Zero-inflated Models for Different Severity Types in Rural Two-lane Crashes

Subasish Das (Corresponding author)
PhD. Candidate
Systems Engineering
University of Louisiana
Lafayette, LA 70504
Email: sxd1684@louisiana.edu
Phone: 225-288-9875

Xiaoduan Sun, Ph.D. & PE
Professor
Civil Engineering Department
University of Louisiana
Lafayette, LA 70504
Email: xsun@louisiana.edu
Phone: 337-739-6732
Fax: 337-739-6688

Word Count: 6,000 including 2 Figures and 4 Tables

Submitting to the 94th TRB Annual Meetings for Presentation and Publication under
Safety Data, Analysis, and Evaluation (ANB20)

ABSTRACT

This research aims to investigate the application of zero-inflated models for different severity types in rural two-lane highway crashes. These roadways carry one-third of the total vehicle miles traveled (VMT) and have experienced a considerably high percentage of fatal crashes in Louisiana. A careful analysis indicates that a wide variety of factors appear to be associated with the crash dynamic of rural two-lane highways. The roadway variables include segment length, pavement width and type, shoulder type, and traffic volume. Crashes recorded from 2004 to 2011, of which 1,780 were fatal, and 36,569 resulted in injuries, were analyzed. It is found that there are a large number of highway segments which contain no crashes under the recorded years. To tackle this issue, zero-inflated models, zero-inflated Poisson (ZIP) models and zero-inflated negative binomial (ZINB) models, have been developed for crash frequencies of different severity types. The researchers of this study have used the qualitative values of the variables to develop the model for convenient interpretation. The results showed that specific categories of traffic flow, segment length, pavement type and width, and shoulder type were found to be statistically significant variables for total, injury, and property damage only (PDO) crashes. Two additional associations are: 1) wider shoulder and pavement width reduced the likelihood of crash occurrence, and 2) roadways with gravel-top pavements were inclined towards crash proneness. The findings of this paper will help highway professionals improve the safety outcome of rural two-lane roadways.

Key words: rural two-lane highways, count data modeling, over dispersion, severity type, zero-inflated models.

INTRODUCTION

Highway safety is a crucial issue in Louisiana. The State of Louisiana controls 60,937 miles of public road serving nearly 105,000 vehicle miles a day, and consisting of 46,959 miles of rural roads and 13,941 miles of urban roads. Nearly 58,000 miles of undivided rural roadways are two-lane in nature [1]. Each year, approximately 150,000 crashes occur, over 90,000 of which are on the state-maintained highway system. In 2013, 703 people were killed and 70,658 were injured in highway crashes in Louisiana. Rural two-lane highways in this state carry one-third of the total vehicle miles traveled (VMT) and have experienced a considerably high percentage of fatal crashes. In 2012, approximately 35% of fatal crashes and 36% of fatalities in the entire state occurred on rural two-lane highways [2].

The conservative method of traffic safety research is to establish relationships between the roadway characteristics and crash occurrence. It includes a wide-ranging exhibition of research areas and the most prominent of them is exploratory analysis of crash frequency data. In recent years attention has been increased at determining the key association factors affecting the injury severity outcome in traffic crashes. Count-data modeling methods are widely used for crash frequency analysis as the number of crashes on roadway segment per unit of time is a non-negative integer. Traditionally, highway safety analyses have used Poisson or negative binomial distributions to model crash counts for different levels of crash severity. Crashes recorded from 2004 to 2011, of which 1780 were fatal, 36,569 were injury crashes, and 48,996 resulted no injuries, were analyzed in this study. A careful observation indicates that there are a large number of highway segments which contain no crashes under the recorded years. Zero-inflated models, zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB), have been developed in this study for crash frequencies of different severity types. These models effectively handle data characterized by an excessive amount of zeroes. The researchers of this study used the qualitative values of the variables to develop the model for convenient interpretation.

LITERATURE REVIEW

In recent literature, it has been suggested that traffic crashes can effectively be modeled by assuming a dual-state data-generating procedure which implies that geometric properties exist in one of two states—perfectly safe and unsafe. As a result, the ZIP and ZINB are two models that have been applied to account for the excessive zeroes frequently observed in crash count data. From the start, zero-inflated models have been widely popular among transportation safety researchers [4-7].

Zero-inflated models have been used in traffic safety studies to modeling crashes for different applications: single and multi-vehicle crashes on rural two-lane roads [3, 4, and 8]; single vehicle crashes in rural roadways [6], and vehicle-pedestrian crashes on urban and suburban areas [4]. In these studies, usage of the zero-inflated models has been justified by the test statistic of the Vuong test. The authors usually assumed that crashes must follow a dual-state process, with the exception of Miaou (1994). Miaou et al. first used ZIP structure for traffic crash analysis [3]. Shankar et al. presented an empirical review into the applicability of zero-inflated count data modeling to roadway segment crash frequencies [4]. The findings show that the ZIP structure models are sufficient enough to justify the model. A study by Lee et al. used zero-inflated count models and nested logit models for developing crash frequency models and severity models. The findings also showed significant potential in applying these two techniques to single vehicle crash analysis [5]. In their study, Shankar et al. employed an empirical inquiry into the predictive modeling of crashes involving pedestrians and motorized traffic on roadways. Empirical models based on ZIP were presented and discussed in terms of their applicability to pedestrian crash phenomena [7]. The results showed that ZIP is effective enough to provide explanatory insights into the causality behind pedestrian-traffic crashes. In their paper, Lord et al. attempted to provide defensible guidance on how to appropriate model crash data. They used ZIP and ZINB to account for the dominance of excessive zeroes observed in crash count data [8].

However, comparison of the traditional Poisson and negative binomial models with the ZIP and ZINB models for the frequency of different severity types has yet to be applied in traffic crash analysis research. In this study, we have applied ZINB and ZIP distributions to the eight years (2004-2011) of count dataset from Louisiana rural two-lane highways. In place of using continuous variables, this study uses the categorical recoding of the continuous variables. This study has applied the technique to total, injury and PDO crashes to evaluate the significance of the roadway categorical variables. These results provide robust support for the notion that the usage of the qualitative roadway factors in negative binomial modeling is adequate for developing the predictive model for rural two-lane highways.

BACKGROUND

Count Data Models

To deal with the data and methodological issues associated with crash-frequency data, a wide variety of methods have been applied over the years. As crash-frequency data are non-negative integers, the application of the standard ordinary least-squares regression (which assumes a continuous dependent variable) is not appropriate. Given that the dependent variable is a non-negative integer, most of the recent thinking in the field has used the Poisson regression model as a starting point. If the discrete random variable X is Poisson distributed with intensity or rate parameter λ, where $\lambda > 0$, then X has probability mass function (pmf)

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}; \quad k = 0,1,2,3... \tag{1}$$

This pmf is widely used to model many naturally occurring events where X represents the “number of events per unit of time or space”. It’s important to note that X takes only nonnegative integer values [9].

Zero-inflated Models

Although the Poisson model has served as a starting point for crash-frequency analysis for several decades, researchers have often found that crash data exhibit characteristics that make the application of the simple Poisson regression (as well as some extensions of the Poisson model) problematic. In such a case, a modified version of a regular *Poi*(λ) distribution, known as the ZIP distribution, becomes useful.

Let X_i be the number of crashes on roadway section i in some specified time period and let π_i be the probability that roadway section i will exist in the zero-crash state. Thus 1 -

π_i is the probability that a zero-crash observation actually follows a true Poisson distribution.

The ZIP distribution with parameters π_i and λ_i has the following probability mass function:

$$
\begin{aligned}
P(X_i = 0) &= \pi_i + (1-\pi_i)e^{-\lambda_i}; \quad k = 0 \\
P(X_i = k) &= (1-\pi_i)\frac{\lambda_i^{k} e^{-\lambda_i}}{k!}; \quad k > 0
\end{aligned}
\tag{2}
$$

Here, $0 \leqslant \pi_i \leqslant 1$ and $\lambda_i \geqslant 0$. Henceforth, the probability mass function in (2) will be referred to as the ZIP(π, λ) distribution. The parameter λ_i gives the extra probability thrust at the value 0. Note that when $\pi_i = 0$, then ZIP(π_i, λ_i) reduces to Poi(λ_i). The probability π may be set as a constant or may depend on regressors via a binary outcome model such as logit or probit.

The equation (2) can be viewed as a finite mixture model with two components. The mixture weights for the two components are π_i and $1 - \pi_i$. The mean and variance of ZIP(π_i, λ_i) are given as follows:

$$
\begin{aligned}
E(X_i) &= \lambda_i(1-\pi_i) \\
Var(X_i) &= \lambda_i(1-\pi_i)(1+\lambda_i\pi_i)
\end{aligned}
\tag{3}
$$

The ZINB regression model follows a similar formulation with events, $X = (X_1, X_2, \ldots, X_n)$, being independent

$$
\begin{aligned}
P(X_i = 0) &= \pi_i + (1-\pi_i)\left(\frac{\delta}{\delta+\lambda_i}\right)^{\delta}; \quad k = 0 \\
P(X_i = k) &= (1-\pi_i)\left(\frac{\Gamma(\delta+k)\gamma_i^{\delta}(1-\gamma_i)^{k}}{k!\,\Gamma(\delta)}\right); \quad k > 0
\end{aligned}
\tag{4}
$$

where,

$\delta = \dfrac{1}{\alpha}$ [α is the dispersion parameter]

$\gamma_i = \dfrac{\delta}{\delta + \lambda_i}$

It's important to note that the dispersion parameter, α, relaxes the Poisson assumption that requires the mean to be equal to the variance by letting $Var(X_i) = E(X_i)[1 + \alpha E(X_i)]$.

The ZIP and ZINB regressions directly model the zeroes in the structural portion of the model. ZIP and ZINB models are generally considered as mixture models in which the complete distribution of the outcome is approximated by mixing two component distributions. The basic idea is to assume a logistic regression model for the *'zero, and not zero'* aspect of the consequence and either a Poisson or negative binomial distribution for the count portion in the model. ZIP and ZINB are well suited for the models in which there are two procedures and where the factors of the two procedures vary [9].

METHODOLOGY

Data Preparation

The source of traffic crash data was the Louisiana Department of Transportation and Development (LADOTD) crash database. The data was obtained in computer-ready form, which included coded information on reported crashes that occurred on the state highways in Louisiana. The coded information for each crash contains important attributes describing the conditions that contributed to the collision and the outcome. The final count dataset was prepared from the DOTD section data. The important roadway factors considered in the study include segment length, pavement type and width, shoulder type and width, and annual average daily traffic (AADT). These were categorized into subclasses from the original records as shown in Table 1.

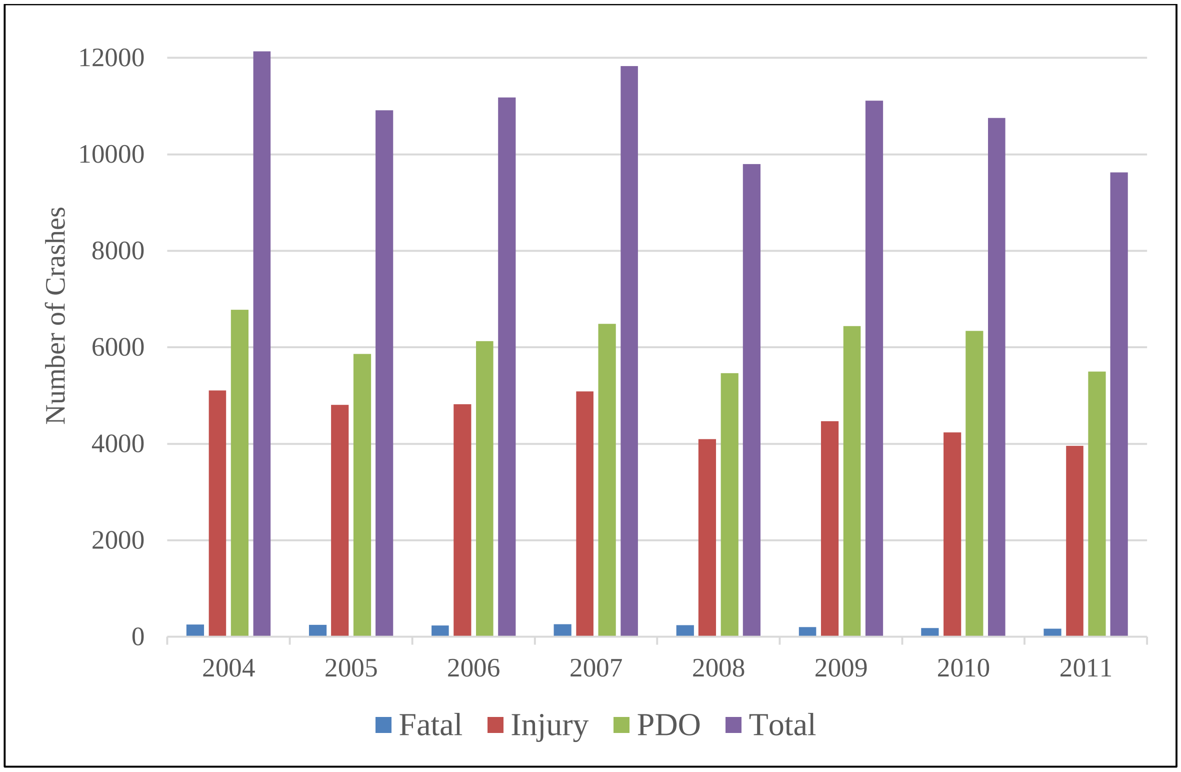

FIGURE 1 Crash frequencies of different severity types.

There are a total of 7,779 rural two-lane roadway segments in each year's crash dataset. The key variables available in the current dataset, related to roadway geometrics, were considered here. LADOTD maintained crash data doesn't have details on other roadway geometrics like vertical and horizontal curve degree, deflection angle, and percentage of gradient. The segment length varies from 0.01 to 27.5 miles, with an average

TABLE 1 Percentage of Crash Frequencies by Key Variables

Category	Total	Fatal	Injury	PDO
SECTION_LENGTH				
0.00-0.50	5.31%	2.87%	4.78%	5.79%
0.51-1.00	6.91%	5.00%	6.67%	7.16%
1.01-2.00	11.96%	10.45%	11.61%	12.28%
2.01-3.00	13.76%	13.43%	13.86%	13.69%
3.01-4.00	12.96%	13.03%	13.34%	12.68%
4.00 above	49.10%	55.22%	49.74%	48.40%
ADT				
0-2000	32.49%	36.12%	33.61%	31.52%
2001-6000	45.94%	46.07%	45.91%	45.96%
6001-10000	14.95%	12.42%	14.61%	15.29%
10001-20000	6.56%	5.34%	5.82%	7.16%
20000 above	0.06%	0.06%	0.05%	0.06%
SHOULDER_TYPE				
Shoulder < 6 ft.	59.11%	58.43%	59.69%	58.70%
Shoulder > 6 ft.	40.02%	41.01%	39.57%	40.32%
Curb and Gutter	0.82%	0.51%	0.69%	0.93%
No Info.	0.05%	0.06%	0.05%	0.05%
PAVEMENT_TYPE				
Bituminous Concrete	92.67%	93.93%	92.44%	92.80%
Bituminous	6.43%	5.76%	6.73%	6.20%
PCC Concrete	0.73%	0.20%	0.65%	0.77%
Gravel	0.07%	0.00%	0.06%	0.13%
No Info.	0.10%	0.11%	0.11%	0.09%
PAVEMENT_WIDTH				
Wide	54.64%	51.80%	53.97%	55.25%
Narrow	44.63%	47.58%	45.29%	44.03%
Very Wide	0.73%	0.62%	0.74%	0.72%

of 2.26 miles. The pavement width varies from 18 to 38 ft, with an average of 22 ft. The shoulder width varies from 0 to 21 ft, with an average of 4.2 ft. The AADT value varies from 0 to 24100, with an average of 2,447. The originally defined crash types (fatal, severe, moderate, complaint and PDO) have been re-categorized into the four groups (total, fatal, injury and PDO). The percentages of the crash frequencies by key variables are listed in Table 1.

The frequency of crash severity from 2004 to 2011 is illustrated in Figure 1. The highest number of crashes happened in 2007. PDO and Injury crashes are more frequent than severe or fatal crashes with a sudden decline visible in 2008.

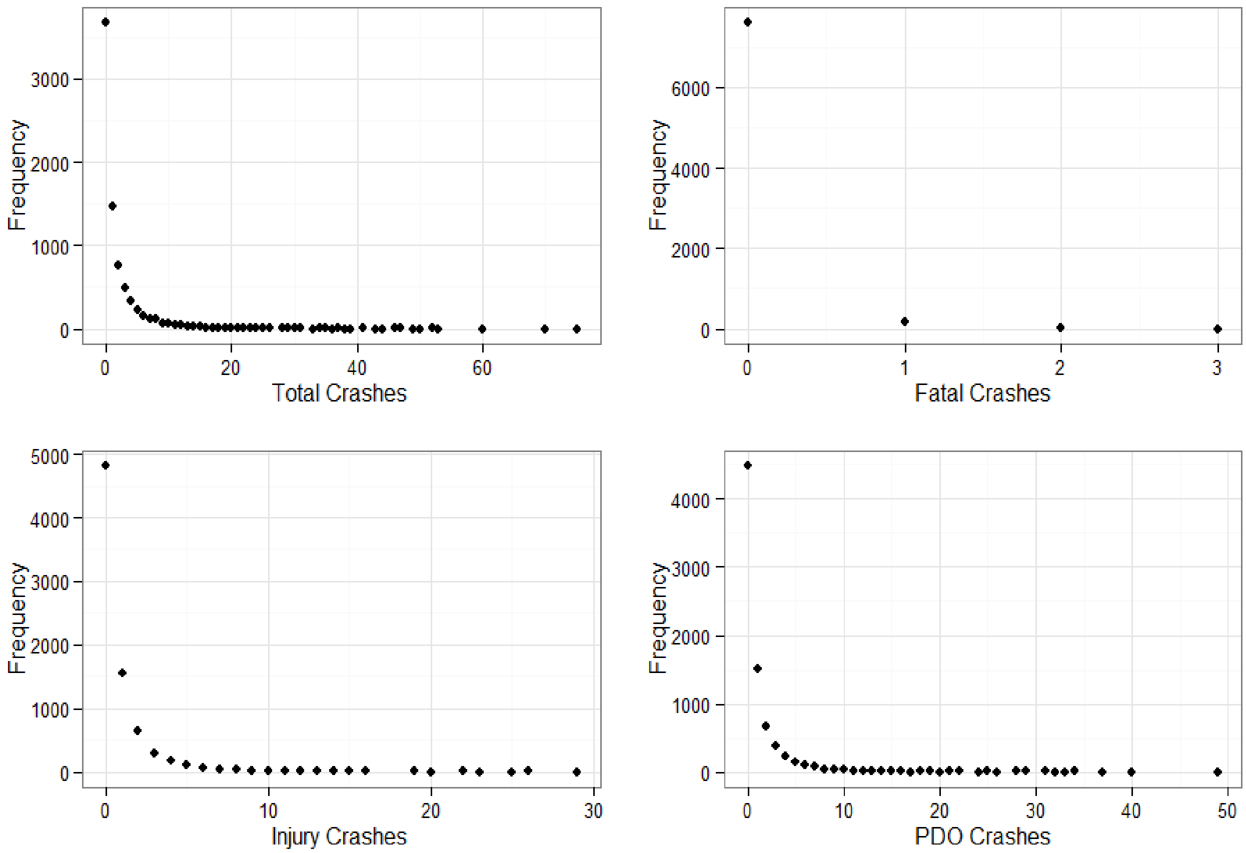

FIGURE 2 Crash frequency of different count of crashes per segment.

In the DOTD control section databases, there are 7,779 control sections in rural two-lane highways in each year's dataset. There are a significant number of highway segments where no crashes occurred in the eight years of period (2004-2011). Figure 2 illustrates the crash frequencies of the segments for different types of crash severities.

Modeling Results

The main objective of modeling with several variables simultaneously was to permit greater insight into the relative effects of the different roadway geometric variables on crashes. It is also important to know that there are a large number of variables (some of them are redundant in model development) apart from traffic flow and length that might contribute to crashes. The variable selection is based on extensive literature review and principle component analysis of the preliminary dataset. Modeling was undertaken for three stages of traffic severities—total, injury and PDO crashes. We have used Poisson, negative binomial, ZIP and ZINB models for all three different datasets. The model development in this paper was performed by using open source statistical "R Version 3.02" software [10].

The coefficients for both the non-zero-crash state and the zero-crash state were found to be statistically significant and of plausible sign. The results of the ZIP and ZINB models are shown in Table 2 and Table 3.

TABLE 2 Zero-Inflated Poisson (ZIP) Coefficients

	Total Crashes				Injury Crashes				PDO Crashes			
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Count model coefficients (Poisson with log link)												
(Intercept)	-0.996	0.123	-8.079	**0.000**	-1.981	0.282	-7.028	**0.000**	-1.270	0.176	-7.225	**0.000**
SECTION_LENGTH0.51-1.00	0.422	0.063	6.676	**0.000**	0.442	0.156	2.829	0.005	0.389	0.095	4.107	**0.000**
SECTION_LENGTH1.01-2.00	0.868	0.056	15.474	**< 2e-16**	0.889	0.139	6.388	**0.000**	0.857	0.084	10.238	**< 2e-16**
SECTION_LENGTH2.01-3.00	1.280	0.055	23.209	**< 2e-16**	1.350	0.135	9.989	**< 2e-16**	1.163	0.083	14.066	**< 2e-16**
SECTION_LENGTH3.01-4.00	1.556	0.056	28.008	**< 2e-16**	1.662	0.136	12.260	**< 2e-16**	1.435	0.083	17.255	**< 2e-16**
SECTION_LENGTH4.00 above	2.066	0.052	39.753	**< 2e-16**	2.167	0.131	16.550	**< 2e-16**	1.988	0.078	25.522	**< 2e-16**
ADT2001-6000	0.927	0.021	43.285	**< 2e-16**	0.920	0.039	23.854	**< 2e-16**	0.907	0.031	29.047	**< 2e-16**
ADT6001-10000	1.537	0.028	54.429	**< 2e-16**	1.554	0.049	31.928	**< 2e-16**	1.539	0.040	38.455	**< 2e-16**
ADT10001-20000	1.870	0.038	49.247	**< 2e-16**	1.819	0.066	27.528	**< 2e-16**	1.860	0.052	35.596	**< 2e-16**
ADT20000 above	2.843	0.175	16.293	**< 2e-16**	2.653	0.294	9.020	**< 2e-16**	2.922	0.223	13.094	**< 2e-16**
PAVEMENT_WIDTHVery Wide	-0.367	0.123	-2.978	**0.003**	-0.142	0.208	-0.681	0.496	-0.533	0.180	-2.952	**0.003**
PAVEMENT_WIDTHWide	0.083	0.020	4.236	**0.000**	0.065	0.034	1.938	**0.053**	0.062	0.027	2.252	**0.024**
SHOULDER_TYPENo Info.	-0.552	0.288	-1.920	0.055	-0.300	0.534	-0.561	0.575	-0.454	0.473	-0.961	0.337
SHOULDER_TYPEShoulder < 6 ft.	0.236	0.121	1.958	0.050	0.272	0.285	0.956	0.339	0.090	0.172	0.525	0.600
SHOULDER_TYPEShoulder > 6 ft.	0.040	0.120	0.331	0.741	0.086	0.285	0.301	0.764	-0.105	0.172	-0.610	0.542
Zero-inflation model coefficients (binomial with logit link)												
(Intercept)	1.917	0.291	6.576	**0.000**	2.450	0.539	4.547	**0.000**	2.131	0.353	6.033	**0.000**
SECTION_LENGTH0.51-1.00	-1.075	0.142	-7.598	**0.000**	-1.226	0.313	-3.914	**0.000**	-1.032	0.188	-5.496	**0.000**
SECTION_LENGTH1.01-2.00	-1.421	0.127	-11.223	**< 2e-16**	-1.688	0.270	-6.249	**0.000**	-1.233	0.160	-7.722	**0.000**
SECTION_LENGTH2.01-3.00	-1.707	0.132	-12.945	**< 2e-16**	-2.084	0.263	-7.932	**0.000**	-1.781	0.171	-10.411	**< 2e-16**
SECTION_LENGTH3.01-4.00	-1.738	0.138	-12.632	**< 2e-16**	-2.045	0.258	-7.937	**0.000**	-1.791	0.173	-10.330	**< 2e-16**
SECTION_LENGTH4.00 above	-2.164	0.119	-18.129	**< 2e-16**	-2.376	0.230	-10.339	**< 2e-16**	-2.006	0.148	-13.588	**< 2e-16**
ADT2001-6000	-0.562	0.084	-6.715	**0.000**	-0.522	0.127	-4.101	**0.000**	-0.614	0.095	-6.457	**0.000**
ADT6001-10000	-0.656	0.138	-4.757	**0.000**	-0.579	0.190	-3.047	**0.002**	-0.577	0.146	-3.942	**0.000**
ADT10001-20000	-0.767	0.220	-3.481	**0.000**	-0.850	0.304	-2.801	**0.005**	-0.982	0.239	-4.107	**0.000**
ADT20000 above	-12.709	1.2E+03	-0.011	0.991	-13.504	1.8E+03	-0.007	0.994	-13.288	1.4E+03	-0.009	0.993
PAVEMENT_WIDTHVery Wide	-0.499	4.2E-01	-1.202	0.230	0.344	5.0E-01	0.692	**0.489**	-0.685	5.2E-01	-1.322	0.186
PAVEMENT_WIDTHWide	-0.385	0.087	-4.437	**0.000**	-0.453	0.131	-3.465	**0.001**	-0.421	0.097	-4.357	**0.000**
SHOULDER_TYPENo Info.	-0.283	7.5E-01	-0.378	0.705	0.464	1.0E+00	0.446	0.656	-0.193	0.941	-0.205	0.837
SHOULDER_TYPEShoulder < 6ft.	-0.384	0.280	-1.375	0.169	-0.544	0.553	-0.983	0.325	-0.282	0.343	-0.821	0.411
SHOULDER_TYPEShoulder > 6ft.	-0.651	2.8E-01	-2.328	**0.020**	-0.778	5.5E-01	-1.404	0.160	-0.537	0.342	-1.568	0.117
PAVEMENT_TYPEBituminousConcrete	-0.549	0.096	-5.745	**0.000**	-0.668	0.133	-5.016	**0.000**	-0.638	0.107	-5.982	**0.000**
PAVEMENT_TYPEGravel	0.470	0.359	1.309	0.190	1.393	0.512	2.721	**0.007**	0.328	0.416	0.789	0.430
PAVEMENT_TYPENo Info.	14.038	3.2E+02	0.044	0.965	13.139	3.7E+02	0.035	0.972	13.705	3.4E+02	0.040	0.968
PAVEMENT_TYPEPCC Concrete	-0.902	3.2E-01	-2.788	**0.005**	-0.882	4.5E-01	-1.971	**0.049**	-0.349	3.3E-01	-1.058	0.290

TABLE 3 Zero-Inflated Negative Binomial (ZINB) Coefficients

	Total Crashes				Injury Crashes				PDO Crashes			
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Count model coefficients (negbin with log link)												
(Intercept)	-1.693	0.189	-8.965	**< 2e-16**	-2.764	0.264	-10.480	**< 2e-16**	-2.218	0.324	-6.840	**0.000**
SECTION_LENGTH0.51-1.00	0.549	0.096	5.687	**0.000**	0.389	0.162	2.404	**0.016**	0.612	0.132	4.627	**0.000**
SECTION_LENGTH1.01-2.00	1.087	0.091	12.010	**< 2e-16**	0.989	0.158	6.278	**0.000**	1.140	0.127	8.975	**< 2e-16**
SECTION_LENGTH2.01-3.00	1.529	0.091	16.870	**< 2e-16**	1.518	0.160	9.517	**< 2e-16**	1.508	0.128	11.758	**< 2e-16**
SECTION_LENGTH3.01-4.00	1.815	0.094	19.210	**< 2e-16**	1.826	0.160	11.395	**< 2e-16**	1.720	0.130	13.263	**< 2e-16**
SECTION_LENGTH4.00 above	2.416	0.086	28.106	**< 2e-16**	2.419	0.155	15.640	**< 2e-16**	2.406	0.120	20.006	**< 2e-16**
ADT2001-6000	0.970	0.041	23.778	**< 2e-16**	0.992	0.044	22.540	**< 2e-16**	0.966	0.048	20.185	**< 2e-16**
ADT6001-10000	1.600	0.062	25.672	**< 2e-16**	1.626	0.066	24.713	**< 2e-16**	1.603	0.072	22.171	**< 2e-16**
ADT10001-20000	2.040	0.101	20.221	**< 2e-16**	1.912	0.101	18.922	**< 2e-16**	2.002	0.108	18.551	**< 2e-16**
ADT20000 above	2.966	0.868	3.416	**0.001**	2.815	0.798	3.529	**0.000**	3.183	0.890	3.575	**0.000**
PAVEMENT_WIDTHVery Wide	-0.190	0.208	-0.913	0.361	0.091	0.272	0.335	0.737	-0.446	0.234	-1.903	0.057
PAVEMENT_WIDTHWide	0.170	0.039	4.398	**0.000**	0.172	0.043	4.039	**0.000**	0.161	0.045	3.554	**0.000**
SHOULDER_TYPENo Info.	-0.318	0.456	-0.696	0.486	-0.089	0.606	-0.147	0.883	-0.427	0.455	-0.939	0.348
SHOULDER_TYPEShoulder < 6 ft.	0.367	0.182	2.017	**0.044**	0.525	0.245	2.139	0.032	0.339	0.311	1.092	0.275
SHOULDER_TYPEShoulder > 6 ft.	0.269	0.182	1.474	0.140	0.436	0.245	1.781	0.075	0.199	0.317	0.628	0.530
Log(theta)	0.327	0.054	6.082	**0.000**	0.601	0.059	10.133	**< 2e-16**	0.301	0.059	5.075	**0.000**
Zero-inflation model coefficients (binomial with logit link)												
(Intercept)	1.290	0.668	1.933	0.053	2.979	0.952	3.127	**0.002**	1.074	2.648	0.406	0.685
SECTION_LENGTH0.51-1.00	-1.895	0.500	-3.790	**0.000**	-5.798	3.069	-1.889	0.059	-1.679	0.540	-3.108	**0.002**
SECTION_LENGTH1.01-2.00	-1.981	0.358	-5.530	**0.000**	-4.100	0.809	-5.065	**0.000**	-1.739	0.438	-3.975	**0.000**
SECTION_LENGTH2.01-3.00	-2.373	0.403	-5.896	**0.000**	-4.551	0.932	-4.881	**0.000**	-2.756	0.484	-5.699	**0.000**
SECTION_LENGTH3.01-4.00	-2.190	0.410	-5.343	**0.000**	-3.985	0.732	-5.447	**0.000**	-3.887	1.160	-3.351	**0.001**
SECTION_LENGTH4.00 above	-2.401	0.308	-7.786	**0.000**	-4.167	0.533	-7.821	**0.000**	-2.751	0.393	-7.003	**0.000**
ADT2001-6000	-0.645	0.255	-2.528	**0.011**	-0.815	0.365	-2.234	**0.025**	-1.245	0.367	-3.389	**0.001**
ADT6001-10000	-0.443	0.367	-1.205	0.228	-1.062	0.513	-2.071	0.038	-0.853	0.600	-1.421	0.155
ADT10001-20000	-0.505	0.559	-0.903	0.367	-3.272	1.587	-2.062	**0.039**	-3.080	3.352	-0.919	0.358
ADT20000 above	-12.709	2.6E+03	-0.005	0.996	-13.504	5.0E+03	-0.003	0.998	-13.288	8.5E+03	-0.002	0.999
PAVEMENT_WIDTHVery Wide	-0.178	7.4E-01	-0.240	0.810	2.387	9.0E-01	2.663	**0.008**	-2.280	6.0E+00	-0.379	0.705
PAVEMENT_WIDTHWide	-0.540	0.257	-2.097	**0.036**	-0.275	0.376	-0.732	0.464	-0.793	0.412	-1.924	0.054
SHOULDER_TYPENo Info.	0.241	1.8E+00	0.136	0.892	2.348	1.9E+00	1.220	0.222	-7.379	31.222	-0.236	0.813
SHOULDER_TYPEShoulder < 6ft.	-0.191	0.642	-0.297	0.767	-0.193	0.851	-0.227	0.821	0.775	2.714	0.285	0.775
SHOULDER_TYPEShoulder > 6ft.	-0.460	6.6E-01	-0.699	0.485	0.168	8.2E-01	0.206	0.837	0.058	2.830	0.020	0.984
PAVEMENT_TYPEBituminousConcrete	-1.081	0.219	-4.945	**0.000**	-2.400	0.496	-4.841	**0.000**	-1.268	0.258	-4.914	**0.000**
PAVEMENT_TYPEGravel	0.658	0.528	1.247	0.212	2.266	0.710	3.190	**0.001**	0.572	0.591	0.968	0.333
PAVEMENT_TYPENo Info.	14.038	2.9E+02	0.049	0.961	13.140	4.9E+02	0.027	0.979	17.709	4.1E+01	0.433	0.665
PAVEMENT_TYPEPCC Concrete	-2.218	1.4E+00	-1.626	0.104	-2.514	1.1E+00	-2.302	**0.021**	-0.562	1.1E+00	-0.530	0.596

The idea of zero-inflated model is simple: it assumes that the outcomes originate from two processes. One process models zero inflation, the second models the non-zero counts using ZIP or ZINB. By observing the values from Tables 2 and 3, we find that the all types of segment length, all ADT values, and wide pavement were significant for the count model part of both models because the associated *p* value of these factors is less than 5%. We remark that the categories of variables, i.e. all types of segment length, low volume ADTs, wide pavement, specific shoulder types and bituminous pavements were statistically significant for the zero-inflated part.

An increasing unit value of these categories was found to reduce the likelihood of rural two-lane highway crash occurrence. For example, the variable *ADT2001-6000* in the ZIP model has a coefficient of *-0.562* for the model developed for total crashes; this category is statistically significant. In ADT categories, when the ADT values were over 20,000, the category was less significant in the ZIP model for total, injury and PDO crashes. But in ZINB model, ADT values were significant for only the 2001-6000 level for total and PDO crashes. Wide pavement is consistent in significance for total, injury and PDO crashes in ZIP models while wide pavement is only significant for total crashes in ZINB model. Shoulder type seems insignificant for both models. An exception is found for one specific shoulder type (shoulder width > 6ft.) which is significant only for total crashes in ZIP model. When the pavement type is bituminous, it is statistically significant in crash reduction for all types of crashes in both models. For the PCC concrete pavements, the significance is not sufficient for fatal and PDO crashes. Pavement with gravel top generally increases the crashes for all types of crashes, especially more significant for injury crashes in both models. The ZIP and ZINB models fail to clearly explain the data of fatal crashes because of excessive amount of zero values; that is why the results for fatal crashes are excluded in the tables.

The Pearson residual values and other statistical output comparison for both models are listed in Table 4.

TABLE 4 Model Comparison

		ZIP Model			**ZINB Model**		
		Total Crashes	Injury Crashes	PDO Crashes	Total Crashes	Injury Crashes	PDO Crashes
Pearson Residuals	Min	-2.832	-2.278	-2.411	-1.114	-1.227	-1.104
	Median	-0.358	-0.338	-0.337	-0.319	-0.363	-0.311
	Max	31.833	23.370	26.111	29.370	26.505	24.689
Iteration (BFGS)		38	40	38	42	64	97
logL		-1.03E+04	-8.32E+03	-1.03E+04	-1.23E+04	-7.95E+03	-9.45E+03
DOF		34	34	34	35	35	35
Theta					1.38	1.82	1.35

From the investigation of the model output, it can be said that wider shoulder and pavement were found to reduce the likelihood of crash occurrence in rural two-lane highways. Gravel-top pavements are inclined to crash proneness according to both of the models. Lower values of AADT is significant for reducing the likelihood of crashes.

Model Validation

Vuong has introduced a test that is a well-suited approach to compare zero-inflated models to the conventional models for counts data [11]. It is based on a comparison of the predicted probabilities of two models that do not nest (e.g., ZIP versus ordinary Poisson, or ZINB versus ordinary negative binomial). A large, positive test statistic provides evidence of the superiority of model 1 over model 2, while a large, negative test statistic is evidence of the superiority of model 2 over model 1. Under the null that the models are indistinguishable, the test statistic is asymptotically distributed standard normal. The Vuong statistics is

$$V = \frac{\bar{\tau}\sqrt{N}}{S} \tag{5}$$

where,

$\bar{\tau} = \ln\left(\frac{pdf_1(.)}{pdf_2(.)}\right)$ [where, τ is the ratio of pdf1(.) is the ZNB/ZIP pdf and pdf2(.) is the pdf of NB/Poisson]

S= Standard deviation

N= Sample size

TABLE 5 Vuong Test Statistic

Severity Types	ZINB versus negative binomial	Vuong Test-Statistic	p-value	ZIP versus Poisson	Vuong Test-Statistic	p-value
Total Crashes	ZINB > NB	2.6410	0.0041	ZIP > Poisson	15.164	0.0000
Injury Crashes	ZINB > NB	2.1639	0.0152	ZIP > Poisson	8.1841	0.0000
PDO Crashes	ZINB > NB	3.4129	0.0003	ZIP > Poisson	12.7071	0.0000

When the test statistic value > 1.96 (the 95% confidence level for the t-test), the ZINB or ZIP model is more significant than traditional negative binomial or Poisson model. From Table 5, we find that the ZIP and ZINB models are showing better performance than conventional Poisson or negative binomial model for total, injury and PDO crashes.

Limitations

The intent of this research is to examine ZIP and ZINB models that could potentially explain crash frequencies on rural two-lane roadway segments for different severity types. Vuong test results indicate that ZIP and ZINB models give better prediction than conventional Poisson and negative binomial models. Lord explained that although zero-inflated models offer improved statistical fit to crash data in many cases, it is argued that the inherent assumption of a dual state process underlying the development of these models is inconsistent with crash data [8]. He also explained in his paper that if the only goal consists of finding the best statistical fit then the zero-inflated models may be appropriate, since they offer improved statistical fit compared to Poisson or negative binomial models. This research aims to utilize ZIP and ZINB models to investigate the significance of the recoded categorical values of the geometric factors for traffic crashes of different severities which has not been done extensively in crash data analysis before. The comparison of the model output clearly distinguishes the influence of the key factors on different severity types. The recoding of the continuous variables to the categorical values was performed for easier interpretation.

One future scope of this research is to introduce non-parametric statistical methods to the extended dataset to compare the statistical significance.

CONCLUSIONS

In this paper, ZIP and ZINB models were estimated to identify the impact of key geometric factors contributing to crashes of different severities. Specifically, our aim was to determine whether the factors contributing to one particular severity were different for other types of severities. The models were developed for all types of crash severity counts occurring on the rural two-lane highway segments of Louisiana for eight years (2004–2011). Based on the test statistic, ZIP and ZINB models provided a better fit than conventional Poisson or negative binomial model for total, injury and PDO crashes. From the modeling results, several categories of segment length, pavement type and width, traffic volume and shoulder type were found to be significant in predicting total, injury and PDO crashes. The findings also confirmed that wider shoulder and pavement were found to be associated with the reduction of the likelihood of crash occurrence on rural two-lane highways. Although only 0.07% of the pavements are gravel-top pavements, but these pavements were found to be associated with crash proneness according to both of the models. Lower values of AADT was significantly associated in reducing the likelihood of crashes. The findings of this study are suggestive but limited as these models were based only on rural two-lane highways in Louisiana.

REFERENCES

[1] Louisiana Crash Data Report. http://datareports.lsu.edu/CrashReportIndex.aspx Accessed July, 2014.
[2] Research and Innovative Technology Administration (RITA).http://www.rita.dot.gov/ Accessed July, 2014.
[3] Miaou, S. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention, Vol. 26 (4), pp. 471–482, 1994.
[4] Shankar, V., Milton, J., and Mannering, F. Modeling accident frequency as zero-altered probability processes: an empirical inquiry. Accident Analysis and Prevention, Vol. 29 (6), pp. 829–837, 1997.
[5] Carson, J., and Mannering, F. The effect of ice warning signs on accident frequencies and severities. Accident Analysis and Prevention, Vol. 33 (1), pp. 99–109, 2001.
[6] Lee, J., and Mannering, F. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accident Analysis and Prevention, Vol. 34 (2), pp. 149–161, 2002.
[7] Shankar, V., Ulfarsson, G., Pendyala, R., and Nebergall, M. Modeling crashes involving pedestrians and motorized traffic. Safety Science 41 (7), pp. 627–640, 2003.
[8] Lord, D., Washington, S. and Ivan, J. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention, Vol. 37, pp. 35–46, 2005.
[9] Beckett, S., Jee, J., Ncube, T., Pompilus, S., Washington, Q., Singh, A., and Pal, N.

Zero-Inflated Poisson (ZIP) Distribution: Parameter Estimation and Applications to Model Data from Natural Calamities' (with). Involve – A Journal of Mathematics. Volume 7(4), 2014.

[10] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[11] Winkelmann, R. Econometric Analysis of Count Data. Springer; 5th edition. April 8, 2008.