MAD, MSPE and LS values indicate a better overall prediction performance for the given model.

Table 6 reports the MAD, MSPE, and LS values of the competing methods for the NB regression model. Bold values in table 6 are the smallest MAD, MSPE, and LS values among selected models. As shown in table 6, except the MAD value in scenario 1, all other goodness-of-fit values indicate that BMA can improve the model prediction accuracy for the test data. The difference in LS of 15.38 (between BMA and full model in scenario 1) can be viewed as an improvement in prediction performance. For example, if the average prediction probability, $-\sum_{y \in D^T} p(y \mid M, D^B)/100$, is 25%, and the corresponding logarithmic score is $-\mathrm{Ln}(0.25) = 1.386$. Then after implementing BMA, the new average prediction probability will be $\exp(-(1.386-\frac{15.38}{100}))=29.2\%$. This means that BMA can predict the number of crashes 4.2%

more accurately than the method using the full model. Although the difference appears small for some of the measures, it is large enough that BMA should be selected over the full model (see a related discussion about GOF and biased models in Lord (2006)). In sum, in predicting the crash frequency of the test data, the proposed BMA model outperforms the conventional models based on MAD, MSPE and LS values. Thus, we conclude that BMA can improve the prediction performance for the NB regression model.

Discussion

In this study, the results showed that BMA can provide better prediction performance than the conventional statistical technique for the NB regression model. The findings suggest that BMA may be an appropriate methodology for predicting crash data; note that BMA should not be used for examining relationships between variables. Thus, further studies are needed to examine the applicability of BMAs to other types of crash model. In previous studies, the conventional statistical techniques were commonly used for traffic accident analysis partially because the built-in functions for crash models are available in many statistical software programs, and usually the analysis results can be easily interpreted and provide clear and valuable information for traffic safety analysts in order to make further inferences. In contrast, the advantage of BMA is that this model overcomes the problem in accounting for model uncertainty by conditioning, not on a single "best" regression model, but on the entire statistical regression model space, and the output of BMA combines inferences and predictions from multiple candidate models. For the NB regression example, a total of 3 models were selected. Another advantage of BMA is that, in the presence of model uncertainty, it can yield prediction performance improvements over single selected models. Despite the above merits of BMA, there are a few limitations associated with this model. First, when the number of explanatory variables in crash data is large, for instance, 20 explanatory variables are included in the analysis, then, the application of the Occam's window method is

TABLE 6 Performance Index Values for Negative Binomial Regression Models

Scenario	Performance Index	Full model	Model with significant variables*	BMA
	MAD	5.82	6.35	5.85
	MSPE	100.63	118.88	91.21
1	LS	292.28	288.89	276.90
	MAD	7.99	8.00	7.76
	MSPE	285.77	289.96	266.41
2	LS	311.65	308.61	298.57
	MAD	8.91	8.40	7.60
	MSPE	357.49	314.37	231.53
3	LS	314.75	312.98	304.16
	MAD	5.48	5.44	5.10
	MSPE	84.19	83.55	64.89
4	LS	281.00	277.65	266.57

* Model with significant variables at a significance level of 0.05.

very time-consuming because there are a total of 2^{20} = 1,048,576 candidates model in the complete model space. Therefore, the efficiency of BMA can be compromised by the number of explanatory variables examined in the analysis. Second, after Occam's window method is applied, in our experience, the number of terms in equation (4) can be reduced to fewer than 20, often to as few as 1 or 2. For example, as illustrated in tables 2 and 4, two or three models are selected based on Occam's window method. This finding may be different from the typical application of BMA and thus understate the value of BMA. To better demonstrate its usefulness, some other statistical models (i.e., Poisson-lognormal, Poisson-Weibull, etc.) for analyzing crash data could be used. Another way to increase model uncertainty is to implement BMA without using Occam's window.

CONCLUSIONS

This paper has documented the application of the Bayesian model averaging approach for predicting motor vehicle crashes. Crash data collected on rural interstate road sections in Indiana were analyzed using the proposed approach. Poisson and NB regression models were used to establish the relationship between traffic accident frequency and highway geometric variables and traffic characteristics. The results of this study revealed that the model uncertainty problem can be solved or at least minimized using BMA; and, in the presence of model uncertainty, the proposed approach can provide better prediction performance than single models selected by conventional statistical techniques for the NB models. This study also presented a new methodology in predicting the traffic accident frequency. For future work, since the crash data used in this study were collected at rural interstate roads, an application of BMA to other types of data would be meaningful. Moreover, it would also be interesting to examine the results of applying BMA to more complex crash prediction models, such as the newly introduced Negative Binomial-Lindley model (Geedipally et al. 2012). Finally, this study did not apply the Markov chain Monte Carlo model composition method to directly approximate the terms in equation (4). The Occam's window method and the Markov chain Monte Carlo model composition method should be compared, and their influence on the modeling results investigated.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Fred Mannering from Purdue University for graciously providing us with the Indiana data.

REFERENCES

Abdelwahab, H. T. and M. A. Abdel-Aty. 2002. Artificial Neural Networks and Logit Models for Traffic Safety Analysis of Toll Plazas. *Transportation Research Record* 1784:115–125.

Anastasopoulos, P.C., A. Tarko, and F. Mannering. 2008. Tobit Analysis of Vehicle Accident Rates on Interstate Highways. *Accident Analysis and Prevention* 40(2):768–775.

Chang, L.Y. 2005. Analysis of Freeway Accident Frequencies: Negative Binomial Regression versus Artificial Neural Network. *Safety Science* 43(8):541–557.

Duan, Q., N.K. Ajami, and S. Sorooshian. 2007. Multimodel ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources* 30(5):1371–1386.

Geedipally, S.R., D. Lord, and S.S. Dhavala. 2012. The Negative-Binomial-Generalized-Lindley Generalized Linear Model: Characteristics and Application using Crash Data. *Accident Analysis & Prevention* forthcoming.

Gibbons, J.M., G.M. Cox, A.T.A. Wood, J. Craigon, S.J. Ramsden, D. Tarsitano, and N.J.M. Crout. 2008. Applying Bayesian averaging to mechanistic models: an example and comparison of methods. *Environmental Modelling and Software* 23(8):973–985.

Good, I.J. 1952. Rational decisions. *Journal of the Royal Statistical Society* Series B 14(1):107-114.

Haleem, K., M. A. Abdel-Aty, and J. Santos. 2010. Multiple Applications of Multivariate Adaptive Regression Splines Technique to Predict Rear-End Crashes at Unsignalized Intersections. *Transportation Research Record* 2165: 33–41.

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

Hilbe, J.M., 2011. *Negative Binomial Regression*, 2nd Edition, Cambridge University Press, Cambridge, UK.

Hoeting, J.A., D. Madigan, A.E. Raftery, and C.T. Volinsky. 1999. Bayesian model averaging: a tutorial. *Statistical Science* 14(4):382–417.

Kass, R.E. and L. Wasserman. 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz Criterion. *Journal of the American Statistical Association* 90(431):928–934.

Li, G. and J. Shi. 2010. Application of Bayesian model averaging in modeling long-term wind speed distributions. *Renewable Energy* 35(6):1192–1202.

Li, X., D. Lord, Y. Zhang, and Y. Xie. 2008. Predicting Motor Vehicle Crashes Using Support Vector Machine Models. *Accident Analysis and Prevention* 40(4):1611–1618.

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., and F.L. Mannering. 2010. The Statistical Analysis of Crash-frequency Data: A Review and Assessment of Methodological Alternatives. *Transportation Research* Part A 44(5):291–305.

Madigan, D. and A.E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89(428):1535–1545.

Madigan, D. and J. York. 1995. Bayesian graphical models for discrete data. *International Statistical Review* 63:215–232.

Malyshkina, N.V., F.L. Mannering, and A.P. Tarko. 2009. Markov Switching Negative Binomial Models: an Application to Vehicle Accident Frequencies. *Accident Analysis and Prevention* 41(2):217–226.

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. and D. Lord. 2003. Modeling Traffic Crash-flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes versus Empirical Bayes. *Transportation Research Record* 1840:31–40.

Oh, J., C. Lyon, S.P. Washington, B.N. Persaud, and J. Bared. 2003. Validation of the FHWA Crash Models for Rural Intersections: Lessons Learned. *Transportation Research Record* 1840:41-49.

Park, B.J. and D. Lord. 2009. Application of Finite Mixture Models for Vehicle Crash Data Analysis. *Accident Analysis and Prevention* 41(4):683–691.

Pei, X., S.C. Wong, and N.N. Sze. 2011. A joint-probability approach to crash prediction models. *Accident Analysis and Prevention* 43(3):1160–1166.

Raftery, A.E. 1995. Bayesian model selection in social research. *Sociological Methodology* 25:111–163.

Raftery, A.E., D. Madigan, and J.A. Hoeting. 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92(437):179–191.

Raftery, A.E., T. Gneiting, F. Balabdaoui, and M. Polakowski. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* 133:1155–1174.

Viallefont, V., A.E. Raftery, and S. Richardson. 2001. Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine* 20:3215–3230.

Washington, S., M. Karlaftis, and F. Mannering. 2011. *Statistical and Econometric Methods for Transportation Data Analysis*. Second edition, Chapman and Hall/CRC, Boca Raton, FL.

Xie, Y., D. Lord, and Y. Zhang. 2007. Predicting Motor Vehicle Collisions Using Bayesian Neural Network Models: An Empirical Analysis. *Accident Analysis & Prevention* 39(5):922–933.

Zou, Y., D. Lord, and Y. Zhang, 2012. Analyzing highly dispersed crash data using the Sichel generalized additive models for location, scale and shape. Working paper.

Lane Width Crash Modification Factors for Curb-and-Gutter Asymmetric Multilane Roadways: Statistical Modeling

THOBIAS SANDO, PH.D., P.E., PTOE
GEOPHREY MBATTA, PH.D.
REN MOSES, PH.D., P.E.

School of Engineering
University of North Florida
1 UNF Drive
Jacksonville, Florida 32256
Phone: 904-620-1142
Fax: 904-620-1391
Email: t.sando@unf.edu

Department of Civil Engineering
FAMU-FSU College of Engineering
2525 Pottsdamer Street, Room 129-B
Tallahassee, FL 32310
Phone: (850) 410-6587
Fax: (850) 410-6587
Email: mbattageo@eng.fsu.edu

Department of Civil Engineering
FAMU-FSU College of Engineering
2525 Pottsdamer Street, Room 129
Tallahassee, FL 32310
Phone: (850) 410-6191
Fax: (850) 410-6142
Email: moses@eng.fsu.edu

ABSTRACT

This study developed lane width crash modification factors (CMFs) for urban curb-and-gutter multilane roadways with asymmetric lanes, i.e., outside lane wider than inside lane. The roadway segments used were urban four-lane with a raised median (4D) and with a two-way left-turn lane (5T). Three crash categories were evaluated: KABCO (Fatal (K), incapacitating-injury (A), non-incapacitating injury (B), possible injury (C) and property damage only crashes (O)), KABC (Fatal (K), incapacitating-injury (A), non-incapacitating injury (B), and possible injury crashes (C)), and PDO (property damage only) crashes.

A cross-sectional method was used as it was the most practical and feasible for this study. Six-year (2004 to 2009) of segment crashes were examined. The analysis involved statistical modeling using the negative binomial model, whose coefficients were used to develop multiplicative CMF equations for a combined effect of variable inside and outside lane width.

In summary, the results show that reducing the inside lane width from 12 ft to 11 ft does not affect estimated crash frequency of four-lane with a raised median segments for all three crash categories, and PDO crashes for four-lane with a two-way left-turn lane segments. However, narrowing the inside lane width appears to be associated with increased estimated KABCO and KABC crashes for four-lane with a two-way left-turn lane sections. The

KEYWORDS: Crash modification factors, lane width, asymmetric roadways

results also suggest that widening the outside lane from the baseline 12 ft causes a reduction in estimated crash frequency for all three crash categories (KABCO, KABC, and PDO) for both four-lane with a raised median and four-lane with a two-way left-turn lane segments.

INTRODUCTION

This paper presents the process that was used to develop crash modification factors (CMFs) for urban multilane curb-and-gutter asymmetric lanes in the state of Florida. The CMFs developed in this study describe change in safety when a typical section measuring 12 ft for both inside and outside lanes is changed to an asymmetric section, also known as an atypical configuration. An example of this would be changing a 12 ft inside and outside lane to an 11 ft inside lane and 13 ft outside lane. Most of these roadway configurations result from retrofitting, by repainting, or widening of the roadway rendering the outside lane a shared lane for bicycles and motor vehicles.

LITERATURE REVIEW

There are several studies that were conducted to develop lane width crash modification factors (CMFs) for two-lane rural highways (Griffin and Mak (1987), Zeeger et al. (1987), Harwood et al. (2000), Harwood et al. (2003), and Harkey et al. (2008)). These studies are also cited in the *Highway Safety Manual*, HSM (2010). They were all conducted on two-lane rural highways. Separate CMFs were reported for roadways with average annual daily traffic (AADT) less than 400 motor vehicles per day and for roadways with AADT greater than 2000 motor vehicles per day. These CMFs indicate that widening of lanes reduce a specific set of related accident types, namely single-vehicle run-off-road accidents, multiple-vehicle head-on, opposite-direction sideswipe, and same-direction sideswipe collisions (Harkey et al. (2008) & Highway Safety Manual (2010)). This decrease was relative to 12 ft lane width, which was considered the base line of comparison.

In another lane width study, Lord and Benneson (2007) developed lane width CMFs for two-lane rural highway frontage roads for the state of Texas. Rural frontage roadways differ from rural two-lane roadways because they have restricted access along at least one side of the road, a higher percentage of turning traffic, and periodic ramp-frontage-road terminals with yield control. The results showed increased crash frequency as lane width decreased from 12 ft to 9 ft.

CMFs reported in the Highway Safety Manual (2010) for rural multilane roadways were developed by the study that was conducted by Harkey et al. for the National Cooperative Highway Research Program (NCHRP) (2008). An expert panel was used to develop CMFs for rural multilane roadways. Lane width CMFs developed by modeling crashes for multilane highways are absent. Furthermore, it is clear that CMFs reported in the HSM (2010*)* for rural multilane (both divided and undivided) highways may not apply to urban multilane roadways. This is due to the difference in traffic operations and level of activities surrounding urban highways.

In a recent lane width study, Potts et al. (2007) investigated the relationship between lane width and safety for roadway segments on urban and suburban arterials. The study by Potts et al. did not develop CMFs. The study did not find any indication of safety risk on urban and suburban arterials when lane width narrower than 12 ft was used.

Based on the summary of the literature review, two main observations need special attention. First, the average lane width was used in all

previous studies that developed CMFs for lane width. While averaging may apply to symmetric lane configurations, such as 12 ft inside lane and 12 ft curb lane, it may be too simplistic for asymmetric sections, which have wider curb lanes and narrow inside lanes. Second, all existing CMFs for lane widths were developed for rural highways. None of the CMFs reported in previous studies were developed to specifically address the safety consequences of lane width in urban roadways. These two observations are the motivation of this study as it employs individual lane measurements instead of the average of aggregated lane width and focuses on urban segments, helping to fill the knowledge gap that exists in lane width CMF development.

Rural and urban highways differ in their cross-sectional geometric configurations. Rural highways have shoulders while urban highways considered in this study have curb-and-gutter. The shoulder provides room for road users to veer to the right if they are on the outside lane to avoid a crash while curb-and-gutter roadways causes a constraint for lateral movement to the right of the travel lanes. Also, bicyclists do not share a lane with motorists on rural highways. They ride on the shoulder to the right of the white stripe. This study presents the analysis of urban wide curb lanes, i.e., outside lanes widths greater than 12 feet, to accommodate bicyclists and motorists on the same lane.

RESEARCH OBJECTIVE

The objective of this study was to evaluate the safety of urban multilane roads with atypical lane width configurations, i.e., outside lane width greater than the standard lane width (12 ft) and narrower inside lane narrower than 12 ft. This objective was accomplished by developing Crash Modification Factors for two types of atypical multilane urban cross sections namely urban four-lane roadways with divided median (4D) and urban four-lane with two-way left-turn lane (5T). Crash Modification Factors quantify the change in expected average crash frequency (crash effect) caused by implementing a particular treatment. The value of Crash Modification Factor below 1 indicates treatment causes crash reduction while Crash Modification Factor greater than 1 indicates that the treatment is expected to result in an increased number of crashes. a Crash Modification Factor of 1 represents no effect on safety.

DATA COLLECTION

Roadway Data

Databases Used

Roadway characteristics inventory (RCI): This database was used to identify the type of road configuration and roadway characteristics including: total surface lane width, number of lanes, shoulder type, and traffic characteristics. All four-lane with a raised median and four-lane with a two-way left-turn lane were filtered and further analyzed using Florida Department of Transportation (FDOT) as-built plans.

FDOT scanned copy of as-built plans: FDOT archives scanned copies of as-built plans for state maintained roadway projects. The database has most of the roadway plans for completed projects and projects that are under construction. The advantage of as-built plans over RCI database is that they show individual lane width while roadway characteristics inventory database shows the total surface width. From the as-built plans, using roadway ID obtained from roadway characteristics inventory database, four-lane with a raised median and four-lane with a two-way left-turn lane multilane roadways with asymmetric lanes were verified

by examining individual lane width. Additional data obtained from the as-built plans include individual lane width, type of median, number of driveways, number of median openings, and approximated segment length.

Comparison Sites

Comparison sites were roadways with standard lane width of 12 ft for both inside and outside lanes. They were obtained by using a technique suggested by Bonneson and Pratt (2008). At first, sites adjacent to selected asymmetric lanes on the same roadway (figure 1a) were chosen to ensure that the pairs were homogenous to the asymmetric segments. However, it was not possible to get sufficient data using this technique. Therefore, the selection was expanded to consider parallel (figure 1b) and intersecting (figure 1c) the selected asymmetric segments. Parallel and intersecting segments were considered only if their roadway characteristics were similar to the paired asymmetric segments and had a comparable average annual daily traffic (AADT). The attributes used for the selection of comparison sites were number of lanes, median type, posted speed limit, degree of curve, type of shoulder (curb), and type of onstreet parking.

Verification of Roadway Geometric Information for Asymmetric and Comparison Segments

Two issues emerged when reviewing the as-built drawings. First, it was discovered that most of as-built plans are not updated regularly. Second, there was inconsistency in the way the curb lane was measured. Therefore, as-built plans measurements were verified by performing field measurements for segments with asymmetric lanes and comparison sites. A total of 918 road segments were verified. After field verification, 454 segments (49.5% of all segments) were dropped as their characteristics differed those recorded on RCI and as-built drawings, hence did not qualify for analysis. A minimum of 100 segments is recommended for modeling (Agrawal and Lord, 2006). After field verification, both 4D and 5T were found to have enough segments for modeling with a total of 224 and 240 segments, respectively, for both segments with asymmetric lanes and comparison segments.

Crash Data

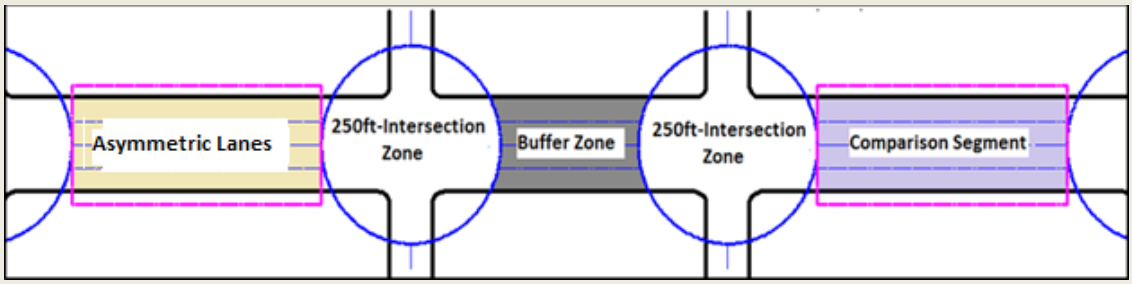
Statewide crash data was obtained from crash analysis reporting (CAR) database, an electronic repository of crashes maintained by FDOT. The data was from 2004 to 2009. The location of each crash was linearly referenced to the Florida Department of Transportation roadway system using the milepost system indexed by the roadway identification number (Roadway ID). Data was filtered to remain with mid-block crashes only. All crashes that occurred within a radius of 250 ft from the center of intersections were discarded.

DATA ANALYSIS

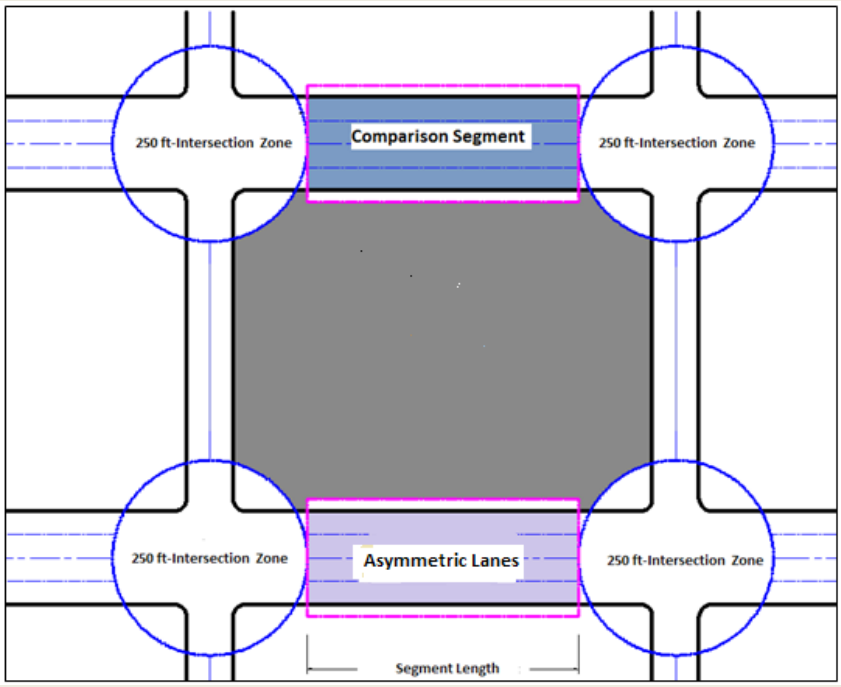
Crash Rate Analysis for Segments With Asymmetric Lanes Configuration

Categories of outside lane width were formed by grouping ranges of lane widths as follow: 11.8 ft–12.2 ft formed a 12 ft category; 12.3 ft–12.7 ft formed a 12.5 ft category; 12.8 ft–13.2 ft formed a 13 ft category; 13.3 ft–13.7 ft formed a 13.5 ft category; and 13.8 ft–14.2 ft formed a 14 ft category. It should be noted that for the 12 ft category of outside lane width, the inside lane width was also12 ft (comparison sites). However, for all other lane width categories, the inside lane width was fixed to 11 ft. These categories were used for an explanatory analysis whose results are reported in table 1.

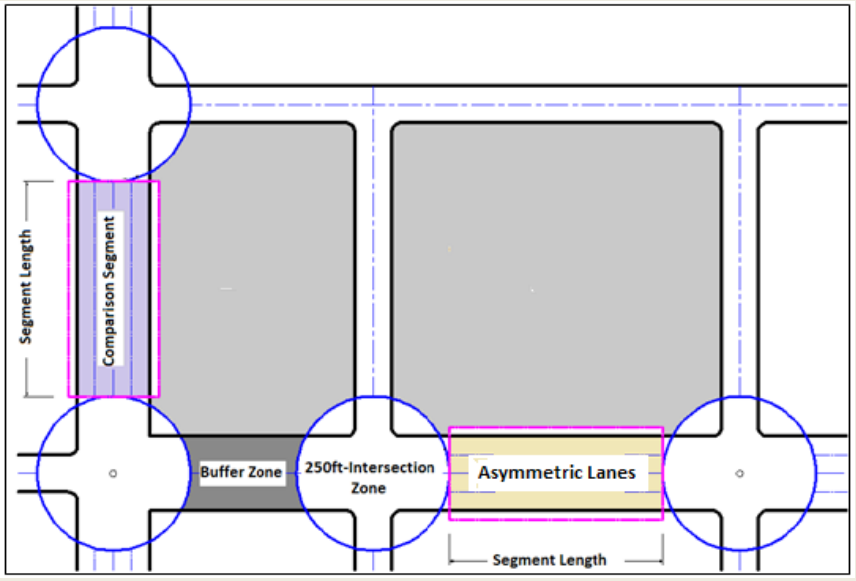
FIGURE 1 Criteria Used in Selection of Comparison Segments



(a)



(b)



(c)

As can be seen in table 1, crash categories are described using three acronyms, i.e., KABCO, KABC, and PDO crashes, derived from the Highway Safety Manual naming convention. KABCO stands for fatal (K), incapacitating (A), non-incapacitating (B), possible injury (C) and PDO (O) crashes while KABC represents fatal (K), incapacitating (A), non-incapacitating (B), and possible injury (C) crashes. PDO is used for crashes that result in property damage only.

Figure 2 is a graphical representation of the results shown in table 1, depicting the relationship between crash rate per million vehicle miles (mvm) and the outside lane width. The two graphs presented in figure 2 show an increase of crashes when outside lane width increased from 12 ft (with inside lane width of 12 ft) to 12.5 ft (with an inside lane of 11 ft). There is a discernible pattern of decreased crash rate as the outside lane width is increased from 12.5 ft to 14 ft with a fixed inside lane width of 11 ft. This trend was observed for all three crash categories, i.e., KABCO, KABC, and PDO crashes.

Statistical Modeling

Selection of the Statistical Model

Two regression count models used to analyze crash data are Poisson and Negative Binomial. Poison regression distribution requires the mean and variance of the dependent variable to be equal. For most crash data, the variance of the crash frequency exceeds the mean and, in such case, the data would be overdispersed. The Highway Safety Manual (2010) specifically calls for the use of the Negative binomial model in lieu of Poisson model because the degree of overdispersion in a negative binomial model is represented by a statistical parameter, known as the *overdispersion parameter* that is estimated along with the coefficients of the regression equation. The larger the value of the overdispersion parameter, the more the crash data vary as compared to a Poisson distribution with the same mean.

Selection of the Function

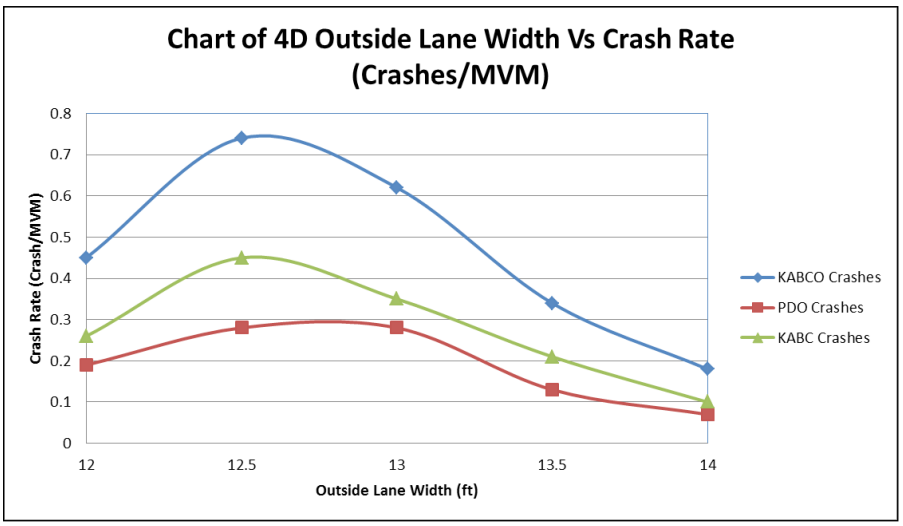
The first step toward development of predictive models is the selection of the functional

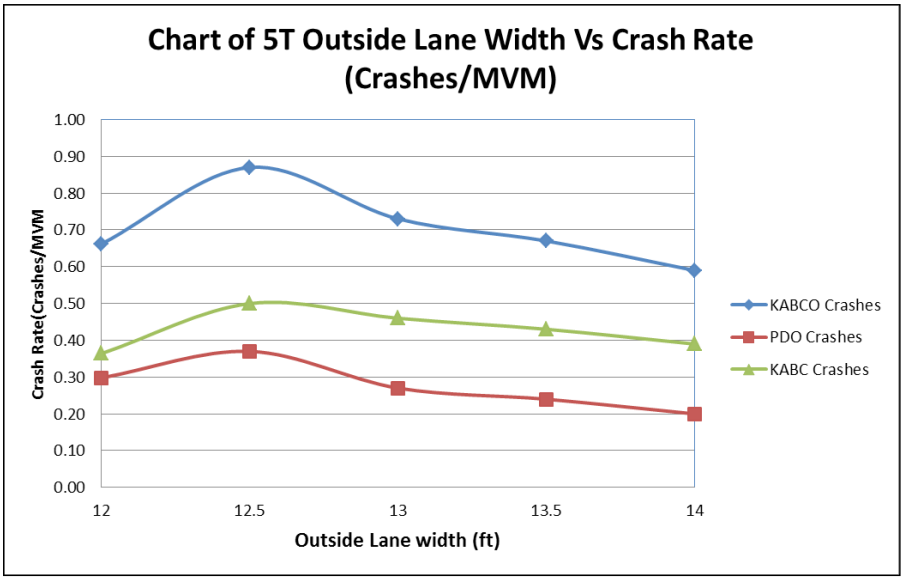
TABLE 1 Crashes Rate for Different 4D Outside Lane Width Categories

Inside lane width (ft)	Outside lane width (ft)	Exposure (mvm)	KABCO crashes	PDO crashes	KABC crashes	KABCO crashes/mvm	PDO crashes/ mvm	KABC crashes/ mvm
12	*12.0	1631.24	739	313	426	0.45	0.19	0.26
11	12.5	297.81	219	84	135	0.74	0.28	0.45
11	13.0	595.56	372	164	208	0.62	0.28	0.35
11	13.5	482.45	162	63	99	0.34	0.13	0.21
11	14.0	254.45	45	19	26	0.18	0.07	0.10
Crashes rate for different 5t outside lane width categories								
12	*12.0	634.31	420	189	231	0.66	0.30	0.36
11	12.5	215.94	187	79	108	0.87	0.37	0.50
11	13.0	209.46	153	56	97	0.73	0.27	0.46
11	13.5	88.16	59	21	38	0.67	0.24	0.43
11	14.0	120.84	71	24	47	0.59	0.20	0.39

*Comparison sites with inside lane width of 12.0. All other categories have inside lane width of 11.0 ft.

FIGURE 2 Graphs of Outside Lane Width and Crashes Rate by Severities





form. Normally, the function is determined empirically after several runs of different variable combinations which correlate the dependent variable (outcome variable) to the model covariates. Different functions were considered and fitness of resulting models was assessed. After several trials of different combination of variables, the function based on Negative Binomial (NB) model presented as equation 1 was selected.

$$\mu_i = \beta_o L_i (ADT)^{\beta_1} e^{\sum_{i=2}^{n} x_{ji}\beta_i} \quad (1)$$

Equation 1 was simplified to provide a linear relationship between the dependent variable and covariates by taking natural logarithm on both sides. The resulting formula is presented as equation 2.

$$\ln(\mu_i) = \ln(\beta_o) + \ln(L_i) + \beta_1 \ln(ADT) + \sum_{i=2}^{n} x_{ji}\beta_i \quad (2)$$

Where

$AADT$ = is an average annual daily traffic over six years of study period

L_i = segment length

μ_i = mean number of crashes for six year period for site i

$x_1, x_2, \ldots\ldots x_n$ = explanatory variables

$\beta_o, \beta_1, \ldots\ldots, \beta_n$ = regression coefficients to be estimated

Selection of Explanatory Variables

Previous studies have found that roadway cross-section variables such as lane width, median width, median type, grade, segment length, and degree of curve have contribution to occurrences of crashes (Zeeger et al. (1987) & Harkey et al. (2008)). Mauga and Kaseko (2010) found median opening density and driveway density to have contributed to the increase in crashes in urban multilane roads. Also, AADT and posted speed limit have been widely reported as important variables in crash modeling (Harkey et al. (2008), HCM (2010), & Mauga and Kaseko (2010)). In this study, two main explanatory variables—AADT and segment length were considered to be key variables that relate number of crashes to predictors. In addition, inside and outside lane width were considered as study variables and were given equal importance as key variables. Due to the nature of sites used for this study, other variables including posted speed limit (mph), median width (ft), degree of curve (degree), and driveway density (number of driveway/0.1 mile) and median opening density were also added in the model. Number of median opening density was found to be irrelevant for four-lane with a two-way left-turn lane configuration as the configuration does not restrict turning at any point. However, it was important for four-lane with a raised median configuration as turning to access adjacent properties is only through median openings.

It is worth noting that in previous studies, driveway density and median opening density had been expressed in terms of number of driveways/mile or number of openings/mile. However, for this study, segment lengths were ranged from 0.01 mile to 0.52 mile for four-lane with a two-way left-turn lane configuration with 0.1 mile being a mean value. In order to avoid having high values for driveway density and median openings, the mean segment length of 0.1 mile was used to scale the median openings and driveway density.

Negative Binomial (NB) Model Selection and Evaluation

The negative binomial model was developed to analyze the influence of the independent variables on three response variables i.e., KABCO, KABC, and PDO crashes. Model results were tested at 0.05 level of significant. All insignificant variables were removed to form a reduced model. A reduced model was re-run and tested again at the same level of significant.

Thereafter, a comparison of the full and reduced models was performed using two information-theoretic approach indicators i.e., Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The general criterion for comparison is that the model with a smaller value of AIC and BIC is considered to be better. The values of BIC and AIC for the reduced models were smaller than those of the full models for all three response variables (KABCO, KABC, and PDO) crashes. The reduced model was then selected for all three response variables.

Model Results for 4D Segments

Model results for four-lane with a raised median segments are reported in table 2. The results revealed an increase in KABCO, KABC, and PDO crashes as outside lane widths is decreased. The increase in crashes was significant when tested at 95% confidence level with *p*-values of 0.010, 0.044 and 0.0153 for the

TABLE 2 Results for Urban Four-Lane Roadway With Divided Median (4D)

4D Segments—metadata for the KABCO, KABC, and PDO crashes

Variable	Mean	Standard deviation	Minimum	Maximum
AADT	37,510	7,383	25,100	52,500
Segment length (length)	0.17	0.1	0.01	0.64
Outside lane width (ft)	12.63	0.7	12	14
Inside lane width (ft)	11.68	0.4	11	12
Median opening density	0.68	1.3	0	10.6
Driveway density (drive way/0.1mile)	1.16	2.1	0	19

KABCO crashes

Parameter	Estimate	Standard error	p-values	Comment
Intercept	-33.1729	4.8553	<.0001	Significant
Log of AADT	3.7903	0.4586	<.0001	Significant
Log of length	0.3505	0.126	0.0054	Significant
Outside lane width (ft)	-0.3591	0.1395	0.0101	Significant
Median opening density	0.1713	0.0716	0.0167	Significant
Deviance (value/df):	1.12			
Over-dispersion parameter k:	1.51			
BIC:	1198.95			
AIC:	1177.82			
Pearson $\chi 2$ (value/df):	1.04			

KABC crashes

Intercept	-31.8083	5.321	<.0001	Insignificant
Log of AADT	3.5618	0.5011	<.0001	Significant
Log of length	0.3944	0.1391	0.0046	Significant
Outside lane width (ft)	-0.3113	0.1546	0.044	Significant
Median opening density	0.1921	0.0802	0.0165	Significant
Deviance (value/df):	1.05			
Over-dispersion parameter k:	1.71			
BIC:	1002.31			
AIC:	981.98			
Pearson $\chi 2$ (value/df):	1.01			

PDO crashes

Intercept	-38.6478	5.6048	<.0001	Insignificant
Log of AADT	4.2327	0.5342	<.0001	Significant
Log of length	0.3038	0.1444	0.0354	Significant
Outside lane width (ft)	-0.3743	0.1544	0.0153	Significant
Median opening density	0.1584	0.0736	0.0315	Significant
Deviance (Value/df):	1.01			
Over-dispersion parameter k:	1.51			
BIC:	869.59			
AIC:	849.26			
Pearson $\chi 2$ (value/df):	1.03			

outside lane width. The effect of the inside lane width was insignificant, therefore the coefficient was removed. Also, the increase in median opening density resulted in the increase of KABCO, KABC, and PDO crashes. This was evident as the *p*-values of 0.0167, 0.0165, and 0.0315 for KABCO, KABC, and PDO crashes were observed. These results were consistent to the Mauga and Kaseko (2010) study which observed the increase in injury crashes with an increase in median opening density.

Model Results for Four-Lane With a Two-Way Left-Turn Lane Segment

Table 3 presents the model results for four-lane with a two-way left-turn lane segments. According to the results reported in table 3, both KABCO and KABC crashes increased with reduced lane width for both lanes (inside and outside). The results were significant at 95% confidence level. For KABC crashes, *p*-values of 0.0184 and 0.0294 for inside and outside lane, respectively, were observed while for KABCO crashes, *p*-values for inside and outside were 0.0493 and 0.0106, respectively.

Both KABCO and KABC crashes were significantly correlated to driveway density. The increase in driveway density resulted in the increase in KABCO and KABC crashes. *P*-values of 0.0334 and 0.0007 for KABCO and KABC crashes, respectively, were observed. This finding is consistent to the results reported by Mauga and Kaseko (2010) which observed the increase in injury crash rate as driveway densities were increased. However, with respect to PDO crashes, the inside lane width and driveway densities were found to be insignificant not only at 95%, but also at 90% confidence level.

The influence of AADT was found to be significant for all three response variables (KABCO, KABC and PDO crashes). The model yielded *p*-values of 0.0001, 0.0001 and 0.0001 for KABCO, KABC and PDO crashes, respectively, for four-lane with a raised median segments. *P*-values of 0.0001, 0.0001, 0.0111, for KABCO, KABC and PDO crashes, respectively, were observed for four-lane with a two-way left-turn lane segments. In all three cases for the four-lane with a two-way left-turn lane segments, the model coefficient for segment length was approximately 1.000, therefore was used as an offset variable.

Developing Crash Modification Factors

Method Used

The Highway Safety Manual (2010) provides a list of methods that can be used for developing Crash Modification Factors. The most preferred methods are controlled experiments and the Empirical Bayes method using the before-and-after data. Due to the difficulty in obtaining the exact date that a treatment was implemented, the before-and-after analysis was not feasible for this study. Another method recommended by the Highway Safety Manual (2010), i.e., the cross-sectional method was therefore adopted as it does not require the "before" period data for analysis. Instead, it employs the treatment and comparison sites of "after" period data for analysis. It is the same method that was used by Lord and Bonneson (2007) to estimate Crash Modification Factors for rural frontage roads in Texas. The method estimates Crash Modification Factors by using coefficients developed from regression models, for this case, coefficients reported in tables 2 and 3. Crash Modification Factors for each specific response variable follow an exponential relationship shown in equation 3.

$$CMF_i = e^{\beta_i[x_i - y_i]} \quad (3)$$

TABLE 3 Results for Urban Four-Lane Roadway With TWLTL (5T)

5T Segments—metadata for the KABCO, KABC, and PDO crashes

Variable	Mean	Standard deviation	Minimum	Maximum
AADT	22,078	7,118	7,480	43,929
Segment length (length)	0.10	0.07	0.01	0.52
Outside lane width (ft)	12.60	0.7	12	14
Inside lane width (ft)	11.50	0.5	11	12
Median width (ft)	12.10	1	10	14
Driveway density (drive way/0.1mile)	5.00	3	0	24

KABCO crashes

Parameter	Estimate	Standard error	p-values	Comment
Intercept	7.9121	6.8456	0.2478	Insignificant
Log AADT	1.0190	0.2412	<.0001	Significant
Driveway density	0.0504	0.0237	0.0334	Significant
Outside lane width (ft)	-0.5887	0.2305	0.0106	Significant
Inside Lane width (ft)	-0.6318	0.3214	0.0493	Significant
	Deviance (value/df):	1.11		
	Over-dispersion parameter k:	1.07		
	BIC:	999.53		
	AIC:	979.56		
	Pearson χ2 (value/df):	1.38		

KABC crashes

Intercept	5.6102	6.7447	0.4055	Insignificant
Log AADT	1.1963	0.2523	<.0001	Significant
Driveway density	0.0823	0.0243	0.0007	Significant
Outside lane width (ft)	-0.4978	0.2285	0.0294	Significant
Inside lane width (ft)	-0.7452	0.3162	0.0184	Significant
	Deviance (value/df):	1.09		
	Over-dispersion parameter k:	0.87		
	BIC:	795.99		
	AIC:	815.96		
	Pearson χ2 (value/df):	1.40		

PDO crashes

Intercept	-1.2237	3.9480	0.7566	Insignificant
Log AADT	0.9114	0.3587	0.0111	Significant
Outside lane width (ft)	-0.4092	0.1835	0.0258	Significant
	Deviance (value/df):	0.90		
	Over-dispersion parameter k:	2.14		
	BIC:	676.39		
	AIC:	659.75		
	Pearson χ2 (value/df):	1.14		

Where

x_i= range of values or a specific value investigated (e.g. lane width, etc.)

y_i = baseline conditions or average conditions for the variable

β_i= regression coefficient

Lane Width Crash Modification Factors

The lane width of 12 ft for both inside and outside lanes was considered as a base condition. Since CMFs are multiplicative factors when used to predict crash frequencies, based on the results presented in tables 2 and 3, and equation 3, the resulting CMFs were derived as:

- *four-lane with a raised median Crash modification functions*

CMF for KABCO crashes

$$CMF_{KABCO} = e^{-0.36[x_{outside}-12]} \quad (4)$$

CMF for KABC crashes

$$CMF_{KABC} = e^{-0.31[x_{outside}-12]} \quad (5)$$

CMF for PDO crashes:

$$CMF_{PDO} = e^{-0.37[x_{outside}-12]} \quad (6)$$

- *four-lane with a two-way left-turn lane Crash modification functions*

CMF for KABCO crashes:

$$CMF_{KABCO} = e^{-0.59[x_{outside}-12]}. e^{-0.63[x_{inside}-12]} \quad (7)$$

CMF for KABC crashes:

$$CMF_{KABC} = e^{-0.50[x_{outside}-12]}. e^{-0.75[x_{inside}-12]} \quad (8)$$

CMF for PDO crashes:

$$CMF_{PDO} = e^{-0.41[x_{outside}-12]} \quad (9)$$

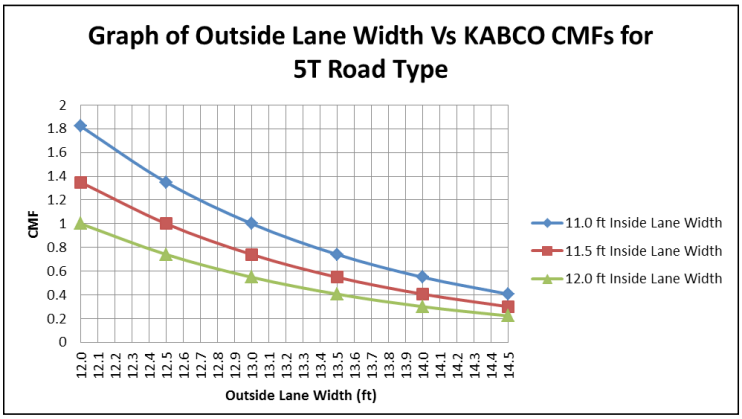
It can be noted that the coefficient of the inside lane width is not included in the Crash Modification Factors of all three crash categories for four-lane with a raised median segments (equations 4 to 6) and Crash Modification Factor of PDO crashes for four-lane with a two-way left-turn lane segments (equation 9). This is because the inside lane width was not significant for these particular cases as explained in previous sections.

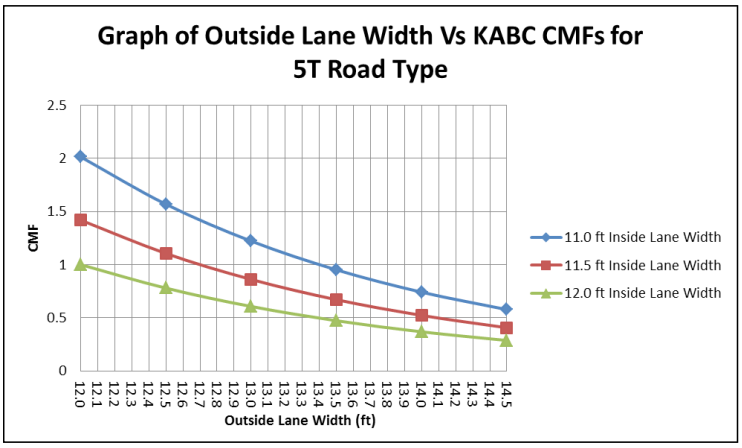
Crash Modification Factor Curves and Interpretation

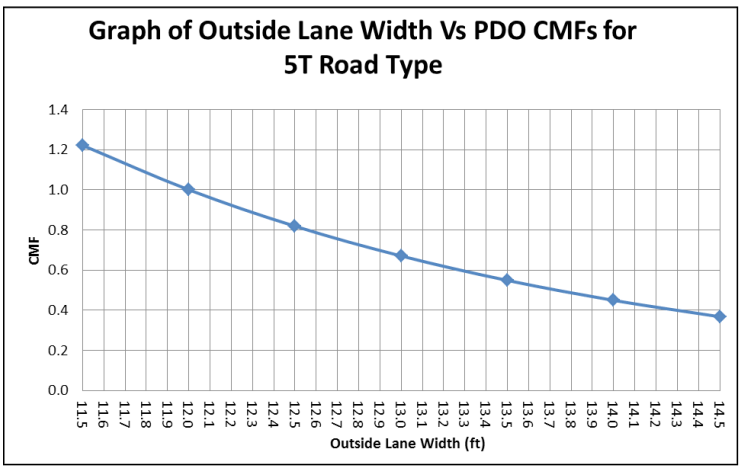
Figures 3 and 4 show Crash Modification Factor curves for KABCO, KABC, and PDO crashes for four-lane with a two-way left-turn lane segments and four-lane with a raised median segments, respectively. The six curves were developed by substituting lane widths of 11 ft and 12 ft for the inside lane and varying lane widths of 12.5 to 14 for the outside lane, in equations 4 through 9. The base CMF of 1.00 corresponds to the segments with the inside and outside lane width of 12 ft each.

When considering different combination of inside and outside lane widths, the following observations were made. For four-lane with a two-way left-turn lane configuration (figure 3), apart from the base condition (12 ft for both inside and outside lanes), Crash Modification Factor of 1.0 for KABCO crashes was observed for a combination of 11 ft inside and 13 ft outside lane and a pair of 11 ft inside and 13 ft outside lane width. The combination of 11.5 ft and 13 ft resulted to Crash Modification Factor of 0.75, which indicates reduced estimated average KABCO crash frequency in comparison to the base condition. For KABC crashes, the combination of 11 ft inside/13 ft outside lane width and 11.5 ft inside/12.5 ft outside lane width yielded Crash Modification Factors greater than 1.00, which indicates an increase in estimated KABC crashes. On the other hand, the combination of 11.5

FIGURE 3 Graphs of CMFs for Four-Lane With a Two-Way Left-Turn Lane Segments



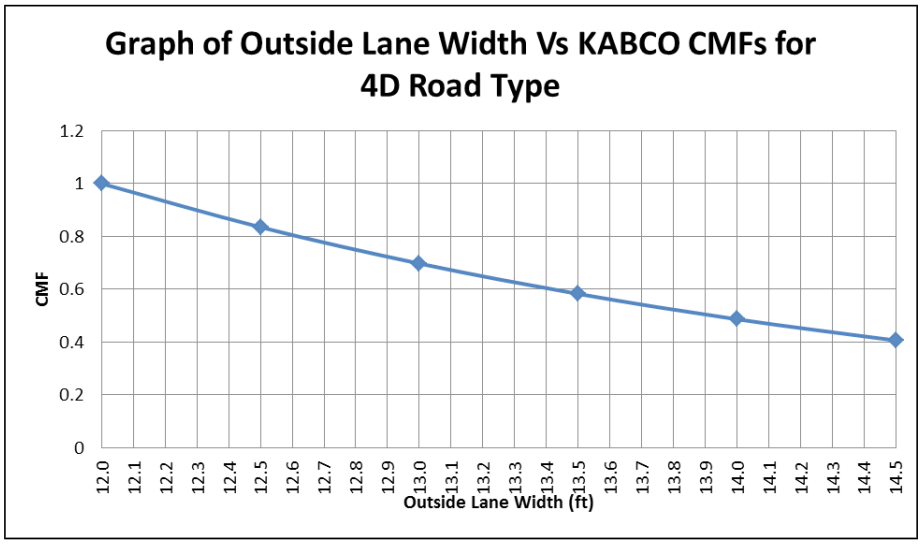


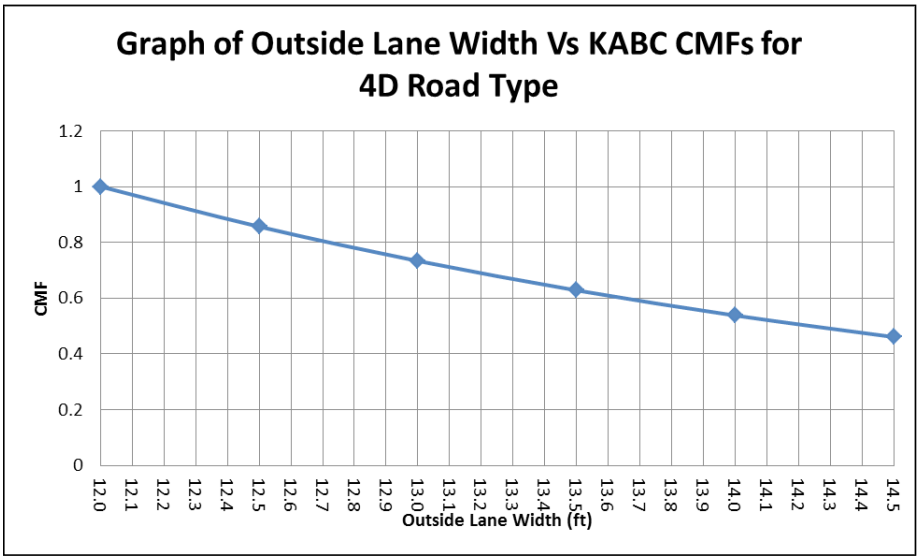


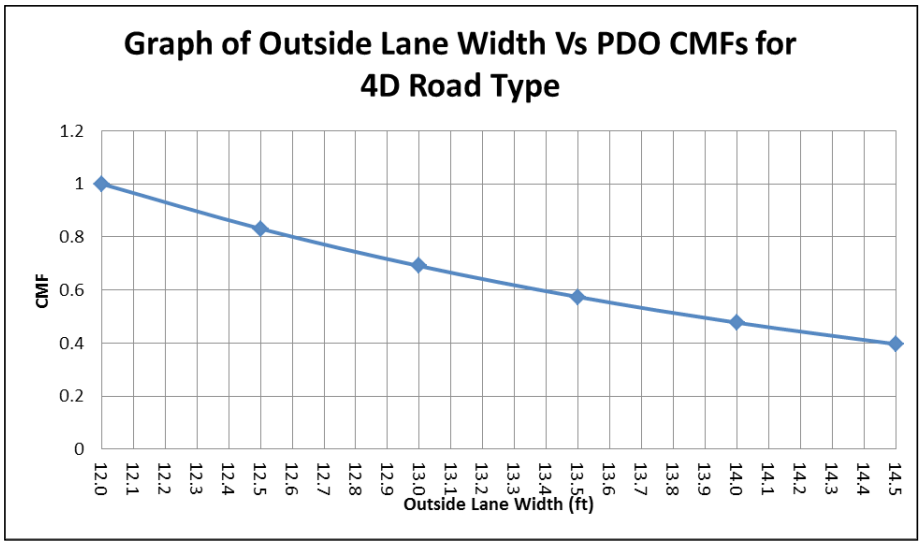
ft and 13 ft resulted in a CMF Crash Modification Factor smaller than 1.00, which indicates a reduction in estimated KABC crashes. With respect to PDO crashes, it was outside lane width which had an effect on Crash Modification Factor. As the width increased to greater than 12 ft the Crash Modification Factor was less than one indicating reduction in PDO crashes.

For four-lane with a raised median configuration (figure 4), 60% reduction for KABCO and PDO crashes was observed as the Crash Modification Factor decreased from 1.0 to 0.4 as the outside lane width increases from 12 ft to 14.5 ft. The crash reduction of 66% was observed for KABC crashes as the outside lane width widens from 12 ft to 14.5 ft.

FIGURE 4 Graph of CMFs for four-lane with a raised median segments







SUMMARY

This study developed lane width crash modification factors for asymmetric urban multilane roadways. The roadway segments used were urban four-lane with a raised median (4D) and with two-way left-turn lane (5T). In total, the analysis reported in this study used 25 centerline miles of four-lane with a two-way left-turn lane segments and 39 centerline miles of four-lane with a raised median roadways.

Development of Crash Modification Factors followed a protocol described by the Highway Safety Manual (2010). The cross-sectional method was used. Negative binomial regression models were used to model the relationship between crash frequency and model variables. Variables considered in modeling included driveway density, median opening density, posted speed limit, inside lane width, outside lane width, median width, segment length, and average annual daily traffic (AADT). Six years (2004–2009) of segment crashes were examined. Three crash categories were evaluated: KABCO (Fatal (K), incapacitating-injury (A), non-incapacitating injury (B), possible injury (C) and property damage only crashes (O)), KABC (Fatal (K), incapacitating-injury (A), non-incapacitating injury (B), and possible injury crashes (C)), and PDO (property damage only) crashes.

The results of the safety analysis are summarized in table 4. These values are calculated using equations 4 through 9. A Crash Modification Factor of 1.00 indicates no influence in causing crashes while Crash Modification Factors smaller and greater than 1.00 indicate that a change of a variable from a base value causes a decrease and increase in crashes, respectively. According to the results depicted in table 4, for four-lane with a raised median segments, the effect of inside lane width is insignificant, indicating that the decrease of lane width from 12 ft to 11 ft does not cause an increase in crash frequency. According to the results, crashes decrease as the outside lane width is increased from 12 ft. This decrease is seen on all types of crashes analyzed in this study, i.e., KABCO, KABC, and PDO.

For four-lane with a two-way left-turn lane sections, the results show an increase in crashes as the inside lane width is reduced to 11 ft while the outside lane width is increased to 12.5 ft. This trend was observed for both KABCO and KABC crashes, but not for PDO crashes. However, the combination of 11.5 ft or more for the inside lane and 13 ft for the outside lane width resulted in the decrease in crashes for KABCO and KABC crashes. Crash Modification Factors for PDO crashes were found to be independent of the inside lane width, but dependent of outside lane width. Relative to outside lane width of 12 ft, the Crash Modification Factors for PDO crashes were found to decrease as the outside lane width increased.

As stated above, for four-lane with a raised median segments, narrowing the inside lane from 12 ft to 11 ft did not result in an increase in crash frequency for any of the three types of crashes. Also, for four-lane with a two-way left-turn lane segments, the decrease in inside lane width was not significant for PDO crashes. It was only significant for KABCO and KABC crashes, hence higher values of Crash Modification Factors for KABCO and KABC crashes for four-lane with a two-way left-turn lane. As far as four-lane with a two-way left-turn lane segments are concerned, higher Crash Modification Factor values for KABCO and KABC crashes might have been attributed to the type of median and might have less to do with the inside lane width. Having higher values of Crash Modification Factors for KABCO

TABLE 4 Comparison of CMFs for 4D and 5T When Inside Lane Width is Fixed to 11 ft While Outside Lane Width Varies

	[4D CMF] (5T CMF) Ratio (5T CMF)/(4D CMF)					
Outside lane width range (ft)	11.8-12.2	12.3-12.7	12.8-13.2	13.3-13.7	13.8-14.2	14.3-14.7
Outside lane width (ft)	12	12.5	13	13.5	14	14.5
	[1.00]	[0.84]	[0.70]	[0.58]	[0.49]	[0.41]
	(1.88)	(1.40)	(1.04)	(0.77)	(0.58)	(0.43)
CMF for KABCO crashes	1.88	1.67	1.49	1.32	1.18	1.05
	[1.00]	[0.86]	[0.73]	[0.63]	[0.54]	[0.64]
	(2.12)	(1.65)	(1.28)	(1.00)	(0.78)	(0.61)
CMF for KABC crashes	2.12	1.92	1.75	1.59	1.44	0.95
	[1.00]	[0.83]	[0.69]	[0.57]	[0.48]	[0.40]
	(1.00)	(0.81)	(0.66)	(0.54)	(0.44)	(0.36)
CMF for PDO crashes	1.00	0.98	0.96	0.95	0.92	0.9

and KABC crashes (total crashes) on roads with a two-way left-turn lane is consistent with studies by Mauga and Kaseko (2010) and 15 studies reviewed by Gluck et al. (1999). These studies reported crash reduction that range from 3% and 57% for KABCO crashes on roads with raised median in comparison to segments with two-way left-turn lane. Mauga and Kaseko (2010) also found a decrease of 21% on KABC crashes for roads with raised median in comparison to those with two-way left-turn lane.

Table 4 also shows the ratio between the Crash Modification Factors developed for four-lane with a raised median and four-lane with a two-way left-turn lane segments with a fixed inside lane of 11 ft while outside lane width varied from 12.5 ft to 14.5 ft. The results revealed that with respect to KABCO crashes, the Crash Modification Factor for four-lane with a two-way left-turn lane segments, when the inside lane width is 11 ft and the outside lane width is 12 ft is 1.88 times that of four-lane with a raised median segments. The ratio decreases as the outside lane width increases from 12.5 ft to 14.5ft, where the four-lane with a two-way left-turn lane CMF is 1.05 times that of four-lane with a raised median segments. A similar trend was observed for KABC crashes as the ratio decreased from 2.12 to 0.95 as the outside lane width increased from 12 ft to 14.5 ft while keeping the inside lane width constant at 11 ft. As can be seen in table 4, for PDO crashes, the ratio of Crash Modification Factors for four-lane with a raised median segments to Crash Modification Factors for four-lane with a two-way left-turn lane segments is smaller than 1.0, indicating that for PDO crashes, a higher crash reduction is expected for four-lane with a two-way left-turn lane segments than for four-lane with a raised median segment when the outside lane width is widened while keeping the inside lane fixed at 11 ft.

When comparing a typical 12 ft inside and a 12 ft outside through lane width segment (a total of 24 ft) with an asymmetric segment of an 11 ft inside lane and a 13 ft outside through lane (also, a total of 24 ft), the results in table

4 show that a four-lane with a raised median asymmetric lane configuration would result in fewer crashes (See highlighted cells—CMFs of 0.70, 0.73, and 0.69 for KABCO, KABC, and PDO crashes, respectively). For four-lane with a raised median configurations, given a total of 24 ft pavement width for both lanes, the results presented in table 4 indicate that restriping a roadway 12 ft to an 11 ft inside and a 13 ft outside through lane would result in a decrease in crashes. For four-lane with a two-way left-turn lane sections, the results are mixed, showing a slight increase for KABCO and KABC crashes (CMFs of 1.04 and 1.28, respectively) and a reduction of PDO crashes (CMF of 0.66), when a typical roadway is retrofitted to an 11 ft inside and a 13 ft outside through lane, respectively. The results also show that as the width of outside lane increases, for both four-lane with a raised median and four-lane with two-way left-turn lane configurations, crashes decrease.

RECOMMENDATIONS FOR FURTHER STUDY

This study is not without limitations. The most preferred methods for developing Crash Modification Factors are controlled experiments and observational before-and-after studies. This study used a cross-sectional method. A before-and-after method would have given more robust results but was not practical or feasible as exact dates when standard 12 ft lanes were retrofitted to create asymmetric lanes could not be obtained.

The results of this study are not without bias. The Highway Safety Manual protocol calls for use of homogeneous segments for obtaining crash modification factors. Hence segments tend to be shorter, rendering a small number of crashes per segment, potentially causing higher dispersion of data. Although sites were selected randomly from around the state of Florida, many potential sites were dropped from analysis because there was no homogeneous comparison sites, i.e., sites with similar variables except for a few variables considered in the model.

Lane width Crash Modification Factors for urban roadways do not exist. Therefore, there were no existing Crash Modification Factor equations to compare the results with. The robustness of CMFs developed by statistical modeling is improved by using homogeneous sites, i.e., sites with similar properties, whereas the only variables are AADT, segment length, and the treatment variable, for this case, lane width. This was not practical as it was not possible to get sufficient segments with similar properties such as the posted speed limit, median opening density, and driveway density. Also, due to limited data, area type was not considered as a variable. A much wider study is recommended, which will develop lane width separate Crash Modification Factors for residential, industrial, and central business district areas. This study did not model the effect of truck percentage due to lack of accurate data for truck traffic at studied sites. Future studies should consider truck percentage as it might have significant contribution to crash occurrences. Last but not least, further research is needed to calibrate the developed Crash Modification Factors to make them useful elsewhere other than Florida.

REFERENCES

Agrawal, R. and D. Lord. Effects of Sample Size on Goodness-of Fit Statistic and Confidence Intervals of Crash Prediction Models Subjected to Low Sample Mean Values. In Transportation Research Record: Journal of the Transportation Research Board, No. 1950, *Transportation Research Board of the National Academies*, Washington, D.C., 2006, pp. 35-43.

Bonneson. A. J., and. P. M. Pratt. A Procedure for Developing Accident Modification Factors from Cross

Section Data. Paper No. 08-0323. Present at the 87th Annual Meeting of the Transportation Research Board, Washington, D.C., January, 2008.

Gluck, J., H. S. Levinson, and V. Stover. *Impacts of Access Management Techniques*. NCHRP Report 420: Transportation Research Board, National Research Council,Washington, D.C., 1999.

Griffin, L. I., and K. K. Mak. *The Benefits to Be Achieved from Widening Rural, Two-Lane Farm-to-Market Roads in Texas*. Report No. IAC (86-87)-1039, Texas, Transportation Institute, College Station, TX., 1987.

Harkey, D.L., S. Raghavan, B. Jongdea, F.M. Council, K. Eccles, N. Lefler, F. Gross, B. Persaud, C. Lyon, E. Hauer, and J. Bonneson. *Crash Reduction Factors for Traffic Engineering and ITS Improvement*. NCHRP Report No. 617, Transportation Research Board, Washington, DC., 2008.

Harwood, D.W., E.R.K. Rabbani, K. R. Richard, H. W. McGee, and G. L. Gittings. *Systemwide Impact of Safety and Traffic Operations Design Decisions for 3R Projects*. NCHRP Report No 486, Transportation Research Board of the National Academies, Washington, D.C., 2003.

Harwood, D.W., F. M. Council, E. Hauer, W. E. Hughes, and A. Vogt. *Prediction of the Expected Safety Performance of Rural Two-Lane Highways*. Report No. FHWA-RD-99-207. Federal Highway Administration, U.S. Department of Transportation, December 2000.

Highway Safety Manual. Transportation Research Board of the National Academies, 2010.

Lord, D., and J. Bonneson. Development of Accident Modification Factors for Rural Frontage Road Segments in Texas. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2023, Transportation Research Board of the National Academies, Washing-ton, D.C., 2007, pp. 20-27.

Mauga, T., and M. Kaseko. Modeling and Evaluating Safety Impacts of Access Management Features in the Las Vegas, Nevada, Valley. *In Transportation Research Record: Journal of the Transportation Research Board*, No. 2171, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 57-65.

Potts, I., Harwood, D., and K. Richard. Relationship of Lane Width to Safety for Urban and Suburban Arterials. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2023, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 63-82.

Zegeer, C.V., D.W. Reinfurt, J. Hummer, L. Herf, and W. Hunter. *Safety Effects of Cross-Section Design for Two-Lane Roads*. Publication FHWA-RD-87-008. FHWA, U.S. Department of Transportation, 1987.

A Multidimensional Clustering Algorithm for Studying Fatal Road Crashes

BARAK FISHBAIN
OFFER GREMBEK

Department of Environmental, Water and Agricultural Engineering, The Faculty of Civil & Environmental Engineering, Technion-Israel Institute of Technology Haifa 32000, Israel

Safe Transportation Research and Education Center, University of California, Berkeley, CA 94720-7374, USA

ABSTRACT

Road fatalities are rare outcomes of events that occur in a small time-space region. Although the exact chain of events for each fatality is unique, there are inherent similarities between road fatalities. The science of road safety is dedicated to identifying such similarities, mainly using statistical analysis tools. Researchers typically analyze patterns that emerge over space, such as hot-spot studies, or patterns that emerge over time, such as before-after studies. Traffic research enumerates 84 parameters that characterize a road fatality. A vast number of papers have tried to find the correlation between one or two parameters. In those studies quite often the contribution of other factors is omitted. In this research we utilize a clustering graph theoretic method, known as graph-cuts, for segmenting a very large crash dataset (i.e., all fatal car crashes in the last 2, 5, or 10 years), while incorporating all available crash information into the process. The analysis of the clusters allows one to find subtle trends and significant causes for traffic fatalities. With this method, for example, we have found high correlation between hit-and-run and pedestrians fatalities, which was overlooked by previous studies. An additional output of the research is a full description of the typical fatality, thus all factors that characterized the representative crash in a cluster.

INTRODUCTION

The study of road crashes and their causes affects many fields, such as vehicle design, road design, transportation planning, law enforce-

KEYWORDS: traffic safety, vehicles crashes, graph cuts, clustering, fatal accidents

ment, policy making, and actuarial science. The identification and analysis of factors affecting the frequency and severity of crashes is a fundamental step in all of the aforementioned applications and disciplines. Traffic crashes are expected to form clusters in the spatiotemporal space as many of the crashes' contributing factors exhibit spatial and temporal patterns (McGuigan 1981; Plug et al. 2011; Shino 2008; Steenberghen et al. 2004). For example, collision frequency is typically tied to traffic volumes, which present spatial and temporal patterns (Eleni et al. 2007). Spatial correlation is typically attributed to highway infrastructure, its access and egress points and its deficient design and maintenance (Black 1991). In this study we apply data cluster analysis for identifying the factors affecting the frequency of fatal crashes.

The spatial distribution of crashes has been investigated mainly in an attempt to identify high collision concentration locations (i.e., hot spots) (McGuigan 1981; Steenberghen et al. 2004). Many studies have utilized point pattern analysis for this task (Bailey and Gatrell 1995; Cressie 1993; Diggle 1983; Fotheringham et al. 2000). This cohort of methods aims at determining whether an observed distribution of point events results from a random pattern or whether there is an underlying mechanism affecting these crashes. These methods include quadrant analysis (Shino 2008), nearest-neighbor analysis (Getis and Franklin 2010), Kernel Density Estimation (KDE) (Xie and Yan 2008), and K-function techniques (Yamada and Thill 2004). Each of these methods produces good results, but suffers from inherent deficiencies. Most of the traditional quadrat and nearest-neighbor analyses (e.g., Bailey and Gatrell 1995; Cressie 1993; Diggle 1983; Fotheringham et al. 2000) do not regard the fact that a traffic crash is an event that is constrained by the transportation network. Shino (Shino 2008) coped with this problem by driving the transportation network into mutually exclusive subnetworks (i.e., network-based quadrats). The complexity of finding a single quadrat is a function of the entire network's size. For n_e representing the number of links and n_p representing the number of nodes in the network, the complexity of finding a single quadrat is given by $O(n_e log n_p) + O(n_e + n_p)$ (Aho et al. 1983, Shino 2008). Furthermore, the problem of finding a set of subnetworks that covers the entire network is essentially the Set Cover Problem (SCP), which is known to be computationally intractable (i.e., NP-Hard) (Alon et al. 2006; Karp 1972; Shino 2008). Hence, the problem is computationally intense and consequentially is not suitable for applications where networks of more than a few dozens of miles are considered (Ang et al. 2012; Shino 2008). KDE methods drape a grid of equal-sized cells over the transportation network and perform the analysis in each cell. However, these methods disregard the density of the network links—hence, the number of road sections in each grid cell. This in return biases the results (Steenberghen et al. 2004). The K-Function method produces a global measure, which does not reveal the clusters' locations. Steenberghen et al. (Steenberghen et al. 2010) suggest a moving segment approach for solving this problem. The moving segment measures, at each point of the road network, the level of risk over all the road sections given a specific distance. To allow for the consideration of the network constraints, the distance is measured along the network. The main criticism of spatial methods is that they consider only location-related information. Hence they omit temporal and other data from the analysis.

Of the numerous factors that play a role in collisions, the temporal factors (i.e., time of day, day of the week, and year) are significant determinants. Many different methodological approaches for modeling occurrence of traffic crashes have been developed (Lord and Mannering 2010). Loosely speaking, methods for temporal analysis of traffic crashes occurrence can be categorized into two (Abdel-Aty and Pande 2007): methods which investigate crashes' frequency over long periods of time (Golob and Recker 2003; Hauer 1986; Miaou and Lum 1993; Persaud 1991); and techniques for real-time crash analysis, which determine the probability of crash occurrence in real time (Golob and Recker 2004; Lee et al.). Temporal analysis often addresses specific questions. The distribution of fatal crashes before and after the annual daylight savings time (DST) changes in spring and fall was examined to identify the increased risk to pedestrians in darkness (Sullivan and Flannagan 2001). In this study crashes were taken from a one-hour clock window three weeks before and three weeks after the transitions in both the spring and fall. The key assumption was that traffic conditions are the same in the weeks immediately before and after the changeover to DST, as traffic is principally governed by clock time rather than sunset time. Observed differences in crash levels between these two periods were attributed to the difference in ambient light level, and therefore can be used to quantify the effect of light in fatal crashes. While the research has shown a significant increase in risk to pedestrians, the method is limited to a short span of time. A simple visual exploratory approach to examine the relationship between fatal pedestrian crashes and time of day, day of the week, and time of the year showed that the first two hours of darkness typically presented the greatest frequency of pedestrian fatal collision (Griswold et al. 2011). The temporal correlation of fatal crashes with the drivers' age and alcohol consumption were also investigated within the scope of this study. Road traffic crash study in Christchurch, New Zealand, showed a comparative increase in rates during morning rush hour and during school run, the times students are traveling to and from school - 08.00–09.00 and 15.00–15.30 in this specific study (Kingham et al. 2011). It is important to note that this study failed to find spatial patterns in the data set. However, the search for patterns was held separately for the temporal and the spatial factors.

Other studies, aimed at finding an association between one or more crash attributes, employ various statistical regression analyses. The most common models are multiple linear regression, Poisson regression and Negative Binomial (NB) regression. NB and Poisson models were used for selecting the variables with the highest impact (e.g., Wong et al. 2007) and for predicting crashes (e.g., Lord and Mannering 2010). This work is conducted under a univariate modeling regime, where each variable is a scalar value. Recent work has extended these frameworks to multivariate models, in which each variable is a vector, typically describing different crash parameters (e.g., Ma and Kockelman 2006; Miaou and Song 2005). However, Both Poisson and NB models apply strong assumption regarding the crash data: the Poisson model requires the variance-to-mean ratio of the crash data to be about 1, and both Poisson and NB models require the crash data to be uncorrelated in time. As there is a correlation in time (e.g., Griswold et al. 2011), both models are limited.

Finite mixture/Markov switching models are also a common tool for examining heterogeneous populations (Frhwirth-Schnatter 2006). In these models, the underlying assumption is that the overall data are generated from sev-

eral distributions that are mixed together 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 to highway safety as they offer considerable potential for providing important new insights into the analysis of crash data. However, these models are quite complex to estimate (Lord and Mannering 2010).

Machine learning techniques were also used for analyzing contributing factors to crashes. Neural network models, such as Back Projection Neural Network (BPNN), have been recently used for modeling and predicting crash data (Abdel-Aty and Pande 2007; Mussone et al. 1999; Xie et al. 2007). These models often over-fit the data, especially when the sample size is small (Vogt and Bared 1998). Methods based on Bayesian Neural Networks (BNN) overcome the over-fitting problem (e.g., Xie et al. 2007) and were found to be more efficient than NB models for predicting crashes. Support Vector Machine (SVM) methods (Li et al. 2008; Zhang and Xie 2008) were suggested for the crash analysis task as well. SVM-based methods present less over-fitting and can be generalized in a simpler fashion than BPNN or BNN (Li et al. 2008). SVM, however, is computationally complex, and analyzing a large data set becomes impracticable.

This paper presents a two-step methodology, which avoids most of the deficiencies associated with the traditional data analysis methods. The method does not apply any assumptions on the data and does not require any tuning stage. A similar notion was recently presented by Depaire et al. (Depaire et al. 2008). In their study, the accidents were grouped by accident type, accident location type, driver age, road type, and vehicle type. While this study excluded analysis subjectivity and estimated a model stochastically, the set of parameters was preselected, which weakens the results of this set of features.

In the method presented here, at the first stage a graph-theory based method is employed to identify the set of road fatalities that are the least different (the typical fatality). By doing this we remove some of the noise originating from the most different fatalities. To obtain this we utilize clustering graph theoretic methods, known as graph-cuts (Ford and Fulkerson 1956) for segmenting crash data-sets. The segmentation is done using all available crash data. Incorporating all the information at hand allows one to find associations between the various factors, which may not have been identified otherwise. At the second stage, each group is analyzed by applying conventional statistical techniques. In this second stage both intra-group and inter-group analyses are conducted. Intra-group analysis allows for finding subtle trends and patterns as the groups are more homogeneous than the entire data set. Inter-group analysis facilitates the examination of the differences between the groups and effective identification of interactions.

METHOD

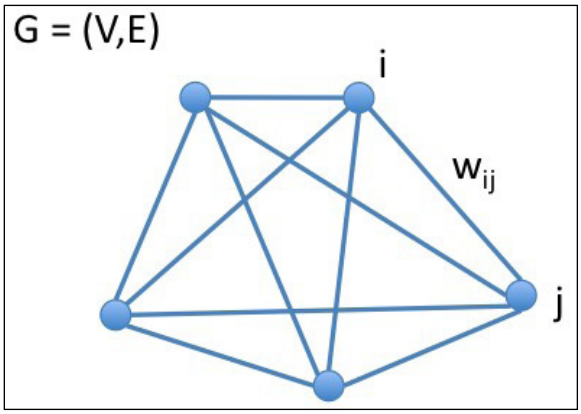
Graph Representation of Traffic Crash Data

A graph theoretical framework is suitable for segmentation and grouping problems of multi-dimensional data in general, and for multivariate crashes data in particular. The segmentation problem is presented on an undirected complete graph $G=(V,E)$, where V is the set of crashes and E are the set of edges connecting two crashes. Each edge carries a weight associated with the similarity of the two crashes it connects (See figure 1). The similarity weight w_{ij}, which is associated with the edge connect-

ing nodes i and j increases as the two crashes i and j are perceived to be more similar. Low values of w_{ij} are interpreted as dissimilarity.

FIGURE 1 Example of a Complete, Undirected Graph With Edge Weights

Each node corresponds to a crash, and edge weights reflect the similarity between two crashes.



The similarity between two vertices i and j in the graph is determined by a function that takes as input a feature or observation vector x_i (for vertex i) and another, x_j, for vertex j. In our application, x_i is the i^{th} crash's feature vector (e.g., date, time, number of involved vehicles, number of fatalities, manner of collision, driver's alcohol consumption, number of lanes, etc.). The function outputs a single real-valued number, where larger numbers indicate a higher degree of similarity between i and j. Depending on the properties of the feature vectors, a variety of similarity functions can be used. The most commonly used similarity metrics are measures of correlation, such as Pearson's correlation or Kendall's τ rank correlation (Kendall 1938). Euclidean distance or l^2-norm is a common measure of *dissimilarity*, where larger distances correspond to greater degrees of dissimilarity. For a, b, and c user-defined parameters, the Gaussian similarity function transforms the Euclidean distance measure to a similarity function as follows:

$$S_{ij} = S(x_i, x_j) = aexp\{-b\|x_i - x_j\|^c\} \quad (1)$$

We opted to use this similarity function due to its intuitive behavior and flexibility, however, any alternative monotonically increasing function in $\|x_i - x_j\|$ may be used in its place.

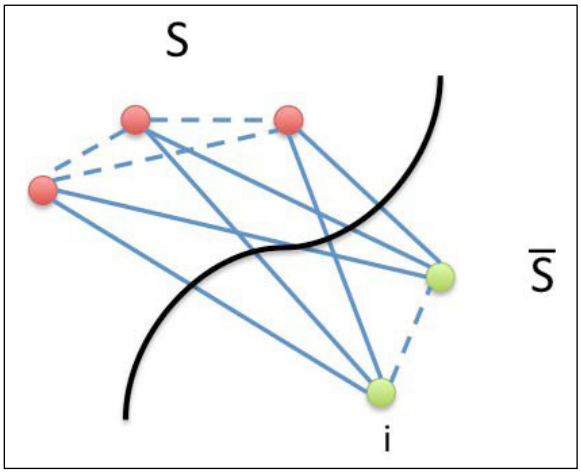
Cuts on a Graph

We now provide a formal definition for cuts or partitions on a graph, which is a major component of our proposed method.

A bipartition of a graph is called a *cut*, $(S, \bar{S}) = \{\{[i,j] | i \in S, j \in \bar{S}\}\}$, where $\bar{S}$ is the complement of S, $(\bar{S} = V \backslash S)$. This is illustrated in figure 2. The *capacity of a cut* $(S, \bar{S})$ is defined as the sum of the weights of all edges between S and $\bar{S}$, thus edges that have one endpoint in S and the other in $\bar{S}$. In the example given in figure 2, these edges are represented in solid lines. Formally the capacity of the cut is given by:

FIGURE 2 Example of a Cut on an Undirected, Complete Graph

The cut is indicated by the dark black line that partitions the node set V into two disjoint sets: S and $\overline{S}$. The capacity $C(S,\overline{S})$ of the cut is the sum of the weights of the edges that cross the cut (i.e., the sum of the weights of all edges that have one endpoint in S and one in $\overline{S}$.



$$C(S,\bar{S}) = \sum_{i\in S, j\in \bar{S}} w_{ij} \quad (2)$$

More generally, for any pair of sets $A, B \subseteq V$, we define the set of edges going between these two sets as $(A,B) = \{[i,j] | i \in A, j \in B\}$, and the capacity of (A,B) is

$$C(A,B) = \sum_{i\in A, j\in B} w_{ij} \quad (3)$$

Consequentially, the capacity of a set, $A \subset V$ is given by:

$$C(A,A) = \sum_{i,j\in A} w_{ij} \quad (4)$$

and is denoted by $C(A)$.

Finally, the total sum of the weights of edges from nodes in $S \subset V$ to all nodes in the graph, V, is denoted by *d(S)* and is referred to as the volume of the set:

$$d(S) = C(S,V) = \sum_{i\in S, j\in V} w_{ij} \quad (5)$$

Using the notation above, one can easily see that:

$$d(S) = C(S,V) = C(S,\bar{S}) + C(S) \quad (6)$$

and

$$C(S) = C(V) - d(\bar{S}) \quad (7)$$

where *C(V)* is the capacity of the graph—hence, the sum of all edge weights in *G(V,E)*, which for a given graph, *G(V,E)*, is constant.

The segmentation problem can be formulated in many ways. One common formulation is to find a partition into two sets which minimizes the similarity between crashes that are associated with different groups. This problem is the minimum-cut problem (Ford and Fulkerson 1956), which oftentimes results in an unbalanced partition (Shi and Malik 2000). Shi and Malik tried to overcome this unbalance problem, by maximizing the dissimilarity between groups and the similarity within a group (Shi and Malik 2000). This is achieved by finding a partition to two nonempty disjoint sets minimizing the quantity called the Normalized Cut (NC):

$$NC(S,\bar{S}) = \frac{C(S,\bar{S})}{d(S)} + \frac{C(S,\bar{S})}{d(\bar{S})} \quad (8)$$

This objective function drives the segment *S* and its complement to be approximately of equal size. However, there is no efficient way for finding the partition *S* that minimizes equation (8). Hence, this objective is NP-hard (by reduction from set partitioning (Shi and Malik 2000). In this paper we use a slightly different quantity, *NC"*:

$$NC''(S,\bar{S}) = \frac{C(S,\bar{S})}{C(S)} + \frac{C(S,\bar{S})}{C(\bar{S})} \quad (9)$$

The set $\subset V$, which minimizes equation (8) also minimizes equation (9). As shown below:

$$\frac{C(S,\bar{S})}{C(S)} + \frac{C(S,\bar{S})}{C(\bar{S})} = \frac{C(S,\bar{S})}{C(V) - d(\bar{S})} + \frac{C(S,\bar{S})}{C(V) - d(S)} \quad (10)$$

As *C(V)* is constant, for a given *G(V,E)*; and as $C(V) > d(S)$ for all nonempty $S \subset V$, equation (9) is equivalent to equation (8). This means that the set *S*, which minimizes equation (8) also minimizes equation (9) and finding it is also NP-hard.

The *NC"* problem finds a bi-partition of the graph. The extension of this objective to partition the graph into *K* segments is given by:

$$NC''_K = \min_{S_1,S_2,S_3,\cdots,S_K} \left(\sum_{i=1}^{K} \frac{C(S_i,\bar{S}_i)}{C(S_i)} \right) \quad (11)$$

For solving the multigroup segmentation, we extend a stochastic approximation scheme, originally suggested by Karger and Stein (Karger and Stein 1996) for solving the minimum-cut problem, for solving the multi-segmentation $NC"_K$ problem.

Stochastic Approximation for the Normalized Cut Problem

This algorithm is iterative. At each step one edge e_{ij}, connecting two vertices, i and j, is randomly selected. The selected edge, e_{ij}, is removed and the vertices i and j are merged into one *meta-node* k. The edges going from any node v in the remaining of the graph to either i or j are replaced with edges connecting v and the new meta-node k. This edge's weight is then given by:

$$w_{k,v} = w_{j,v} + w_{i,v} \tag{12}$$

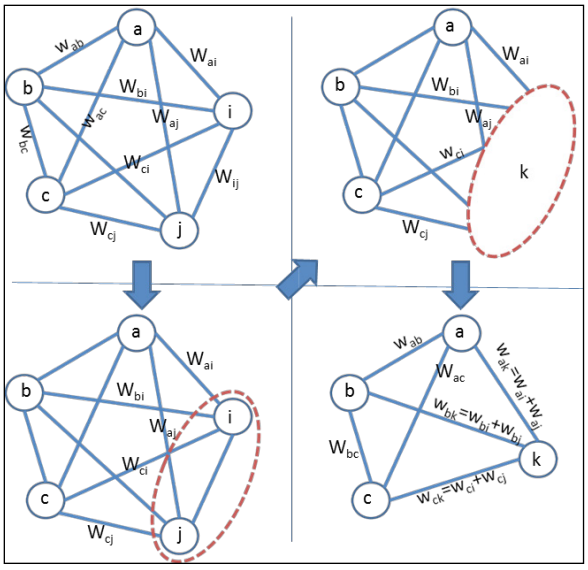
The rest of the graph remains unchanged. An example of an edge contraction is given in figure 3. The algorithm ends when there are K nodes (or meta-nodes) left in the graph. Because the algorithm has a stochastic component in it, the entire process is repeated several times, and the graph partitioning that gives the minimum value of equation (11) is chosen.

The merging algorithm is based on the idea that because the number of edges that reside on the cut is small, a randomly chosen edge is unlikely to be part of the cut. In a try to increase the chances that the optimal minimum of (11) is reached, the probability, $P(e_{ij})$, of choosing an edge e_{ij} is inversely proportional to the maximum value of the NC" objective of the two nodes it connects, hence:

$$P(e_{ij}) \propto \frac{1}{max\{\frac{C(S_i,\overline{S_i})}{C(S_i,S_i)}, \frac{C(S_j,\overline{S_j})}{C(S_j,S_j)}\}} \tag{13}$$

FIGURE 3 Merging Step

Merging nodes *i* and *j* into node *k*



DATA

Crash Data

To illustrate the suggested method, we examine 1,573 fatal collisions in California during 2008. The crash data was extracted from the Fatality Analysis Reporting System (FARS), a comprehensive surveillance system of U.S. fatal collisions maintained by the National Highway Traffic Safety Administration (NHTSA). The 1,573 fatalities were chosen arbitrarily by ordering them according to the recorded time of the crash and taking every other crash.

For each fatality, the FARS system holds more than 125 data elements. These elements are reported on four standard forms (Accident, Vehicle, Driver and Person) that include detailed information about persons and vehicles involved in the crash, and the physical environment in which the crash occurred (NHTSA 2004). While the algorithm is not restricted by the number of features incorporated into the crashes' feature vectors, we focus here on all 28 features included in the Accident file (see table 8). This subset consists of the features that describe the crash and its possible contributing factors; time related factors; and road infrastructure descriptors. The objective here is to identify causal attributes of fatal crashes; post-crash attributes were omitted from the analysis. Additional administrative attributes such as case id and milepost were also removed.

Similarity Measures

For computing the distance of two feature vectors, as a measure for similarity (inverse proportion—see Section 2.1), we consider the nature (or type) of the different variables. For all types of variables the distance is normalized so the maximum possible distance is 1. The distance between any two crashes is computed by summing both crashes' features' distances.

The distance of time-related variables (month, day of the month, hour and minute) between two crashes is computed in a cyclic fashion. Hence, the distance between a crash that occurs on the 31st to a crash that took place on the 1st, is the same as the distance between the latter and a crash on the 2nd. Similarly, the distance between a crash that is reported at 11 pm and a crash at midnight is the same as the distance between midnight and 1am.

Out of the 28 features considered in this study, three categorical variables were grouped due to their apparent association. For these variables the possible reported values are first grouped into a small number of groups. If two crashes' values fall into the same group, then the distance is 0, otherwise it is 1. For example for manner of collision, we determine 5 categories: collision with a fixed object, collision with other road user, car malfunction, driver's fault and non-collision. For hit and run we designate two groups: hit and run has occurred (hit and run after collision with motor vehicle in transport, after collision with pedestrian, after collision with parked/stopped car, etc.) and no hit and run. Days of the week are grouped into four groups: weekdays (Tuesday through Thursday), Monday, Friday and weekend (Saturday and Sunday). The reason for this grouping is that Monday, Friday and weekends exhibit different crash patterns (e.g., Griswold et al. 2011).

For all other variables, if they are the same then the distance is 0, if they are different then the distance is 1. By doing this we avoid hard questions such as whether the distance between 1-lane road and 2-lanes road is the same as the distance between 2-lanes to 3-lanes roads. Other similar questions are the distance between different roadway surface conditions and number of fatalities.

As described in Section 2.1 the weights on the graph represent the similarity between two crashes rather than the distance. To shift from distance to similarity, we use equation (1), where a = 1, b = 0.1 and c = 1. These values were selected after a binary search and they provide the lowest value for NC''_K,equation (11).

RESULTS

Each crash in the data set is characterized by 28 features grouped into three types:

"what" variables: include manner of collision, harmful event, number of fatalities, number of people involved, drunk drivers, number of pedestrians, number of involved vehicles in transport, total number of vehicles and whether it was a hit and run.

"when" variables: include date, time, and light conditions.

"road" variables: speed limit, number of lanes, roadway alignment and profile, traffic flow, relation to junction and roadway, surface condition, traffic control device, and construction zone indication.

We present here four different analyses, all using the proposed method. The difference between the analyses is the features that are included in the feature vectors. We run the analysis when the feature vectors consist only of the "what", "when" and "road" variables, as well as a model incorporating all variables.

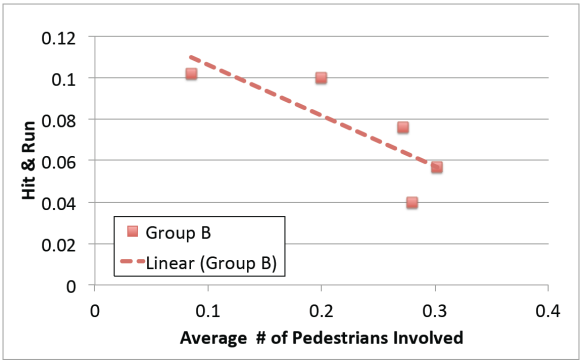
For executing the clustering one has to compute the similarity matrix and then find the minimum cut, which in the scheme presented here, is an iterative process. The runtime is on average 25 μ-seconds per sample per feature for creating the similarity matrix, and 28 seconds per iteration for computing the cut, regardless of the feature vector size. Thus, for 1,573 samples with feature vectors of length 28 (all features) it takes about 880 seconds for computing the similarity matrix and about 3 hours for 300 iterations, which were found to be sufficient for the clustering algorithm here.

Analysis of the "What" Variables

The segmentation of the 1,573 fatalities, considered in this study, into 12 clusters is presented in table 1. For each cluster, the table presents its size and the mean and mode (most frequent) values for each of the variables. The latter is important for descriptive features (i.e., manner of collision and harmful event) for which averaging makes little sense.

The presentation of the crash data in this fashion reveals an interesting phenomenon. For the analysis, let us consider clusters 1 and 2 of table 1 as one group, A, and clusters 3,4,5 and 6 as group B. This grouping procedure allows to pinpoint what are the features that separate the clusters of group B from the ones in A. Examining group B's characteristics shows an inverse proportional relation between the number of crashes involving pedestrians and the hit and run frequency. This is illustrated in figure 4, where each of the four clusters of group B are projected on a two-dimensional space—with the percentage in the cluster of Hit & Run accidents as one dimension and the percentage of accidents with pedestrians involved as the second. Hence, we have characterized crashes in group A by observing what makes this group different with respect to group B. Group A consists of 1,179 crashes out of 1,573 accidents that were analyzed. If we take the characteristics of the crashes in group A to represent the typical (or common) crash, we have found a behavioral relation between frequency of pedestrian involvement in crashes and the hit and run percentage.

FIGURE 4 Trend Analysis of "What" Variables of Group B, Which Consists of the 3rd , 4th , 5th, and 6th Largest Clusters



It would be reasonable to assume that the impact in a fatal accident when two cars are involved is much larger than a fatal crash when a vehicle and a pedestrian are involved, as in the former the passengers have the car's frame and its safety means to protect them. Therefore it is more probable that a car can flee the scene after being involved in a fatal accident with a pedestrian than after a fatal crash with another car. While intuitive, this association has not been investigated. Studies that tried to characterized hit and run crashes (e.g., Rifaat and Chin 2007; Tay et al. 2009; and Tay et al. 2008) found many other factors, such as roadway functional class, routes, traffic flow, types of roadway section, speed limit, traffic control device, functioning of traffic control device, lighting condition, roadway alignment and roadway profile, as important determinants. In these studies all accidents data were

TABLE 1 "What" Variable Segmentation

For each cluster, the mean and the most common value (mode) of each characteristic are presented: The cluster size (size), Manner of Collision (MAN COLL), Harmful Event (HARM EV), number of fatalities (Fatals), number of persons involved (PERSONS), number of drunk drivers (DRUNK), number of pedestrians involved (PEDS), number of vehicles in transport involved (VE FORMS), number of total vehicles involved (VE TOTAL) and Hit and Run (HIT RUN)

		Man. coll.		Harm ev.		Fatals		Persons		Drunk		Peds.		Ve. forms		Ve. total		Hit run	
	Size	mean	mode	mean	mode	mean	mode	mean	mode	mean	mode	mean	mode	mean	mode	mean	mode	mean	mode
1	741	1.644	0	16.717	12	1.089	1	2.825	2	0.418	0	0.238	0	1.571	1	1.642	1	0.085	0
2	438	1.189	0	17.217	12	1.091	1	2.614	2	0.436	0	0.237	0	1.505	1	1.543	1	0.080	0
3	192	0.672	0	17.557	8	1.042	1	2.250	1	0.427	0	0.302	0	1.302	1	1.365	1	0.057	0
4	92	1.217	0	14.239	12	1.120	1	2.761	2	0.380	0	0.272	0	1.489	1	1.543	1	0.076	0
5	59	2.966	0	13.814	12	1.051	1	3.458	2	0.322	0	0.085	0	1.966	2	1.966	2	0.102	0
6	25	0.520	0	19.720	12	1.440	1	3.280	2	0.400	0	0.280	0	1.760	1	1.920	1	0.040	0
7	10	0.000	0	32.000	42	1.000	1	1.400	1	0.800	1	0.200	0	1.000	1	1.000	1	0.100	0
8	8	2.875	1	12.000	12	1.125	1	3.375	3	0.375	0	0.000	0	2.375	2	2.375	2	0.125	0
9	4	1.000	0	17.000	12	1.000	1	3.000	1	0.500	0	0.250	0	1.750	1	1.750	1	0.250	0
10	2	4.500	2	12.000	12	1.000	1	3.000	3	0.000	0	0.000	0	3.000	3	3.000	3	0.500	0
11	1	1.000	1	12.000	12	2.000	2	4.000	4	1.000	1	1.000	3	2.000	2	2.000	2	0.000	0
12	1	7.000	7	14.000	14	1.000	1	2.000	2	0.000	0	1.000	1	1.000	1	2.000	2	0.000	0

considered in the analysis. As a result crashes in group B weaken the connection found here. When applying the segmentation and effectively removing group B crashes, this connection manifests itself.

The results of an analysis to determine whether the link between pedestrians involvement and hit and run frequency holds outside the scope of the 1,573 fatalities investigated in our initial study are presented in figure 5. The analysis was carried out over the entire 2008 FARS data set, which holds 34,017 fatalities. Figure 5 depicts a plot of the frequency of hit and run versus the frequency of pedestrians crashes for each of the 53 states and U.S. territories. It is evident from figure 5 that this association between pedestrians and hit and run does hold.

For quantitative analysis of the results we apply analysis of variance (ANoVA) (Hogg and Ledolter 1987) on the segmentation results of the entire data-set—once for the clustering results and once for an arbitrary random clustering. For each "what" variable, ϑ_i , we compare the F-statistics and its corresponding p-value for the null-hypotheses between the clustering results and a random clustering assignment (with the same clusters size). The smaller the p-value for a given ϑ_i, the less homogeneous the different clusters with respect to this variable. In the comparison here, a feature with higher p-value in the *NC"* clustering (than the random clustering) is salient in the clustering process. Hence, the clusters are characterized by different values of ϑ_i and are more homogeneous with respect to this feature. It is important to note that the features, which were found salient in the clustering process, came out over a large number of runs (thousands). The results of the ANoVA are presented in table 2. Hit-and-run and the number of fatalities are found to be the most salient features in the segmentation. This reinforces the findings above. An additional parameter that was found to be salient is the number of vehicles involved, which also can differentiate between fatal accident with or without pedestrians' involvement.

Analysis of the "When" Variables

FIGURE 5 Frequency of Hit & Runs as a Function of the Frequency of Pedestrian Crashes

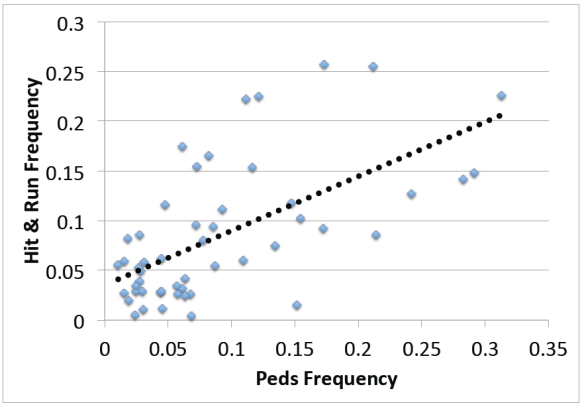


TABLE 2 Analysis of Variance (ANOVA) Results for "What" Variables Segmentation

	NC′		Random	
	F-stats	P-V	F-stats	P-V
Man. coll.	1.02	0.42	1.1	0.36
Harm ev.	1.03	0.42	0.74	0.87
Fatals	0.44	0.78	1.22	0.30
Persons	0.88	0.6	0.53	0.93
Drunk	2.06	0.08	1.63	0.16
Peds.	0.92	0.47	0.78	0.56
Ve. forms	1.10	0.36	1.93	0.08
Ve. total	0.46	0.8	1.97	0.07
Hit run	0.14	0.98	1.23	0.29

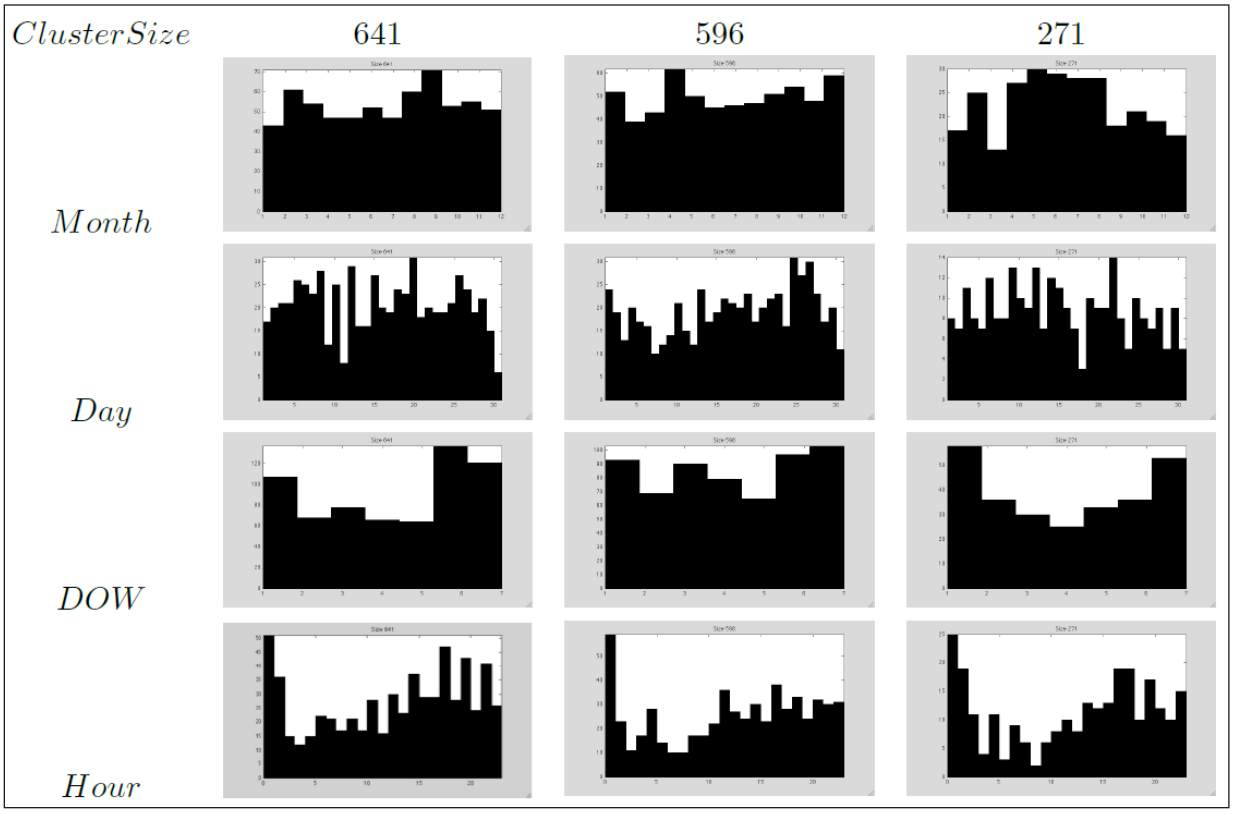
Table 3 details the segmentation results of the 1,573 fatalities considered in this study by their "when" attributes. In order to allow better understanding of the results, figure 6 depicts the histograms of the month, day of the month, day of the week and the reported hour of the

TABLE 3 "When" Variable Segmentation

For each cluster, the mean and the most common value (mode) of each characteristic are presented: The cluster size (size), YEAR, MONTH, DAY of the Month, Day Of the Week (DOW), HOUR, MINUTE, light condition (LGT) and Weather (WEATHER)

	Year		Month		Day		Dow		Hour		Minutes		LGT	Weather
Size	mean	mode	mean	mode	mean	mode	mean	mode	mean	mode	mean	mode	mode	mode
641	2008	2008	6.64	9	15.64	20	4.26	6	14.49	18	30.83	45	1	1
596	2008	2008	6.64	4	16.84	25	4.10	7	14.35	1	29.31	30	1	1
271	2008	2008	6.39	5	15.38	22	3.96	1	15.29	2	31.17	15	1	1
27	2008	2008	7.93	8	17.96	18	4.70	6	17.96	17	28.30	10	1	1
21	2008	2008	7.29	1	13.52	7	4.05	1	12.90	3	25.86	20	1	1
10	2008	2008	7.90	11	17.20	22	4.70	7	14.70	23	12.20	2	2	1
4	2008	2008	8.75	7	9.50	8	4.00	2	14.75	6	43.50	20	3	1
3	2008	2008	9.00	8	10.33	2	1.67	1	10.33	6	18.67	2	1	1

FIGURE 6 Histograms of the When Clusters for Month, Day of the Month (Day), Day of the Week (DOW), and Hour of Crash



crash for the three largest clusters. The evaluation of the histograms suggests that the first cluster (size of 641 fatalities) is characterized by weekend crashes, mainly at night time. This is derived by the DOW and hour histograms. The second largest cluster (size 596) consists of week-day crashes. We conclude this as the third cluster (271 crashes) consists of mainly spring-summer weekend vacation crashes. The third cluster's month histogram shows a significant peak between April and August; its Day histogram has strong periodical peaks that correspond to weekends; and finally its DOW histogram clearly shows that the distribution of the day of the crash, for this cluster, is biased towards the weekend days. Another characteristic that suggest that this cluster is a vocation weekend cluster is that its hour histogram does not have peaks that correspond to the morning or afternoon rush. The hour histogram of the first (largest) two clusters does present these peaks. Another interesting phenomenon that the clusters reveal is that the second week of the month of the first two clusters has fewer numbers of crashes.

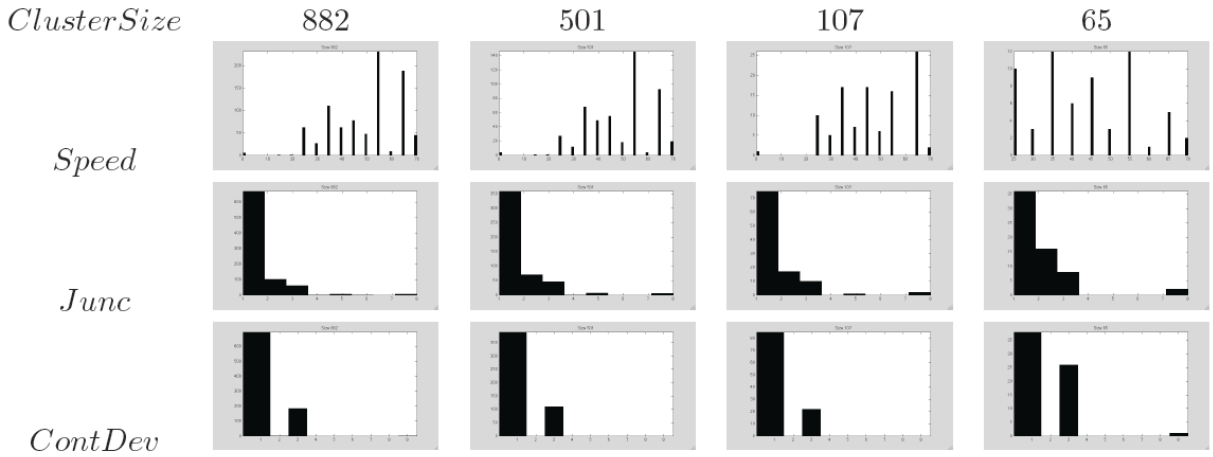
TABLE 4 "Road" Variable Segmentation

For each cluster, the mean and the most common value (mode) of each characteristic are presented: Cluster size (size), Speed Limit (SPEED), number of lanes (LANES), Roadway Alignment (ALIGN), Roadway Profile (Pro- file), Trafficway Flow (FLOW), Relation to Junction (REL JUNC), Relation to roadway (REL ROAD), Surface Type (PAVE TYP), Surface condition (SUR COND), Traffic control Device (TRA CONT), Traffic Control Device Functioning (TCF) and Construction / Maintenance Zone (Zone)

	Speed		Lanes		Align		Profile		Flow		Rel. junc.	
Size	mean	mode	mean	mode	mean	mode	mean	mode	mean	mode	mean	mode
882	50.558	55	2.822	2	1.339	1	1.615	1	2.141	1	1.724	1
501	49.553	55	2.784	2	1.299	1	1.715	1	2.076	1	1.904	1
107	47.196	65	2.804	2	1.299	1	1.579	1	2.187	1	1.738	1
65	45.200	35	2.662	2	1.154	1	1.338	1	2.062	1	2.231	1
10	46.000	45	2.600	2	1.100	1	1.200	1	2.200	2	1.600	1
5	54.000	55	2.200	2	1.600	2	1.400	1	1.200	1	1.000	1
2	47.500	40	2.000	2	1.500	1	1.000	1	1.500	1	2.000	1
1	25.000	25	2.000	2	1.000	1	2.000	2	1.000	1	1.000	1

	Rel. road		Pave type		Sur. cond.		Tra. cont.		TCF		Zone	
Size	mean	mode	mean	mode	mean	mode	mean	mode	mean	mode	mean	mode
882	2.398	1	1.906	2	1.129	1	3.582	0	0.638	0	0.023	0
501	2.399	1	1.982	2	1.044	1	3.307	0	0.669	0	0.030	0
107	2.084	1	1.841	2	1.159	1	2.037	0	0.626	0	0.084	0
65	1.862	1	1.908	2	1.092	1	5.569	0	1.338	0	0.015	0
10	1.600	1	1.900	2	1.100	1	3.700	0	1.200	0	0.100	0
5	2.000	1	2.000	2	1.200	1	16.000	0	1.200	0	0.000	0
2	1.000	1	2.000	2	1.000	1	1.500	0	1.500	0	0.000	0
1	1.000	1	2.000	2	1.000	1	0.000	0	0.000	0	0.000	0

FIGURE 7 Histograms of the Road Clusters for Speed Limit (*Speed*), Relation to Junction (*Junc*) and Control Device (*ContDev*)



Having this information in hand, one can further evaluate the crashes within each cluster in different aspects. Such aspects may be the severity of the crash, the drivers' characteristics, and the type of cars involved. As the crashes within a cluster are more homogeneous than the entire data set, this analyses are expected to provide new insights.

Analysis of the "Road" Variables

The clustering results of the crashes by the "road" variables are given in table 4 and figure 7. The three largest clusters can be characterized as highway crashes. This notion is based on the speed histograms which, for the first three clusters, are shifted towards the higher speed limits.

The notion is also supported by the histograms of the relation to junction and control device, where the fourth cluster presents larger numbers than the first three in noninterchange-junctions or at traffic signals. Having the data clustered prior to any analysis, as illustrated here, will allow for finding more subtle trends as the crash groups are homogeneous and the noise is filtered out. Table 5 presents the ANoVA results for the road variables. The same variables that were found to play a significant role in the segmentation process in the analysis above are shown to be salient in the ANoVA results as well. The dominant features are the number of lanes, road profile, speed limit and relation to roadway. These findings support the observation that this clustering distinguishes between highway and non-highway accidents

TABLE 5 Analysis of Variance (ANOVA) Results for "Road" Variables Segmentation

	NC'		Random	
	F-stats	P-V	F-stats	P-V
Speed	0.40	0.96	0.64	0.83
Lanes	0.81	0.58	1.86	0.07
Align	1.11	0.33	1.19	0.31
Profile	0.09	0.96	1.16	0.33
Flow	3.33	0.00	0.29	0.94
Rel. junc.	1.43	0.15	0.85	0.58
Rel. road	1.02	0.42	1.65	0.01
Pave typ.	1.03	0.39	0.43	0.79
Sur. cond.	0.87	0.51	0.34	0.91
Tra. cont.	3.38	0.02	3.08	0.03
TCF	1.84	0.04	1.79	0.04
Zone	0.56	0.64	0.60	0.61

TABLE 6 Typical Road Fatality

Variable		Entire data	Segment 1	Segment 2	Segment 3	Segment 4
Size		1573	422	413	334	176
Month	mean	6.65	6.79	6.30	6.75	6.98
	mode	9	11	2	7	9
	std	3.42	3.52	3.46	3.38	3.26
Day	mean	16.04	16.32	16.19	16.12	14.74
	mode	27	15	10	23	7
	std	8.74	8.68	9.12	8.40	8.54
Hour	mean	14.66	15.00	14.51	14.74	15.20
	mode	18	18	22	17	20
	std	14.71	14.15	13.90	15.78	15.96
Minute	mean	30.09	30.05	29.15	31.22	31.08
	mode	30	20	30	45	35
	std	20.12	19.83	19.43	20.75	20.04
Ve. total	mean	1.59	1.56	1.58	1.70	1.42
	mode	1	1	1	1	1
	std	0.89	0.82	0.85	1.03	0.69
Persons	mean	2.72	2.63	2.70	3.07	2.38
	mode	2	2	2	2	2
	std	2.20	1.63	2.07	3.18	1.46
Peds.	mean	0.24	0.27	0.22	0.25	0.23
	mode	0	0	0	0	0
	std	0.53	0.54	0.50	0.63	0.46
Harm ev.	mean	16.82	16.20	17.35	15.95	17.70
	mode	12	12	12	12	12
	std	12.40	11.47	12.69	12.00	13.18
Man. coll.	mean	1.41	1.46	1.20	1.66	1.02
	mode	0	0	0	0	0
	std	3.31	2.28	2.08	5.78	2.05
Rel. junc.	mean	1.80	1.71	1.64	1.87	1.79
	mode	1	1	1	1	1
	std	3.22	1.88	1.99	2.37	2.20
Rel. road	mean	2.35	2.24	2.68	1.99	2.14
	mode	1	1	1	1	1
	std	5.14	4.96	6.94	1.53	1.52
Flow	mean	2.12	2.09	2.18	2.12	2.11
	mode	1	1	1	1	1
	std	1.55	1.67	1.39	1.57	1.60
Lanes	mean	2.80	2.84	2.78	2.85	2.80
	mode	2	2	2	2	2
	std	1.35	1.43	1.28	1.37	1.45
Speed	mean	49.75	48.49	50.54	49.75	49.68
	mode	55	55	55	55	55
	std	14.68	14.52	14.62	15.43	14.78

continued on next page

TABLE 6 Typical Road Fatality *(continued)*

Variable		Entire data	Segment 1	Segment 2	Segment 3	Segment 4
Align	mean	1.32	1.33	1.27	1.31	1.28
	mode	1	1	1	1	1
	std	0.94	0.94	0.78	0.94	0.92
Profile	mean	1.63	1.54	1.77	1.56	1.65
	mode	1	1	1	1	1
	std	1.85	1.67	2.07	1.77	1.94
Pave typ.	mean	1.93	1.97	1.92	1.88	1.99
	mode	2	2	2	2	2
	std	0.87	0.89	1.00	0.59	1.12
Sur. cond.	mean	1.10	1.12	1.09	1.07	1.19
	mode	1	1	1	1	1
	std	0.47	0.50	0.33	0.30	0.94
Tra. cont.	mean	3.51	3.25	2.75	4.13	3.60
	mode	0	0	0	0	0
	std	9.39	8.30	8.23	10.61	9.09
TCF	mean	0.68	0.75	0.55	0.79	0.65
	mode	0	0	0	0	0
	std	1.28	1.35	1.16	1.32	1.24
Hit run	mean	0.08	0.09	0.07	0.07	0.10
	mode	0	0	0	0	0
	std	0.58	0.61	0.56	0.48	0.70
LGT	mean	1.92	1.88	1.96	1.94	1.95
	mode	1	1	1	1	1
	std	1.01	1.02	0.96	1.00	1.20
Zone	mean	0.03	0.03	0.04	0.01	0.04
	mode	0	0	0	0	0
	std	0.25	0.24	0.32	0.08	0.20
Fatals	mean	1.09	1.05	1.10	1.12	1.10
	mode	1	1	1	1	1
	std	0.41	0.23	0.33	0.58	0.54
DOW	mean	4.15	4.27	4.08	4.21	3.88
	mode	7	6	6	7	1
	std	2.14	2.10	2.16	2.17	2.17
Drunk	mean	0.42	0.43	0.45	0.43	0.37
	mode	0	0	0	0	0
	std	0.64	0.65	0.64	0.61	0.65
Ve. forms	mean	1.54	1.52	1.52	1.62	1.37
	mode	1	1	1	1	1
	std	0.84	0.80	0.82	0.91	0.59
Weather	mean	1.11	1.08	1.11	1.11	1.15
	mode	1	1	1	1	1
	std	0.57	0.36	0.59	0.59	0.82

TABLE 7 Analysis of Variance (ANOVA) results for all variables

	NC		Random	
	F-stats	P-V	F-stats	P-V
Month	0.82	0.62	1.07	0.38
Day	0.86	0.69	0.89	0.64
Hour	1.37	0.11	1.01	0.44
Minute	0.94	0.6	0.99	0.49
Ve. total	0.47	0.8	0.81	0.54
Persons	1.05	0.4	1.2	0.26
Peds	1.39	0.22	1.64	0.15
Harm ev.	0.9	0.64	1.03	0.42
Man. coll.	1.14	0.33	1.29	0.24
Rel. junc.	1.14	0.33	1.83	0.05
Rel. road	0.95	0.48	1.61	0.11
Flow	0.4	0.88	1.1	0.36
Lanes	1.09	0.37	0.72	0.65
Speed	1.95	0.02	0.71	0.75
Align	2.73	0.07	0.88	0.41
Profile	0.47	0.7	0.57	0.63
Pave typ.	0.7	0.59	1.45	0.22
Sur. cond	0.52	0.8	1.11	0.35
Tra. cont.	1.66	0.07	0.46	0.94
TCF	4.23	0.01	0.65	0.58
Hit run	0.99	0.42	1.25	0.29
LGT	0.67	0.65	0.59	0.71
Zone	0.12	0.95	0.32	0.81
Fatals	1.41	0.23	0.66	0.62
DOW	0.99	0.43	1.02	0.41
Drunk	0.78	0.55	0.61	0.66
Ve. forms	0.66	0.65	0.64	0.67
Weather	0.6	0.7	1.65	0.15

Analysis of All Variables

The segmentation procedure creates homogeneous clusters of crashes. Inter- and intra-analysis of these clusters allows for observing subtle trends and small changes. In previous sections the analysis was carried out when different groups of features out of the 28 crash characteristics were considered. In this section we provide the analysis when all 28 parameters are incorporated into the process. Table 6 details the mean, mode and standard deviation (std) of each of the 28 parameter for the entire data set as well as for each of the 4 largest clusters. Evaluating the standard deviations shows that there are several features that exhibit significantly lower values for one or more clusters than the standard deviation of the entire data set. Examples are number of persons for the 4th segment, manner of collision in segments 1, 2 and 4, relation to the junction in all segments, road profile (segments 1 and 3), number of fatalities (segments 1 and 2), number of vehicles involved (segment 4) and weather (Segment 1). Following the statistical law of large numbers, had the clusters were assembled randomly the standard deviation would have been

higher for the smaller subsets. Thus a lower standard deviation value for a specific feature of a certain cluster (with respect to the entire data set) signifies that this feature is more homogenous in the cluster. The same findings are obtained through the analysis of variance (ANoVA), which is presented in table 7. Hence the same set of features presents higher P-values for the *NC"* clustering comparing to the random-clustering values. These are the features that characterized the cluster and form its common fatality. This information can be then used for further study for extracting new connections between features or to single out a cluster which is more suitable for investigation of a certain phenomenon. Using these lines of research for analyzing crash data is expected to result in new insights and observations.

CONCLUSIONS

Road fatality analysis pose a great challenge as traffic crashes are rare outcomes of events that are confined to a small time-space region. This problem increases when a subset of crashes is considered (e.g., pedestrians or fatal crashes). In this paper we propose to apply a cluster analysis of the crashes prior to any other analysis. Then, each cluster is separately analyzed both internally and with respect to the other clusters. Those inter- intra-relations of the cluster point out trends phenomena and characteristics that are less prominent in the entire data set, and therefore are often overlooked. This procedure is known as mixed models and several such analyses were presented in traffic accident analysis (e.g., Depaire et al. 2008; Lord and Mannering 2010).

For the cluster analysis, we employ a graph-theory-based segmentation algorithm. The algorithm is based on the well-known normalized-cut optimization criterion. The solution is sought through a stochastic approximation scheme. This scheme is a novel extension of a method for solving the minimum-cut problem.

The cluster analysis often results in finding subtle trends and significant causes for traffic fatalities. For example, the method has found a correlation between hit-and-run and pedestrians fatalities, which was not identified by previous studies. An additional output of the research is a description of the typical fatality, which is a result of the segmentation analysis done when all factors that characterized a crash are considered.

Future research may expand the analysis presented here so other features that are recorded in the FARS data set are considered and naturally for analyzing other data sets. An emerging research field in the traffic safety arena is naturalistic driving studies (Guo and Fang 2012; Klauer et al. 2006). These studies are based on collecting continuous streams of different data. The data collected sums up to a sheer amount, which calls for exploratory tools such as the scheme presented here. Due to the efficiency of the presented mechanism, it likely to highlight new insights in the examined big-data and may lead to new findings in highway safety.

TABLE 8 Accident File Data Fields (C indicates a categorical variable)

Short name	Description
Man. coll. (C)	Manner Of Collision
Harm ev. (C)	First Harmful Event
Fatals	Number of Fatalities In Crash
Persons	Number of Person Forms Submitted
Drunk	Number of Drunk Drivers in Crash
Peds.	Number of Non-Motorist Forms Submitted
Ve. forms	Number of Vehicle Forms Submitted, Motor Vehicles in Transport
Ve. total	Number of Vehicle Forms Submitted, Total-Includes Motor Vehicles Not in Transport
Hit run	Hit-And Run
Year	Crash Year
Month	Crash Month
Day	Crash Day of the Month
DOW	Day of Week
Hour	Crash Hour
Minute	Crash Minute
LGT (C)	Light Condition
Weather(C)	Atmospheric Conditions
Speed	Speed Limit
Lanes	Number of Travel Lanes Align (C)
Align (C)	Roadway Alignment
Profile (C)	Roadway Profile
Flow	Trafficway Flow
Rel. junc. (C)	Relation To Junction
Rel. road (C)	Relation To Roadway PAVE TYP (C)
TYP (C)	Roadway Surface Type
Sur. cond. (C)	Roadway Surface Condition
TCF (C)	Traffic Control Device Functioning
Tra. cont. (C)	Traffic Control Device
Zone (C)	Construction/Maintenance Zone

ACKNOWLEDGMENTS

The authors would like to thank John E. Baumler for his great help and contribution in devising and implementing the stochastic approximation scheme for normalized cut.

The first author was partially funded by the New-York Metropolitan and the Technion's Security Science and Technology research funds, The German-Israeli Foundation for Scientific Research and Development (GIF) Young Scientist Program, the Technion Center of Excellence in Exposure Science and Environmental Health and the CITI-SENSE project of the 7th European Framework Program (FP7/2007-2013), under grant agreement No. 30824.F. The second author acknowledges the support of the UC Berkeley Safe Transportation Research and Education Center (SafeTREC).

REFERENCES

Abdel-Aty, M., and A. Pande, *Crash data analysis: Collective vs. individual crash level approach*, Journal of Safety Research 38 (2007), no. 5, 581 – 587.

Aho, A., J. Hopcroft, and J. Ullman, *Data structure and algorithms*, Boston: Addison-Wesley, 1983.

Alon, N., D. Moshkovitz, and S. Safra, *Algorithmic construction of sets for k-restrictions*, ACM Trans. Algorithms (2006), no. 2, 153–177.

Ang, Q., A. Baddeley, and G. Nair, *Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology*, Scandinavian Journal of Statistics 39 (2012), no. 4, 591-617.

Bailey, T.C., and A.C. Gatrell, *Interactive spatial data analysis*, Longman, Harlow, 1995.

Black, W.R. *Highway accidents: a spatial and temporal analysis*, Transportation Research Record 1318 (1991), 75–82.

Cressie, N., *Statistics for spatial data*, John Wiley & Sons, New York., 1993.

Depaire, B., G. Wets, and K. Vanhoof, *Traffic accident segmentation by means of latent class clustering*, Accident Analysis & Prevention 40 (2008), no. 4, 1257 – 1266.

Diggle, P.J. *Spatial analysis of spatial point patterns*, Academic Press, New York., 1983.

Ford, L.R., and D.R. Fulkerson, *Maximal flow through a network*, Canadian Journal of Math. 8 (1956), no. 3, 339 – 404.

Fotheringham, A., C. Brunsdon, and M. Charlton, *Quantitative geography: Perspectives on spatial data analysis,* Sage Publication, London., 2000.

Frhwirth-Schnatter, S., *Finite mixture and markov switching models*, Springer Series in Statistics, Springer, New York, 2006.

Getis, A., and J. Franklin, *Second-order neighborhood analysis of mapped point patterns, Perspectives on Spatial Data Analysis* (Luc Anselin and Sergio J. Rey, eds.), Advances in Spatial Science, Springer Berlin Heidelberg, 2010, pp. 93–100.

Golob, T.F., and W.W. Recker, *Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions*, Journal of Transportation Engineering 129 (2003), no. 4, 342–353.

Golob, T.F., and W.W. Recker, *A method for relating type of crash to traffic flow characteristics on urban freeways*, Transportation Research Part A: Policy and Practice 38 (2004), no. 1, 53 – 80.

Griswold, J., B. Fishbain, S. Washington, and D.R. Ragland, *Visual assessment of pedestrian crashes*, Accident Analysis & Prevention 43 (2011), no. 1, 301 – 306.

Guo, F., and Y. Fang, *Individual driver risk assessment using naturalistic driving data*, Accident Analysis & Prevention (2012).

Hauer, E. *On the estimation of the expected number of accidents*, Accident Analysis & Prevention 18 (1986), no. 1, 1– 12.

Hogg, R. V., and J. Ledolter, *Engineering statistics*, New York: MacMillan, 1987.

Karger, D.R., and C. Stein, *A new approach to the minimum cut problem*, J. ACM 43 (1996), no. 4, 601–640.

Karp, R. M., *Complexity of computer computations*, Reducibility Among Combinatorial Problems, p. 85103, New York: Plenum, 1972.

Kendall, M.G., *A new measure of rank correlation*, Biometrika 30 (1938), no. 1-2, 81.

Kingham, S., C.E. Sabel, and P. Bartie, *The impact of the school run on road traffic accidents: A spatio-*

temporal analysis, Journal of Transport Geography 19 (2011), no. 4, 705 – 711.

Klauer, S.G., T.A. Dingus, V.L. Neale, JD Sudweeks, and DJ Ramsey, *The impact of driver inattention on near crash/crash risk: An analysis using the 100-car naturalistic driving study data*, Tech. report, 2006.

Lee, C., B. Hellinga, and F. Saccomanno, *Real-time crash prediction model for application to crash prevention in freeway traffic*, Transportation Research Record 1840 (2007), 67–77.

Li, X., D. Lord, Y. Zhang, and Y. Xie, *Predicting motor vehicle crashes using support vector machine models*, Accident Analysis & Prevention 40 (2008), no. 4, 1611 – 1618.

Lord, D., and F. Mannering, *The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives*, Transportation Research Part A: Policy and Practice 44 (2010), no. 5, 291 – 305.

Ma, J., and K.M. Kockelman, *Bayesian multivariate Poisson regression for models of injury count, by severity*, Transportation Research Record: Journal of the Transportation Research Board 1950 (2006), no. 1, 24–34.

McGuigan, D.R.D. *The use of relationships between road accidents and traffic flow in black-spot identification*, Traffic Engineering and Control 22 (1981), 448 – 453.

Miaou, S., and H. Lum, *Modeling vehicle accidents and highway geometric design relationships*, Accident Analysis & Prevention 25 (1993), no. 6, 689 – 709.

Miaou, S., and J.J. Song, *Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence*, Accident Analysis & Prevention 37 (2005), no. 4, 699 – 720.

Mussone, L., A. Ferrari, and M. Oneta, *An analysis of urban collisions using an artificial intelligence model*, Accident Analysis & Prevention 31 (1999), no. 6, 705 – 718.

National Highway Traffic Safety Administration, Fatality analysis reporting system: Fatal crash data overview, U.S. Department of Transportation Publication Washington, DC, 2004, DOT HS 809726.

Persaud, B.N., *Estimating accident potential of Ontario road sections*, Transportation Research Record 1327 (1991), 4753.

Plug, C., J.C. Xia, and C. Caulfield, *Spatial and temporal visualization techniques for crash analysis*, Accident Analysis & Prevention 43 (2011), no. 6, 1937 – 1946.

Rifaat, S. M., and H. C. Chin, *Accident severity analysis using ordered Probit* model, Journal of Advanced Transportation 41 (2007), no. 1, 91–114.

Shi, J.B., and J. Malik, *Normalized cuts and image segmentation*, IEEE Transactions On Pattern Analysis and Machine Intelligence 22 (2000), no. 8, 888–905.

Shino, S., *Analysis of a distribution of point events using the network-based quadrat method*, Geographical Analysis 40 (2008), no. 4, 380–400.

Steenberghen, T., K. Aerts, and I. Thomas, *Spatial clustering of events on a network,* Journal of Transport Geography 18 (2010), no. 3, 411 – 418

Steenberghen, T., T. Dufays, I. Thomas, and B. Flahaut, *Intra-urban location and clustering of road accidents using GIS: a Belgian example*, International Journal of Geographical Information Science 18 (2004), no. 2, 169-181.

Sullivan, J.M., and M.J. Flannagan, *Characteristics of pedestrian risk in darkness*, http://hdl.handle.net/2027.42/49450, 2001, UMTRI-2001-33, PB2002-101216.

Tay, R., U. Barua, and L. Kattan, *Factors contributing to hit-and-run in fatal crashes*, Accident Analysis & Prevention 41 (2009), no. 2, 227 – 233.

Tay, R., S. M. Rifaat, and H. C. Chin, *A logistic model of the effects of roadway, environmental, vehicle, crash and driver characteristics on hit-and-run crashes*, Accident Analysis & Prevention 40 (2008), no. 4, 1330 – 1336.

Vlahogianni, E. I., M. G. Karlaftis, and J. C. Golias, *Spatio-temporal short-term urban traffic volume forecasting using genetically optimized modular networks*, Computer-Aided Civil and Infrastructure Engineering 22 (2007), no. 5, 317–325.

Vogt, A., and J. Bared, *Accident models for two-lane rural segments and intersections*, Transportation Research Record: Journal of the Transportation Research Board 1635 (1998), no. 1, 18–29.

Wong, S.C., N.N. Sze, and Y.C. Li, *Contributing factors to traffic crashes at signalized intersections in Hong Kong*, Accident Analysis & Prevention 39 (2007), no. 6, 11071113.

Xie, Y., D. Lord, and Y. Zhang, *Predicting motor vehicle collisions using bayesian neural network models: An empirical analysis*, Accident Analysis & Prevention 39 (2007), no. 5, 922 – 933.

Xie, Z., and J. Yan, *Kernel density estimation of traffic accidents in a network space*, Computers, Environment and Urban Systems 32 (2008), no. 5, 396 – 406.

Yamada, I., and J. Thill, *Comparison of planar and network K-functions in traffic accident analysis*, Journal of Transport Geography 12 (2004), no. 2, 149 – 158.

Zhang, Y., and Y. Xie, *Forecasting of short-term freeway volume with <i>v</i>-support vector machines*, Transportation Research Record: Journal of the Transportation Research Board 2024 (2008), no. 1, 92–9

Volume 10 Number 1, 2014
ISSN 1094-8848

JOURNAL OF TRANSPORTATION AND STATISTICS

CONTENTS