

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

4
5 Jian Xu

6 Graduate Student Researcher
7 Southeast University, China
8 jianxu@utexas.edu
9

10 Kara M. Kockelman

11 (Corresponding author)

12 Professor and William J. Murray Jr. Fellow

13 Department of Civil, Architectural and Environmental Engineering

14 The University of Texas at Austin

15 kcockelm@mail.utexas.edu

16 Phone: 512-471-0210

17 Yiyi Wang

18 Assistant Professor

19 Civil Engineering Department

20 Montana State University

21 yiyi.wang@ce.montana.edu
22

23 Presented at the 93rd Annual Meeting of the Transportation Research
24 Board.

25 Word count: 5277 + 4 Figures + 3 Tables = 7,027 word-equivalents

26 **ABSTRACT**

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

40
41 **Keywords:** traffic safety; crash count modeling; land use; demographics, zero inflated negative
42 binomial models; Poisson models

1

2 **1. INTRODUCTION**

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

12

13 Many statistical methods already exist to rigorously and reliably forecast crash counts and
14 severities as a function of multiple covariates (see, e.g., Kulmala, 1995; Poch and Mannering
15 1996; Abdel-Aty and Radwan, 2000; Ma and Kockelman, 2006, Ma et al., 2008; Quddus et al.,
16 2010; Wang et al., 2011). Prior studies tend to rely on homogenous highway segments (e.g.,
17 Aguero-Valverde and Jovanis 2008) or zone-level crash frequencies (using states, Census tracts,
18 or other topologies -- e.g., Wang et al. 2011), while controlling for variables like vehicle-miles
19 traveled (VMT), curve lengths, degree of curvature, speed limits, and truck shares. Typically
20 neglected covariates include local land use and demographic conditions, climate, and hospital
21 access; so these are controlled for here.

22

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

28

29 **2. LITERATURE REVIEW**

30

31 Various styles of observational units have been used in the crash-count modeling literature,
32 including counties (Miaou et al., 2003; Aguero-Valverde and Jovanis 2006), regions
33 (Washington et al., 1999), districts (Jones et al., 2008), English wards (Noland and Quddus,
34 2004), census tracts (Wang and Kockelman, 2013), and roadway segments (Ma and Kockelman,
35 2006, Ma et al., 2008). Each topology has advantages and disadvantages, and different
36 aggregations of data can lead to somewhat different results. Here, Texas' extensive highway
37 system is divided into hundreds of thousands of homogenous segments, with attributes of
38 curvature, surface width, speed limit, lane count, and so forth constant.

39

40 Commonly used control variables are also reflected here. These include traffic characteristics
41 (e.g., VMT, annual average daily traffic [AADT], and speed limit) and roadway design features
42 (e.g., surface width and horizontal and vertical alignment details [as used in Poch and Mannering
43 1996, Abdel-Aty and Radwan 2000, Ma and Kockelman 2006, Wang et al. 2011]). Other factors,
44 including average rainfall and local land use attributes are also used, to test their predictive
45 powers. Brijs et al. (2008) studied the effects of weather conditions on daily crashes for three

1 large cities in Netherlands in 2001, and the results show that rainfall, temperature and city-
2 specific estimates were highly significant with respect to the number of crashes. Kim and
3 Yamashita (2002) modeled crashes as a function of several land use variables, and they found
4 that the highest crash frequencies occurred near commercial or business properties. In addition,
5 local demographic features for each segment's Census tract were used to predict recent crash
6 counts (as done, to some extent, in Graham and Glaister, 2003; Agüero-Valverde and Jovanis,
7 2006; Kim et al., 2006; and Quddus, 2008).

8
9 A basic specification for crash count prediction is the Poisson model (see, e.g., Jovanis and
10 Chang, 1986; Miaou, 1994), which does not allow for latent heterogeneity across seemingly
11 identical observational units (resulting in an equi-dispersion assumption, where the expected
12 crash rate equals its variance for each unit), unlike the similarly tractable NB approach (Abdel-
13 Aty and Radwan, 2000; Lord, 2000). Other examples include Li et al.'s (2007) Bayesian
14 approach to rank roadway segments by crash risk, Ma and Kockelman's (2008) multivariate
15 Poisson-lognormal model (MVPLN), allowing for simultaneity across counts (by crash severity),
16 and Park and Lord's (2009) finite mixture model (to capture heterogeneity and overdispersion).
17 Wang et al. (2011) used Bayesian estimation for their spatially mixed logit model, combining
18 both crash frequency and severity, and Qin and Reyes (2011) proposed a quantile regression (QR)
19 method for depicting the relationship between a family of conditional quantiles and site
20 covariates.

21
22 Spatial econometric modeling techniques are relatively complex but gaining traction; they enable
23 both spatial heterogeneity and spatial dependence. Standard spatial methods for continuous
24 responses (such as the spatial autoregressive model [SAR] and the spatial error model [SEM])
25 have been employed to analyze zone-level crash counts (e.g., counties and cities, [Baller et al.,
26 2001]), and Bayesian estimation techniques for hierarchical methods have proven a valuable
27 alternative for more discrete/integer counts, at smaller geographical levels (e.g., along short road
28 segments and census tracts [e.g., Wang and Kockelman 2013, Li et al. 2007]). Miaou et al. (2003)
29 used a conditional autoregressive (CAR) structure to illustrate the existence of spatial
30 autocorrelation among adjacent roadway segments along Texas' rural two-lane highways. And
31 Wang et al. (2009) used a series of Poisson-based CAR models to examine the role of traffic
32 congestion on British expressways' crash counts. Due to matrix inversion challenges with larger
33 data sets, the likelihood functions of spatial models are often intractable or their estimation
34 sequence non-convergent (Wang and Kockelman 2013), so evaluation of very large count-data
35 sets are not yet feasible with spatial models

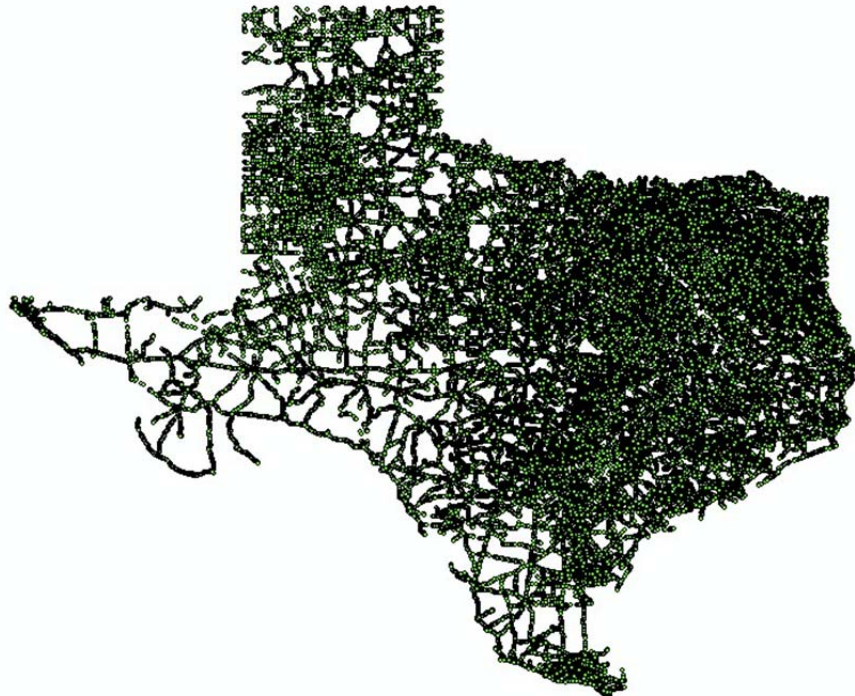
36
37 Many segments and zones demonstrate zero crash counts in any given year, especially for severe
38 (less common) crash counts. Zero-inflated (ZI) models (as used by Miaou, 1994; Zamani and
39 Ismail, 2010; Shankar et al. 2003; Qin et al. 2004; Lord and Geedipally, 2011; and Yan et al.
40 2012) help reflect settings where some locations may never experience crashes. Lord et al. (2005)
41 have provided guidance on how to appropriately model relationships between road safety and
42 traffic exposure, emphasizing a comparison of ZIP and the ZINB models. More recently, Lord
43 and Geedipally (2011) are recommending the negative binomial-Lindley model (NB-L), where
44 "Lindley" refers to a type of two-parameter distribution whose long-term mean never equals zero,
45 but can handle a preponderance of zeros (per site or observational unit), while still maintaining

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

3 4 **3. DATA SETS USED**

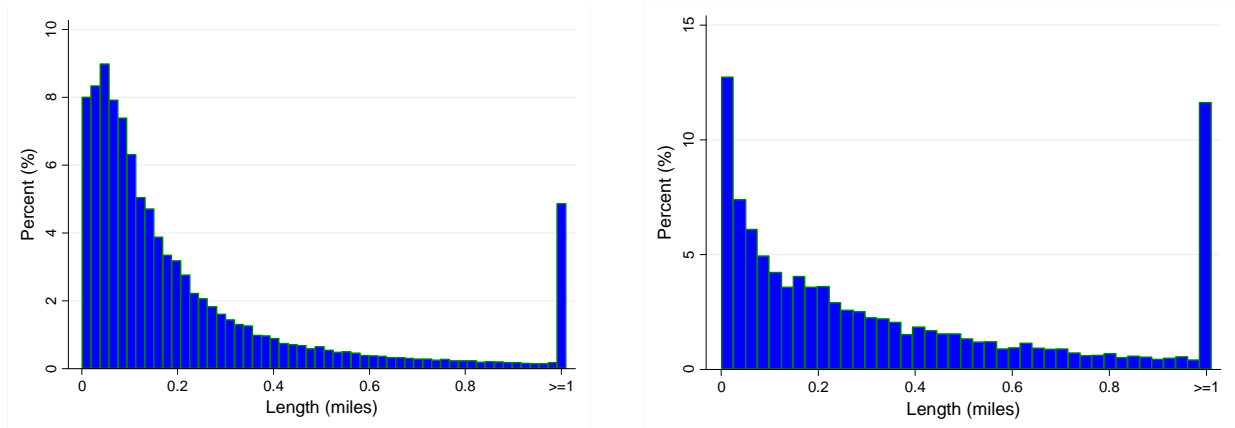
5
6 The Texas Department of Transportation (TxDOT) manages and maintains approximately
7 80,000 centerline-miles of highways, including roughly 7,000 edge-miles of one-way frontage
8 roads alongside the state's freeway corridors. In this study, year 2010 highway, land use, and
9 reported crash data were used to examine the associations between crash counts and various
10 contributing factors along Texas' highways and frontage roads.

11
12 The Texas DOT's Crash Record Information System (CRIS), Road-Highway Inventory Network
13 (RHiNo), and horizontal curve (GEO-HINI) databases were spatially matched/mapped to one
14 another along with local rainfall, land use and demographic attributes. Highways and frontage
15 roads were split into homogenous segments based on consistency/constancy in geometric
16 characteristics (e.g., curvature, surface width, speed limit, lanes number, and AADT) following
17 the merge of the RHiNo and GEO-HINI data sets. No curvature information for frontage roads is
18 provided, because the GEO-HINI data set only pertains to Texas' mainlanes. CRIS crash details
19 were matched to segments using control section-milepoint information. Other variables were
20 mapped to the segments using the ArcGIS toolbox. As a result, 277,510 (highway) mainlane
21 segments (for a total of 72,994 centerline miles) and 15,781 frontage road segments (totaling
22 7,041 edge miles) were identified for analysis, as shown in Figure 1.



23
24 Figure 1. Locations of Homogenous Highway and Frontage Road Segments for Analysis of
25 Texas Crashes
26

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



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

5
 6
 7 The 2010 CRIS data sets contain 243,388 crashes on the state-/TxDOT-maintained network (plus
 8 another 227,794 crashes that occurred on other public roadways across Texas). Among these
 9 243,338 crashes, 209,053 contained control-section and milepoint details that place them along
 10 mainlanes and frontage roads, allowing them to be geocoded onto the set of 350,504
 11 homogenous highway segments. The remaining 34,335 crashes (14.10 percent of the on-system
 12 set) contain no location information or occurred at on- and off-ramps or connectors and detours
 13 and so were removed. For example, 6,224 of the mainlane crashes and 4,291 of the frontage road
 14 crashes were missing appropriate control section-milepost address information. Of course, if
 15 there is any bias in the locations of removed crashes and/or missing crash records (e.g., police
 16 may have a harder time defining crash locations in rural areas than in urban areas, and property-
 17 damage-only (PDO) and less injurious crashes often go unreported (Aguero-Valverde and
 18 Jovanis 2008), there will be some bias in results. These are regular issues in crash-plus-network
 19 databases, however, so such biased are likely to exist in many, perhaps most, analyses already
 20 published.

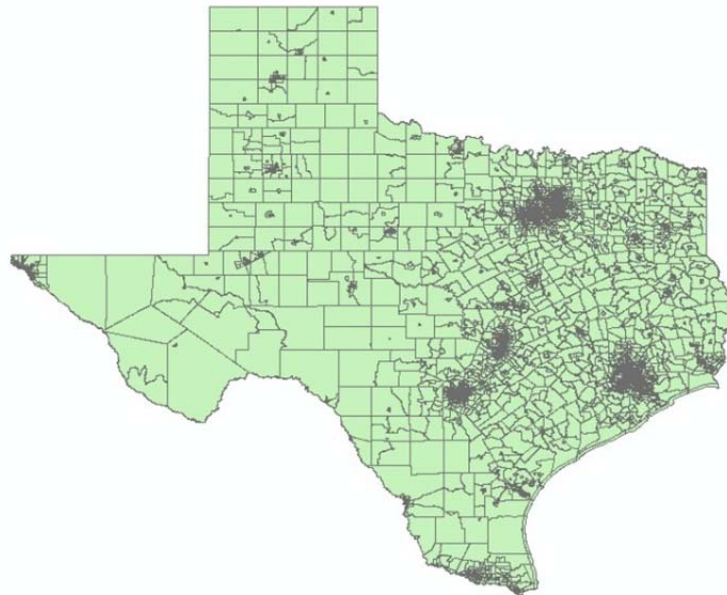
21
 22 In this study, the response or dependent variables are the numbers of crashes by type/severity
 23 level, on each homogenous highway segment. The five severity types are No Injury/PDO,
 24 Possible Injury, Non-Incapacitating, Incapacitating Injury, and Fatal crash counts. As noted
 25 earlier, covariates define highway design, traffic attributes, land use information, climate and
 26 access factor, local demographics, with summary statistics provided in Table 1.

27
 28 **Highway Design Variables:** Highway design decisions have important impacts on a link's use
 29 (flow volumes and truck shares), its speeds, crash counts, and crash severities. Such relationships
 30 have been widely investigated in the past (see, e.g., Poch and Mannering 1996; Lord, 2000; Ma
 31 and Kockelman, 2006, Ma et al., 2008; Wang et al., 2011). Here, a number of design variables
 32 were used (as control variables directly or to construct other covariates, like VMT); these include
 33 Average Shoulder Width (using both inside and outside shoulders), Number of Lanes, Median
 34 Width (not including shoulders or other drivable area), Curve Length, Degree of Curve (i.e.,

1 angle subtended by 100 feet of curve arc), and an Indicator for (the presence of) Curvature. Table
2 1 lists these attributes and their summary statistics.

3
4 **Traffic Attributes:** Traffic characteristics also play a critical role in crash outcomes and
5 prediction. Average Daily Traffic (ADT) estimates (from nearby vehicle-count samples) describe
6 traffic intensity and congestion (using an ADT- or volume-to-capacity variable). VMT is a key
7 crash exposure term (since crash counts closely scale with VMT, everything else constant), and
8 is simply the product of ADT, segment length, and 365 (days per year). In addition, Speed Limit
9 and Percentage of Single and Combo Truck ADT were also included in this study because large
10 trucks are involved in a disproportionately small fraction of the total crashes but a large fraction
11 of fatalities (Vadlamani et al. 2011).

12
13 **Land Use Information:** Information on actual land use designations (e.g., parcels in residential
14 vs. industrial vs. commercial and other uses) could not be obtained across the entire state, but
15 Census data on population demographics and jobs counts per census tract (population density
16 and jobs density) are readily available, along with other variables, like indicators for Rural,
17 Small Urban, Large Urban and Urbanized settings. The data of year 2010 population
18 demographics and jobs counts were obtained from US Census Bureau. Centroids of highway
19 segments were matched to their closest Census tract (measured using Euclidean distances to tract
20 centroids) using ArcGIS's spatial join routine. All count data were normalized by census tract
21 areas (in square miles). The spatial distribution of census tracts in Texas is illustrated in Figure 3.
22

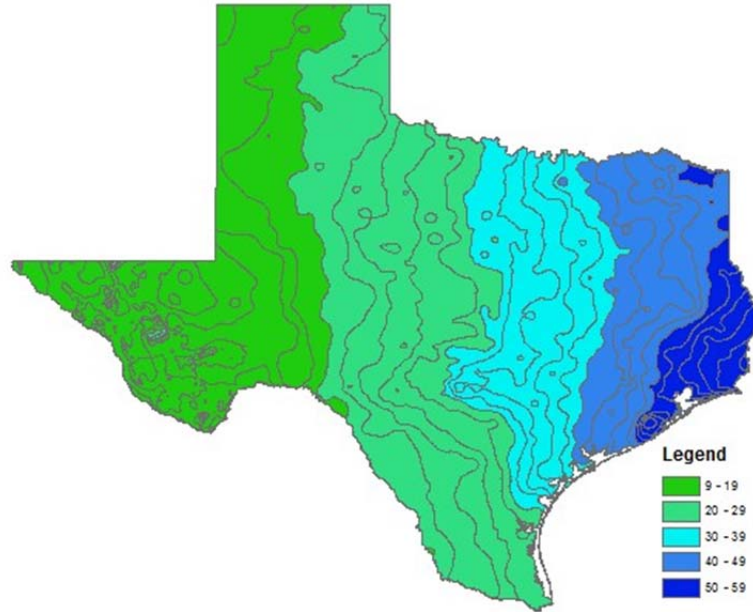


23
24 Figure 3. Census Tracts across Texas
25

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

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

7



8

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

10

11 **Demographic Variables:** Relatively recently, demographic characteristics have come into use
12 for analysis of crash counts (e.g., Graham and Glaister [2003]; Kim et al. [2006]; Quddus et al.
13 [2008]). This work relies on U.S. 2010 Census tract data (as measured using Euclidean distances
14 from segment midpoint to the nearest tract centroid, for density variables of persons of low and
15 high education (that is, less than a high school diploma versus beyond a bachelor's degree), and
16 commuting workers. The average age of people and average income per capita in census tract
17 were also obtained from US Census Bureau. Such variables provide more information on the
18 types of users in the vicinity of the highway, who may be using it regularly, in vehicles or
19 possibly on foot. They also give one a greater sense of the land use setting, versus simply
20 population and jobs density variables.

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								
<i>Crash Count in 2010</i>	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								
<i>Exposure Variable</i>								
VMT (vehicle-miles)	1604.1	7292.9	0	559789	1674.6	4720.6	0.01	103484
<i>Highway Design</i>								
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	/	/	/	/
<i>Traffic Attributes</i>								
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
<i>Land Use Information</i>								
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
<i>Climate & Access Factors</i>								

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			

1
2 **Notes:** ¹Total 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
3 5000 people, 5000 to 49,9993 people, 50,000 to 199,999 people, and 200,000+ people, respectively. ⁶The distances refer to each segment's centroid to the
4 nearest centroid of MMSA.
5

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 $\varepsilon_i \sim \text{gamma}(\gamma, \gamma)$. 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 $i = 1, 2, \dots, n$, 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):

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

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 variables¹ 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).

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

3
4 The association between crash exposure (VMT) and crash rates is estimated to be non-linear
5 (with estimated exponents of $\alpha = 0.6759$ for mainlanes and $\alpha = 0.3273$ for frontage roads), with
6 crash rates effectively falling as VMT rises. It may be that higher VMT levels often correspond
7 to more congested traffic conditions, with reduced operating speeds and lower speed differentials,
8 somewhat dampening the likelihood of collision. Interpretations of other parameter estimates are
9 provided below.

10
11 **Highway Design:** As shown in Table 2 and 3, the Number of Lanes and Curve Length are
12 estimated to increase the average crash rate of mainlanes, while the other highway design
13 variables exhibit negative associations with average crash rates. This is expected since the
14 likelihood of collision is in general higher along curves than that along straight segments,
15 whereas the crash rates may be lower along curves (as drivers tend to be more cautious), a result
16 consistent with the findings of Wang et al (2009).

17
18 It is interesting to find that Average Shoulder Width is negatively associated with crash rates
19 along mainlanes, but positively associated with frontage road rates. Such differences may be due
20 to the different effects of access control used on such facilities.

21
22 **Traffic Attributes:** Not surprisingly, ADT per Lane, % Single and Combo Truck ADT, and
23 Speed Limit were found to be statistically significant. ADT per Lane is estimated to have
24 positive effects on crash rates along mainlanes and frontage roads, consistent with prior studies
25 (e.g., Quddus et al., 2010) and presumably due to greater congestion, which puts vehicles at
26 tighter headways and higher probability of crashing. Interestingly, higher Speed Limits in both
27 models is associated with lower *total* crash rates, presumably due to better roadway designs and
28 less crash-prone locations being assigned higher speed limits; however, this can come with more
29 severe crashes, as found in Ma et al. (2008). As shown in Table 3, the Speed Limit variable is
30 estimated to have a very slight positive effect on fatality rates here; while higher Speed Limit
31 roadways enjoy better design and safer conditions overall, crashes that do occur tend to be more
32 injurious and deadly (see, e.g., Kockelman et al. [2006]). Finally, Single and Combo Truck
33 percentages are associated with lower crash rates, presumably due to greater regulation of and
34 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.

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

Variables	Mainlanes		Frontage Roads	
	Total (%)	Fatal (%)	Total (%)	Fatal (%)
<i>Highway Design</i>				
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	/	/
<i>Traffic Attributes</i>				
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
<i>Land Use Information</i>				
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
<i>Climate & Access Factors</i>				
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
<i>Demographics</i>				
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

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

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

11
12 Indicator variables also entered the model to measure the effects of urbanization, with Rural and
13 Small Urban settings offering negative effects (on crash rates) and the Large Urban setting
14 indicator having a positive impact; such results imply that higher levels of urbanization come
15 with greater crash rates (in large part due, no doubt, to a more complex operating environment,
16 with more interchanges, driveways or intersections per mile, for example – along with
17 congestion impacts [already controlled for here via the AADT-per-lane variable, for example]).
18 As shown in Table 3, mean crash rates tend to fall by 5.32% when land use converts to a rural
19 setting and rise by 1.30% when the reference case (an urbanized setting) becomes a large urban
20 setting.

21
22 **Climate and Access Factors:** As revealed in Tables 2 and 3, rainfall is estimated to be positively
23 associated with crash rates, but only slightly (in a practical sense). As discussed previously,
24 distances to hospitals, schools, and MMSAs are rarely considered as covariates in the crash
25 modeling literature, and here they yield negative impacts in the first (ZI or logistic) process, as
26 shown in Table 2. Results also suggest that shorter distances (greater proximity) come with
27 higher crash rates (probably due to these distances proxying for [the inverse of] land use
28 intensity), but, as expected, positive associations exist for fatal-crash rates (presumably due to
29 more severe collision impacts at higher speeds and time lost in transporting crash victims to an
30 emergency room), similar to the impacts of the Speed Limit variable.

31
32 **Demographics:** As shown in Table 2, all of demographic variables are statistically significant,
33 and some are practically significant, according to Table 3's results. Higher-income locations
34 having lower crash rates is consistent with the World Health Organization's (Global status report
35 on road safety 2013) report (where low- and middle-income countries have higher road traffic
36 fatality rates, per VMT, compared to higher-income countries). Higher Average Age is also
37 estimated to come with lower total and fatal crash rates, thanks presumably to more driving
38 experience and more caution by older drivers. Interestingly, both education extremes (below high
39 school and beyond a bachelor's degree) are estimated to have positive effects on crash rates.

40 41 **CONCLUSIONS**

42
43 This paper developed a series of ZINB and ZIP crash prediction models, based on the crash
44 inventory for 277,510 mainlane, homogenous-roadway segments and 15,781 frontage-road
45 homogenous segments in Texas, while controlling for highway design, traffic attributes, land use,
46 climate and access factor, and local demographics. Expected percentage changes in total and

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

10
11 The ZINB model specification adopted here allows for excess zeros and overdispersion in the
12 crash count data sets (with marked improvement in goodness-of-fit, as compared to ZIP and NB
13 models) but does not capture the spatial autocorrelation and/or spatial dependencies that are
14 likely to exist across neighboring segments. While sample size would have to be dramatically
15 reduced from what was used here (due to computing limitations for rigorous spatial regression
16 techniques), such tools should be tested on these data, along with covariates on pavement
17 roughness, topography, sight distances, and other potentially valuable variables for crash
18 prediction and design decisions.

19 20 **ACKNOWLEDGEMENTS**

21
22 The authors thank the Texas Department of Transportation (TxDOT) for providing the CRIS and
23 geometric data sets, as well as Dr. David Maidment and Mr. Graham James for offering rainfall
24 data set and RHINO Geodatabase, respectively. The authors appreciate Dr. Srinivas Reddy
25 Geedipally provided SAS code and administrative support of Ms. Annette Perrone.

26 27 **REFERENCES**

- 28
29 Abdel-Aty, M., Radwan, E., 2000. Modeling traffic accident occurrence and involvement.
30 *Accident Analysis and Prevention*, 32 (5), 633–642.
- 31 Aguero-Valverde, J., and P. P. Jovanis, 2006. Spatial analysis of fatal and injury crashes in
32 Pennsylvania. *Accident Analysis & Prevention*, 38 (3), 618-625.
- 33 Aguero-Valverde, J. and P.P. Jovanis, 2008. Analysis of road crash frequency with spatial
34 models. *Transportation Research Record*, No. 2061, 55-63.
- 35 Baller, R. D., L. Anselin, S. F. Messner, G. Deane, D. F. Hawkins, 2001. Structural covariates of
36 U.S. county homicide rates: incorporating spatial effects. *Criminology*, 39 (3), 561-588.
- 37 Brijs, T., Karlis, D., Wets, G., 2008. Studying the effect of weather conditions on daily crash
38 counts using a discrete time-series model. *Accident Analysis & Prevention*, 40(3), 1180-1190.
- 39 Global status report on road safety 2013.
40 http://www.who.int/violence_injury_prevention/road_safety_status/2013/en/index.html

- 1 Graham, D.J., Glaister, S., 2003. Spatial variation in road pedestrian casualties: the role of urban
2 scale, density and land-use mix. *Urban Studies*, 40 (8), 1591–1607.
- 3 Jansakul, N., and J.P. Hinde, 2008. Score tests for extra-zero models in zero-inflated negative
4 binomial models, *Communications in Statistics - Simulation and Computation*, 38(1), 92-108.
- 5 Jones, A.P., R. Haynes, and V. Kennedy et al., 2008. Geographical variations in mortality and
6 morbidity from road traffic accidents in England and Wales. *Health & Place*, 14 (3), 519-535.
- 7 Jovanis, P., Chang, H.L., 1986. Modeling the relationship of accidents to miles traveled.
8 *Transportation Research Record*, 1068, 42–51.
- 9 Kim, K., Yamashita, E.Y., 2002. Motor vehicle crashes and land use: empirical analysis from
10 Hawaii. *Transportation Research Record*, 1784, 73–79.
- 11 Kim, K., Brunner, I.M., Yamashita, E.Y., 2006. The influence of land use, population,
12 employment and economic activity on accidents. *Transportation Research Record*, 1953, 56–64.
- 13 Kockelman, K., CRA International, 2006. Safety Impacts and Other Implications of Raised
14 Speed Limits on High-Speed Roads. NCHRP 17-23 Final Report, for the Transportation
15 Research Board, Washington, D.C.
- 16 Kulmala, R., 1995. Safety at rural three-and four-arm junctions: development and application of
17 accident prediction models. *VTT Publications*. Espoo: Technical Research Center at Finland.
- 18 Lee, J., Mannering, F.L., 2002. Impact of roadside features on the frequency and severity of run-
19 off-road accidents: an empirical analysis. *Accident Analysis & Prevention*, 34 (2), 349–361.
- 20 Li, L., Zhu, L., Daniel Z. S., 2007. A GIS-based Bayesian approach for analyzing spatial–
21 temporal patterns of intra-city motor vehicle crashes. *Journal of Transport Geography*, 15(4),
22 274-285.
- 23 Lord, D., 2000. The Prediction of Accidents on Digital Networks: Characteristics and Issues
24 Related to The Application of Accident Prediction Models. Ph.D. Dissertation, Department of
25 Civil Engineering, University of Toronto, Toronto.
- 26 Lord, D., Simon P. Washington, John N. Ivan, 2005. Poisson, Poisson-gamma and zero-inflated
27 regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident*
28 *Analysis & Prevention*, 37(1), 35-46.
- 29 Lord, D., Geedipally S.R., 2011. The negative binomial–Lindley distribution as a tool for
30 analyzing crash data characterized by a large amount of zeros. *Accident Analysis and Prevention*,
31 43(5), 1738-42.

- 1 Ma, J., Kockelman K.M., 2006. Bayesian multivariate Poisson regression for models of injury
2 count, by severity. *Transportation Research Record*, 1950, 24-34.
- 3 Ma, J., Kockelman K.M., Damien, P., 2008. A multivariate Poisson-Lognormal regression model
4 for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis &*
5 *Prevention*, 40(3), 964-975.
- 6 Miaou, S., 1994. The relationship between truck accidents and geometric design of road sections:
7 Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, 26 (4), 471–482.
- 8 Miaou, S., J. J. Song, and B. Mallick, 2003. Roadway traffic crash mapping: A space-time
9 modeling approach. *Journal of Transportation and Statistics*, 6 (1), 33-57.
- 10 Noland, R. B., and M. A. Quddus, 2004. A spatially disaggregate analysis of road casualties in
11 England. *Accident Analysis & Prevention*, 36 (6), 973-984.
- 12 Park, B.J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis.
13 *Accident Analysis & Prevention*, 41(4), 683-91.
- 14 Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection accident frequencies.
15 *Journal of Transportation Engineering*, 122(2), 105–113.
- 16 Qin, X., Ivan, J.N., Ravishankar, N., 2004. Selecting exposure measures in crash rate prediction
17 for two-lane highway segments. *Accident Analysis & Prevention*, 36 (2), 183–191.
- 18 Qin, X. and Reyes, P., 2011. Conditional quantile analysis for crash count data. *Journal of*
19 *Transportation Engineering*, 137 (9), 601-607.
- 20 Quddus, M.A., Wang, C., and Ison, S.G., 2010. Road traffic congestion and crash severity:
21 econometric analysis using ordered response models. *Journal of Transportation Engineering*,
22 136(5), 424-435.
- 23 Quddus, M.A., 2008. Modeling area-wide count outcomes with spatial correlation and
24 heterogeneity: An analysis of London crash data, *Accident Analysis & Prevention*, 40(4), 1486-
25 1497.
- 26 Shankar, V.N., Ulfarsson, G.F., Pendyala, R.M., Nebergal, M.B., 2003. Modeling crashes
27 involving pedestrians and motorized traffic. *Safety Science*, 41 (7), 627–640.
- 28 TxDOT, 2013. Comparison of motor vehicle traffic deaths, vehicle miles, death rates, and
29 economic loss: 2003-2012. Information contained in this report represents reportable data
30 collected from Texas Peace Officer's Crash Reports (CR-3) received and processed by the
31 Department as of May 27, 2013.

- 1 Vadlamani, Sravani, Erdong Chen, Soyoung Ahn and Simon Washington, 2011. Identifying
2 large truck hot spots using crash counts and PDOEs. *Journal of Transportation Engineering*,
3 137(1), 11–21.
- 4 Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypothesis.
5 *Econometrica*, 57, 307–333.
- 6 Wang, C., Quddus, M.A., and Stephen Ison, 2009. The effects of area-wide road speed and
7 curvature on traffic casualties in England. *Journal of Transport Geography* 17(5), 385-395.
- 8 Wang, C., Quddus, M.A., Ison, S.G., 2011. Predicting accident frequency at their severity levels
9 and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis*
10 *& Prevention*, 43(6), 1979-1990.
- 11 Wang, Y., Kockelman K.M., 2013. A conditional-autoregressive count model for pedestrian
12 crashes across neighborhoods. *The 92nd Annual Meeting of the Transportation Research Board*.
- 13 Washington, S., J. Metarko, I. Fomunung, R. Ross, F. Julian, and E. Moran, 1999. An Inter-
14 Regional Comparison: Fatal Crashes in the Southeastern and Non-Southeastern United States:
15 Preliminary Findings. *Accident Analysis & Prevention*, 31 (1-2), 135-146.
- 16 Yan, X., B. Wang, M. An, C. Zhang, 2012. Distinguishing between rural and urban road segment
17 traffic safety based on zero-inflated negative binomial regression models. *Discrete Dynamics in*
18 *Nature and Society*, 10.1155.
- 19 Zamani, H., Ismail, N., 2010. Negative binomial–Lindley distribution and its application.
20 *Journal of Mathematics and Statistics*, 6 (1), 4–9.