ELSEVIER

Accident Analysis and Prevention 38 (2006) 1137–1150

ACCIDENT
ANALYSIS
&
PREVENTION

www.elsevier.com/locate/aap

Temporal and spatial analyses of rear-end crashes at signalized intersections

Xuesong Wang, Mohamed Abdel-Aty*

Department of Civil & Environmental Engineering, University of Central Florida, Orlando, FL 32816-2450, USA

Received 19 February 2006; received in revised form 10 April 2006; accepted 29 April 2006

Abstract

In this study, the generalized estimating equations with the negative binomial link function were used to model rear-end crash frequencies at signalized intersections to account for the temporal or spatial correlation among the data. The longitudinal data for 208 signalized intersections over 3 years and the spatially correlated data for 476 signalized intersections which are located along different corridors were collected in the state of Florida. The modeling results showed that there are high correlations between the longitudinal or spatially correlated rear-end crashes. Some intersection related variables are identified as significantly influencing rear-end crash occurrences at signalized intersections. Intersections with heavy traffic on the major and minor roadways, having more right and left-turn lanes on the major roadway, having a large number of phases per cycle (indicated by the left-turn protection on the minor roadway), with high speed limits on the major roadway, and in high population areas are correlated with high rear-end crash frequencies. On the other hand, intersections with three legs, having channelized or exclusive right-turn lanes on the minor roadway, with protected left-turning on the major roadway, with medians on the minor roadway, and having longer signal spacing have a lower frequency of rear-end crashes.

Keywords: Rear-end crashes; Signalized intersections; Temporal correlation; Spatial correlation; Negative binomial; Generalized estimating equations

1. Introduction

Rear-end crashes occur when the front of a vehicle strikes the rear of a leading vehicle. They are common in road networks. In the U.S., there were approximately 1.89 million rear-end crashes in 2004 (constitute about 30.5% of all police-reported crashes) resulting in 2083 fatal crashes and 555,000 injury crashes (National Highway Traffic Safety Administration, 2006). Rear-end crashes are the leading crash type occurring at signalized intersections. They represent 40.2 percent of all reported intersection crashes based on the crash history of 1531 signalized intersections in the state of Florida (Abdel-Aty et al., 2005b), and 42% in another study (Federal Highway Administration [FHWA], 2004). Considering most unreported crashes are rear-end, the actual percentage of rear-end crashes are even higher, which means that rear-end crashes are a real problem at signalized intersections.

The rear-end crashes at signalized intersections result in a huge cost to society in terms of death, injury, lost productivity, and property damage. From 2002, the research has been conduced in the state of Florida to identify the crash profiles for the major intersection types considering geometric design features, traffic control and operational features, and traffic characteristics (Abdel-Aty et al., 2005b). Including in the study are the statistics of rear-end crashes for the major intersection types, which could be used as reference values to assist in identifying intersections with high numbers of rear-end crashes. The data collected in the research were used to examine the crash type and the crash severity (Abdel-Aty and Keller, 2005; Abdel-Aty et al., 2005a). The purpose of this study is to further investigate the safety effect of intersection related variables on rear-end crash occurrence in order to develop efficient countermeasures to reduce their occurrence at signalized intersections.

Many studies have investigated rear-end crashes by considering the driver or vehicle related factors. From the driver's perspective, Kostyniuk and Eby (1998) found that the action of the driver in the leading vehicle was the dominant contributing factor for a rear-end crash (i.e., the leading vehicle stopped unexpectedly or did not move when it should have).

* Corresponding author. Tel.: +1 407 8235657; fax: +1 407 8234676.
E-mail address: mabdel@mail.ucf.edu (M. Abdel-Aty).

0001-4575/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2006.04.022

ITS Joint Program Office (1999) identified that driver inattention, following too close, and distraction were primary causes for approximately 92% of rear-end collisions. Singh (2003) found that there was an association between driver's age and driver's role (striking/struck) in a rear-end crash, as was of an association between gender of the young driver and driver's role.

The steering and braking performance of different types of vehicles are also critical in the avoidance of crashes; differences between vehicles in braking performance are responsible for many rear-end crashes (Strandberg, 1998). Moreover, the size of the leading vehicle may influence the behavior of the following driver. Graham (2001) reported that light truck vehicles (LTV) make it impossible for drivers in smaller vehicles to see the traffic ahead of them. Therefore, driver's visibility significantly affects the chance of being involved in a rear-end collision when the leading vehicle stops suddenly. Abdel-Aty and Abdelwahab (2004) found that driver's visibility and inattention are the largest factors in a rear-end collision of a regular passenger car striking an LTV.

However, in the above studies, only driver and vehicle factors were addressed; therefore, deficiencies related to roadway and traffic factors could not be identified. The specific road environment conditions of signalized intersections could play a significant role in rear-end crashes and they may contain all kinds of non-driver and non-vehicle related factors such as intersection geometric design features, traffic control and operational features, and traffic characteristics. For example, it is well accepted that installing a signal might cause an increase in rear-end crashes because of the cyclical stopping of the traffic stream (Roess et al., 2004). Therefore, some studies investigated rear-end crashes focusing on signalized intersections and including intersection related factors (Mitra et al., 2002; Poch and Mannering, 1996; Yan et al., 2005).

Yan et al. (2005) investigated certain rear-end crashes at signalized intersections (two-vehicle involved rear-end crashes and both vehicles proceeded straight) using binary logistic regression models. Several intersection related factors were included (e.g., division, number of lanes at crash site, and speed limit). The logistic regression can investigate each crash or crash involvement, which is better for exploring driver, vehicle and specific crash conditions; however, since the dichotomy-dependent variable of rear-end crash (represented by "1") versus other crash (represented by "0") was used, the modeling results should be interpreted carefully as rear-end crashes compares to other crashes.

The frequency model, which can model the number of rear-end crashes on intersection related factors is better for examining the safety effect of intersection related factors. Poch and Mannering (1996) fitted a rear-end crash frequency model at the approach level (four observations per intersection per year) for 63 four-legged intersections (including signalized and unsignalized intersections) over 7 years (1987–1993) using the Negative Binomial regression. Mitra et al. (2002) fitted a rear-end crash frequency model at the roadway level (two observations per intersection per year) for 52 four-legged signalized intersections in Singapore over 8 years (1992–1999); in addition to the comparably low percentage of rear-end crashes among the data (which is only 15%; it is around 40% in the U.S. as aforementioned), the intersection rear-end crashes were also disaggregated by year and by roadway, which cause extra zeros among the data; therefore, the zero-inflated Poisson (ZIP) model was used to account for the excess zeros. The approach or roadway level models are better able to relate the number of rear-end crashes to specific approach and/or roadway characteristics; however, disaggregating of the crashes by roadway or approach will give rise to "site correlation" and cause excess zeros.

Common to both frequency studies is the use of the longitudinal rear-end crash data; however, the temporal correlation among the longitudinal crash data was not accounted for in the models. A likelihood ratio test was used to test the temporal effect on the estimated coefficients between the models based on the full sample and the subsets (e.g., different years). However, the correlation among the data will affect standard errors and is a major concern for correlated data. There are serious problems arising when basic count data models (e.g., Poisson and negative binomial) are used for longitudinal data, since basic count data models assume the dependent variables are independent. For longitudinal data, the error structures become a mixture of random between-intersection errors and highly correlated within-intersection errors.

The spatial correlation is another important issue in analyzing rear-end crashes. The signalized intersections along a certain corridor, especially for those in close proximity, will affect each other in many aspects: several adjacent signalized intersections along a certain corridor will share a high percentage of the same traffic since corridors usually serve relatively long trips between major points; adjacent intersections along a corridor probably have similar types of land use and roadway design; the coordination in signals along a corridor will promote platooning of vehicles crossing intersections, and this coordination may reduce rear-end crashes due to reducing the probability of having to stop at each signal (FHWA, 2004). The use of basic models for spatially correlated data may produce biased estimators and invalid test statistics (Abdel-Aty and Wang, 2006). To avoid the spatial correlation among the data, Poch and Mannering (1996) used a small subsample of the total number of intersections; however, in order to examine the spatial effect on rear-end crashes, there is a need to look at the spatial relationship for signalized intersections along a corridor rather than treat each intersection as an isolated entity.

Rear-end crash frequencies at intersections are count data, the negative binomial regression possesses most of the desirable statistical properties in describing adequately random, discrete, nonnegative, significantly overdispersed, and typically sporadic vehicle crashes at intersections (Chin and Quddus, 2003). Having multiple observations on the same units allows us to control certain unobserved characteristics of intersections or intersection clusters when using panel data models (Wang et al., 2006; Abdel-Aty and Wang, 2006). Generalized estimating equations (GEEs) provide an extension of generalized linear models (GLMs) to the analysis of temporally or spatially clustered data, which can account for the correlation among the observations for a given intersection or an intersection cluster, which is proven to be a robust modeling procedure for temporally or spatially correlated crash data (Abdel-Aty and Wang, 2006; Lord and Persaud, 2000; Wang et al., 2006).

In summary, there have been many studies investigating rear-end crash occurrence; however, most of these studies have focused on the driver or vehicle characteristics. The frequency analysis of rear-end crashes is able to examine the safety effect of intersection related factors; however, the existing rear-end frequency models at approach or roadway levels do not account for the potential temporal, site, or spatial correlation among the data. There is no work on the intersection level for rear-end crashes if the data have temporal or spatial correlations. The objective of this study is to predict and describe the temporally or spatially correlated rear-end crash frequencies at the intersection level using the GEE with the negative binomial as the link function.

2. Data preparation

In order to explore the temporal and spatial correlations and identify the significant variables influencing the rear-end crash occurrence at signalized intersections, one needs to select a variety of intersections possessing different characteristics in geometry and traffic. Restricted by the data availability, different data were used for temporal and spatial analyses. For the temporal analysis, a total number of 208 four-legged signalized intersections in Brevard and Seminole Counties were selected in suburban areas, and a total number of 476 signalized intersections along 41 principle and minor arterials were selected in Orange, Brevard, and Miami-Dade Counties in the state of Florida for the spatial analysis.

For the temporal analysis, the necessary data needed to be collected over the study period including intersection geometric design features, traffic control and operational features, traffic characteristics, and crash data for the same intersections. It is difficult to obtain all this information over a long period, and therefore data for three recent years (2000, 2001, and 2002) were collected and used. The yearly traffic volume data on major and minor roadways for all 208 intersections for 3 years were provided by the traffic engineering departments in each county.

In order to examine the spatial correlation of rear-end crashes among the intersections, the sequences of 476 intersections along 41 corridors were identified automatically by using the geocoded GIS map. If the distance between intersections along a certain corridor is very long and the number of intersections along a corridor is extremely large, the intersections were then divided into sub-clusters in which intersections group together. The distance between intersections is considered for grouping them. The input data are *x*–*y* coordinates for each intersection generated from the GIS map. In this analysis, the SAS MODECLUS procedure is used for cluster analysis. It begins with each intersection in a separate cluster. Then find the nearest intersection with a greater estimated density for each. Compute an approximate *p*-value for each cluster by comparing the estimated maximum density in the cluster with the estimated maximum density on the cluster boundary. The least significant cluster is joined with a neighboring cluster repeatedly until all remaining clusters are significant. The number of clusters per corridor varied from 1 to 7. The number of intersections in clusters varied from 1 to 13; the data are unbalanced. Intersections within a cluster are spatially correlated, and intersections from different clusters are assumed to be statistically independent. The traffic volume data on major and minor roadways for all 476 intersections were extracted.

Intersection geometric design features, traffic control and operational features in the study period were extracted from the intersection traffic planning and design diagrams provided by the counties. Hundreds of drawings were then individually examined and identified. The geometric design features of the intersection include: number of through, left, and right lanes for each approach; the presence of exclusive turn lanes at each approach; and the presence of a median at each approach. The traffic control and operational features include: speed limit for each approach and signal timing. Intersection location type was obtained from the standard crash report for intersections with crashes and from the FDOT Roadway Characteristics Inventory (RCI) database for intersections without crashes. It is worth mentioning that most intersection related variables are first inputted at the approach level. As an intersection level crash frequency analysis, the approach level variables are then aggregated into the roadway level (major and minor roadways).

The rear-end crashes that occurred at the intersections were collected by retrieving the Crash Analysis Reporting (CAR) system for state road intersections (at least one intersecting roadway is a state road) and by using the county maintained crash database for county road intersections. The crashes considered in the analyses were rear-end crashes occurring within 250 feet of the intersection milepost and labeled ‘at intersection’ or ‘influenced by intersection’ for crash site location. For the temporal analysis, the annual rear-end crashes and the traffic volume data vary from year to year from 2000 to 2001. For spatial analysis, the rear-end crash frequency is the number of rear-end crashes in two years (1999, 2000).

Note that almost each jurisdiction has a reporting threshold so that crashes are officially reported only if they involve some degree of injury or, in the absence of injury, a specified amount (in terms of dollars) of property damage. In the state of Florida, police report injury crashes and some of the property damage only (PDO) crashes on long forms. Other non injury crashes are sometimes reported on short forms, which are not coded into the state electronic databases. Since many of the rear end crashes are PDO crashes, and then it is expected that some of them to be reported on short forms. The crashes reported on short forms for some counties were obtained, but they are not consistently available for all selected counties. To be consistent and comparable with other studies, the long form rear-end crashes are used in our analyses. Abdel-Aty et al. (2005a) looked into the quality and completeness of the crash data and the effect that incomplete data has on the final results by using the tree-based regression. They found that for rear-end, right-turn and sideswipe crashes, the important factors are fairly consistent between the models created by complete (reported on long and short forms) and restricted datasets (reported only on long forms).

The type of the right-turn lanes (channelized, exclusive, or shared) for major and minor roadways for the longitudinal data, and the upstream and downstream signal spacings (segment

Table 1
Descriptive statistics for the temporally correlated data

Variable	Mean	Minimum	Maximum	S.D.
Number of rear-end crashes per year for intersection	2	0	31	3.1
Number of through lanes on major roadway	3.8	2	6	1.1
Number of through lanes on minor roadway	2.4	2	6	0.8
Number of exclusive right-turn lanes on major roadway	0.5	0	2	0.7
Number of exclusive right-turn lanes on minor roadway	0.5	0	2	0.7
Number of left-turn lanes on major roadway	2.2	2	4	0.6
Number of left-turn lanes on minor roadway	2.2	2	4	0.5
Type of right-turn lanes on major roadway (equal 2 if channelized; equal 1 if exclusive; equal 0 if shared with through lane)	0.4	0	2	0.6
Type of right-turn lanes on minor roadway (equal 2 if channelized; equal 1 if exclusive; equal 0 if shared with through lane)	0.4	0	2	0.6
Number of left-turn lanes on major roadway (equal 1 if more than 2; equal 0 if less than or equal to 2)	0.1	0	1	0.4
Number of left-turn lanes on minor roadway (equal 1 if more than 2; equal 0 if less than or equal to 2)	0.2	0	1	0.4
Angle of intersection (degree)	84.7	48	90	9.8
Median on major roadway (equal 1 if with median; equal 0 if without median)	0.7	0	1	0.5
Median on minor roadway (equal 1 if with median; equal 0 if without median)	0.5	0	1	0.5
Left-turn protection on major roadway (equal 1 if both approaches are protected; equal 0 if one or none of approaches is protected)	0.9	0	1	0.3
Left-turn protection on minor roadway (equal 1 if both approaches are protected; equal 0 if one or none of approaches is protected)	0.6	0	1	0.5
Speed limit on major roadway (mph, the values in parentheses are in km/h)	42.6 (68.5)	20 (32.2)	55 (88.5)	5.6 (9.01)
Speed limit on minor roadway (mph, the values in parentheses are in km/h)	33.8 (54.4)	20 (32.2)	50 (80.5)	6.2 (9.98)
ADT on major roadway in each year (vehicles/day)	27.452	1625	68460	13196
ADT on minor roadway in each year (vehicles/day)	12.258	1140	63477	8418
Location type (equal 2 for suburban area with population higher than 2500; equal 1 for suburban area with population less than 2500)	1.3	1	2	0.5

length) for each intersection for the spatial analysis were identified and measured by retrieving the software Google Earth (2005). The Google Earth provides high-resolution aerial and satellite imagery and other geographic information; thereby searching and viewing the imagery for a specific intersection are easy.

The sample covers various types of intersections in geometric design features, in traffic control and operational features, in traffic characteristics, and in crashes. The roadway level variable will be first included into the model without any transformation. For categorical variables, if a certain level has no sufficient observations and is not significant in the model, it will be combined with another level; and if two nearest levels have similar coefficients and similar level of significance, they will be combined into one level. The summary statistics of variables are presented in Tables 1 and 2 for temporal and spatial analyses, respectively.

3. Methodology

This section describes how the generalized estimating equations (GEE) accounts for the correlation among the temporally or spatially correlated crash data for the intersection level rear-end crash frequency models. The type III analysis which can be used to identify variables' relative significance is explained and followed by the introduction of model assessment techniques of cumulative residuals test and marginal R-square statistic.

3.1. *Modeling temporal or spatial correlation in GEEs*

The GEE comes from specifying a known function of the marginal expectation of the dependent variable as a linear function of covariates, assuming that the variance is a known function of the mean, and in addition, specifying a "working" correlation matrix for the observations for each cluster (Liang and Zeger, 1986; Zeger and Liang, 1986).

Let y_{ij} represent the jth observation on the ith subject, for $i = 1,2,\ldots,K$ and $j = 1,2,\ldots,n_i$. For the temporal analysis, y_{ij} represent the annual rear-end crash frequency occurred at intersection i in year j, and the numbers of repeated observations for each intersection are fixed and do not vary among intersections in our analysis. There are K intersections. For the spatial study, y_{ij} represent the two-year rear-end crash frequency for intersection j in cluster i. Define K as the total number of clusters, and n_i is the number of intersections in cluster i; the number of intersections per cluster varies and the data are unbalanced and there are $\sum_{i=1}^{K} n_i$ total intersections. In following, "subject" is used to represent "intersection" for the temporal analysis and to represent "intersection cluster" for the spatial analysis.

Let the vector of rear-end crash frequency for the ith subject be $Y_i = (y_{i1}, \ldots, y_{\text{in}_i})'$ with corresponding means $\mu_i = (\mu_{i1}, \ldots, \mu_{\text{in}_i})'$ and V_i is an estimator of the covariance matrix of Y_i. Suppose $x_{ij} = (x_{ii1},\ldots,x_{ijp})'$ denote a $p \times 1$ vector of explanatory variables associated with y_{ij}.

The GEE for estimating β is an extension of the GLMs to the correlated data. The link function and linear predictor setup is

Table 2
Descriptive statistics for the spatially correlated data

Variable	Mean	Minimum	Maximum	S.D.
Number of rear-end crashes in 2 years at intersection	7.6	0	55	8.9
Intersection configuration (equal 2 if three-legged; equal 1 if four-legged)	1.2	1	2	0.4
Number of through lanes on major roadway	4.3	2	8	1.2
Number of through lanes on minor roadway	2.6	1	6	1.0
Number of exclusive right-turn lanes on major roadway	0.5	0	2	0.7
Number of exclusive right-turn lanes on minor roadway	0.4	0	2	0.7
Number of left-turn lanes on major roadway	1.9	0	4	0.7
Number of left-turn lanes on minor roadway	1.9	0	4	0.7
Right-turn lanes on major roadway (equal 1 if at least one approach has an exclusive right-turn lane; equal 0 if no exclusive right-turn lane)	0.3	0	1	0.5
Right-turn lanes on minor roadway (equal 1 if at least one approach has an exclusive right-turn lane; equal 0 if no exclusive right-turn lane)	0.4	0	1	0.5
Number of left-turn lanes on major roadway (equal 1 if more than 2; equal 0 if less than or equal to 2)	0.1	0	1	0.3
Number of left-turn lanes on minor roadway (equal 1 if more than 2; equal 0 if less than or equal to 2)	0.1	0	1	0.3
Median on major roadway (equal 1 if with median; equal 0 if without median)	0.7	0	1	0.5
Median on minor roadway (equal 1 if with median; equal 0 if without median)	0.4	0	1	0.5
Left-turn protection on major roadway (equal 1 if at least one approach has protected left-turn lane; equal 0 if no protected left-turn lane)	0.7	0	1	0.5
Left-turn protection on minor roadway (equal 1 if at least one approach has protected left-turn lane; equal 0 if no protected left-turn lane)	0.9	0	1	0.4
Speed limit on major roadway (mph, the values in parentheses are in km/h)	40.8 (65.6)	25 (40.2)	55 (88.5)	6.2 (9.98)
Speed limit on minor roadway (mph, the values in parentheses are in km/h)	27.8 (44.7)	25 (40.2)	40 (64.4)	4.8 (7.72)
ADT on major roadway in the study period (vehicles/day)	38367	3500	96000	16726
ADT on minor roadway in the study period (Vehicles/day)	19269	1633	68133	10894
The distance to the nearest signal along a corridor for an intersection (feet, the values in parentheses are in meters)	1304.2 (397.4)	61.0 (18.6)	17107.0 (5213.1)	1492.4 (454.8)
The average distance of upstream and downstream segments of an intersection along corridor (feet, the values in parentheses are in meters)	1812.8 (552.4)	138.5 (42.2)	17952.5 (5470.8)	1809.8 (551.5)

as regular GLMs and is given by

$$S(\beta)=\sum_{i=1}^{K}\frac{\partial\mu_t'}{\partial\beta}V_i^{-1}(Y_i-\mu_i(\beta))=0 \qquad (1)$$

Since $g(\mu_{ij})=x'\beta$, where g is the link function. The $p\times n_i$ matrix of partial derivatives of the mean with respect to the regression parameters for the ith subject is given by

$$\frac{\partial\mu_t'}{\partial\beta}=\begin{bmatrix}\dfrac{x_{i11}}{g'(\mu_{i1})} & \cdots & \dfrac{x_{\mathrm{in}_i1}}{g'(\mu_{in_i})}\\ \vdots & & \vdots\\ \dfrac{x_{i1p}}{g'(\mu_{i1})} & \cdots & \dfrac{x_{\mathrm{in}_ip}}{g'(\mu_{\mathrm{in}_i})}\end{bmatrix} \qquad (2)$$

The covariate matrix of Y_i is specified as the estimator $V_i=\phi A_i^{1/2}R_i(\alpha)A_i^{1/2}$, where A_i is a $n_i\times n_i$ diagonal matrix with $v(\mu_{ij})$ as the jth diagonal element. V_i can be different from subject to another, but generally is to specify the same form of V_i for all subjects. $R_i(\alpha)$ is a $n_i\times n_i$ working correlation matrix that is fully specified by the vector of parameters α. Liang and Zeger (1986) have suggested several possible working correlation structures:

(1) Independent $R_i(\alpha)$

The independent correlation structure assumes that repeated observations for a subject are independent. In this case, the GEE estimates are the same as the regular GLM. However, their standard errors are different because the GEE procedure still accounts for the correlation by operating at the cluster level.

$$\mathrm{Corr}(y_{ij},y_{ik})=\begin{cases}1 & j=k\\ 0 & j\neq k\end{cases},\quad \text{e.g.,}$$

$$R_{3\times3}=I_{3\times3}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} \qquad (3)$$

(2) Exchangeable $R_i(\alpha)$

The exchangeable working correlation makes constant the correlations between any two observations within a subject.

$$\mathrm{Corr}(y_{ij},y_{ik})=\begin{cases}1 & j=k\\ \alpha & j\neq k\end{cases},\quad \text{e.g.,}$$

$$R_{3\times3}=\begin{bmatrix}1&\alpha&\alpha\\ \alpha&1&\alpha\\ \alpha&\alpha&1\end{bmatrix} \qquad (4)$$

where $\hat{\alpha}=(1/(N^*-p)\phi)\sum_{i=1}^{K}\sum_{j<k}e_{ij}e_{ik}$, $N^*=0.5\sum_{i=1}^{K}n_i(n_i-1)$. The dispersion parameter ϕ is esti-

mated by $\widehat{\phi} = (1/(N-p))\sum_{i=1}^{K}\sum_{j=1}^{n_i} e_{ij}^2 e_{ij}$ and e_{ij} are the Pearson residuals.

(3) Autoregressive (AR-1) $R_i(\alpha)$

AR-1 weighs the correlation between two observations for a subject by their separated gab (order of measure). As the distance increases the correlation decreases.

$$\mathrm{Corr}(y_{ij}, y_{i,j+t}) = \alpha^t,\ t = 0, 1, 2, \ldots, t_i - j \quad \text{e.g.,}$$

$$R_{3\times 3} = \begin{bmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{bmatrix} \tag{5}$$

where $\hat{\alpha} = (1/(M-p)\phi)\sum_{i=1}^{K}\sum_{j\le n_i-1} e_{ij}e_{ij+1}$ and $M = \sum_{i=1}^{K}(t_i - 1)$.

(4) Unstructured $R_i(\alpha)$

It assumes there are different correlations between any two observations for a subject.

$$\mathrm{Corr}(y_{ij}, y_{ik}) = \begin{cases} 1 & j = k \\ \alpha_{jk} & j \neq k \end{cases}, \quad \text{e.g.,}$$

$$R_{3\times 3} = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & 1 & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & 1 \end{bmatrix} \tag{6}$$

where $\hat{\alpha} = (1/(n-p)\phi)\sum_{i=1}^{K} e_{ij}e_{ik}$.

The estimation of the working correlation structures for unbalanced data can use the *all available pairs* method (SAS Institute Inc., 2004), in which all non-missing pairs of data are used in the moment estimators of the working correlation parameters.

The model-based estimator of $\mathrm{Cov}(\hat{\beta})$ is then given by $\sum_m(\hat{\beta}) = I_0^{-1}$, where $I_0 = \sum_{i=1}^{K}(\partial\mu_i'/\partial\beta)V_i^{-1}(\partial\mu_i/\partial\beta)$. It is worth mentioning that multicollinearity is an obvious phenomenon for intersection safety analysis: larger intersections usually have more traffic volume, higher speed limit, more left-turn lanes, etc. But the multicollinearity does not violate any assumption and would not cause the estimators to be biased, inefficient, or inconsistent, and does not affect the forecasting performance of the model (Ramanathan, 1995). The "problem" is that it will lead to higher standard errors. For our data, correlations among independent variables did not have high values, and there was no observation that the estimated coefficients were drastically altered when variables were added or dropped. Furthermore, the coefficients in the estimated models were significant and had meaningful signs and magnitudes.

3.2. Model assessment

The GEE estimates are obtained when a quasi-likelihood technique is used; therefore, the goodness-of-fit tests for the basic negative binomial regression are not valid for the GEE negative binomial.

If the data are balanced (e.g., panel data without missing observations), Lin et al. (2002) present a graphical and numerical cumulative residuals method based on the cumulative sums of residuals for checking the link function of GEEs. For a GEE model, the distribution of the stochastic processes under the assumed model can be approximated by the distribution of certain zero-mean Gaussian processes whose realizations can be generated by simulation. Each observed residual pattern could then be compared, both graphically and numerically, with a number of realizations from the Gaussian process. Both the maximum absolute value of the observed cumulative sum and the p-value for a Kolmogorov-type supremum test can be calculated. Like the raw residual plot, if the model is correct, the residuals are centered at zero and the plot of the residuals against any coordinate should exhibit no systematic tendency. The cumulative residuals test is used to assess the GEE models in the temporal analysis.

The cumulative residuals test is not suitable for unbalanced data. Zheng (2000) introduced a simple extension of R^2 statistics for GEE models as "marginal R^2", which is calculated by

$$R^2 m = 1 - \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{ij})^2} \tag{7}$$

where $\bar{y}_{ij}$ is the marginal mean rather than the cross-sectional mean. The marginal R^2 is interpreted as the amount of variance in the response variable that is explained by the fitted model. The marginal R^2 statistics are used to assess the GEE models in both temporal and spatial analyses of rear-end crashes.

3.3. Type III analysis

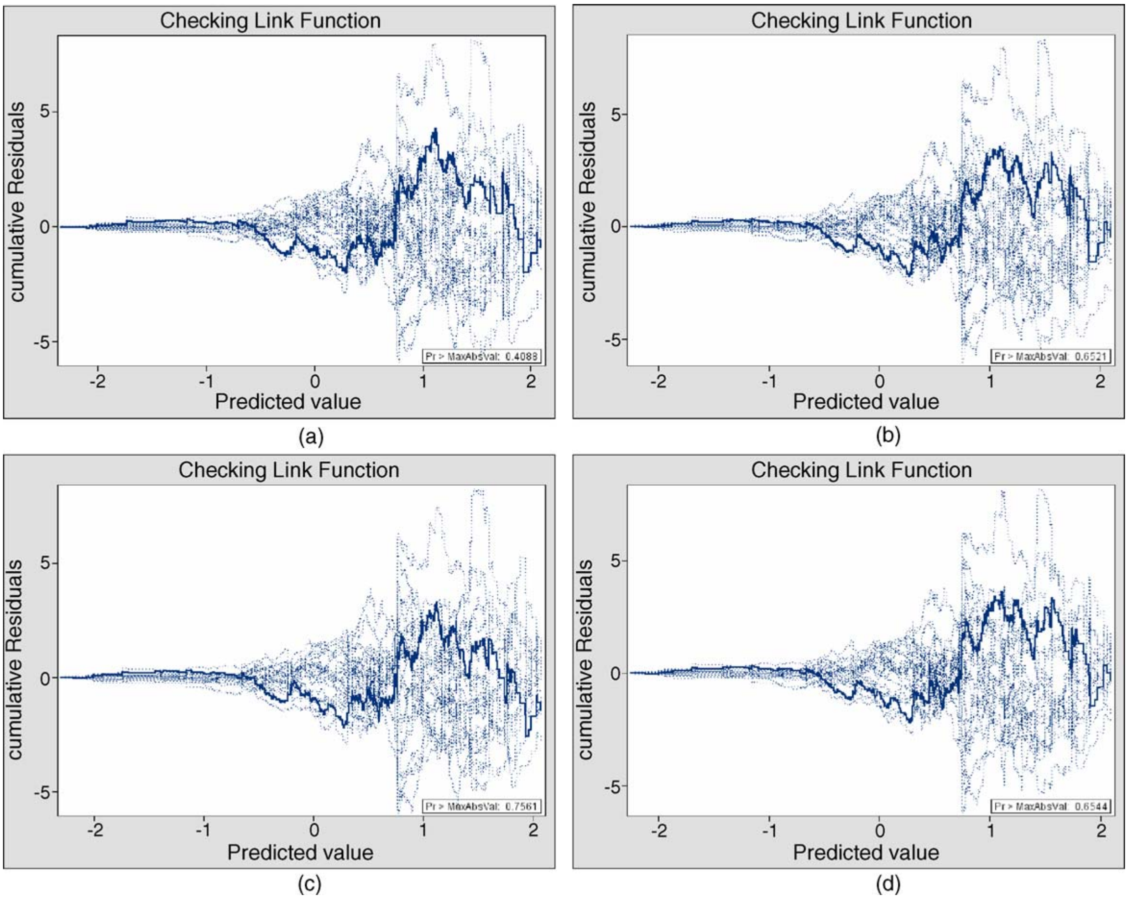
The type III analysis can be used to identify variables' relative significance. The type III χ^2-value for a particular variable is the difference between the generalized score statistic for the model with all the variables included and the generalized score statistic for the model with this variable excluded. The hypothesis tested in this case is the significance of this variable given that all the other variables are in the model. The small p-value indicates that the effect of this variable is highly significant (SAS Institute Inc., 2004).

4. Estimation results

The intersection level rear-end crash frequencies were modeled using the Generalized Estimating Equations (GEEs) with the Negative Binomial link function for the data with temporal or spatial correlation separately. The different correlation structures suggested by Liang and Zeger (1986) were explored. Only the variables that are significant in at least one model among basic Negative Binomial and GEE Negative Binomial models were included. The type III analysis was used to identify the relative significance of the variables in the models. Models were evaluated by both the cumulative residuals test and the marginal R^2 statistic for the temporal analysis and by the marginal R^2 statistic for the spatial analysis.

Table 3
Model estimates for the temporally correlated rear-end crashes

Parameter	MLE estimates (S.E.)		GEE negative binomial estimations							
			Independent		Exchangeable		Autoregression		Unstructured	
	Coeff.	S.E. (*p*-value)	Coeff.	S.E. (*p*-value)	Coeff.	S.E. (*p*-value)	Coeff.	S.E. (P-value)	Coeff.	S.E. (P-value)
Intercept	−13.1993	1.4512 (<0.0001)	−13.1993	1.5115 (<0.0001)	−12.7856	2.044 (<0.0001)	−13.1113	1.9477 (<0.0001)	−12.7678	2.0625 (<0.0001)
Logarithm of ADT on major roadway	0.5567	0.1055 (<0.0001)	0.5567	0.1139 (<0.0001)	0.5509	0.1549 (0.0004)	0.5739	0.1474 (<0.0001)	0.5515	0.1563 (0.0004)
Logarithm of ADT on minor roadway	0.574	0.0993 (<0.0001)	0.574	0.108 (<0.0001)	0.5335	0.1458 (0.0003)	0.5416	0.1386 (<0.0001)	0.5307	0.1471 (0.0003)
Type of right-turn lanes on minor roadway										
Channelized	−0.7276	0.2128 (0.0006)	−0.7276	0.2166 (0.0008)	−0.7085	0.2966 (0.0169)	−0.6862	0.2801 (0.0143)	−0.6854	0.299 (0.0219)
Exclusive	−0.3213	0.1298 (0.0133)	−0.3213	0.1308 (0.014)	−0.3089	0.1792 (0.0848)	−0.3103	0.1694 (0.0669)	−0.3115	0.181 (0.0853)
Shared with through lane	0	–	0	–	0	–	0	–	0	–
Number of left-turn lanes on major roadway										
More than 2	0.521	0.1638 (0.0015)	0.521	0.1633 (0.0014)	0.5382	0.2238 (0.0162)	0.5022	0.2117 (0.0177)	0.5314	0.2259 (0.0186)
Less than or equal to 2	0	–	0	–	0	–	0	–	0	–
Left-turn protection on minor roadway										
Both approaches are protected	0.5198	0.1143 (<0.0001)	0.5198	0.1187 (<0.0001)	0.5291	0.1626 (0.0011)	0.5266	0.1537 (0.0006)	0.5276	0.1641 (0.0013)
One or none of approaches is protected	0	–	0	–	0	–	0	–	0	–
Location type										
Suburban area with population higher than 2500	0.4281	0.1069 (<0.0001)	0.4281	0.1077 (<0.0001)	0.4288	0.1476 (0.0037)	0.4218	0.1396 (0.0025)	0.4376	0.149 (0.0033)
Suburban area with population less than 2500	0	–	0	–	0	–	0	–	0	–
Median on minor roadway										
With median	−0.2301	0.1122 (0.0403)	−0.2301	0.1127 (0.0412)	−0.2265	0.1545 (0.1426)	−0.2171	0.146 (0.137)	−0.2208	0.1559 (0.1567)
Without median	0	–	0	–	0	–	0	–	0	–
Speed limit on major roadway (mph)	0.0575	0.0105 (<0.0001)	0.0575	0.0111 (<0.0001)	0.0576	0.0152 (0.0001)	0.0583	0.0144 (<0.0001)	0.0575	0.0153 (0.0002)
Dispersion	0.872	0.0938	1.0311	–	1.0341	–	1.0273	–	1.0364	–
Summary statistics										
Number of intersections (number of clusters)	208		208		208		208		208	
Number of continuous years (cluster size)	3		3		3		3		3	
Number of observations	624		624		624		624		624	
Sum of initial rear-end/total crashes	1275/2754		1275/2754		1275/2754		1275/2754		1275/2754	
Maximum absolute value	2.4536		4.2498		3.5416		**3.2801**		3.5871	
Pr > MaxAbsVal	0.1083		0.4088		0.6521		**0.7561**		0.6544	
Marginal R^2 statistics	–		0.1799		0.1800		**0.1813**		0.1795	



(a)

(b)

(c)

(d)

Fig. 1. Model assessment and comparison for GEEs with different correlation structures in temporal analysis: (a) with independence correlation structure; (b) with exchangeable correlation structure; (c) with autoregression correlation structure (AR-1); and (d) with unstructured correlation structure. The *p*-values (Pr > MaxAbsVal) pertain to the supremum test with 10,000 realizations.

4.1. Modeling temporal correlated rear-end crashes

The GEE estimates for the annual rear-end crash frequencies with the different correlation structures (independent, exchangeable, autoregression, and unstructured) are reported in Table 3. For comparison, the basic negative binomial estimations are also reported. The estimated coefficients for the negative binomial and the GEE negative binomial with the independent correlation structure are exactly the same as expected. The GEE models have slightly higher estimated standard errors than the negative binomial model because not accounting for the temporal correlation will under represent the standard errors (Lord and Persaud, 2000). The four correlation structures have produced unequal coefficients, which show the effect of the different correlation structures in the analysis.

The assessment and comparison of the GEE models with the different correlation structures are performed using the cumulative residuals test, which can assess the models graphically and numerically (Lin et al., 2002). The cumulative residual plots for the GEE models with the different correlation structures are drawn using SAS ODS graphic techniques (SAS Institute Inc., 2004) as shown in Fig. 1. The observed cumulative residuals are represented by the heavy lines, and the simulated curves are represented by the light lines. The *p*-values (Pr > MaxAbsVal) are computed based on a sample of 10,000 simulated residual paths as shown in the lower-right corner on each plot. A comparison of the cumulative residual plots shows that the GEE model with an autoregression correlation structure is the best model

Table 4
Estimated working correlation structures for the temporally correlated rear-end crashes

Year	2000	2001	2002
Independent correlation structure			
2000	1.0000	0.0000	0.0000
2001	0.0000	1.0000	0.0000
2002	0.0000	0.0000	1.0000
Exchangeable correlation structure			
2000	1.0000	0.4349	0.4349
2001	0.4349	1.0000	0.4349
2002	0.4349	0.4349	1.0000
Autoregression correlation structure (AR-1)			
2000	1.0000	0.4454	0.1984
2001	0.4454	1.0000	0.4454
2002	0.1984	0.4454	1.0000
Unstructured correlation structure			
2000	1.0000	0.3941	0.4327
2001	0.3941	1.0000	0.5225
2002	0.4327	0.5225	1.0000

Table 5
Type III analysis for the temporally correlated rear-end crashes

Main variables	DF	MLE type III analysis (p-value)	GEE model type III analysis: χ^2 (p-value)			
			Independent	Exchangeable	Autoregression	Unstructured
Logarithm of ADT on major roadway	1	24.01 (<0.0001)	15.58 (<0.0001)	15.62 (<0.0001)	15.91 (<0.0001)	15.7 (<0.0001)
Speed limit on major roadway (mph)	1	36.71 (<0.0001)	11.47 (0.0007)	11.25 (0.0008)	10.82 (0.001)	11.19 (0.0008)
Location type	1	20.84 (<0.0001)	9.85 (0.0017)	9.89 (0.0017)	9.26 (0.0023)	9.88 (0.0017)
Left-turn protection on minor roadway	1	24.26 (<0.0001)	9.08 (0.0026)	9.69 (0.0019)	8.91 (0.0028)	10.16 (0.0014)
Logarithm of ADT on minor roadway	1	8.76 (0.0031)	6.62 (0.0101)	5.63 (0.0177)	5.5 (0.019)	5.75 (0.0165)
Type of right-turn lanes on minor roadway	2	14.92 (0.0006)	6.23 (0.0444)	5.63 (0.06)	5.38 (0.0679)	5.51 (0.0636)
Number of left-turn lanes on major roadway	1	1.48 (0.2238)	4.47 (0.0346)	4.67 (0.0307)	4.1 (0.043)	4.57 (0.0326)
Median on minor roadway	1	2.73 (0.0985)	1.97 (0.1605)	1.87 (0.172)	1.68 (0.1948)	1.8 (0.1793)

with no systematic tendency and the highest p-value (0.7561). The maximum absolute value of its observed cumulative sum is 3.2801. Since we used 10,000 realizations in the supremum test, the p-value 0.7561 means that out of 10,000 realizations from the null distribution, 75.61% have maximum cumulative residuals greater than 3.2801. The GEE model with an autoregression structure also has a higher marginal R^2-value (0.1803) as shown in Table 3.

Since the number of repeated observations for each intersection is three, the estimated working correlation is a symmetric matrix and its dimension is three with one in each diagonal position as shown in Table 4. The autoregression structure assumes that the correlations between the multiple observations for a certain intersection will decrease as the time-gap increases. For example, it is 0.4454 for each successive two years and 0.1984 for the years 2000 and 2002. These correlations indicate that the temporal correlation should be accounted for in the longitudinal crash data. The conclusion that the GEE autoregression model has better goodness-of-fit is consistent with the theory that autoregression structure is specifically appropriate for time-dependent data structures.

The significant variables in Table 3 can be classified into four types: traffic characteristics, intersection geometric design features, traffic control and operational features, and location types. The logarithms of traffic volumes on the major and minor roadways are found to be significant (p-values < 0.0001); they both have positive coefficients (0.5739 and 0.5416)[1], which indicate the higher the traffic volumes the larger the number of rear-end crashes. The left-turn lanes are critical for intersection operation and rear-end crash occurrence; more rear-end crashes occurred with a higher number of left-turn lanes on the major roadway (coeff. = 0.5022, p-value < 0.0177). The number of approaches with protected left-turning is directly related to the number of phase per cycle; increasing the number of phases will increase rear-end crashes at intersections, which is indicated by the positive coefficient (0.5266) of the dummy variable of having protected left-turn lanes for both approaches on the minor roadway. Compared to the shared right-turn lane, the channelized or exclusive right-turn lanes on the minor roadway reduce rear-end crashes, which is indicated by the negative coefficients (−0.6862 and −0.3103, respectively). The intersection with a median on the minor roadway has a lower number of rear-end crashes (coeff. = −0.2171). The high posted speed limit on the major roadway is significant for rear-end crashes (p-value < 0.0001). The selected intersections are located in suburban areas with different population levels; the positive coefficient for the dummy variable of having a higher population (0.4218) shows that intersections located in high population areas are associated with high rear-end crashes.

To examine the relative significance of the explanatory variables, the type III analysis was performed for all the variables included in the models as shown in Table 5. The results show that the ADT on the major roadway is the most significant variable ($\chi^2 = 15.91$), and followed by the speed limit on the major roadway, location type, left-turn protection on the minor roadway, ADT on the minor roadway, number of left-turn lanes on the major roadway, and then median on the minor roadway. Among the traffic control and operational features, the speed limit is the most significant variable in the model ($\chi^2 = 10.82$).

4.2. Modeling spatially correlated rear-end crashes

The GEE models with a negative binomial link function for rear-end crash frequency in two years were fitted for independent, exchangeable, and autoregression correlation structures and the associated estimates are reported in Table 6. The unstructured correlation structure has been tried, but it failed to converge. All non-missing pairs of data are used in the moment estimators of the working correlation parameters. For our data, the number of response pairs for estimating correlation is less than or equal to the number of regression parameters especially for clusters with the extra large size (e.g., the clusters with the size larger than 8). The estimated coefficients and standard errors for the negative binomial regression and the GEE models with exchangeable and autoregression correlation structures are different. The three correlation structures have produced different coefficients and standard errors, which show the effect of different correlation structures in the analysis.

The estimated working correlation structures are presented in Table 7. The correlation structures are symmetric. Since the cluster size varies from 1 to 13, the dimension of the correlation matrix is 13 with one in each diagonal position. The correlation

[1] The estimates for the GEE model with the Autoregression correlation structure are used for variable interpretation. To avoid redundancy, it is not repeated for each variable.

Table 6
Model estimates for the spatially correlated rear-end crashes

Parameters	MLE estimate (S.E.)		GEE negative binomial estimate					
			Independent		Exchangeable		Autoregression	
	Coeff.	S.E. (*p*-value)	Coeff.	S.E. (*p*-value)	Coeff.	S.E. (*p*-value)	Coeff.	S.E. (*p*-value)
Intercept	−7.7926	1.1152 (<0.0001)	−7.7926	1.2275 (<0.0001)	−7.6801	1.4784 (<0.0001)	−4.5202	1.2874 (0.0004)
Logarithm of ADT on major roadway	0.48	0.1265 (0.0001)	0.48	0.1386 (0.0005)	0.4888	0.1566 (0.0018)	0.3854	0.1338 (0.004)
Logarithm of ADT on minor roadway	0.5124	0.1038 (<0.0001)	0.5124	0.1152 (<0.0001)	0.4449	0.114 (<0.0001)	0.2826	0.0848 (0.0009)
Intersection configuration								
Three-legged	−0.427	0.1163 (0.0002)	−0.427	0.1263 (0.0007)	−0.3486	0.1283 (0.0066)	−0.2873	0.0949 (0.0025)
Four-legged	0	–	0	–	0	–	0	–
Right-turn lanes on major roadway								
At least one approach has an exclusive right-turn lane	0.2607	0.1107 (0.0186)	0.2607	0.1186 (0.0279)	0.2213	0.1234 (0.0731)	0.2215	0.0972 (0.0226)
No exclusive right-turn lane	0	–	0	–	0	–	0	–
Right-turn lanes on minor roadway								
At least one approach has an exclusive right-turn lane	−0.4243	0.1028 (<0.0001)	−0.4243	0.108 (<0.0001)	−0.3859	0.1038 (0.0002)	−0.1288	0.0778 (0.0977)
No exclusive right-turn lane	0	–	0	–	0	–	0	–
Number of left-turn lanes on major roadway								
More than 2	0.5804	0.1517 (0.0001)	0.5804	0.1587 (0.0003)	0.6263	0.156 (<0.0001)	0.3802	0.1233 (0.002)
Less than or equal to 2	0	–	0	–	0	–	0	–
Left-turn protection on major roadway								
At least one approach has protected or partially protected left-turn lane	−0.3254	0.1301 (0.0124)	−0.3254	0.1461 (0.0259)	−0.3717	0.1462 (0.011)	−0.1763	0.1078 (0.1021)
No protected left-turn lane	0	–	0	–	0	–	0	–
Left-turn protection on minor roadway								
At least one approach has protected or partially protected left-turn lane	0.3969	0.1039 (0.0001)	0.3969	0.1111 (0.0004)	0.4377	0.1097 (<0.0001)	0.2435	0.078 (0.0018)
No protected left-turning movement	0	–	0	–	0	–	0	–
Median on minor roadway								
With median	−0.2356	0.1032 (0.0224)	−0.2356	0.1094 (0.0312)	−0.1828	0.1099 (0.0962)	−0.1527	0.083 (0.0658)
Without median	0	–	0	–	0	–	0	–
Logarithm of the average distance of upstream and downstream segments of an intersection along corridor	−0.1408	0.063 (0.0253)	−0.1408	0.0669 (0.0352)	−0.094	0.0712 (0.1864)	−0.1246	0.0595 (0.0362)
Speed limit on major roadway (mph)	0.0196	0.0097 (0.0423)	0.0196	0.0099 (0.048)	0.0216	0.0102 (0.0344)	0.0131	0.0081 (0.1054)
Dispersion	0.6044	0.0527	1.0792	–	1.0886	–	1.058	–
Summary statistics								
Number of corridors	41		41		41		41	
Number of intersections	476		476		476		476	
Number of clusters	–		116		116		116	
Minimum cluster size/maximum cluster size	–		1/13		1/13		1/13	
Sum of initial rear-end/total crashes	3620/8731		3620/8731		3620/8731		3620/8731	
Marginal R^2 statistics	–		0.4069		0.4560		0.4591	

Table 7
Estimated working correlation structures for the spatially correlated rear-end crashes

Intersection #	1	2	3	4	5	6	7	8	9	10	11	12	13
Independent correlation structure													
1	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Exchangeable correlation structure													
1	1.0000	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833
2	0.1833	1.0000	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833
3	0.1833	0.1833	1.0000	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833
4	0.1833	0.1833	0.1833	1.0000	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833
5	0.1833	0.1833	0.1833	0.1833	1.0000	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833
6	0.1833	0.1833	0.1833	0.1833	0.1833	1.0000	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833
7	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	1.0000	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833
8	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	1.0000	0.1833	0.1833	0.1833	0.1833	0.1833
9	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	1.0000	0.1833	0.1833	0.1833	0.1833
10	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	1.0000	0.1833	0.1833	0.1833
11	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	1.0000	0.1833	0.1833
12	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	1.0000	0.1833
13	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	0.1833	1.0000
Autoregression correlation structure (AR-1)													
1	1.0000	0.6316	0.3990	0.2520	0.1592	0.1005	0.0635	0.0401	0.0253	0.0160	0.0101	0.0064	0.0040
2	0.6316	1.0000	0.6316	0.3990	0.2520	0.1592	0.1005	0.0635	0.0401	0.0253	0.0160	0.0101	0.0064
3	0.3990	0.6316	1.0000	0.6316	0.3990	0.2520	0.1592	0.1005	0.0635	0.0401	0.0253	0.0160	0.0101
4	0.2520	0.3990	0.6316	1.0000	0.6316	0.3990	0.2520	0.1592	0.1005	0.0635	0.0401	0.0253	0.0160
5	0.1592	0.2520	0.3990	0.6316	1.0000	0.6316	0.3990	0.2520	0.1592	0.1005	0.0635	0.0401	0.0253
6	0.1005	0.1592	0.2520	0.3990	0.6316	1.0000	0.6316	0.3990	0.2520	0.1592	0.1005	0.0635	0.0401
7	0.0635	0.1005	0.1592	0.2520	0.3990	0.6316	1.0000	0.6316	0.3990	0.2520	0.1592	0.1005	0.0635
8	0.0401	0.0635	0.1005	0.1592	0.2520	0.3990	0.6316	1.0000	0.6316	0.3990	0.2520	0.1592	0.1005
9	0.0253	0.0401	0.0635	0.1005	0.1592	0.2520	0.3990	0.6316	1.0000	0.6316	0.3990	0.2520	0.1592
10	0.0160	0.0253	0.0401	0.0635	0.1005	0.1592	0.2520	0.3990	0.6316	1.0000	0.6316	0.3990	0.2520
11	0.0101	0.0160	0.0253	0.0401	0.0635	0.1005	0.1592	0.2520	0.3990	0.6316	1.0000	0.6316	0.3990
12	0.0064	0.0101	0.0160	0.0253	0.0401	0.0635	0.1005	0.1592	0.2520	0.3990	0.6316	1.0000	0.6316
13	0.0040	0.0064	0.0101	0.0160	0.0253	0.0401	0.0635	0.1005	0.1592	0.2520	0.3990	0.6316	1.0000

estimated by exchangeable structure is 0.1833. The autoregression structure has a maximum correlation of 0.6313 for any two successive intersections along a corridor; the correlation between intersections decreases as the spacing between intersections increases.

The marginal R^2-values are reported in Table 6. The GEE model with autoregression structure has a slightly higher marginal R^2-value (0.4591), which indicates that the autoregression structure could be the appropriate structure for spatial correlation. Ballinger (2004) suggests that the decisions about the correlation structures should be guided first by theory. For the spatial correlation of signalized intersections along corridors, as the spacing between intersections increases, it is reasonable to assume the correlation between them decreases, which is consistent with the autoregression approach.

Turning to the significant variables presented in Table 6, traffic volumes on major and minor roadways are still the most significant variables, which are similar to the temporal analysis. Among the intersection geometric design features, the number of legs, the presence of exclusive right-turn lanes on both roadways, and the number of left-turn lane on major roadway are significant to rear-end crash occurrence. The left-turn protection on major roadway will reduce rear-end crashes, while the protection on minor roadway will increase rear-end crash occurrence, which is indicated by the coefficients −0.1763 and 0.2435, respectively. The intersection with a median on the minor roadway has a lower number of rear-end crashes. The posted speed limit on the major roadway is significant for rear-end crashes. The average length of upstream and downstream segments of an intersection, the distance to the nearest signals along a corridor for an

Table 8
Type III analysis for the spatially correlated rear-end crashes

Main variables	DF	MLE type III analysis (p-value)	GEE model type III analysis: χ^2 (p-value)		
			Independent	Exchangeable	Autoregression (AR-1)
Logarithm of ADT on major roadway	1	14.21 (0.0002)	8.47 (0.0036)	8.72 (0.0032)	11.34 (0.0008)
Intersection configuration	1	12.9 (0.0003)	12.21 (0.0005)	7.27 (0.007)	9.98 (0.0016)
Logarithm of ADT on minor roadway	1	23.34 (<0.0001)	14.21 (0.0002)	10.63 (0.0011)	9.97 (0.0016)
Number of left-turn lanes on major roadway	1	15.29 (<0.0001)	10.38 (0.0013)	13.06 (0.0003)	9.94 (0.0016)
Left-turn protection on minor roadway	1	14.13 (0.0002)	7.85 (0.0051)	8.99 (0.0027)	8.51 (0.0035)
Logarithm of the average distance of upstream and downstream segments of an intersection along corridor	1	4.97 (0.0258)	3.77 (0.0523)	1.79 (0.1808)	5.94 (0.0148)
Right-turn lanes on major roadway	1	5.55 (0.0185)	3.18 (0.0744)	2.94 (0.0861)	5.11 (0.0237)
Left-turn protection on major roadway	1	6.41 (0.0114)	5.08 (0.0242)	7.65 (0.0057)	4.07 (0.0437)
Right-turn lanes on minor roadway	1	16.74 (<0.0001)	13.49 (0.0002)	11.43 (0.0007)	3.73 (0.0536)
Median on minor roadway	1	5.17 (0.023)	3.52 (0.0605)	2.13 (0.1444)	3.09 (0.0786)
Speed limit on major roadway (mph)	1	4.1 (0.043)	3.05 (0.0805)	3.96 (0.0466)	2.95 (0.0859)

intersection, and their logarithm transformations are included in the model alternatively, the logarithm of the average length of upstream and downstream segments of an intersection is identified to be the most significant variable (p-value = 0.0362). The negative sign for this factor (coeff. = −0.1246) indicates that the effect of both neighboring signals (not just upstream or downstream segment) decreases as the distance increases. The type III analysis is presented in Table 8 and all explanatory variables are sorted by their relative significance based on the GEE model with the autoregression structure.

5. Summary and conclusions

This study investigated the temporal and spatial correlation for longitudinal data and intersection clusters along corridors for the rear-end crashes at signalized intersections. The intersection level rear-end crash frequency model is capable of identifying the intersection related significant factors by modeling the relationship between the numbers of rear-end crashes and the intersection geometric design features, traffic control and operational features, and traffic characteristics. Note that many minor rear-end crashes (no injury and under a specified amount of property damage) are not reported in almost each jurisdiction. To be consistent and comparable with other studies, only the state maintained rear-end crashes (long form) were used in our analyses.

The data for 208 signalized intersections over 3 years and 476 signalized intersections which are located along different corridors were collected in the state of Florida. The data are temporally or spatially correlated. The use of basic models for such correlated data may produce biased estimators and invalid test statistics. The intersection level rear-end crash frequencies were modeled using the generalized estimating equations (GEEs) with a negative binomial link function for temporal or spatial correlation separately, and the different working correlation structures (independent, exchangeable, autoregressive, and unstructured) have been explored. The GEE autoregression models assuming that the correlations between the multiple observations for a certain intersection or intersection cluster will decrease as the time or space gap increases are better for either temporal or spatial correlated rear-end crashes.

In the temporal analysis, it was found that the estimated correlation is 0.4454 for each successive two years and 0.1984 for the years 2000 and 2002. The estimates have been modified when considering the temporal correlation. In order to have consistent estimation, the temporal correlation should be considered for the panel data by using panel data models (e.g., GEE) especially this correlation is large. In the spatial analysis, the estimated correlation is 0.6316 for two nearest intersections along corridors, which is relatively high. Similarly, it shows that the model estimates will change when considering the spatial correlation. From the statistical point of view, this spatial correlation should be accounted for in order to have consistent estimation for the intersections which are not isolated.

As mentioned before, there are two studies investigating rear-end crash frequencies which focus on signalized intersections and including intersection related factors (Poch and Mannering, 1996; Mitra et al., 2002). Both studies used panel data and disaggregated crashes by approach or roadway; however, the potential temporal correlation and "site correlation" among the disaggregated data were not accounted for. In order to avoid potential spatial correlation among the data, Poch and Mannering (1996) selected a small portion of intersections, and Mitra et al. (2002) tried to select intersections randomly. The GEE procedure used in our study can account for the correlation and provide efficient parameter estimates for correlated data and produce easily interpretable and communicable results.

Turning to the significant variables, the variables included in this paper can be divided into five types: traffic characteristics, geometric design features, traffic control and operational features, location type, and corridor level factors. Poch and Mannering (1996) found that intersection volume, number of signal phases, left-turn protection, area types, roadway types, speed limit, grade, and sight distance are significant to affect rear-end crash occurrence. Mitra et al. (2002) included intersection volume, wide median, number of phases, left-turn protection, surveillance camera, and signal control (adaptive or not) in their analysis.

For traffic characteristics, instead of using total traffic volume at intersections in previous studies, this study showed the logarithm transformation of traffic volumes on major and minor roadways are the better functional forms for traffic volume in rear-end crash frequency model.

For geometric design features, this paper found that the number and the types of right-turn lanes on minor roadway, the number of right-turn lanes on major roadway, the number of left-turn lane on major roadway, median on minor roadway, and intersection configuration (3 or 4 legs) are significant to effect rear-end crash occurrence. The purpose of providing a right-turn lane is to increase operational efficiency and improve safety by removing turning vehicles from through lanes; compared to the shared right-turn lane, the channelized and exclusive right-turn lanes on the minor roadway reduce rear-end crashes. Three-legged intersections tend to exhibit lower rear-end crashes than four-legged intersections. The numbers of right and left-turn lanes on the major roadway are used as surrogate variables for the magnitude of right and left-turning volumes. The higher the number of turning lanes on the major roadway the more rear-end crashes occur. The presence of medians on the minor roadway was found to reduce rear-end crashes; in comparison, Mitra et al. (2002) found wide median (>2 m) will increase rear-end crashes.

Among the traffic control and operational features, this paper confirmed that left-turn protection on the major roadway is associated with lower risks of rear-end crashes found in the previous analysis (Mitra et al., 2002). However, this paper found that protecting left-turning movement on minor roadways will increase rear-end crashes. The number of approaches with protected left-turn lanes is directly related to the number of phases per cycle; increasing the number of phases increases rear-end crashes at intersections. Poch and Mannering (1996) also found eight-phase will increase rear-end crashes. The safety advantage of traffic signal control is to reduce the frequency and severity of certain types of crashes, e.g., angle, which tend to be severe, while the disadvantage is that the left-turn protection will cause an increase in rear-end crashes (Roess et al., 2004). This paper also confirmed that the high speed limit on the major roadway is related to more rear-end crashes reached by Poch and Mannering (1996).

For the temporal analysis, the selected intersections are located in suburban areas with different population levels; and we found that intersections located in high population areas are associated with high rear-end crash frequency. Location type is found to be significant to effect rear-end crashes by Poch and Mannering (1996). But surprisingly, it was found that intersection in central business district has lower rear-end crashes in that study.

In the spatial analysis, it was found that there is high correlation between the nearest intersections along a certain corridor, and as the space gap between intersections increases, the correlation decreases. The average distance to the neighboring signals along corridors is identified to be significant to affect rear-end crash occurrence. These findings indicate that intersections along corridors affect each other and should not be considered in isolation. From the safety point of view, the intersections along corridors should be well coordinated (signal and spacing) in order to reduce rear-end crashes.

Acknowledgements

The authors wish to acknowledge the comments and suggestions of the anonymous referees. Their recommendations resulted in a substantially improved paper. We also acknowledge the financial support of the Florida Department of Transportation.

References

Abdel-Aty, M., Abdelwahab, H., 2004. Modeling rear-end collisions including the role of driver's visibility and light truck vehicles using a nested logit structure. Accident Anal. Prev. 36, 447–456.

Abdel-Aty, M., Keller, J., 2005. Exploring the overall and specific crash severity levels at signalized intersections. Accident Anal. Prev. 37 (3), 417–425.

Abdel-Aty, M., Keller, J., Brady, P., 2005a. Analysis of the types of crashes at signalized intersections using complete crash data and tree-based regression. Transport. Res. Rec.: J. Transport. Res. Board 1908, 37–45.

Abdel-Aty, M., Lee, C., Wang, X., Nawathe, P., Keller, J., Kowdla, S., Prasad, H., 2005b. Identification of intersections' crash profiles/patterns, FDOT Final Report.

Abdel-Aty, M., Wang, X., 2006. Crash estimation at signalized intersections along corridors: analyzing spatial effect and identifying significant factors. In: Proceedings of the 85th Annual Meeting of the Transportation Research Board, Washington D.C., 2006.

Ballinger, G.A., 2004. Using generalized estimating equations for longitudinal data analysis. Org. Res. Methods, 7.

Chin, H.C., Quddus, M.A., 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accident Anal. Prev. 35, 253–259.

Federal Highway Administration, 2004. Signalized Intersections: Informational Guide (Rep. No. FHWA-HRT-04-091). Washington, D.C., USDOT, FHWA, 2004.

Google Inc., 2005. Google Earth [Computer Software]. Retrieved July 17, 2005. from http://earth.google.com/.

Graham, J., 2001. Civilizing the sport utility vehicle. Issues Sci. Technol. 17 (2), 57–62.

ITS Joint Program Office, 1999. Problem Area Descriptions: Motor Vehicle Crashes – Data Analysis and IVI Program Emphasis. Washington, DC.

Kostyniuk, L., Eby, D., 1998. Exploring Rear-End Roadway Crashes from the Driver's Perspective. Human Factors Division, Transportation Research Institute, Michigan University, Ann. Arbor.

Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Lin, D.Y., Wei, L.J., Ying, Z., 2002. Model-checking techniques based on cumulative residuals. Biometrics 58, 1–12.

Lord, D., Persaud, B., 2000. Accident prediction models with and without trend: application of the generalized estimating equations (GEE) procedure. Transport. Res. Rec. 1717, 102–108.

Mitra, S., Chin, H.C., Quddus, M.A., 2002. Study of intersection accidents by maneuver type. Transport. Res. Rec. 1784, 43–50.

National Highway Traffic Safety Administration, 2006. Traffic safety facts 2004: a compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system, 2004 Final Edition. Washington, DC.

Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection-accident frequencies. J. Transport. Eng. 122, 105–113.

Ramanathan, R., 1995. Introductory Econometrics with Applications. The Dryden Press, Fort Worth, TX.

Roess, G.P., Prassas, E.S., McShane, W.R., 2004. Traffic Engineering, 3rd ed. Pearson Prentice-Hall.

SAS Institute Inc., 2004. SAS OnlineDoc® 9.1.2. SAS Institute Inc., Cary, NC.

Singh, S., 2003. Driver attributes and rear-end crash involvement propensity, national highway traffic safety administration, Report No. DOT HS 809 540.

Strandberg, L., 1998. Winter braking tests with 66 drivers, different tires and disconnectable ABS. Paper presented at International Workshop on Traffic Accident Reconstruction, Tokyo.

Wang, X., Abdel-Aty, M., Brady, P., 2006. Crash estimation at signalized intersections: significant factors and temporal effect. In: Proceedings of the 85th Annual Meeting of the Transportation Research Board, Washington D.C., 2006.

Yan, X., Radwan, E., Abdel-Aty, M., 2005. Characteristics of rear-end accidents at signalized intersections using multiple logistic regression model. Accident Anal. Prev. 37 (6), 983–995.

Zeger, S.L., Liang, K.Y., 1986. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42, 121–130.

Zheng, B., 2000. Summarizing the goodness of fit on generalized linear models for longitudinal data. Stat. Med. 19, 1265–1275.