Accident Analysis and Prevention 49 (2012) 439–448

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

On the significance of omitted variables in intersection crash modeling

Sudeshna Mitra[a,*], Simon Washington[b]

[a] Civil Engineering Department, Indian Institute of Technology, Kharagpur, West Bengal 721302, India
[b] School of Civil Engineering and Built Environment, Faculty Science and Engineering and Centre for Accident Research and Road Safety, Faculty of Health, Queensland University of Technology, 2 George St GPO Box 2434, Brisbane, Qld 4001, Australia

ARTICLE INFO

Article history:
Received 24 December 2010
Received in revised form 6 March 2012
Accepted 9 March 2012

Keywords:
Omitted variables
Spatial variables
Signalized intersection safety
Motor vehicle crashes
Crash modeling
Negative binomial
Traffic safety

ABSTRACT

Advances in safety research – trying to improve the collective understanding of motor vehicle crash causes and contributing factors – rest upon the pursuit of numerous lines of research inquiry. The research community has focused considerable attention on analytical methods development (negative binomial models, simultaneous equations, etc.), on better experimental designs (before–after studies, comparison sites, etc.), on improving exposure measures, and on model specification improvements (additive terms, non-linear relations, etc.).

One might logically seek to know which lines of inquiry might provide the most significant improvements in understanding crash causation and/or prediction. It is the contention of this paper that the exclusion of important variables (causal or surrogate measures of causal variables) cause omitted variable bias in model estimation and is an important and neglected line of inquiry in safety research. In particular, spatially related variables are often difficult to collect and omitted from crash models – but offer significant opportunities to better understand contributing factors and/or causes of crashes.

This study examines the role of important variables (other than Average Annual Daily Traffic (AADT)) that are generally omitted from intersection crash prediction models. In addition to the geometric and traffic regulatory information of intersection, the proposed model includes many spatial factors such as local influences of weather, sun glare, proximity to drinking establishments, and proximity to schools – representing a mix of potential environmental and human factors that are theoretically important, but rarely used. Results suggest that these variables in addition to AADT have significant explanatory power, and their exclusion leads to omitted variable bias. Provided is evidence that variable exclusion overstates the effect of minor road AADT by as much as 40% and major road AADT by 14%.

1. Introduction and background

Motor vehicle crash research spans a wide range of topics including analytical method improvements (such as use of count data models over ordinary least square techniques to explain random, discrete, sporadic crash occurrences; negative binomial specifications to address commonly overdispersed nature of crash data and simultaneous equation models to model endogeneity in explanatory variables, etc.), improved experimental designs (before–after studies to capture the safety benefit of implemented countermeasures, comparison sites to compare the trend in crash occurrence with sites with similar geometric and traffic characteristics, etc.), assessing alternative exposure metrics, and model specification improvements (additive terms, non-linear relations, etc.). While these topics are critical and relevant, comparatively little research has focused on the identification and inclusion of traditionally excluded or omitted variables in crash models. In particular, variables related to spatial factors are typically unavailable in crash databases and as a result have not been examined in great detail. Spatial effects have been modeled (e.g. spatial correlations), but attempts to understand and quantify the nature of these spatial effects is lacking. Studies by Anastasopoulos and Mannering (2009) and El-Basyouny and Sayed (2009) have presented random parameter models, which are able to minimize the effect of unobserved heterogeneity from potentially omitted variables within subjects (sites, intersections, etc.). While this elegant and appropriate econometric approach deals directly with omitted variable bias, it does not provide insight into what factors are missing and contributing to the unobserved heterogeneity.

There are many factors ranging from unobserved human characteristics, vehicle related attributes, and site specific factors which can influence road traffic crash occurrence, both at the segment level and at the intersection level – directly or indirectly. For typically data availability reasons, these variables are commonly omitted from crash prediction models. While the problem of

* Corresponding author. Tel.: +91 3222 283400; fax: +91 3222 282254.
E-mail addresses: sudeshna@civil.iitkgp.ernet.in (S. Mitra), simon.washington@qut.edu.au (S. Washington).

0001-4575/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2012.03.014

estimation bias due to omitted variables is of concern both for models of *road segments* and *intersections*, intersection crash models are examined in this study due to relatively poor explanatory power of intersection crash prediction models compared to the crash prediction models for road segments – as evidenced in safety literature review (Bauer and Harwood, 1996; Griebe, 2005). Moreover, intersections are operationally complex locations within the transportation system and as such warrant focused attention. At-grade intersections typically reveal that largest number of crashes (on a per unit of exposure basis) within a transportation system, with relatively large numbers of conflicts, property damage crashes, injuries, and fatal crashes.

Safety performance functions (SPFs) are developed and used to understand and predict crashes as a function of exposure (traffic volumes) and other factors. There have been considerable analytical improvements in the past few decades focusing on the appropriate specification of SPFs, including variable selection and model specification. Researchers have rejected linear regression models (Joshua and Garber, 1990; Miaou and Lum, 1993; Miaou, 1994) in favor of count and modified count models, such as zero-inflated or hurdle models (Shankar et al., 1997; Mitra et al., 2002; Kumara and Chin, 2003; Lord et al., 2004), although some questioned the use of the appropriateness of inflated models (Lord et al., 2004, 2006). Researchers have also debated the merits of panel data models (Shankar et al., 1998; Chin and Quddus, 2003), finite mixture model (Park and Lord, 2009), Markov switching models (Malyshkina et al., 2009), and so on. There is also a considerable number of studies (Hauer et al., 1988; Maher and Summersgill, 1996; Lord and Persaud, 2000; Kim and Washington, 2006; Kim et al., 2007) that have focused on specification issues with count regression models. A comprehensive recent review of the breadth and depth of methods is provided in Lord and Mannering (2010).

While it is acknowledged that accurate model predictions rely on the choice of appropriate mathematical relationships between variables and correct distributional assumptions, the selection of a comprehensive and 'correct' set of independent variables is arguable *as* important in terms of statistical properties, and *more* important in terms of interpretation and practical significance. Omission of important variables introduces bias in model parameters, and will lead to incorrect inference (Washington et al., 2010). The analysis of omitted variable bias is an important field of study in econometric research; however, in transportation safety research, the consequences of omitting relevant variables have been relatively unexplored. The primary reason for its scarcity in the literature is the practical constraint on data availability, which severely limits the number and type of variables that can be included in crash models. While in many recent studies traffic volume is the primary predictor of crash occurrence and explains the majority of variance in crashes – it is not always clear what the relevant exposure metrics should be, as pointed out by Smeed (1949) in his seminal paper "Some Statistical Aspects of Road Safety Research". In this paper, Smeed examined various types of exposure metrics such as the effect of lighting conditions, traffic volume, roadway geometric design elements, vehicle characteristics, effect of population types as well as safety propaganda effects. Thus, while traffic volume is an important factor in road safety, other geometric, environmental, and spatial factors play significant roles in influencing crash risk. Whilst the safety effects of many factors have been examined in prior research, including the effects of various geometric designs, weather effects such as rainfall, snow etc. on road segment crashes (Shankar et al., 1995), the effects of numerous spatially related factors have received relatively sparse attention. To fill this gap, the current study focuses on some important omitted variables – spatial location specific variables, and also incorporates the effects of geometric design and traffic regulatory/control related variables to compute estimation bias in coefficients of major and minor road traffic volumes.

In summary the purpose of this study is twofold:

1. To assess reasonableness of the safety effects of the spatial factors and their contribution on model estimation; and
2. To estimate the amount and direction of bias in coefficient estimates of commonly included variables, and the consequence of omission in overall prediction.

To achieve these objectives, two different models are developed, one with traffic volumes from major and minor-roads as the only exogenous variables, and a second one with a host of spatial variables in addition to commonly included geometric and traffic factors. The results of these two models are compared to test the significance of the spatial variables, their effect on crash occurrence, any evidence of bias due to the omission of important variables, and the overall improvement in model predictive capability. Marginal effects are also computed to assess the effect of various factors, including the spatial variables on crash occurrence.

2. Data description and development

The outcome measure in this analysis is total accident frequency, modeled as a function of spatial, operational, and geometric features. The remainder of this section describes how data were collected and processed, followed by a section that briefly reviews the statistical methods employed.

2.1. Data collection and processing

Variables used in modeling were obtained from six different sources, including: (a) crash data; (b) geometric data; (b) traffic volume or exposure data; (c) information about traffic control parameters; (d) spatial characteristics; (e) weather related factors; and (f) demographic data. The sites examined in the study are signalized intersections in the City of Tucson, Arizona. Intersections types include four-legged and T-junctions. After screening for the availability of traffic volume data, a sub sample to 291 signalized intersections were used. The derivation and processing of the six data sources is now described.

2.1.1. Crash data

Crash data examined were obtained from the Accident Location Identification Surveillance System (ALISS) database maintained since 1975 by the Arizona Department of Transportation (ADOT). The ALISS database contains all of the micro-level information about reported crashes, such as crash type, severity, time of occurrence, crash location and description of site, vehicle maneuvers prior to the crash, direction of movement of the vehicle prior to the crash, information about the people involved in the crash (both driver and passenger), as well as vehicle information. As with most crash databases, these data suffer from possible under-reporting of minor property damage crashes. Data on crashes that occurred from 2001 to 2004 at 291 signalized intersections in City of Tucson were collected and analyzed. Crashes were categorized as intersection-related crashes if they occurred within the curb-line limits of the intersection or if they occurred within the influence area of the intersection, defined to be within 250 ft along any leg of the intersection (from the intersection center point as has been done in previous similar studies by Bauer and Harwood, 1996; Kim and Washington, 2006). A summary of total crash is given in Table 1. Five intersections among 291 recorded zero total accidents over the 4-year period.

Table 1
Summary statistics of variables used in the model.

Variables used in variable name	Variable description	Mean	Std. deviation	Maximum	Minimum
TotCrash	Total intersection crash	53.305	44.449	246	0
Phase	Number of signal phases at the intersection	2.825	0.875	4.00	1.00
ADTMAJ	Average daily traffic from major-road	31,511.892	13,457.927	66,364.29	1617.86
LNADTMJ	Log of average daily traffic from major road	10.236	0.559	11.10	7.39
ADTMIN	Average daily traffic from minor-road	13,704.486	11,275.519	49,296.43	0.00
LNADTMN	Log of average daily traffic from minor-road	9.105	1.194	10.81	0.00
LFTMAJ	Presence of left-turn lane in major direction (1 if present, otherwise 0)	0.945	0.228	1.00	0.00
LFTMIN	Presence of left-turn lane in minor direction (1 if present, otherwise 0)	0.911	0.286	1.00	0.00
RTMAJ	Presence of right-turn lane in major direction (1 if present, otherwise 0)	0.340	0.475	1.00	0.00
RTMIN	Presence of right-turn lane in minor direction (1 if present, otherwise 0)	0.522	0.500	1.00	0.00
MediaWd	Width of median (ft)	2.476	2.864	9.43	0.00
SpdMAJ	Posted speed in major direction (mph)	37.844	5.568	55.00	25.00
SpdMIN	Posted speed in minor direction (mph)	32.062	7.299	55.00	0.00
DwnGrdMAJ	Presence of downhill grade in major direction	0.024	0.153	1.00	0.00
DwnGrdMIN	Presence of downhill grade in minor direction	0.014	0.117	1.00	0.00
EleSc1M	Presence of elementary school within 1 mile of intersection (1 if present, 0 otherwise)	0.948	0.221	1.00	0.00
EleScHM	Presence of elementary school within half mile of intersection (1 if present, 0 otherwise)	0.584	0.494	1.00	0.00
EleScQM	Presence of elementary school within quarter mile of intersection (1 if present, 0 otherwise)	0.213	0.410	1.00	0.00
MidSc1M	Presence of middle school within 1 mile of intersection (1 if present, 0 otherwise)	0.715	0.452	1.00	0.00
MidScHM	Presence of middle school within half mile of intersection (1 if present, 0 otherwise)	0.275	0.447	1.00	0.00
MidScQM	Presence of middle school within quarter mile of intersection (1 if present, 0 otherwise)	0.072	0.259	1.00	0.00
HigSc1M	Presence of high school within 1 mile of intersection (1 if present, 0 otherwise)	0.698	0.460	1.00	0.00
HigScHM	Presence of high school within half mile of intersection (1 if present, 0 otherwise)	0.340	0.475	1.00	0.00
HigScQM	Presence of elementary school within quarter mile of intersection (1 if present, 0 otherwise)	0.162	0.369	1.00	0.00
ColUni1M	Presence of college or university within 1 mile of intersection (1 if present, 0 otherwise)	0.392	0.489	1.00	0.00
ColUniHM	Presence of college or university within half mile of intersection (1 if present, 0 otherwise)	0.134	0.341	1.00	0.00
ColUniQM	Presence of college or university within quarter mile of intersection (1 if present, 0 otherwise)	0.034	0.182	1.00	0.00
Pubs1M	Number of pubs within 1 mile of intersection	6.619	9.168	31.00	0.00
Pub1Mind	Presence of pubs within 1 mile of intersection (1 if present, 0 otherwise)	0.756	0.430	1.00	0.00
PubsHM	Number of pubs within half mile of intersection	2.643	5.669	23.00	0.00
PubHMInd	Presence of pubs within half mile of intersection (1 if present, 0 otherwise)	0.457	0.499	1.00	0.00
PubsQM	Number of pubs within quarter mile of intersection	0.852	2.420	23.00	0.00
PubQMInd	Presence of pubs within quarter mile of intersection (1 if present, 0 otherwise)	0.271	0.445	1.00	0.00
Pubs5M	Number of pubs within 5 mile of intersection	70.646	31.773	118.00	0.00
Pub5Mind	Presence of pubs within 5 mile of intersection (1 if present, 0 otherwise)	0.983	0.130	1.00	0.00
AvgPrcp	Average precipitation near the intersection	902.575	56.080	1064.333	51.67
AvgRDay	Average number of rainy day near the intersection	48.061	4.411	56.080	4.41
POPTOT	Total population near the intersection	1256.124	1318.717	11,190.00	168.00
POPURB	Total population in urban area near the intersection	1217.478	1287.602	11,100.00	168.00
POP00_15	Total population from age 0 to 15 years old near the intersection	293.591	420.906	3450.00	0.00
POP16_64	Total population from age 16 to 64 years old near the intersection	821.052	859.494	7322.00	82.00
POP65	Total population over 65 years old near the intersection	141.473	119.552	618.00	5.00

2.1.2. Geometric data

Geometric data for the signalized intersections were obtained from a City of Tucson maintained database. This database provides information about the names of the cross roads of the intersections, the unique intersection IDs, and the direction of major-roads, which helps to identify the major and minor approaches. It also contains information such as number of lanes, the presence, type, and width of the left-turn bays, widths of the through lanes, and widths of the right-turn bays, width and type of medians, approach speeds, and approach grades. Finally, it contains information about pedestrian crossings, presence of reversible lanes, and unusual configurations such as one-way streets. A list of geometric variables used in the study is shown in Table 1.

2.1.3. Traffic data

The only traffic control information available from the city was the number of intersection signal phases. Traffic regulatory factors included speed limits of major and minor-roads. Acquiring quality traffic volume data presented a significant challenge. Annual traffic volume data for each intersection for each year were unavailable. As a result, estimated average traffic volumes, using data from 2001 to 2003 were used, resulting in measurement errors in volume estimates, an unfortunately all too common limitation associated with these data. Summary statistics of traffic control and volume related variables are also found in Table 1.

2.1.4. Spatial variables

The spatial variables employed in the modeling are intended to represent commonly omitted factors in crash models, as discussed previously. The data are designed to capture the potential spatial effects of weather (e.g. sun glare, rainfall) and driver behavior (e.g. drunk driving). Of course, weather effects are not only spatial but also temporal in nature. While the temporal and directional nature of sun glare on drivers was taken into account by considering the time of crash occurrence and direction of travel, the temporal variation in rainfall was ignored and aggregated, assuming to have reasonably modest impact within a desert city like Tucson, Arizona. The spatial variables are meant to reflect what might be argued as truly causal effects such as sun glare, which can occlude driver visibility sufficiently to cause a crash. A surrogate measure meant to reflect high probability of impaired driving is measured through a surrogate by measuring proximity to drinking establishments. The underlying assumption of this variable is that it is more likely to observe an impaired driving crash in close proximity to establishments that serve alcohol for on premises consumption. It is presumed in this research that the spatial variables are imperfect measures of the underlying causal mechanisms, potentially capturing to varying degrees the causal underlying processes.

2.1.5. Sun glare

Among the different types of spatial variables developed in this research, adequately capturing the effect of sun-glare was the most

significantly challenging. Andrey et al. (2001) reported that Sun glare is a potential environmental factor having strong effect on driving performance; however, very few researchers have investigated the effect of sun and windshield glare on crash occurrence. A study conducted in the UK by Broughton et al. (1999) focused on the effect of daylight change on crash occurrence, while a study by Flahaut (2004) examined the effect of sun glare on crash occurrence. Sun glare is most problematic during early or late hours of the day, presumably within an hour after sunrise and before sunset, when the sun is on the immediate horizon. As one might expect, weather, trees, and hills will play a role on the effect of sun glare, as well as the direction of vehicle travel. Also, time of year plays an important role in the intensity of glare from sunlight and the 'critical' hour in which glare is a potential issue. In Tucson, AZ, sun glare is especially bad in early fall and early spring when the sun rises almost exactly east and sets almost exactly west, because the Tucson road network is on a N–S E–W grid system, is relatively flat, and lacks clouds for about 350 days per year. Given these complexities, it is not surprising that the potential effect of sun glare has not previously been examined or understood.

Fortunately, mathematical expressions are available to estimate the intensity of glare from a light source. Without going into great detail, glare is measured as an equivalent veil luminance and depends on the angle of glare, the angle made by the light source, and the direction of vision. In the case of sun glare, this angle is the angle of sun as seen by the motorist. The sun's position relative to an observer depends on the latitude and the longitude of the observer's position on earth and the time of year, and the effect of sun glare varies by time of day across months throughout the year. To assess the potential role of glare in crashes, it was necessary to distinguish between those crashes that might occur during morning and evening potential glare periods from crashes that occur during periods without potential glare. Moreover, only drivers of vehicles on particular routes will be exposed to potential glare effects, thus both time and route specific groups needed to be distinguished. To accomplish this categorization of crashes, 'critical' times for potential glare, a varying 'glare window' was developed and linked with the time and day and route of each crash. The methodology for developing this metric is now described.

The National Oceanic and Atmospheric Association (NOAA) provides sunrise and sunset times during 365 days of the year for various cities in the USA. Using this source, sunrise and sunset times were obtained for 12 months in Tucson, where hour 'windows' immediately after sunrise and prior to sunset are calculated and shown in Table 2. The total number of crashes during these specified intervals was calculated separately. Then for each intersection, two different crash counts were considered: (1) total crashes during the non-glare time period, i.e. over 22 h of time per day for the 4-year period; and (2) crashes during the 2 h; i.e. the morning and the evening glare period per day for 4 year period. In addition, indicator variables were created to indicate if crashes were observed during glare or non-glare periods for each intersection. As stated previously, sun-glare occurs on an eastbound approach at sunrise and a westbound approach at sunset – and so travel direction was taken into account. Finally, an offset variable is used to take into account the unequal period of observation, i.e. 22 h without potential glare and 2 h time periods with potential glare. Using this methodological approach, the time of observation of crashes is used as a proxy of exposure to glare related crashes. In contrast, a superior approach would be to measure or estimate traffic volumes during glare and non-glare times to use as exposure metrics associated with these crashes. The absence of traffic volume data during these two time periods restricted the ability to adopt this approach, leading to the use of a time proxy of exposure.

Table 2
Typical morning and evening sun glare window.

Months	Morning window (AM)	Evening window (PM)
January	7.30–8.30	4.30–5.30
February	7.30–8.30	5–6
March	6.45–7.45	5.30–6.30
April	6–7	5.55–6.55
May	5.30–6.30	6–7
June	5.20–6.20	6.30–7.30
July	5.25–6.25	6.40–7.40
August	5.40–6.40	6.20–7.20
September	6.10–7.10	5.45–6.45
October	6.30–7.30	5–6
November	6.50–7.50	4.30–5.30
December	7.20–8.20	4.20–5.20

2.1.6. School location

Two additional potential spatial factors examined were school zones and alcohol dispensing locations such as bars and pubs. To process school-zone related data, first the GIS layer containing all types of schools in Arizona was obtained. This layer contained the geographic location of schools along with other attributes such as the type of school, name of school, and school addresses. From this layer several new layers were created based on the school types including elementary, middle, and high schools, as well as colleges and universities. Each of these layers was then mapped to the intersection GIS layers to develop indicator variables for intersections to denote proximity. Indicator variables were created such that intersections that fell within ¼ mile, ½ mile and 1 mile radii of each school type were coded – resulting in 3 (radii) × 4 (school types) = 12 indicator variables. These indicator variables were created within the GIS platform using the "nearest-neighbor" analysis algorithm. This algorithm 'looked' for the presence of specific types of schools within a specified search radius around each intersection in the dataset and created tables showing the distance of the schools from each intersection. From these tables indicator variables were created, with the intent to empirically explore potential spatial effects of schools zones and alcohol dispensing establishments. Summary statistics of these variables are shown in Table 1.

2.1.7. Location of drinking establishments

While the locations of schools were obtained from the GIS layer, finding the locations of drinking establishments was more difficult. First, the availability of a GIS layer containing all alcohol dispensing locations did not exist. Second, the available GIS tiger files (extracts of selected geographic and cartographic data from the United States Census Bureau's TIGER or Topologically Integrated Geographic Encoding and Referencing System and is used to support mapping of points, lines, polygons on census maps) show the locations of establishments with liquor licenses; however, these may not be locations where people consume alcohol. For example, many supermarkets possess liquor licenses, however it is unlikely that alcohol purchased at these locations will be consumed on site. To deal with this problem, addresses of bars and pubs in Tucson were identified from the yellow pages, which provide business listings along with addresses and phone numbers. Then a geocoding service within GIS was used to locate those addresses on an Arizona street map and a new layer was created showing the location of bars and pubs in Tucson. This layer was then used to develop a 'proximity to bars' indicator variables representing the presence of bars and pubs within ¼ mile, ½ mile, 1 mile and 5 miles of the intersections. A search radius of 5 miles was used based on the assumption that drunk drivers might drive a considerable distance before they become involved in a crash and it is not known a priori from theoretical or empirical knowledge whether the effects might be localized or randomly spread over the transportation network. Also, drinking locations are generally clustered in a region, and it

is not known whether the number of bars near a location or just the presence of bars affect safety. Consequently, two types of variables for each search radius were created: (1) total number of bars and pubs within a search radius, and (2) the presence or absence of drinking locations within a search radius. The summary statistics of these variables are shown in Table 1.

2.1.8. Weather data

To account for potential weather effects, weather data at various weather stations in the State of Arizona were obtained from National Oceanic and Atmospheric Association (NOAA). While the database provides information about hours of sunlight, cloudiness, temperature, precipitation and snow fall, information about precipitation was only available for most of the weather stations and for this reason only precipitation related variables were developed. The database contained daily precipitation at different weather stations for all days of the month and for 12 months of the year along with the recording time of precipitation. However, some weather stations had recorded data for less than 12 months of a year, thus only the available month's data were considered in the analysis. Two different precipitation variables were created: *total average precipitation per year* and *Average number of rainy days per year*.

After weather stations were located using their latitude and longitude within GIS platform, a "nearest-neighbor" analysis was performed. However, large search radii were chosen because relatively few weather stations were available within the analysis region. Radii were selected such that all of the intersections were assigned weather attributes. A summary of the weather related variables are shown in Table 1.

2.1.9. Demographic data

Finally, the socio-demographic attributes of the local population living near intersections were taken into account by using census track population data. The census database contains distribution of population in a region in GIS layers. This population GIS layer was joined with intersection locations to identify the distribution of population around intersections – indicating a local driving population. A summary of demographic data is shown in Table 1.

3. Statistical model selection and development

Poisson and Negative Binomial (NB) panel models were used to capture unobserved intersection-specific effects. Fixed effects models address unobserved intersection-specific effects by using indicator variables for specific intersection. An alternative is to use a random effects model which assumes that the unobserved intersection effects are distributed across the population of intersections according to some pre-determined distribution (such as normal distribution). Preferring this latter approach as a more flexible formulation, random effects negative binomial (RENB) models were applied. Additional background on the Poisson, NB and RENB models is provided in Washington et al. (2010) and Cameron and Trivedi (1998).

Poisson and Negative Binomial (NB) panel models were used to capture unobserved intersection-specific effects. Although count data modeling typically begins with a Poisson regression specification, often data are over dispersed (i.e. variance > mean) and thus a NB models are more suitable. The NB generally used for such cases as it includes a gamma-distributed error term in Poisson mean to account for a wide range of unobserved heterogeneity such as omission of relevant variables, measurement error, or just the intrinsic randomness in count data. However, it is always recommended to give due care to correctly specify Poisson mean, as much as possible and not just resort to the NB over dispersion parameter to account for all misspecification and unobserved heterogeneity. As shown by Mitra and Washington (2007), in the presence of a well-defined mean function, the extra-variance structure of NB generally becomes insignificant, which helps in better inference by reducing standard error of estimation and forming a narrow confidence interval. To capture unobserved site specific heterogeneity Shankar et al. (1998) used random effects negative binomial (RENB) model to find factors that influence median cross-over accidents. Miaou et al. (2003) also used random effects model in Bayesian framework to model area level crash prediction and mapping. Among the other methods to account for heterogeneity, works by Anastasopoulos and Mannering (2009) and El-Basyouny and Sayed (2009) are noteworthy. These authors used random parameter count models to account for individual heterogeneity. The advantages of the random parameter model are that all parameters in the mean function are allowed to vary randomly and that cross sectional data is sufficient to capture such randomness. On the other hand fixed effects and random effects models are applied on panel data. The major difference between random effects and random parameter models is that – a random effects model is a special case of a random parameter model where the constant term is *only* modeled as a random parameter as opposed to all parameters in mean function. In keeping, both random parameters and random effects models are specified and compared in this research. Additional background on random parameter models and their estimation details may be found in Greene (2007). For further detail on Poisson, NB and RENB models reader are referred to Washington et al. (2010) and Cameron and Trivedi (1998).

As mentioned previously one of the aims of this research is to assess the potential importance of a host of typically omitted spatial factors on intersection crash occurrence. As mentioned by Cameron and Trivedi (1998), the general consequence of measurement errors in a non-linear count model are similar to heterogeneity that results from over-dispersion. Measurement errors in regressors – such as poorly measured traffic volume data – can arise in various ways. For example, there could be multiplicative or additive errors in the measurement of exogenous variables as well as errors due to *omission* of relevant and important covariates – the focus here. In the case of a linear model with ordinary least square (OLS) estimated parameters, omitted variables cause bias in coefficient estimates of included variables and estimators are not consistent and are biased (Washington et al., 2010). In the case of non-linear count models, the omission of relevant variables can be interpreted as unobserved heterogeneity *only* in cases where the omitted regressors are uncorrelated with included covariates. Suppose that included covariates are X and omitted covariates are Z; and β and γ are the vectors of parameters associated with X and Z respectively. Then the expected number of crashes can be written as:

$$\mu_i|X_i, Z_i = \exp(X_i'\beta + Z_i'\gamma) = \exp(X_i'\beta)\exp(Z_i'\gamma) = \exp(X_i'\beta)u_i \quad (1)$$

Here $u_i = \exp(Z_i'\gamma) = \exp(\varepsilon_i)$ is algebraically similar to the error component or the gamma heterogeneity term included while developing the NB model. Hence, the consequences of omitting important variables are essentially equivalent to those due to unobserved heterogeneity, i.e. over-dispersion and loss of efficiency of the pseudo maximum likelihood estimator (Cameron and Trivedi, 1998).

To test the potential significance and importance of omitted spatial variables, two different model specifications are estimated and compared. In the baseline model only traffic volume related variables are included, while in a second model traffic volumes as well as geometric and spatial variables are included. If X is a vector of commonly included variables, and Z is a vector of commonly omitted factors, then two different mean functions are specified and compared, whereby:

$$\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_k Z_{ik} \quad (2)$$

and

$$\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_k Z_{ik} \quad (3)$$

where β's are vector of covariates associated with covariate vector X and γ is a vector of covariates associated with Z. To test the significance of omitted variables Z, the hypotheses are:

$$H_0: \quad \gamma = 0 \quad \text{and} \quad H_1: \quad \gamma \neq 0 \quad (4)$$

The significance of the coefficients is assessed using a t-test, where the test statistic is

$$t = \frac{\gamma'}{se(\gamma')} \quad (5)$$

which is approximately student t-distributed. Recall that γ' is a vector of coefficient estimates of γ from Eq. (3), and $se(\gamma')$ is the vector of estimated standard errors of the coefficients. A significance level of 0.05 is used for the t-test for this study.

If the test results indicate that the omitted variables are significant and the model with omitted variables offer significantly improved fit, then an estimate of the change in coefficient estimates for β's from Eqs. (2) and (3) is worthwhile. This change in coefficients is calculated as a percentage change in the estimates, with a higher estimate of β from Eq. (2) indicating a positive bias, and a lower estimate of β compared to Eq. (3) indicating a negative bias.

To measure the overall goodness of fit, the log-likelihood ratio index (ρ^2) is calculated. As given in Washington et al. (2010), the ρ^2 statistic is

$$\rho^2 = 1 - \frac{L(\boldsymbol{\beta})}{L(0)} \quad (6)$$

where $L(\beta)$ is the log-likelihood value of the fitted model at convergence, and $L(0)$ is the initial log-likelihood with all parameters set to zero except the constant term. The value of ρ^2 is bounded by 0 and 1, where a perfect model will have a likelihood function equal to one.

Finally, to determine the relative impact of various independent variables on crash frequency, marginal effects are computed. In non-linear regression models such as Poisson or NB, the marginal effects provide an estimated change in the number of crashes given a unit change in the independent variables, and is calculated as the partial derivative, $\partial\mu_i/\partial x$, where μ_i is the expected number of crashes and x is one of the independent variables for which the marginal effect is computed. In this study a comparison of marginal effects gives an indication of the relative importance of the independent variables in influencing intersection crashes.

4. Results and discussion

The statistical models described previously are estimated using LIMDEP econometric software. The "best" models were selected among numerous competing models, considering the usual attributes such as model goodness of fit (GOF), theoretical appeal of variables, agreement with expectation, etc. While traffic volume variables are always included in the models, two different candidate models are developed with and without spatial variables-enabling assessment of the relative impacts of the spatial factors and estimates of bias caused by variable omission.

4.1. Total crashes at intersections

The estimation results for total crash models are presented in Tables 3 and 4, showing the results for the RENB and random parameter negative binomial specifications for the two specifications respectively. The random parameter model results reveal that the effect of the variables "posted speed on the minor road" and "population between 0 and 15" are to be random and rest of the variables to be fixed across intersections. The overall fit of the random parameter model was slightly better with $\rho^2 = 0.717$ compared to $\rho^2 = 0.715$ in the RENB model. While the coefficient estimates from these models are similar and comparable, the major differences are in coefficient estimates of the effect of "posted speed on the minor road" and "total population between 0 and 15" near to the intersection. As discussed in Anastasopoulos and Mannering (2009), a random parameter should be retained in the specification when the standard deviation of the parameter density is statistically different than zero. If a parameters' estimated standard deviation is not statistically different from zero, then the parameter is fixed across intersections. In the random parameters specification the variables "posted speed on the minor road" and "total population between 0 and 15 near the intersection" were found to be significant, suggesting that the effect of these variables varies across intersections.

The full model with traffic, geometric, weather, spatial and demographic covariates revealed a total of fourteen significant variables in estimating total expected intersection crashes. Similar to prior studies and not surprisingly, traffic volumes are the most important and reliable predictors of intersection crashes and are positively associated with total crashes (p-values <0.0001 for major and minor-roads), reflecting increased crash risk with increasing exposure. The number of signal phases is positively associated with crashes (p-value <0.0001). While the effects of number of intersection signal phases have been investigated in previous studies (Chin and Quddus, 2003; Bauer and Harwood, 1996; Poch and Mannering, 1996; Mitra et al., 2002; Wang and Abdel-Aty, 2006) a clear and established relationship between safety and the number of signal phases has not emerged. It is presumed as Poch and Mannering (1996), Mitra et al. (2002) and Chin and Quddus (2003) have postulated, that a higher number of signal phases indicates greater complexity of traffic movements within an intersection, and that crash risk is elevated immediately following phase changes. The presence of turn lanes was not found to be statistically significant except for left-turn lanes on the major-roads (p-value <0.0001). The presence of a left-turn lane on the major-road is associated with a higher number of crashes at intersections, which is not surprising given the possible endogeneity of this variable (Kim and Washington, 2006), the complexities involved with isolating this effect, and the mixed results found in prior research. Wang and Abdel-Aty (2006) observed similar effects for rear-end crashes at signalized intersections.

The effects of the posted speed limit on intersection approaches are interesting. The effect of posted speed along the minor road was best fit with a normally distributed random parameter, with a mean 0.0196 and standard deviation 0.0009. This indicates that an increase in minor road posted speed will generally increase intersection crash frequency but with a varying magnitude across intersections. While higher minor-road speed limits are associated with increased intersection crashes (p-value <0.0001), higher major-road speed limits are negatively associated (p-value <0.0001). These findings should be considered in tandem. It is likely that the effects of posted speed are residual effects of differences between actual and posted speeds, as well as the effects of different design standards of facilities associated with posted speeds. The results seem to suggest that large design speed differences between major and minor roads are more problematic than roads with similar design speeds; however, more work is needed to verify this conclusion.

The remaining significant variables in the model are related to spatial, demographic, and weather effects proximal to intersections. The presence of colleges or universities within half of a mile of the intersection (p-value <0.001) is associated with an increase in total crashes. It is not surprising that this variable has a positive

Table 3
Results for total intersection crashes.

Variables	RENB model with simulated MLE			Random parameter NB model		
	Estimated coefficient	*t*-Statistic	*p*-Value	Estimated coefficient	*t*-Statistic	*p*-Value
Constant	−10.567	−14.535	<0.001	−10.008	−12.002	<0.0001
Standard deviation of random parameters	0.6451	93.280	<0.001	0.6057	89.004	<0.0001
Log of AADT on the major-road	0.6136	36.741	<0.001	0.6126	36.323	<0.0001
Log of AADT on the minor-road	0.1855	24.926	<0.001	0.1873	25.390	<0.0001
Number of phases at the intersection	0.1714	22.28	<0.001	0.2239	29.392	<0.0001
Left-turn lane indicator (1 if at least one left-turn lane on the major-road, 0 otherwise)	0.5393	15.305	<0.001	0.6178	17.928	<0.0001
Posted speed on the major road (mph)	−0.0241	−13.286	<0.001	−0.0253	−14.062	<0.0001
Posted speed on the minor road (mph)	0.0250	18.014	<0.001	0.0196	13.739	<0.0001
Standard deviation of random parameters				0.0009	5.712	<0.0001
College or university within half mile indicator (1 if at least college or university within ½ mile of the intersection, 0 otherwise)	0.3948	24.291	<0.001	0.3843	23.797	<0.0001
Bars within quarter mile indicator (1 if at least one bar or pub within ¼ mile of the intersection, 0 otherwise)	0.2936	22.200	<0.001	0.3186	24.092	<0.0001
Bars within five miles indicator (1 if at least one bar or pub within 5 miles of the intersection, 0 otherwise)	2.1704	9.545	<0.001	3.4455	9.489	<0.0001
Total population between age 0 and 15 near the intersection	0.0016	11.026	<0.001	0.0004	8.764	<0.0001
Standard Deviation of random parameters				0.0006	33.119	<0.0001
Total population between age 16 and 64 near the intersection	−0.0013	−16.157	<0.001	−0.0004	−15.011	<0.0001
Average annual precipitation near intersection	−0.0019	−11.916	<0.001	−0.0018	−11.266	<0.0001
Average annual number of rainy days near intersection	−0.0226	−15.534	<0.001	−0.0287	−19.436	<0.0001
Glare indicator (1 if crash occurred during glare period, 0 otherwise)	2.2095	14.184	<0.001	2.202	7.842	<0.0001
Number of observations	582			582		
Number of groups	291			291		
Log-likelihood at zero	−6738.01			−6738.01		
Log-likelihood at convergence	−1916.08			−1909.02		
ρ^2	0.715			0.717		

Table 4
Results for total intersection crashes with traffic volume only.

Variables	RENB model with simulated MLE			Random parameter NB model		
	Estimated coefficient	*t*-Statistic	*p*-Value	Estimated coefficient	*t*-Statistic	*p*-Value
Constant	−7.0284	−13.867	<0.001	−7.0284	−13.867	<0.0001
Standard deviation of random parameters	0.086	2.427	0.0152	0.086	2.427	0.0152
Log of AADT on the major-road	0.7149	15.008	<0.001	0.7149	15.007	<0.0001
Standard deviation of random parameters				0.0001	1.031	0.4756
Log of AADT on the minor-road	0.3102	30.556	<0.001	0.3102	30.556	<0.0001
Standard deviation of random parameters				0.0001	1.021	0.4835
Number of observations	582			582		
Number of groups	291			291		
Log-likelihood at zero	−6738.01			−6738.01		
Log-likelihood at convergence	−2935.03			−2935.03		
ρ^2	0.564			0.564		

effect on total crash occurrence, since colleges and universities are locations typically associated with higher proportions of inexperienced drivers, multiple modes of travel, complex motor vehicle movements, and impaired persons. The proximity of bars and pubs near an intersection is also significant and is associated with an increase in total predicted intersection crashes. The findings suggest that the effects of bars and pubs on total crashes are either localized within a quarter mile of intersections (p-value <0.0001) or are far away within 1–5 miles of the intersections (p-value <0.0001). This finding, though not conclusive, suggests that the proximity of drinking establishments nearby to intersections increases the number of expected crashes, presumably due to a greater-than-average number of intoxicated pedestrians, bicyclists, and drivers near these locations. It is possible, however, that the 'bar effect' is picking up some other land-use related unknown effects – an effect that is correlated with the presence of alcohol dispensaries. It remains as future work to investigate the effects of BAC level on crash occurrence and/or the proximity of alcohol related crashes with respect to these establishments, but disaggregate data were not available for this study.

The socio-demographic patterns proximal to intersections are associated with total crash frequencies. Intersections near to proportionately larger populations between 0 and 15 years of age (p-value <0.0001) revealed greater numbers of crashes, and intersections near to proportionately larger populations aged between 16 and 64 observed fewer crashes (p-value <0.0001). However, the variable "population between 0 and 15 years" resulted in a normally distributed random parameter, with a mean 0.0004 and standard deviation 0.0006, indicating a random effect across intersections. The effect for the population group above 65 years of age was not significant.

Finally, the weather variables such as annual average precipitation (p-value <0.0001) and annual average number of rainy days (p-value <0.0001) had negative relationships with crash occurrence. Rain in Tucson is quite rare, with about 12 in. of rainfall per year on average. Also, the rainy season is summer, when a significant portion of the driving population leaves Tucson (many people with second homes in Arizona leave during summer). Thus, it is likely that for these Tucson data the rain effect is confounded with other exposure metrics and is not generalizeable to other regions;

however, rain effects may be revealed by other datasets and should be examined further.

The indicator variable for the effect of glare (p-value <0.0001) is positively correlated with crash occurrence. The finding suggests that the presence of sun glare increases traffic crashes – obstructing the normal visibility to drivers after sunrise for east bound traffic and before sunset for west bound traffic when the sun is near the horizon. However, the potential limitation of this finding is in how this variable was coded – it is possible that this variable may be partially capturing the effect of congestion, since peak times of glare coincide with peak period traffic (see Table 2). However, given that only eastbound traffic movements were considered for sunrise effects and westbound traffic were considered for sunset effects, directional effects typically associated with peak periods would be offset. Thus, it is quite likely that the observed effects of glare represent to some degree an increased crash risk during periods of glare conditions.

A comparison of the total crash model to the model with traffic volumes as the sole predictor of crashes shows a lesser overall fit as expected, of about $\rho^2 = 0.564$ for both RENB and random parameter model. However, as was previously identified, the model with geometric and spatial variables reveals ρ^2 of 0.717 and 0.715 in RENB and random parameter models respectively. These comparisons suggest that spatial and geometric variables explain a significant portion of the variability in crashes at intersections. The goodness of fit of these models also indicates that in the absence of geometric and spatial variables, both the random parameter and the RENB model performed almost same, at least for the Tucson dataset examined here.

A comparison of the two models – one with geometric, traffic operation, and spatial significant variables included (Table 3) and one with only traffic volumes (Table 4) reveals some important differences. The presence of spatial and geometric design related factors changed the magnitudes of the coefficient estimates of major and minor road traffic volumes. With other factors omitted, the estimated effects of major road and minor road traffic volumes on intersection crashes gives coefficient estimates of 0.7149 for major road and 0.3102 for minor road traffic from both the RENB and random parameter models (this is due to the fact that the standard deviation of the parameter estimates for major and minor road traffic volumes in the random parameter model are not significant, treating them as fixed parameters). In contrast these estimates are 0.6126 and 0.1873 respectively for major and minor road AADT in full the model with relevant variables as shown in Table 3. The comparison of marginal effects reveals essentially the same trend with estimates of 7.64 for major road and 2.33 for minor road from the random parameter model, where these values are 8.92 and 3.87 for major road and minor road traffic from the AADT only model. This difference clearly indicates that in the absence of omitted relevant variables, the coefficient estimates and the marginal effect estimates for major road and minor road volumes are biased upwards by an average of 14% and 40% respectively. These biased coefficients have clear implications for road safety – reflecting that omitting variables results in biased estimation of model parameters, and that crash predictions are inaccurate and incorrect.

To measure the relative influence of various independent variables on intersection safety, marginal effects for both fully specified model and AADT only model are calculated and shown in Table 5 along with 95% confidence intervals. Marginal effects are estimates of the change in the dependent variables due to unit changes in the independent variables, with all other variables computed at their means. In case of continuous independent variables the marginal effect is the partial derivative taken at mean, but in case of a categorical variable it is the partial derivative for a discrete change of the dummy variable from 0 to 1. A comparison of marginal effects reveals that the indicator variable for glare has the largest effect. While this variable may partially capture congestion effects, there is ample evidence provided here that glare plays a significant role in intersection crashes.

In agreement with numerous prior studies, traffic volumes on major and minor approaches to intersections are influential and positively correlated with crash occurrence. Since the log of major and minor road traffic volumes were used in modeling, the marginal effects for the variable LNADTMJ or log of major-road AADT and LNADTMN or log of minor-road AADT from fully specified model are 7.64 and 2.33 respectively. These effects translate to a 0.24 increase in intersection crashes for every thousand increase in major road AADT and an increase of 0.17 intersection crashes per 1000 increase in minor road AADT. Thus intersection crashes increase 1.4 times faster per AADT on the major compared to an equivalent AADT increase on the minor road. However, the marginal effects from the AADT only model reveal values of 8.92 and 3.87 for LNADTMJ and LNADTMN respectively. These values translate to a 0.28 increase in intersection crashes for every thousand increase in major road or minor road AADT. Thus, with spatial and geometric factors omitted, both major road and minor road traffic have same effect on intersection crashes. The number of signal phases was positively correlated with total crashes, reflecting that larger numbers of signal phases are required at busy and complex intersections with higher crash risk. The marginal effect suggests an increase of 2.22 crashes per unit increase in the number of signal phases.

Among geometric design elements, the presence of turn lanes revealed a significant influence on safety. Numerous prior studies have observed mixed effects of left-turn lanes on safety, in part

Table 5
Marginal effects of covariates for total crash model.

Variables	Random parameter fully specified model		Random parameter AADT only model	
	Marginal effect	95% Confidence interval	Marginal effect	95% Confidence interval
Log of AADT on the major-road	7.64	(5.68–9.69)	8.92	(5.51–11.54)
Log of AADT on the minor-road	2.33	(1.26–3.38)	3.87	(2.57–5.46)
Number of phases at the intersection	2.22	(1.14–3.36)		
Left-turn lane indicator[a]	4.89	(1.88–7.76)		
Posted speed on the major road	−0.267	(−0.536 to −0.022)		
Posted speed on the minor road	0.278	(0.098–0.476)		
College or university within half mile indicator[a]	3.79	(0.487–7.09)		
Bars within quarter mile indicator[a]	4.28	(1.73–6.84)		
Bars within five miles indicator[a]	11.83	(10.02–13.76)		
Total population between age 0 and 15 near the intersection	0.0069	(0.0041–0.0079)		
Total population between age 16 and 64 near the intersection	−0.0041	(−0.0078 to −0.0005)		
Average annual precipitation	−0.022	(−0.042 to −0.002)		
Average annual number of rainy days	−0.271	(−0.503 to −0.040)		
Glare indicator[a]	27.20	(25.71–29.49)		

[a] Indicator variables.

due to the potential endogeneity of this variable. The results here suggest that the presence of left-turn lanes is associated with 4.89 more intersection crashes compared to intersections without left-turn lanes. The presence of other turning lanes did not found to have an estimated impact on intersection crashes. The marginal effects of posted speed on the major road and minor roads are −0.267 and 0.278, indicating that an increase of 10 mph posted speed along major road will decrease intersection crashes by 2.67, whereas the same increase on the minor road will increase crashes by 2.78. As discussed in Oh et al. (2004), the safety impact of actual versus posted speeds can be different, and thus this finding is confounded with omitted operational and design factors. The results are not independent and must be interpreted together, with trade-offs being made between major and minor approach posted speeds. Moreover, the posted speed of a minor road is found to be positively correlated with minor road AADT (correlation coefficient is 0.489), which indicates that omission of this variable may result in biased estimates of minor road AADT effects.

While the impact of geometric and traffic factors are well studied in prior traffic safety research, the impacts of spatial factors have to date been relatively unexamined. Fundamental theory justifies the desire to examine spatial effects, but ecological fallacy and confounding may play a role in the modeling results. While a variety of statistically significant spatial effects are revealed, the results are in need of validation by other researchers and improved spatial variable formulation.

A total of 12 spatial variables were examined to capture possible effects of elementary, middle, and high schools as well as colleges and universities on signalized intersection crash risk. The analysis reveals that the presence of colleges and universities within a half mile of intersections is associated with a 3.79 factor increase in total crashes.

While blood alcohol concentration (BAC) is known to influence traffic safety, this study investigated the spatial effects of nearby drinking establishments and intersection crash risk. Intersections with bars and pubs within a quarter mile observed about 4.28 more total crashes than otherwise similar intersections. The spatial scale of this variable, however, might be confounded with other urban effects such as the presence of tall buildings, complex intersections, many modes of travel, etc., and thus may suffer from ecological correlation. The effect of population near an intersection indicates an increase of 6.9 crashes per 1000 increase in population aged 0–15 years, and a reduction in total crashes of 4.1 per 1000 increase in the population group 16–64.

Inspection of the effect of precipitation on intersection safety indicated that total crashes are inversely related with increases in both annual average precipitation and annual average number of rainy days. The average marginal effects are −0.02 and −0.27 for average precipitation and average number of rainy days, i.e. with a 1 in. increase in average precipitation and a one additional rainy day, there was an observed reduction in total intersection crashes by 0.02 and 0.27 respectively. While precipitation also has revealed mixed effects on safety as identified by Shankar et al. (1995) and Zhang and Holm (2004), the results here suffer from highly aggregated precipitation data and a decreased driving population during the rainy season in the sample data. More research is needed to more adequately examine the effect of precipitation on intersection crashes.

5. Conclusions and recommendations

While prior research has examined the effect of spatial factors at the county or census-track levels, this effort uniquely incorporates spatial variables in intersection crash models. Confounding and ecological correlation effects cannot be ruled out in some of the spatial factors examined, and so conclusions are meant to stimulate further inquiry rather than representing unequivocal evidence on the potential importance of omitted spatial factors. Given these caveats, the following conclusions are offered:

1 Empirical evidence provided here suggests that spatial effects may be important to understanding crash risk at intersections. Logic dictates that the omission of spatial effects on roads would yield similar results and also represents an opportunity for further research.
2 The sample here provides empirical evidence only, and the results reflect the outcome in one jurisdiction and cannot be generalized to other locations. Thus, the magnitudes and in some cases direction of effects may not be generalizable to other locations or jurisdictions.
3 Omission of spatial variables result in biased estimates of retained variables. Results from this study indicate that coefficient estimates of both major road and minor road traffic volumes are biased upwards, i.e. overstated in the absence of other intersection specific factors. The amount of bias is as much as 40% for minor road AADT and 14% for major road AADT. These results highlight the importance of including important omitted variable – spatial and otherwise, but do not suggest that the bias found here would be of similar magnitude or direction in other locations.
4 Intersections near special traffic generators, such as schools and universities; and clusters of drinking establishments observed higher numbers of crashes (after controlling for AADT, lanes, etc.); hence attention should be given to capture the effects of such locations while comparing candidate intersections for safety improvements.
5 Periods of extreme sun glare can drastically reduce the ability to safely operate a motor vehicle. Some drivers routinely commute during periods of extreme sun glare. Some road networks and weather patterns are more susceptible to sun glare related problems, particularly eastbound morning travel and westbound afternoon travel during periods with a clear weather. It should not be surprising that roadways carrying a significant portion of traffic during these times would observe greater numbers of crashes, and that glare is as important as found here. The finding requires further validation with further studies, and countermeasures to reduce sun glare related crashes should be examined.
6 The importance of omitting spatial factors cannot be overstated. Many aspects of roadway safety management may suffer from failing to consider important and relevant spatial variables. Perhaps most importantly, hot spots might be identified for geometric or other improvements when in fact crash counts are elevated due to unobserved spatial effects. An engineer might improve an intersection without realizing benefits, since the elevated crash counts are unrelated to geometric or operational features. In modeling, only correlated variables (with spatial variables) included in a model can 'explain' the effects of the omitted spatial variables – thus over- or under-stating the effect of included variables. In short, omitting spatial variables will lead to inaccurate estimates of safety.

Based on these conclusions, some recommendations and directions for future research are:

1 As the collective understanding of road safety is to improve the existing knowledge about safety, the quality and quantity of data must be improved along with improvements in analytical methods. Spatial factors appear to influence safety, are intuitive, and point to the need to broaden our collective view of factors that should be considered in safety prediction.

2 Spatial factors provide a way to introduce behavioral factors into crash risk prediction and understanding. Many of the spatial factors considered here suggest that behavioral considerations are important to understanding road safety.
3 While much was learned through this research, improvements for follow on research are necessary. Sun-glare exposure should be independent of volumes during times of optimal sun glare and non-glare conditions. DUI crashes should be culled from the data and examined in relation to bars, in addition to total crashes. The proximity of bars and pubs with DUI crashes would strengthen the belief that this spatial factor is indeed capturing the effect of intoxicated drivers. Unfortunately, reporting rates of alcohol and drug use may prohibit an improved analysis – none the less an improved understanding is possible.

Acknowledgements

The authors thank the anonymous reviewers for their insightful comments and suggestions that helped improve the quality of the paper.

References

Anastasopoulos, P.Ch., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41 (1), 153–159.
Andrey, J., Mills, B., Vandermolen, J., 2001. Weather Information and Road Safety, Institute for Catastrophic Loss Reduction, Toronto, Ontario, Canada. Paper Series-No. 15.
Bauer, K.M., Harwood, D.W., 1996. Statistical Models of At-grade Intersection Accidents. Federal Highway Administration.
Broughton, J., Hazelton, M., Stone, M., 1999. Influence of light level on the incidence of road casualties and the predicted effect of changing 'Summertime'. Journal of Royal Statistical Society. Series A 162 (2), 137–175.
Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge, UK.
Chin, H.C., Quddus, M.A., 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accident Analysis and Prevention 35, 253–259.
El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accident Analysis and Prevention 41 (5), 1118–1123.
Flahaut, B., 2004. Impact of infrastructure and local environment on road unsafety: logistic modeling with spatial autocorrelation. Accident Analysis and Prevention 36 (6), 1055–1066.
Greene, W., 2007. Limdep, Version 9.0. Econometric Software Inc, Plainview, NY.
Griebe, P., 2005. Accident prediction models for urban roads. Accident Analysis and Prevention 35 (2), 273–285.
Hauer, E., Ng, J.C., Lovell, J., 1988. Estimation of safety at signalized intersections. Transportation Research Record 1185, 48–61.
Joshua, S.C., Garber, N.J., 1990. Estimating truck accident rate and involvements using linear and Poisson regression models. Transportation Planning and Technology 15 (1), 41–58.
Kim, D., Lee, Y., Washington, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. Accident Analysis & Prevention 39 (1), 125–134.
Kim, D., Washington, S., 2006. The significance of endogeneity problems in crash models: an examination of left-turn lanes in intersection crash models. Accident Analysis and Prevention 38 (6), 1094–1100.
Kumara, S.P., Chin, H.C., 2003. Modeling accident occurrence at signalized intersections with special emphasis on excess zeros. Traffic Injury Prevention 4 (1), 53–57.
Lord, D., Persaud, B.N., 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure. Transportation Research Record 1717, 102–108.
Lord, D., Washington, S., Ivan, J., 2004. Poisson, Poisson-gamma, and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 37 (1), 35–46.
Lord, D., Washington, S., Ivan, J., 2006. Further notes on the application of zero-inflated models in highway safety. Accident Analysis and Prevention 39 (1), 53–57.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice 44 (5), 291–305.
Maher, J., Summersgill, I., 1996. A comprehensive methodology for the fitting of predictive accident models. Accident Analysis and Prevention 28 (3), 281–296.
Malyshkina, N.V., Mannering, F.L., Tarko, A.P., 2009. Markov switching negative binomial models: an application to vehicle accident frequencies. Accident Analysis and Prevention 41 (2), 217–226.
Miaou, S., Lum, H., 1993. A Statistical Evaluation of the Effects of Highway Geometric Design on Truck Accident Involvements. Transportation Research Record 1407, TRB. National Research Council, Washington, DC, pp. 11–23.
Miaou, S., 1994. The relationship between truck accidents and geometric design of road section: Poisson versus negative binomial regression. Accident Analysis and Prevention 26 (4), 471–482.
Miaou, S., Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping: a space-time modeling approach. Journal of Transportation and Statistics 6 (1), 33–57.
Mitra, S., Chin, H.C., Quddus, M.A., 2002. Study of intersection accidents by maneuver type. Transportation Research Record 1784, 43–50.
Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 39 (3), 459–468.
Oh, J., Washington, S.P., Choi, K., 2004. Development of accident prediction models for rural highway intersections. Transportation Research Record 1897, 18–27.
Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis and Prevention 41 (4), 683–691.
Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection accident frequencies. Journal of Transportation Engineering 122 (2), 105–113.
Shankar, V.N., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accident Analysis and Prevention 27 (3), 371–389.
Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability process: an empirical inquiry. Accident Analysis and Prevention 29 (6), 829–837.
Shankar, V., Richard, A., Milton, J., Mannering, F., 1998. Evaluating Median Crossover Likelihoods with Clustered Accident Counts: An Empirical Inquiry Using the Random Effects Negative Binomial Model. Transportation Research Record 1635, TRB. National Research Council, Washington, DC, pp. 44–48.
Smeed, R.J., 1949. Some statistical aspects of road safety research. Journal of the Royal Statistical Society Series A 112 (1), 1–34.
Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analysis of rear-end crashes at signalized intersections. Accident Analysis and Prevention 38 (6), 1137–1150.
Washington, S., Karlaftis, M., Mannering, F., 2010. Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed. Chapman & Hall, Boca Raton.
Zhang, L., Holm, P., 2004. Identifying and Assessing Key Weather-related Parameters and their Impacts on Traffic Operations using Simulation. Federal Highway Administration Turner Fairbanks Highway Research Center.