MODELING CRASH AND FATALITY COUNTS ALONG MAINLANES AND FRONTAGE ROADS ACROSS TEXAS: THE ROLES OF DESIGN, THE BUILT ENVIRONMENT, AND WEATHER

Jian Xu
Graduate Student Researcher
Southeast University, China
jianxu@utexas.edu

Kara M. Kockelman
(Corresponding author)
Professor and William J. Murray Jr. Fellow
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
kkockelm@mail.utexas.edu
Phone: 512-471-0210

Yiyi Wang
Assistant Professor
Civil Engineering Department
Montana State University
yiyi.wang@ce.montana.edu

Presented at the 93rd Annual Meeting of the Transportation Research Board.
Word count: 5277 + 4 Figures + 3 Tables = 7,027 word-equivalents

ABSTRACT

Traffic safety is a top priority for most transportation agencies and many governments. In this study, the geometric details of Texas' extensive highway network were mapped to a variety of traffic, demographic, and built environment variables, including land use, truck volumes, traffic intensity, local population and jobs density, rainfall, income, and education levels. A zero-inflated negative binomial (ZINB) model was used to allow for excess zeros and over-dispersion, and was statistically preferred to the zero-inflated Poisson (ZIP) and negative binomial (NB) models, thanks to lower prediction errors and more robust parameter inference. Estimation results show how crash frequencies and fatality rates clearly rise with local jobs and population densities (as proxies for land use intensities), as well as rainfall. Interestingly, speed limits and distances to the nearest hospitals have negative associations with segment-based crash rates (everything else constant) but, as expected, (slightly) positive associations with fatality rates (presumably due to more severe collision impacts at higher speeds and time lost in transporting crash victims).

Keywords: traffic safety; crash count modeling; land use; demographics, zero inflated negative binomial models; Poisson models

1. INTRODUCTION

Despite increases in vehicle ownership and often vehicle-miles traveled (VMT), fatal and other crash counts have fallen significantly in the United States over the past decade. Likely reasons behind this trend are improvements in roadway and vehicle designs, along with elevated safety awareness by roadway users. Although the safety trend in Texas mirrors that at the national level over the past few years, fatality rates still hover at a relatively high share of crashes (0.70%), with 3,399 lives lost on Texas roadways in 2012 (TxDOT 2013). To better understand the genesis of such crashes and fatalities, this work uses econometric models and new covariates to examine a wide variety of possible factors and provide useful information to network designers and policy makers.

Many statistical methods already exist to rigorously and reliably forecast crash counts and severities as a function of multiple covariates (see, e.g., Kulmala, 1995; Poch and Mannering 1996; Abdel-Aty and Radwan, 2000; Ma and Kockelman, 2006, Ma et al., 2008; Quddus et al., 2010; Wang et al., 2011). Prior studies tend to rely on homogenous highway segments (e.g., Aguero-Valverde and Jovanis 2008) or zone-level crash frequencies (using states, Census tracts, or other topologies -- e.g., Wang et al. 2011), while controlling for variables like vehicle-miles traveled (VMT), curve lengths, degree of curvature, speed limits, and truck shares. Typically neglected covariates include local land use and demographic conditions, climate, and hospital access; so these are controlled for here.

This work investigates the linkage between segment-based crash counts and a variety of relatively unusual factors, using zero-inflated negative binomial regression model (ZINB), as compared to negative binomial (NB) and zero-inflated Poisson (ZIP) models, with crash counts along mainlanes and frontage roads in Texas. The following sections present details of related literature, data sets and methods used, results obtained, and several conclusions.

2. LITERATURE REVIEW

Various styles of observational units have been used in the crash-count modeling literature, including counties (Miaou et al., 2003; Aguero-Valverde and Jovanis 2006), regions (Washington et al., 1999), districts (Jones et al., 2008), English wards (Noland and Quddus, 2004), census tracts (Wang and Kockelman, 2013), and roadway segments (Ma and Kockelman, 2006, Ma et al., 2008). Each topology has advantages and disadvantages, and different aggregations of data can lead to somewhat different results. Here, Texas' extensive highway system is divided into hundreds of thousands of homogenous segments, with attributes of curvature, surface width, speed limit, lane count, and so forth constant.

Commonly used control variables are also reflected here. These include traffic characteristics (e.g., VMT, annual average daily traffic [AADT], and speed limit) and roadway design features (e.g., surface width and horizontal and vertical alignment details [as used in Poch and Mannering 1996, Abdel-Aty and Radwan 2000, Ma and Kockelman 2006, Wang et al. 2011]). Other factors, including average rainfall and local land use attributes are also used, to test their predictive powers. Brijs et al. (2008) studied the effects of weather conditions on daily crashes for three

large cities in Netherlands in 2001, and the results show that rainfall, temperature and city-specific estimates were highly significant with respect to the number of crashes. Kim and Yamashita (2002) modeled crashes as a function of several land use variables, and they found that the highest crash frequencies occurred near commercial or business properties. In addition, local demographic features for each segment's Census tract were used to predict recent crash counts (as done, to some extent, in Graham and Glaister, 2003; Aguero-Valverde and Jovanis, 2006; Kim et al., 2006; and Quddus, 2008).

A basic specification for crash count prediction is the Poisson model (see, e.g., Jovanis and Chang, 1986; Miaou, 1994), which does not allow for latent heterogeneity across seemingly identical observational units (resulting in an equi-dispersion assumption, where the expected crash rate equals its variance for each unit), unlike the similarly tractable NB approach (Abdel-Aty and Radwan, 2000; Lord, 2000). Other examples include Li et al.'s (2007) Bayesian approach to rank roadway segments by crash risk, Ma and Kockelman's (2008) multivariate Poisson-lognormal model (MVPLN), allowing for simultaneity across counts (by crash severity), and Park and Lord's (2009) finite mixture model (to capture heterogeneity and overdispersion). Wang et al. (2011) used Bayesian estimation for their spatially mixed logit model, combining both crash frequency and severity, and Qin and Reyes (2011) proposed a quantile regression (QR) method for depicting the relationship between a family of conditional quantiles and site covariates.

Spatial econometric modeling techniques are relatively complex but gaining traction; they enable both spatial heterogeneity and spatial dependence. Standard spatial methods for continuous responses (such as the spatial autoregressive model [SAR] and the spatial error model [SEM]) have been employed to analyze zone-level crash counts (e.g., counties and cities, [Baller et al., 2001]), and Bayesian estimation techniques for hierarchical methods have proven a valuable alternative for more discrete/integer counts, at smaller geographical levels (e.g., along short road segments and census tracts [e.g., Wang and Kockelman 2013, Li et al. 2007]). Miaou et al. (2003) used a conditional autoregressive (CAR) structure to illustrate the existence of spatial autocorrelation among adjacent roadway segments along Texas' rural two-lane highways. And Wang et al. (2009) used a series of Poisson-based CAR models to examine the role of traffic congestion on British expressways' crash counts. Due to matrix inversion challenges with larger data sets, the likelihood functions of spatial models are often intractable or their estimation sequence non-convergent (Wang and Kockelman 2013), so evaluation of very large count-data sets are not yet feasible with spatial models

Many segments and zones demonstrate zero crash counts in any given year, especially for severe (less common) crash counts. Zero-inflated (ZI) models (as used by Miaou, 1994; Zamani and Ismail, 2010; Shankar et al. 2003; Qin et al. 2004; Lord and Geedipally, 2011; and Yan et al. 2012) help reflect settings where some locations may never experience crashes. Lord et al. (2005) have provided guidance on how to appropriately model relationships between road safety and traffic exposure, emphasizing a comparison of ZIP and the ZINB models. More recently, Lord and Geedipally (2011) are recommending the negative binomial–Lindley model (NB-L), where "Lindley" refers to a type of two-parameter distribution whose long-term mean never equals zero, but can handle a preponderance of zeros (per site or observational unit), while still maintaining

valuable attributes of the traditional NB model. In this study, a ZINB model is used, since the NB-L model requires special programming, and is not yet available in statistical software.

3. DATA SETS USED

The Texas Department of Transportation (TxDOT) manages and maintains approximately 80,000 centerline-miles of highways, including roughly 7,000 edge-miles of one-way frontage roads alongside the state's freeway corridors. In this study, year 2010 highway, land use, and reported crash data were used to examine the associations between crash counts and various contributing factors along Texas' highways and frontage roads.

The Texas DOT's Crash Record Information System (CRIS), Road-Highway Inventory Network (RHiNo), and horizontal curve (GEO-HINI) databases were spatially matched/mapped to one another along with local rainfall, land use and demographic attributes. Highways and frontage roads were split into homogenous segments based on consistency/constancy in geometric characteristics (e.g., curvature, surface width, speed limit, lanes number, and AADT) following the merge of the RHiNo and GEO-HINI data sets. No curvature information for frontage roads is provided, because the GEO-HINI data set only pertains to Texas' mainlanes. CRIS crash details were matched to segments using control section-milepoint information. Other variables were mapped to the segments using the ArcGIS toolbox. As a result, 277,510 (highway) mainlane segments (for a total of 72,994 centerline miles) and 15,781 frontage road segments (totaling 7,041 edge miles) were identified for analysis, as shown in Figure 1.

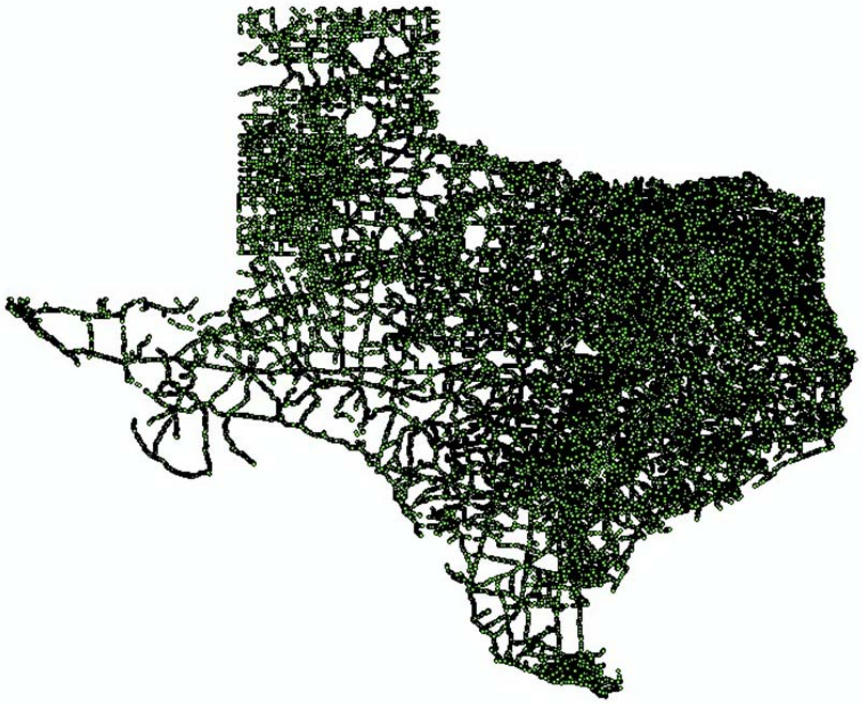

Figure 1. Locations of Homogenous Highway and Frontage Road Segments for Analysis of Texas Crashes

More than ninety percent of these 350,054 segments have lengths of less than 1 mile, as depicted in Figure 2's histograms. Average segment lengths are 0.263 mi on mainlanes and 0.446 mi on frontage roads.

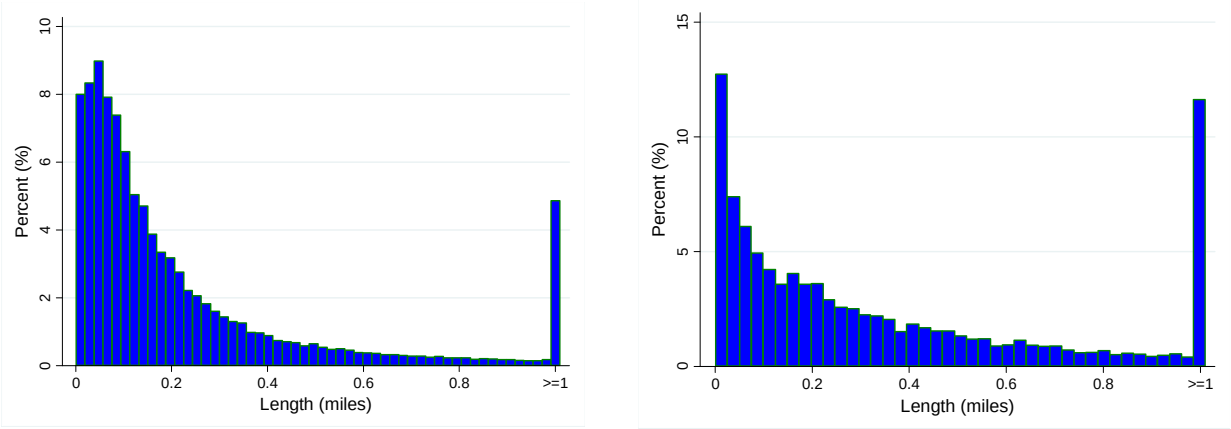

Figure 2. Histograms for Lengths of Mainlane (left) and Frontage Road (right) Segments

The 2010 CRIS data sets contain 243,388 crashes on the state-/TxDOT-maintained network (plus another 227,794 crashes that occurred on other public roadways across Texas). Among these 243,338 crashes, 209,053 contained control-section and milepoint details that place them along mainlanes and frontage roads, allowing them to be geocoded onto the set of 350,504 homogenous highway segments. The remaining 34,335 crashes (14.10 percent of the on-system set) contain no location information or occurred at on- and off-ramps or connectors and detours and so were removed. For example, 6,224 of the mainlane crashes and 4,291 of the frontage road crashes were missing appropriate control section-milepost address information. Of course, if there is any bias in the locations of removed crashes and/or missing crash records (e.g., police may have a harder time defining crash locations in rural areas than in urban areas, and property-damage-only (PDO) and less injurious crashes often go unreported (Aguero-Valverde and Jovanis 2008), there will be some bias in results. These are regular issues in crash-plus-network databases, however, so such biased are likely to exist in many, perhaps most, analyses already published.

In this study, the response or dependent variables are the numbers of crashes by type/severity level, on each homogenous highway segment. The five severity types are No Injury/PDO, Possible Injury, Non-Incapacitating, Incapacitating Injury, and Fatal crash counts. As noted earlier, covariates define highway design, traffic attributes, land use information, climate and access factor, local demographics, with summary statistics provided in Table 1.

Highway Design Variables: Highway design decisions have important impacts on a link's use (flow volumes and truck shares), its speeds, crash counts, and crash severities. Such relationships have been widely investigated in the past (see, e.g., Poch and Mannering 1996; Lord, 2000; Ma and Kockelman, 2006, Ma et al., 2008; Wang et al., 2011). Here, a number of design variables were used (as control variables directly or to construct other covariates, like VMT); these include Average Shoulder Width (using both inside and outside shoulders), Number of Lanes, Median Width (not including shoulders or other drivable area), Curve Length, Degree of Curve (i.e.,

angle subtended by 100 feet of curve arc), and an Indicator for (the presence of) Curvature. Table 1 lists these attributes and their summary statistics.

Traffic Attributes: Traffic characteristics also play a critical role in crash outcomes and prediction. Average Daily Traffic (ADT) estimates (from nearby vehicle-count samples) describe traffic intensity and congestion (using an ADT- or volume-to-capacity variable). VMT is a key crash exposure term (since crash counts closely scale with VMT, everything else constant), and is simply the product of ADT, segment length, and 365 (days per year). In addition, Speed Limit and Percentage of Single and Combo Truck ADT were also included in this study because large trucks are involved in a disproportionately small fraction of the total crashes but a large fraction of fatalities (Vadlamani et al. 2011).

Land Use Information: Information on actual land use designations (e.g., parcels in residential vs. industrial vs. commercial and other uses) could not be obtained across the entire state, but Census data on population demographics and jobs counts per census tract (population density and jobs density) are readily available, along with other variables, like indicators for Rural, Small Urban, Large Urban and Urbanized settings. The data of year 2010 population demographics and jobs counts were obtained from US Census Bureau. Centroids of highway segments were matched to their closest Census tract (measured using Euclidean distances to tract centroids) using ArcGIS's spatial join routine. All count data were normalized by census tract areas (in square miles). The spatial distribution of census tracts in Texas is illustrated in Figure 3.

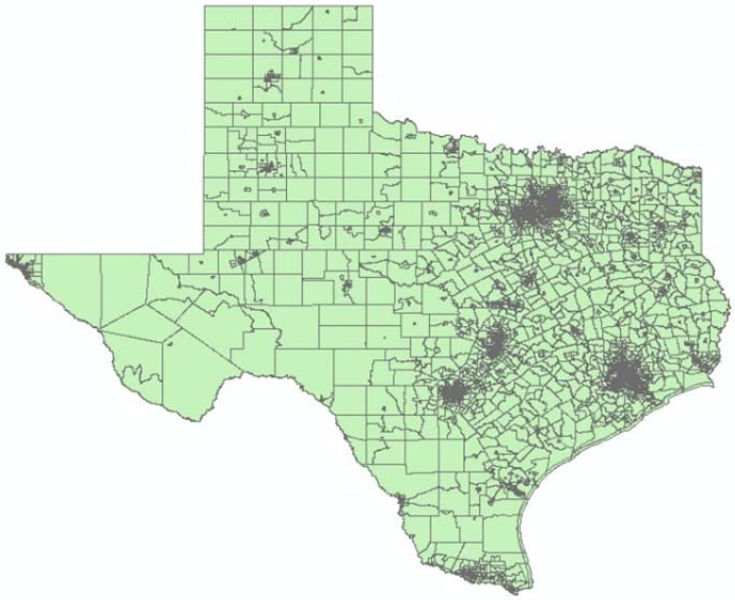

Figure 3. Census Tracts across Texas

Climate and Access Factors: Average annual rainfall values (over the 1961-1990 periods, as shown in Figure 4) were obtained from Texas Natural Resources Information System (TNRIS). It would be best to match year-2010 crash totals to rainfall data for the year 2010, but such information was unavailable. Nevertheless, this too-commonly ignored variable proved valuable in crash prediction. Euclidean distances from each segment's centroid to three interesting sites were also developed, using ArcGIS' spatial join function; these sites are hospitals (hypothesized

here to be important in reducing crash fatalities), schools (an indicator of activity and possibly lowered speeds or more driver caution), and metropolitan and micropolitan statistical areas (MMSAs, which describe access to a developed region or city). Hospital locations were obtained from the Texas Hospital Association, and school locations (for elementary school) were retrieved from Texas State Data Center. Information on MMSA centroid locations comes from the U.S. Census Bureau.

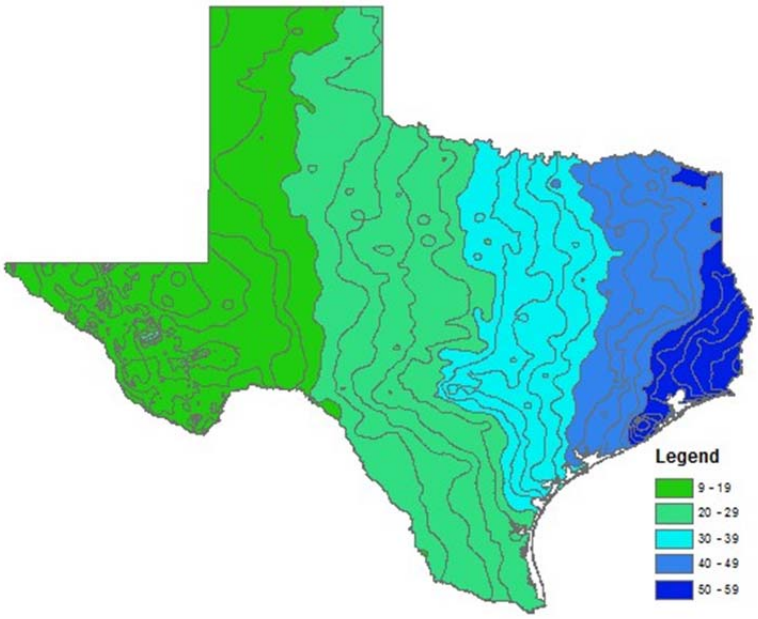

Figure 4. Average Annual Rainfall Values across Texas, in inches (Years 1961-1990)

Demographic Variables: Relatively recently, demographic characteristics have come into use for analysis of crash counts (e.g., Graham and Glaister [2003]; Kim et al. [2006]; Quddus et al. [2008]). This work relies on U.S. 2010 Census tract data (as measured using Euclidean distances from segment midpoint to the nearest tract centroid, for density variables of persons of low and high education (that is, less than a high school diploma versus beyond a bachelor's degree), and commuting workers. The average age of people and average income per capita in census tract were also obtained from US Census Bureau. Such variables provide more information on the types of users in the vicinity of the highway, who may be using it regularly, in vehicles or possibly on foot. They also give one a greater sense of the land use setting, versus simply population and jobs density variables.

1 Table 1. Summary Statistics of Variables for Mainlane and Frontage Road Segments across Texas

Variable Name	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
	Mainlanes				**Frontage Roads**			
Dependent Variables								
Crash Count in 2010	0.6454	3.544	0	387	1.074	4.307	0	80
No Injury Crashes (PDO)	0.4126	2.364	0	242	0.7165	3.012	0	59
Possible Injury	0.125	0.8766	0	116	0.2145	1.002	0	21
Non-Incapacitating	0.0798	0.4670	0	30	0.1122	0.5677	0	22
Incapacitating Injury	0.0212	0.1692	0	10	0.0251	0.1827	0	4
Fatal Crashes	0.0067	0.0844	0	3	0.0051	0.0736	0	3
Totals (across State highway system)[1]	181,305				17,233			
Covariates								
Exposure Variable								
VMT (vehicle-miles)	1604.1	7292.9	0	559789	1674.6	4720.6	0.01	103484
Highway Design								
Average Shoulder Width (feet)	3.964	3.448	0	30	1.184	2.401	0	30
Number of Lanes	2.536	1.130	1	13	2.126	0.390	1	6
Median Width (feet)	6.482	22.15	0	722	/	/	/	/
Indicator for Curvature (1 = yes, 0 = otherwise)	0.320	0.466	0	1	/	/	/	/
Curve Length (miles)	0.0396	0.083	0	3.3	/	/	/	/
Degree of Curve (degrees per 100-ft of arc)	1.357	3.903	0	90.67	/	/	/	/
Traffic Attributes								
ADT per Lane (vehs/day)	1858.7	3353.9	0	48,833	2049.9	2975.7	5	23,530
% Single Truck ADT (%)	8.095	5.672	0	6.52	1.7664	0.2438	0	1.8
% Combo Truck ADT (%)	7.802	8.160	0	9.37	1.3606	0.2315	0	1.4
Speed Limit (mph)	57.44	9.752	5	80	46.476	6.4296	15	75
Land Use Information								
Population Density (persons/sq.mi.)	342.3	1044	4.324	15,035	389.4	1182	9.421	15,035
Jobs Density (employees/sq.mi.)	107.9	743.9	0.821	8163	210.3	854.6	5.482	8163
Rural (1 = yes, 0 = otherwise)[2]	0.793	0.4051	0	1	0.3054	0.4606	0	1
Small Urban (1 = yes, 0 = otherwise)[3]	0.0741	0.2619	0	1	0.0719	0.2584	0	1
Large Urban (1 = yes, 0 = otherwise)[4]	0.0450	0.2073	0	1	0.1394	0.3464	0	1
Urbanized (1 = yes, 0 = otherwise)[5]	0.0878	0.2831	0	1	0.4833	0.4997	0	1
Climate & Access Factors								

Rainfall (average inches per year)	34.43	10.52	9	59	36.24	12.49	9	59
Distance to the Nearest Hospital (miles)	12.30	10.38	0.0067	91.72	6.420	8.747	0.0201	90.89
Distance to the Nearest Elementary School (miles)	6.286	7.462	0.2472	98.42	4.198	8.325	0.4261	56.42
Distances to the Nearest MMSA (miles)[6]	24.84	15.48	0.0175	142.7	17.92	12.00	0.1139	84.38
Demographic Variables								
Average Income per Capita (dollar)	24870	54831	2107	168156	25043	43913	2735	168156
Average Age (year)	33.3	45.6	16.9	55.4	33.6	39.8	17.1	55.2
Pop. Den. of below High School (persons/sq.mi.)	13.81	62.57	0	3601	63.78	131.5	0	2723
Pop. Den. of above College (persons/sq.mi.)	33.38	107.4	0	3672	154.7	234.4	0.0335	3410
Observations	277,510				15,781			

Notes: [1]Total refers to the number of crashes that were matched to segments. [2,3,4,5]Rural, Small Urban, Large Urban, and Urbanized denote places with less than 5000 people, 5000 to 49,9993 people, 50,000 to 199,999 people, and 200,000+ people, respectively. [6]The distances refer to each segment's centroid to the nearest centroid of MMSA.

4. MODEL SPECIFICATION

The negative binomial (NB) regression specification, depicted in Equation 1 (from Miaou [1994]) is valuable for characterizing count-data situations, like traffic crashes (Lord, 2000; Noland and Quddus, 2004; Wang et al, 2009). The crash rate is modeled as a function of the covariates:

$$\lambda_i = VMT_i^{\alpha} \exp\left(\beta_0 + \sum_k x_{ik}\beta_k + \varepsilon_i\right) \quad (1)$$

where *VMT* denotes vehicle-miles traveled along the i^{th} segment (as a measure of crash exposure); the parameter α allows for a potentially non-linear/non-proportional association of crash counts with VMT; β_0 is the intercept term (or constant); β_k denotes coefficient for the k^{th} covariate; x_{ik} indicates the k^{th} covariate for the i^{th} segment; and ε_i is a random error that has a gamma distribution, such that $\varepsilon_i \sim gamma(\gamma,\gamma)$. The probability density function for the NB distribution can be expressed as

$$p(Y_i = y_i) = \frac{\Gamma\left(y_i + \frac{1}{\rho}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\rho}\right)}\left(\frac{\rho\mu_i}{1+\rho\mu_i}\right)^{y_i}\left(\frac{1}{1+\rho\mu_i}\right)^{1/\rho} \quad (2)$$

where Y_i represents crash counts along homogenous highway segment i during a given period (e.g., the year 2010), and $i = 1,2,\ldots,n$, with n denoting the total number of highway segments analyzed. y_i denotes the realization of the random variable Y_i, with mean $E(Y_i) = \mu_i = VMT_i^{\alpha} \exp\left(\beta_0 + \sum_k x_{ik}\beta_k + \varepsilon_i\right)$ and variance $Var(Y_i) = \mu_i + \rho\mu_i^2$. ρ denotes the overdispersion parameter (and the NB collapses to a Poisson specification when $\rho = 0$).

Since highway crashes are generally rare events, analysts frequently have many highway sections with zero reported crashes during the period of interest. Here, 82.0% of the mainlane segments and 79.7% of the frontage road segments had zero (reported/logged and spatially matched) crashes in 2010. But these zeros may simply come from low annual crash rates on all or most segments. It does not mean that overdispersion or zero inflation is present. Model results for the ZINB will illuminate the statistical significance of the parameters defining those features of this relatively flexible model specification. (A Vuong [1989] test also can be used to examine zero-inflation, and was checked here, using STATA software. All results support the notion of zero inflation across many sites.)

This work combines ZI and NB features by turning to the ZINB model specification, as shown in Equation 3 (Jansakul and Hinde, 2008):

$$p(Y_i = y_i) = \begin{cases} \theta_i + (1-\theta_i)(1+\rho\lambda_i)^{-\rho^{-1}}, & \mathrm{y}_i = 0 \\ (1-\theta_i)\dfrac{\rho^{y_i}\lambda_i^{y_i}\Gamma(y_i+\rho^{-1})}{y_i!(1+\rho\lambda_i)^{y_i+\rho^{-1}}\Gamma(\rho^{-1})}, & \mathrm{y}_i = 1,2,\cdots, \end{cases} \quad (3)$$

where Y_i is the response variable, under a dichotomous data-generating processes. The zero counts are captured by the first process with probability θ_i and the second (negative binomial) process with probability $1-\theta_i$, and all non-zero counts are captured by the second process (with probability $1-\theta_i$). In general, the zeros from the first and second processes are called structural zeros and sampling zeros, respectively (Jansakul and Hinde, 2008). The mean and the variance of the ZINB model are $E(Y_i) = (1-\theta_i)\lambda_i = \mu_i$ and $Var(Y_i) = \lambda_i(1-\theta_i)(1+\theta_i\lambda_i+\rho\lambda_i)$. The Poisson, the NB, the ZIP, and the ZINB models are related to each other. For example, if $\rho = 0$, the ZINB model will be reduced to the ZIP model; if $\rho = 0$, $\theta_i = 0$, the ZINB model will collapse to the standard Poisson model.

To involve covariates in both portions of the ZINB model (i.e., the crash rate equation and the logistic equation), the following functions are used here:

$$\ln(\lambda_i) = \alpha^{NB}\ln(VMT_i) + \beta_0^{NB} + \sum_k x_{ik}\beta_k^{NB} + \varepsilon_i^{NB} \quad (4)$$

and

$$\ln\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha^{logit}\ln(VMT_i) + \beta_0^{logit} + \sum_k x_{ik}\beta_k^{logit} + \varepsilon_i^{logit} \quad (5)$$

here α^{NB}, β_0^{NB}, β_k^{NB}, and ε_i^{NB} are same as before, corresponding to the NB component of the ZINB model, while α^{logit}, β_0^{logit}, β_k^{logit}, and ε_i^{logit} are for the logistic component of the ZINB specification.

5. MODEL RESULTS

This section discusses results of model estimation for total crash count (and fatal crash count) on mainlanes and frontage roads, with parameter estimates shown in Table 2. There are fewer (very) statistically significant variables[1] for count prediction along frontage roads than for mainlanes, in large part due, no doubt, to the smaller sample size (of homogeneous frontage road segments). All test statistics suggest that the ZINB model is statically preferred (with p values of 0.000) to the ZIP and NB model specifications, for both roadway types (and both crash count model types).

In order to appreciate the practical significance of various covariates, on both total and fatal crash rates, the expected percentage changes in both types of crash rates, for both types of

[1] All available, starting covariates were maintained in the final models and thus these tabled results, though a few are not statistically significant (and so could be removed, via a process of stepwise deletion or something similar).

facilities (mainlanes and frontage roads) were computed for a one standard-deviation (SD) change in each covariate, as shown in Table 3.

The association between crash exposure (VMT) and crash rates is estimated to be non-linear (with estimated exponents of $\alpha = 0.6759$ for mainlanes and $\alpha = 0.3273$ for frontage roads), with crash rates effectively falling as VMT rises. It may be that higher VMT levels often correspond to more congested traffic conditions, with reduced operating speeds and lower speed differentials, somewhat dampening the likelihood of collision. Interpretations of other parameter estimates are provided below.

Highway Design: As shown in Table 2 and 3, the Number of Lanes and Curve Length are estimated to increase the average crash rate of mainlanes, while the other highway design variables exhibit negative associations with average crash rates. This is expected since the likelihood of collision is in general higher along curves than that along straight segments, whereas the crash rates may be lower along curves (as drivers tend to be more cautious), a result consistent with the findings of Wang et al (2009).

It is interesting to find that Average Shoulder Width is negatively associated with crash rates along mainlanes, but positively associated with frontage road rates. Such differences may be due to the different effects of access control used on such facilities.

Traffic Attributes: Not surprisingly, ADT per Lane, % Single and Combo Truck ADT, and Speed Limit were found to be statistically significant. ADT per Lane is estimated to have positive effects on crash rates along mainlanes and frontage roads, consistent with prior studies (e.g., Quddus et al., 2010) and presumably due to greater congestion, which puts vehicles at tighter headways and higher probability of crashing. Interestingly, higher Speed Limits in both models is associated with lower *total* crash rates, presumably due to better roadway designs and less crash-prone locations being assigned higher speed limits; however, this can come with more severe crashes, as found in Ma et al. (2008). As shown in Table 3, the Speed Limit variable is estimated to have a very slight positive effect on fatality rates here; while higher Speed Limit roadways enjoy better design and safer conditions overall, crashes that do occur tend to be more injurious and deadly (see, e.g., Kockelman et al. [2006]). Finally, Single and Combo Truck percentages are associated with lower crash rates, presumably due to greater regulation of and use of professional drivers in those vehicles.

1 Table 2. Estimation Results of ZINB for Total Crash Count on Mainlanes and Frontage Roads

Variable	Coef.	P>|z|	Coef.	P>|z|	Coef.	P>|z|	Coef.	P>|z|
	Mainlanes				**Frontage Roads**			
	NB Component		Logistic Comp.		NB Component		Logistic Comp.	
Exposure Variable								
Ln(VMT)	0.6759	0.000	-0.6964	0.000	0.3273	0.000	-0.6729	0.000
Highway Design								
Average Shoulder Width	-0.0141	0.000	0.0261	0.000	0.0029	0.849	0.0591	0.008
Number of Lanes	0.0148	0.005	0.0729	0.000	0.2104	0.002	0.2404	0.017
Median Width	-0.0023	0.000	0.0039	0.000	/	/	/	/
Indicator for Curvature	-0.0939	0.000	0.8158	0.000	/	/	/	/
Curve Length	0.0443	0.009	-0.9454	0.031	/	/	/	/
Degree of Curve	-0.0045	0.065	-0.1903	0.000	/	/	/	/
Traffic Attributes								
ADT per Lane	1e-05	0.000	7e-06	0.812	2e-06	0.832	0.0001	0.000
% Single Truck ADT	-0.0139	0.000	-0.0325	0.000	/	/	/	/
% Combo Truck ADT	-0.0028	0.001	-0.0109	0.003	/	/	/	/
Speed Limit	-0.0231	0.000	0.0023	0.379	-0.0225	0.000	0.0044	0.615
Land Use Information								
Population Density	0.0842	0.000	0.3214	0.000	0.2251	0.000	0.0934	0.034
Jobs Density	0.1079	0.034	0.4086	0.092	0.1972	0.000	0.1932	0.655
Rural	-0.4625	0.000	-0.7265	0.000	-0.4502	0.000	-0.1207	0.502
Small Urban	-0.0032	0.890	-0.2541	0.000	-0.1493	0.221	-0.2099	0.304
Large Urban	0.1127	0.000	-0.3006	0.001	-0.1522	0.068	-0.1351	0.334
Urbanized	Base/reference case.							
Climate & Access Factors								
Rainfall	0.0012	0.037	-0.0205	0.000	0.0056	0.000	0.0047	0.000
Distance to the Nearest Hospital	-0.0079	0.000	-0.0022	0.551	-0.0691	0.000	-0.4394	0.010
Distance to the Nearest Elem. School	-0.0255	0.000	-0.0388	0.000	0.0004	0.986	-0.0081	0.817
Distance to the Nearest MMSA	-0.0047	0.000	0.0039	0.043	-0.0029	0.432	-0.0078	0.180
Demographics								
Average Income per Capita	-0.0013	0.000	-0.0009	0.000	-0.0078	0.014	-0.0193	0.057
Average Age	-0.0045	0.000	-0.0234	0.046	-0.0032	0.000	-0.0091	0.013
Pop. Den. of below High School	0.0002	0.000	-0.0045	0.000	0.0005	0.006	-0.0003	0.284
Pop. Den. of above College	0.0009	0.000	0.0004	0.001	0.0003	0.026	-7e-05	0.707
Constant	-3.077	0.000	4.1071	0.000	-1.630	0.000	3.192	0.000
ρ	0.3623		P>|z|: 0.000		0.4844		P>|z|: 0.000	
Vuong test P>z	0.000				0.000			
ZIP LR test P>=chibar2	0.000				0.000			
LR chi2	45906.5				1034.4			
Prob > chi2	0.000				0.000			
Log likelihood	-168087.3				-13591.8			

2 Note: Slashes (/) indicate covariates that are not available in RHiNo and GEO-HINI for the frontage road segments.

Table 3. Expected Percentage Changes in Total and Fatal Crash Rates Corresponding to One Standard Deviation Changes in Variables

Variables	**Mainlanes**		**Frontage Roads**	
	Total (%)	Fatal (%)	Total (%)	Fatal (%)
Highway Design				
Average Shoulder Width	-0.5588	-0.1831	0.0617	-0.0081
Number of Lanes	0.192	0.0003	0.1468	0.0032
Median Width	-0.5856	-0.0938	/	/
Indicator for Curvature	-1.079	-0.0003	/	/
Curve Length	0.042	0.0028	/	/
Degree of Curve	-0.2019	0.0712	/	/
Traffic Attributes				
ADT per Lane	0.3855	0.0017	0.0414	0.0004
% Single Truck ADT	-0.9062	-0.0513	/	/
% Combo Truck ADT	-0.2626	-0.0078	/	/
Speed Limit	**-2.589**	0.0145	-0.1355	0.0047
Land Use Information				
Population Density	**12.48**	1.285	**23.29**	**3.761**
Jobs Density	**39.23**	**2.866**	**48.77**	0.4157
Rural	**-5.316**	-0.4153	-0.2780	-0.2946
Small Urban	-0.0368	-0.0049	-0.0922	-0.0007
Large Urban	1.295	0.3883	-0.0939	-0.0835
Climate & Access Factors				
Rainfall	0.1452	0.0284	0.3155	0.0194
Distances to the Nearest Hospital	-0.9426	0.4156	-0.4429	2.678
Distances to the Nearest Elem. School	-0.7721	0.0013	0.3218	0.0355
Distances to the Nearest MMSA	-0.8363	-0.0004	-0.0277	-0.0001
Demographics				
Average Income per Capita	**-3.482**	-0.1354	**-5.318**	-0.0042
Average Age	**-2.641**	**-8.487**	-0.903	**-3.831**
Pop. Density of below High School	0.144	0.0067	0.0193	0.0012
Pop. Density of above College	1.111	0.0039	0.0199	0.0007

Note: Bolded percentages are to indicate the more practically significant variables' impacts on corresponding crash counts.

Land Use Information: Here, demographic variables are employed as proxies for land use intensity. As shown in Tables 2 and 3, Population Density and Jobs Density are very statistically and practically significant in both models, unlike in Noland and Quddus' (2004) work, which could establish no linkage between land use intensity and crash rates. Higher densities come with higher crash rates, and the Jobs Density variable offers the most practically significant results, with the largest crash-rate percentage change (of +39.23%) in Table 3; this is in part due to the very large standard deviation-to-mean ratio for this covariate (as shown in Table 1), so a one-SD change is a substantial shift. More dense locations are generally more complex to navigate, with more activities and land uses alongside the traveled corridor, with more frequent driveways, interchanges, ramps and intersections, for example.

Indicator variables also entered the model to measure the effects of urbanization, with Rural and Small Urban settings offering negative effects (on crash rates) and the Large Urban setting indicator having a positive impact; such results imply that higher levels of urbanization come with greater crash rates (in large part due, no doubt, to a more complex operating environment, with more interchanges, driveways or intersections per mile, for example – along with congestion impacts [already controlled for here via the AADT-per-lane variable, for example]). As shown in Table 3, mean crash rates tend to fall by 5.32% when land use converts to a rural setting and rise by 1.30% when the reference case (an urbanized setting) becomes a large urban setting.

Climate and Access Factors: As revealed in Tables 2 and 3, rainfall is estimated to be positively associated with crash rates, but only slightly (in a practical sense). As discussed previously, distances to hospitals, schools, and MMSAs are rarely considered as covariates in the crash modeling literature, and here they yield negative impacts in the first (ZI or logistic) process, as shown in Table 2. Results also suggest that shorter distances (greater proximity) come with higher crash rates (probably due to these distances proxying for [the inverse of] land use intensity), but, as expected, positive associations exist for fatal-crash rates (presumably due to more severe collision impacts at higher speeds and time lost in transporting crash victims to an emergency room), similar to the impacts of the Speed Limit variable.

Demographics: As shown in Table 2, all of demographic variables are statistically significant, and some are practically significant, according to Table 3's results. Higher-income locations having lower crash rates is consistent with the World Health Organization's (Global status report on road safety 2013) report (where low- and middle-income countries have higher road traffic fatality rates, per VMT, compared to higher-income countries). Higher Average Age is also estimated to come with lower total and fatal crash rates, thanks presumably to more driving experience and more caution by older drivers. Interestingly, both education extremes (below high school and beyond a bachelor's degree) are estimated to have positive effects on crash rates.

CONCLUSIONS

This paper developed a series of ZINB and ZIP crash prediction models, based on the crash inventory for 277,510 mainlane, homogenous-roadway segments and 15,781 frontage-road homogenous segments in Texas, while controlling for highway design, traffic attributes, land use, climate and access factor, and local demographics. Expected percentage changes in total and

fatal crash rates (for both mainlane facilities and frontage road facilities) corresponding to one-standard-deviation changes in all variables were also computed, to examine the practical significance of covariates. While most covariates are estimated to be statistically significant, few are practically significant. The estimation results show that jobs and population density, as proxies for land use intensity, exert positive effects on crash rates. Moreover, higher levels of urbanization are associated with higher crash rates, while age and income have negative effects. Annual rainfall has slightly positive effects on both crash types of crash rate, while distances to the nearest hospitals and schools have negative associations with total-crash rates and positive correlations with fatality rates.

The ZINB model specification adopted here allows for excess zeros and overdispersion in the crash count data sets (with marked improvement in goodness-of-fit, as compared to ZIP and NB models) but does not capture the spatial autocorrelation and/or spatial dependencies that are likely to exist across neighboring segments. While sample size would have to be dramatically reduced from what was used here (due to computing limitations for rigorous spatial regression techniques), such tools should be tested on these data, along with covariates on pavement roughness, topography, sight distances, and other potentially valuable variables for crash prediction and design decisions.

ACKNOWLEDGEMENTS

The authors thank the Texas Department of Transportation (TxDOT) for providing the CRIS and geometric data sets, as well as Dr. David Maidment and Mr. Graham James for offering rainfall data set and RHiNo Geodatabase, respectively. The authors appreciate Dr. Srinivas Reddy Geedipally provided SAS code and administrative support of Ms. Annette Perrone.

REFERENCES

Abdel-Aty, M., Radwan, E., 2000. Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention*, 32 (5), 633–642.

Aguero-Valverde, J., and P. P. Jovanis, 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis & Prevention*, 38 (3), 618-625.

Aguero-Valverde, J. and P.P. Jovanis, 2008. Analysis of road crash frequency with spatial models. *Transportation Research Record*, No. 2061, 55-63.

Baller, R. D., L. Anselin, S. F. Messner, G. Deane, D. F. Hawkins, 2001. Structural covariates of U.S. county homicide rates: incorporating spatial effects. *Criminology*, 39 (3), 561-588.

Brijs, T., Karlis, D., Wets, G., 2008. Studying the effect of weather conditions on daily crash counts using a discrete time-series model. *Accident Analysis & Prevention*, 40(3), 1180-1190.

Global status report on road safety 2013.
http://www.who.int/violence_injury_prevention/road_safety_status/2013/en/index.html

Graham, D.J., Glaister, S., 2003. Spatial variation in road pedestrian casualties: the role of urban scale, density and land-use mix. *Urban Studies,* 40 (8), 1591–1607.

Jansakul, N., and J.P. Hinde, 2008. Score tests for extra-zero models in zero-inflated negative binomial models, *Communications in Statistics - Simulation and Computation*, 38(1), 92-108.

Jones, A.P., R. Haynes, and V. Kennedy et al., 2008. Geographical variations in mortality and morbidity from road traffic accidents in England and Wales. *Health & Place*, 14 (3), 519-535.

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

Kim, K., Yamashita, E.Y., 2002. Motor vehicle crashes and land use: empirical analysis from Hawaii. *Transportation Research Record,* 1784, 73–79.

Kim, K., Brunner, I.M., Yamashita, E.Y., 2006. The influence of land use, population, employment and economic activity on accidents. *Transportation Research Record,* 1953, 56–64.

Kockelman, K., CRA International, 2006. Safety Impacts and Other Implications of Raised Speed Limits on High-Speed Roads. NCHRP 17-23 Final Report, for the Transportation Research Board, Washington, D.C.

Kulmala, R., 1995. Safety at rural three-and four-arm junctions: development and application of accident prediction models. *VTT Publications*. Espoo: Technical Research Center at Finland.

Lee, J., Mannering, F.L., 2002. Impact of roadside features on the frequency and severity of run-off-road accidents: an empirical analysis. *Accident Analysis & Prevention,* 34 (2), 349–361.

Li, L., Zhu, L., Daniel Z. S., 2007. A GIS-based Bayesian approach for analyzing spatial–temporal patterns of intra-city motor vehicle crashes. *Journal of Transport Geography*, 15(4), 274-285.

Lord, D., 2000. The Prediction of Accidents on Digital Networks: Characteristics and Issues Related to The Application of Accident Prediction Models. Ph.D. Dissertation, Department of Civil Engineering, University of Toronto, Toronto.

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

Lord, D., Geedipally S.R., 2011. The negative binomial–Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. *Accident Analysis and Prevention,* 43(5), 1738-42.

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

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

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

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

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

Park, B.J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention*, 41(4), 683-91.

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

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

Qin, X. and Reyes, P., 2011.Conditional quantile analysis for crash count data. *Journal of Transportation Engineering,* 137 (9), 601-607.

Quddus, M.A., Wang, C., and Ison, S.G., 2010. Road traffic congestion and crash severity: econometric analysis using ordered response models. *Journal of Transportation Engineering*, 136(5), 424-435.

Quddus, M.A., 2008. Modeling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data, *Accident Analysis & Prevention*, 40(4), 1486-1497.

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

TxDOT, 2013. Comparison of motor vehicle traffic deaths, vehicle miles, death rates, and economic loss: 2003-2012. Information contained in this report represents reportable data collected from Texas Peace Officer's Crash Reports (CR-3) received and processed by the Department as of May 27, 2013.

Vadlamani, Sravani, Erdong Chen, Soyoung Ahn and Simon Washington, 2011. Identifying large truck hot spots using crash counts and PDOEs. *Journal of Transportation Engineering*, 137(1), 11–21.

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

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

Wang, C., Quddus, M.A., Ison, S.G., 2011. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis & Prevention*, 43(6), 1979-1990.

Wang, Y., Kockelman K.M., 2013. A conditional-autoregressive count model for pedestrian crashes across neighborhoods. *The 92nd Annual Meeting of the Transportation Research Board.*

Washington, S., J. Metarko, I. Fomunung, R. Ross, F. Julian, and E. Moran, 1999. An Inter-Regional Comparison: Fatal Crashes in the Southeastern and Non-Southeastern United States: Preliminary Findings. *Accident Analysis & Prevention*, 31 (1-2), 135-146.

Yan, X., B. Wang, M. An, C. Zhang, 2012. Distinguishing between rural and urban road segment traffic safety based on zero-inflated negative binomial regression models. *Discrete Dynamics in Nature and Society*, 10.1155.

Zamani, H., Ismail, N., 2010. Negative binomial–Lindley distribution and its application. *Journal of Mathematics and Statistics*, 6 (1), 4–9.