
PERGAMON

Safety Science 41 (2003) 627–640

SAFETY SCIENCE

www.elsevier.com/locate/ssci

Modeling crashes involving pedestrians and motorized traffic

Venkataraman N. Shankar[a,*], Gudmundur F. Ulfarsson[a], Ram M. Pendyala[b], MaryLou B. Nebergall[c]

[a] *Department of Civil and Environmental Engineering, University of Washington, Box 352700, Seattle, WA 98195, USA*

[b] *Department of Civil and Environmental Engineering, University of South Florida, ENB 118, 4202 East Fowler Avenue, Tampa, FL 33620-5350, USA*

[c] *Washington State Department of Transportation, Box 47329, Olympia, WA 98504, USA*

Received in revised form 6 May 2002

Abstract

This paper presents an empirical inquiry into the predictive modeling of crashes involving pedestrians and motorized traffic on roadways. Empirical models based on the negative binomial distribution and mixing distributions, such as the zero-inflated Poisson distribution, are presented and discussed in terms of their applicability to pedestrian crash phenomena. Key modeling issues relating to the presence of excess zeros as well as unobserved heterogeneity in pedestrian crash distributions are addressed. The empirical results show that zero-inflated count distributions, such as the zero-inflated Poisson, are promising methodologies for providing explanatory insights into the causality behind pedestrian-traffic crashes.

1. Introduction

This paper presents an empirical note on the predictive modeling of pedestrian safety. Using data on reported crashes involving pedestrians and motorized traffic in Washington State, the causality of the annual frequency of such crashes is modeled. The need for a pro-active approach that examines the potential for pedestrian-traffic

* Corresponding author. Tel.: +1-206-616-1259; fax: +1-206-543-1543.

E-mail addresses: vns@u.washington.edu (V.N. Shankar), gfu@u.washington.edu (G.F. Ulfarsson), pendyala@eng.usf.edu (R.M. Pendyala), nebergm@wsdot.wa.gov (M.B. Nebergall).

0925-7535/03/$ - see front matter
doi:10.1016/S0925-7535(02)00017-6

crashes is imminent. Designing for pedestrian-friendly roadways in the United States has historically been reactive. Corridors with reported pedestrian-traffic crashes are the usual targets of improvement, with little or no attention to the latent processes that are often at work at locations with unobserved crash potential.

Recently, researchers have studied the relationships between site-design, neighborhood and pedestrian traffic, showing that good pedestrian facilities do attract pedestrians, and that improving suburban pedestrian facilities could significantly influence mode choice. Urban sites with small blocks and extensive sidewalk systems had three times the amount of pedestrian traffic than suburban areas with large blocks and incomplete sidewalk systems (Moudon et al., 1997; Hess et al., 1999). The Federal Highway Administration (FHWA) has published a guide to help planners, designers, and decision makers better consider the safety and mobility of pedestrians in roadway design (Zegeer et al., 2000). Pedestrian crossing at unsignalized intersections remains a risk, particularly to the pedestrians. Site design features that attempt to reduce that risk have been studied by Huang et al. (2000a,b). They found that features, such as improved signs, increased the number of motorists that yielded to pedestrians at crosswalks. Motor-vehicle accidents, involving pedestrians walking along the roadway, were studied by McMahon et al. (1999). They studied geometric factors and operational factors, such as speed limits and the presence of sidewalks, and also socio-economic, neighborhood factors, such as level of unemployment and proportion of single parents. They found that reduced speed and wide, grassy shoulders reduced the probability of an accident, but the neighborhood factors for high unemployment and more single parents increased the probability of accidents occurring. This shows the confounding effect of exposure since these neighborhoods may have more pedestrians but not necessarily less safe pedestrian facilities. The issue of demographic and environmental variables being correlated with pedestrian crashes, was also studied by Lascala et al. (2000). Their results are similar, but add factors, such as gender, education, income, and traffic flow. They also studied the effect of bars in the neighborhood, which are related with more pedestrian accidents where the pedestrian has been drinking. Stevenson et al. (1999) focused on children as pedestrians, and studied three community intervention programs and their effect on child pedestrian safety. National Highway Traffic Safety Administration (NHTSA) is conducting the Pedestrian Crash Data Study, which will examine pedestrian injuries in frontal accidents with late model vehicles (Jarrett and Saul, 1998). What all of these studies explicitly or implicitly suggest is that partial observability effects are at work. That is, either due to self-selection in the locations requiring pedestrian-friendly treatment, or due to simple data survey restrictions, inferring the type of factors that can affect pedestrian crash potential may remain a methodological challenge.

The goals of this paper are to address specific portions of that methodological challenge: (a) identifying performance measures for modeling pedestrian safety, (b) examining the appropriateness of statistical distributions for modeling such measures, and (c) implementing candidate models in the presence of incomprehensive data. The rest of this paper presents in order, methodologies, empirical model structures, design causalities, conclusions and recommendations.

2. Methodology

Several major methodological issues arise in the modeling of pedestrian safety. Foremost, but not limited to, these issues relate to the definition of corridor-level pedestrian safety measures, the modeling frameworks that provide insights on roadway design parameters that can be used for effective control of crashes, and the type of data that is required to implement the models in a regional setting.

2.1. Definition of the performance measure

Crashes involving pedestrians and motorized traffic, hereinafter, called pedestrian crashes, present problems if one were to evaluate them in terms of typical exposure-based methods. Conventional exposure-based methods provide for a measure of safety referred to as "crash rate". Crash rate is defined as the number of crashes per year per a pre-defined unit of pedestrian and traffic demand. Central to the computation of this number is average annual daily traffic (AADT), which for a mature system, such as vehicular traffic networks, is a commonly measured variable and easily accessible from public records. In addition, the relationship also involves pedestrian demand in the form of pedestrian daily traffic (PDT), which suffers from significant limitations. First and foremost, PDT can empirically range from zero in the winter months to a positive count in the summer months. Developing an exposure measure that intends to capture the potential for pedestrian collisions on the basis of zero PDT's creates observational error problems. In addition, a composite measure of exposure involving both AADT and PDT is required. AADT for mature traffic networks could exceed 100,000 vehicles a day, while PDT occurs at a drastically lower order of magnitude (a few hundred to a thousand pedestrians for busy central business district locations). Formulating a relationship accommodating significant shifts in the order of magnitude of exposure can not only be challenging, but rife with parameter estimation problems (for example non-linearity). The safety function, as this relationship is generally called, also varies from location to location. Pedestrian safety functions are virtually unknown and no in-depth study to date has been conducted to investigate this functional form. In that light, implementing a safety function that is common across locations would be highly restrictive and misguided. The non-linearity between crash counts and crash exposure has caused vehicular traffic crash literature to focus on count models (for example, see Shankar et al., 1997). A count modeling framework that explicitly models the count of crashes per pre-defined time period provides for a framework that can give unbiased and consistent parameters. For the aforementioned reasons, we define pedestrian safety performance for a given roadway corridor as a frequency of pedestrian crashes per year. The methodological challenge is then to uncover factors having a significant impact on this frequency measure. Uncovering factors affecting pedestrian crash frequency involves a priori assumptions about its distributional nature. Some methodological issues arise here, and are discussed next.

2.2. The distribution of the performance measure

The pedestrian crash frequency is a direct function of the size of the sampling unit. In statewide safety improvement models, sampling units are usually at the 1-mile scale, primarily for reasons of ease of interpretation and implementation. To be consistent with improvements derived for regular modes of transportation such as passenger cars, buses and trucks, roadway engineers prefer the sampling unit for pedestrian improvements to be at 1-mile scales. Despite the ease of implementation and interpretation that this practice provides, methodological issues remain. A sampling unit of 1 mile of roadway length, will likely yield significant non-zero annual crash counts in a regular motorized vehicle crash context. However, in the context of pedestrian crashes, the likelihood of a large proportion of zero frequencies can remain high. An over-abundance of zeros in the pedestrian crash count distribution poses challenges. The methodological challenge here is to discern the nature of this zero count. On the one hand, the zero counts observed in the sample may reflect true life-time proportions in the statewide network. On the other hand, a portion of the zero counts could arise as a result of partial observability. That is, due to a limited history of observations or a lack of maturation of the pedestrian system, some locations in the sample may be reporting zero counts in the short-term, while in reality, over their lifetime, they could regress to an expected number of crashes greater than the observed zeros in the sample.

In the case partial observability is not an issue, a single-parent count distribution is adequate to model both zero and non-zero pedestrian crash counts. Where excess zero counts are not an issue, and the application of a single-parent distribution is well justified—what has now become—conventional wisdom (see for example, Shankar et al., 1995) dictates that a negative binomial distribution can adequately capture the count process. The negative binomial distribution captures "overdispersion" in crash data. Overdispersion in crash data occurs as a result of unobserved heterogeneity. That is, some locations in the sample deviate from the average location in the sample due to latent effects affecting their causality. Latent effects could arise from unmeasured variables, such as weather effects, human, land use, and vehicle-specific factors. Roadway factors are usually measured in adequate detail; so their impact on crash counts is explicit. Formally, if one models λ_i, the expected number of pedestrian crashes per year per mile, as a function of β, a set of estimable parameters, and x_i, a set of roadway, exposure, and traffic factors, then,

$$\ln\lambda_i = \beta x_i + \varepsilon_i, \tag{1}$$

where $\exp(\varepsilon_i)$ is a gamma distributed variable with mean 1 and variance α. If we condition the number of pedestrian crashes on the gamma-distributed error term, the resulting probability distribution for a given count, n_i, is

$$P(n_i|\varepsilon_i) = e^{-\lambda_i\varepsilon_i}(\lambda_i\varepsilon_i)^{n_i}/n_i!, \tag{2}$$

where n_i is a non-negative integer. Integrating ε_i out of this expression produces the unconditional negative binomial distribution of n_i:

$$P(n_i) = \frac{\Gamma(n_i + \theta)}{\Gamma(\theta) n_i!} \left(\frac{\theta}{\theta + \lambda_i}\right)^{\theta} \left(\frac{\lambda_i}{\theta + \lambda_i}\right)^{n_i}, \tag{3}$$

where $\theta = 1/\alpha$, α, being the overdispersion parameter. A non-zero value for α indicates that the pedestrian count variance-mean ratio for any given roadway section exceeds unity, indicating overdispersion at work. Using the density function shown in (3), one can derive the first and second moments to formalize the variance–mean relationship:

$$\frac{\mathrm{Var}(n_i)}{E(n_i)} = 1 + \alpha E(n_i) \tag{4}$$

A value greater than unity for α clearly indicates the overdispersion of the distribution about the mean, as (4) shows.

In the event partial observability issues become significant in pedestrian crash count phenomena, one has to account for the impact of latent effects differently. As opposed to the single-parent negative binomial distribution as shown above, a mixture distribution may be more appropriate. Formally, one can view the partial observability phenomenon as the product of two latent processes, Z and Y^*, where Z indicates whether a location is permanently in the zero-crash count state and Y^* denotes the count state. We observe neither Z nor Y^*, but only the observed count Y, such that $Y = Z \cdot Y^*$. Determining the latent components can then be viewed as a mixing distribution problem, with Z being modeled as a dichotomous probability and Y^* being modeled as a count probability. In vehicular crash contexts, such distributions have been found to be appropriate (see Shankar et al., 1997). In regular vehicular crash contexts, the effect of partial observability has primarily been attributed to roadway design deviations. The effect of such deviations has been found to—at the least—cause partial observability, and in certain design situations, overdispersion as well. If a mixture distribution is not employed, the negative binomial may capture all overdispersion spuriously, even if the overdispersion were to arise not from latent effects, but from other processes, such as partial observability. The choice of a mixture distribution, however, is not self-apparent. A priori assumptions regarding the density generators are required in formulating the mixture distribution. Formally, let Y_i be the annual number of pedestrian crashes reported for corridor i, and let p_i be the probability that corridor i will exist in the zero-crash state over its lifetime. Thus $1 - p_i$ is the probability that corridor i actually follows a true count distribution in the non-zero state. For our immediate purposes, we assume that this count state follows a Poisson distribution. We note that the Poisson state contributes to zero counts as well. By our assumption, dual density generators of zero counts are at work. That is, the splitting regime, with probability

p_i, and the Poisson process, with probability $1 - p_i$, contribute—in combination—to the apparent excess zero problem. Given this,

$$\mathrm{P}(Y_i = 0) = p_i + (1 - p_i)\mathrm{e}^{-\lambda_i} \tag{5}$$

and

$$\mathrm{P}(Y_i = k) = (1 - p_i)\frac{\mathrm{e}^{-\lambda_i}\lambda_i^k}{k!}, \tag{6}$$

where k is the number of accidents (positive numbers starting from one), with λ_i being the mean. In (5) and (6), the probability of being in the zero-accident state, p_i, is formulated as a logistic distribution such that $\ln(p_i/1 - p_i) = \mathbf{G}_i\gamma$, and λ_i satisfies $\ln(\lambda_i) = \mathbf{H}_i\beta$, where $\mathbf{G}_i$ and $\mathbf{H}_i$ are covariate vectors, and γ and β are coefficient vectors. The covariates that affect the mean, λ_i, of the Poisson state may or may not be the same as the covariates that affect the zero-accident state probability, i.e. p_i. Alternatively, the vectors $\mathbf{G}_i$ and $\mathbf{H}_i$ may be related to each other by a single, real-value parameter, τ. In such a case, a natural parameterization is $\ln(p_i/1 - p_i) = \tau\mathbf{B}_i\beta$. If τ is not significantly different from zero, then the corridor is equally likely to be in the zero or non-zero lifetime state. Combining (5) and (6) provides us with the zero-inflated Poisson (ZIP) model of pedestrian crash frequency. The ZIP model assumes that excess zero mass in crash counts occurs entirely due to partial observability. In a general sense, if partial observability and overdispersion are suspected, negative binomial variants of the ZIP model are plausible. In statistically validating any zero-altered model, one has to distinguish between the count model, such as the Poisson, and the zero-inflated probability model, such as the ZIP. A statistical test to make this distinction has been proposed by Vuong (1989). The Vuong-test is based on the t-statistic and has reasonable power in count-data applications (see Greene, 1994). The Vuong-statistic (V-statistic) is computed as

$$V = \frac{\overline{m}\sqrt{N}}{S_m}, \tag{7}$$

where $\overline{m}$ is the mean of $m = \ln\left[\frac{\mathrm{f}_1(\cdot)}{\mathrm{f}_2(\cdot)}\right]$, where $\mathrm{f}_1(\cdot)$ is the density function of the ZIP distribution and $\mathrm{f}_2(\cdot)$ is the density function of the parent-Poisson distribution, and S_m and N are the standard deviation and sample size respectively. The advantage of using the Vuong-test is that the entire distribution is used for comparison of the means, as opposed to just the excess zero mass. A value greater than 1.96 (the 95% confidence level for the t-test) for the V-statistic favors the ZIP, while a value less than -1.96 favors the parent-Poisson, with values in between 1.96 and -1.96 meaning that the test is inconclusive. The intuitive reasoning behind this test is that if the processes are not statistically different, the mean ratio of their densities should equal one. To carry out the test, both the parent and zero-inflated distributions need to be estimated and tested using a t-statistic. The Vuong-test allows us to distinguish

between the negative binomial and the ZIP model as well. Studies have shown that a Vuong-statistic favoring the ZIP model does so due to reasonable power of that statistic (Greene, 1994). That is, the negative binomial is not a reliable alternative to the ZIP model when the statistic exceeds $+1.96$, even when the overdispersion parameter α (in the negative binomial model) is significant. A favorable Vuong-statistic for the ZIP model should make the ZIP model more plausible, thus avoiding the likelihood of spurious overdispersion as detected by the negative binomial model. Spurious overdispersion is common in count models of crash frequencies. In pedestrian crash contexts, the problem of mis-identification of unobserved heterogeneity can be exacerbated by the similarity in the variance-mean relationships of the negative binomial and the ZIP models. To illustrate this, the variance-mean ratios for the negative binomial and the ZIP model can be compared. The variance-mean ratio for the ZIP model is

$$\frac{\mathrm{Var}(n_i)}{\mathrm{E}(n_i)} = 1 + \left(\frac{p_i}{1-p_i}\right)\mathrm{E}(n_i), \tag{8}$$

while (4) shows a similar form for the negative binomial model. It is self-evident from comparing the two equations that α in (4) can spuriously pick up the effect of the splitting regime, denoted by the ratios of the state probabilities in (8).

2.3. Incorporating pedestrian demand

Either of the modeling techniques discussed previously is reliant on consistent measures of pedestrian exposure. As mentioned previously, pedestrian daily traffic (PDT) is not commonly available in roadway databases. Second, pedestrian demand is seasonal, and suffers from fluctuations that make the PDT highly susceptible to measurement error. In combination with missing values, the measurement error encountered in PDT makes it an untenable choice for exposure. From a statistical standpoint, the importance of errors-in-variables cannot be ignored when the independent variable, such as pedestrian demand, is subject to much higher error and variation. For a detailed exposition on errors-in-independent variables, see for example Johnston (1990). Even the common statistical prescription of "instrumenting" a poorly measured independent variable does not suffice in the pedestrian context. By instrumenting, any stochasticity in the independent variable is—in principle—eliminated and the demand variable is rendered exogenous. This does not, however, smooth out seasonal variations. With consistent seasonal pedestrian demand data being unavailable, coupled with other constraints (such as cumbersomeness of repeated measurements) on the PDT variable, the consequences of using such a variable in a parametric model can be fatal. Parameter estimates can be inconsistent, and significantly biased in small samples, which is usually the case in pedestrian crash databases. To avoid other fatal problems, such as omitted variable bias, surrogates for pedestrian exposure are required. Network characteristics allow roadway engineers to address the exposure issue from the supply side, using measures such as driveway spacing, signal spacing, illumination at intersections, availability of

sidewalks, paved shoulders and availability of median crossings as "instruments" (surrogates) for exposure. Network attributes, combined with land use data, such as zoning characteristics along the corridor, remove any omitted variable bias that could arise in the model.

3. Empirical model structure

3.1. Model estimation

The discussed models were applied to an empirical dataset of reported pedestrian crashes in Washington State. Reported crashes for the years 1991 through 1994 were sampled for corridors for which network attributes, land use attributes, detailed crash data, traffic and roadway factors were consistently available. The sampling of 1-mile pedestrian sections was conducted so as to ensure a representative set on the basis of pedestrian usage on the major roadway types. A total of 440 observations were used in estimating the negative binomial and ZIP models of annual pedestrian crash frequency per mile. Each observation represented a one-mile pedestrian section with multiple years (4 years) of crash history. Estimation of the negative binomial and ZIP models was conducted by standard maximum likelihood methods. In the current sample, the number of positive crash counts was 61 and the number of zeros was 379. The likelihood function for the negative binomial is straightforward, using the log-likelihood:

$$\sum_i \ln[\mathrm{P}(Y_i)] = \sum_i \Bigg[\ln[\Gamma(n_i+\theta)] - \ln[\Gamma(\theta)] - \ln n_i! + \theta \ln\left(\frac{\theta}{\theta+\lambda_i}\right) + n_i \ln\left(\frac{\lambda_i}{\theta+\lambda_i}\right) \Bigg]. \tag{9}$$

For the ZIP model, the probability statement of Y_i (crash frequency) becomes:

$$\mathrm{P}(Y_i) = (1-p_i)\frac{\mathrm{e}^{-\lambda_i}\lambda_i^k}{k!} + Z_i p_i, \tag{10}$$

where $Z_i = 1$ when Y_i is observed to be zero and $Z_i = 0$ for all other values of Y_i. The log-likelihood function is then simply $\sum_i \ln[\mathrm{P}(Y_i)]$. The use of the indicator variable Z_i makes maximization of the log-likelihood function easy and uniform across the entire sample. The log-likelihood function is estimated using the gradient/line search approach proposed by Greene (1994).

The empirical sample consisted of oversampled zero crash counts. In order to account for such sampling bias in the model, corrections are provided in the estimated model to account for the true population share of a "zero-state" location. As a result of zero-oversampling, the parameters of the logit splitting regime (zero versus non-zero accident state) will be inconsistently estimated. A consistent method to estimate the logit splitting regime parameters under oversampling is to use weighted

exogenous sampling maximum likelihood (WESML) estimators. Per Manski and Lerman (1977), the WESML is given by the criterion

$$\max_{\theta\in\Theta}\sum_{n=1}^{N} w(i_n)\ln \mathrm{P}(i_n|z_n,\theta), \tag{11}$$

where $w(i)=\frac{Q(i)}{H(i)}$, $i\in C$, C is the set of states (in our case zero or non-zero crashes), $w(i)$ is a known positive weight used to correct the likelihood function for outcome-based sampling, $Q(i)$ and $H(i)$ being population and sample shares respectively, z_n is the attribute space for corridor n, and $\mathrm{P}(i_n)$ the probability of the ith state for corridor n. Operationalizing this criterion in a zero-inflated Poisson context posed convergence problems, even for constrained estimation of the vector for the zero state, since density functions arising from both the logit splitting regime and the Poisson process are at work. Since the usual exogenous sampling likelihood (ESL) procedure produces consistent estimates for all parameters except the constant term under oversampling of one state (zero-count locations in this case), a reasonably tractable alternative was to adjust the constant in the logit splitting regime (see Ben-Akiva and Lerman, 1985) using the population and sample shares, $Q(i)$ and $H(i)$. This adjustment is simply

$$\alpha-\ln\left[\frac{H_0}{Q_0}\right]-\ln\left[\frac{H_{>0}}{Q_{>0}}\right], \tag{12}$$

where α is the estimated constant for the logistic state in the ZIP model, $\ln\left[\frac{H_0}{Q_0}\right]$ is the adjustment to the constant term in the splitting regime based on the respective sample and population shares of zero-count locations, and $\ln\left[\frac{H_{>0}}{Q_{>0}}\right]$ is the additional adjustment (from the non-zero state) to the zero-state constant against a zero-baseline specification. If the model follows the ZIP() structure, we replace α by $\tau\cdot\omega$ where ω is the estimated constant in the Poisson state and τ is the scalar multiplier. The population share of zero 1-mile pedestrian corridors was determined to be approximately 78%, while the sample share is approximately 89%.

3.2. Empirical findings

In general statistically significant variables affecting pedestrian crash frequency are indicators in nature. Several indicator variables such as "presence of sidewalks", "presence of crosswalks", "presence of overhead crossings", "presence of driveways", "presence of traffic signals", "presence of illumination along the roadway", and "presence of center turn lanes" were examined for possible significance in the pedestrian crash context. These are the classic exposure variables often discussed in the pedestrian literature. In this empirical study, considering that land use indicator variables were also examined in addition to the above-mentioned factors, the statistically significant effects on pedestrian crash causation are somewhat surprising.

Land use indicator variables included factors such as "presence of retail shopping", "presence of schools", "presence of single-family housing", "presence of commercial office", or "presence of multi-family housing",

The empirical findings indicate a combination of threshold effects on pedestrian crash causation rather than continuous effects. While it is possible that such a finding might be an artifact of the dataset at hand, it is intuitively consistent with the behavioral aspect of pedestrian-vehicle accidents. That is, once certain thresholds are met or exceeded, the probability of accidents (in this case both the "state" and count variables) or associated severity significantly changes. The following discussion provides detailed insights into statistically significant threshold effects. Table 1 shows the summary statistics of key variables affecting the pedestrian crash context. Table 2 shows the estimation of the ZIP model of pedestrian crash frequency. Results for the negative binomial model of pedestrian crash frequency are not shown, since the ZIP model was statistically validated. The ZIP specification with separate regressor vectors for the splitting regime and the positive count states was determined to be statistically valid (Vuong-statistic of 2.155), compared to alternative forms including the single-parent negative binomial and the constrained ZIP model. The overdispersion parameter for the negative binomial was marginally significant (t-statistic of 1.32) suggesting that unobserved heterogeneity in the positive-count state is weak. However, it does not eliminate the possibility of overdispersion arising from the excess zero process, which a single-parent Poisson model cannot adequately capture.

Significant main effects affecting pedestrian accident frequency varied between the positive crash count (Poisson) state and the zero (logistic) state. In the positive count state, factors included geometric factors such as presence of center turn lanes and vehicular exposure factors such as average daily traffic. The cross-section variable (presence of two-way turn lane) act as surrogates for pedestrian exposure and demand since roadway and land use development go hand-in-hand. The two-way center, turn lane variable captures the correlation between high-density retail and multi-family housing uses in a corridor. This variable increases the probability of

Table 1
Summary statistics of key attributes for frequency-sample pedestrian corridors

Variable	Minimum	Maximum	Mean	Standard deviation
Accident frequency (per mile per year)	0	4	0.139	0.476
Number of lanes	2	5	2.56	0.95
Signal spacing (km)	0.16	1.6	1.34	0.33
Center turn lane corridors (two-way)	0	1	0.11	
Illuminated corridors (midblocks and intersections)	0	1	0.22	
Illuminated corridors (intersections only)	0	1	0.12	
Corridors with mid-block/non-intersection crosswalks	0	1	0.075	
Corridors with median treatments	0	1	0.14	
Average daily traffic	10,350	98,875	25,425	20,475

positive pedestrian crashes, while corridors with annualized daily vehicular traffic of less than 10,000 vehicles per day decrease that probability. It is acknowledged that the cross-section variables may not be the best instruments in the absence of good land use data; however, these substitutes do not render the specifications unusable, because they are not correlated with the error term.

In the zero-state (logistic), traffic control factors such traffic signal spacing and network supply factors such as illumination were significant. All variables were of intuitively correct sign. If traffic signals are spaced greater than 0.8 km apart, then the probability of that corridor being in the zero pedestrian crash state decreases. This is intuitive since it captures the effect of "block lengths". Longer block lengths would have higher traffic signal spacings, and hence would induce pedestrians to cross the roadway at non-signalized points as opposed to at the traffic signals. Roadway illumination on the other hand increases the probability that zero crashes would occur in the corridor. Land use and environmental factors were insignificant. This finding may in part stem from the coarseness of the land use variables. As opposed to using indicator variables (for example, presence of retail shopping), perhaps density and area measures may be useful. Currently, this type of data is unavailable on a consistent basis across the 39 counties in Washington State.

Table 2
Zero-inflated Poisson-tau (ZIP) model of annual pedestrian accident frequency for 1-mile corridors

Variable	Coefficient	*t*-statistic
Non-zero crash probability state as Poisson function		
Constant	−1.500	−3.949
Center turn lane indicator (1 if center turn lane is a two-way turn lane; 0 otherwise)	1.554	4.280
Traffic volume indicator (1 if average annual daily vehicular traffic on section is less than or equal to 10,000 vehicles per day; 0 otherwise)	−3.189	−3.030
Zero crash probability state as logistic function		
Constant	0.157	0.181
Signal spacing (1 if signals are spaced at 0.8 km or more; 0 otherwise)	−2.021	−2.093
Illumination indicator (1 if roadway lighting exists for the corridor; 0 otherwise)	2.800	2.910
Number of observations	432	
Log likelihood at zero	−236.613	
Log likelihood at convergence	−125.935	
Adjusted ρ^2	0.468	
Vuong statistic	2.155	

As mentioned in the estimation section, to obtain a consistent estimate of the constant term for the splitting regime, adjustments are made using the population and sample shares $Q(i)$ and $H(i)$ to the constant values in the logistic zero state. The adjusted constant in the zero state is computed to be 0.025 and the constant for the non-zero portion of the splitting regime to be 0.693. It should be noted that all non-constant exogenous regressors in the splitting regime are specified for the zero state.

4. Conclusions and recommendations

4.1. Conclusions

This study presents a methodology for detecting statistically significant roadway effects on pedestrian-traffic crashes in the absence of complete land use and exposure data. Pedestrian crashes in corridors (as opposed to intersections) are often incomplete and inaccurate in terms of exposure data such as pedestrian volumes, and land use information surrounding the corridor. Such information acts in a latent manner on pedestrian crash involvement, and if not accounted for, results in overdispersion due to unobserved heterogeneity. The impact of the heterogeneity is of key concern to pedestrian designers, for it can affect the identification of the right set of variables affecting pedestrian crash probabilities. If heterogeneity is assumed to follow the negative binomial distribution as most crash studies would suggest, such an assumption would result in a spurious model for pedestrian crash contexts. Pedestrian corridor crash samples often have excess zero counts as opposed to positive counts, and to ensure a systematic method that captures zero-induced overdispersion correctly, mixture distributions are plausible techniques. This paper examined the negative binomial distribution as well as mixture distributions such as the ZIP model. The findings illustrate that a combination of factors are at play in the pedestrian crash context. Cross-sectional and traffic volume factors appear to have an effect on whether a pedestrian corridor will have an increased probability of non-zero crashes per year. On the other hand, traffic supply and control factors such as traffic signal spacing and illumination appear to have an effect on whether or not a pedestrian corridor will have any crashes at all.

To summarize our empirical findings, we concluded that the ZIP model is most suitable for analyzing pedestrian crash contexts. A combination of exposure variables such as average daily traffic, traffic control factors such traffic signal spacing, network supply factors such as illumination, and presence of center-turn lanes were found to have a statistically significant impact on pedestrian-vehicle crash probabilities.

In the larger context of contemporary findings on pedestrian crash factors, it is well known that pedestrian crossing at unsignalized intersections remains a risk, and the significance of the traffic signal spacing variable in our study corroborates that. In fact our study suggests thresholds for signal spacing (0.8 km) as safe limits. Other prominent studies have focused on the larger sociological context of gender, education, income, single parenthood, alcohol and the confounding effects these may have with pedestrian exposure. In our study we attempted to account for these effects through land use and network variables. Land use variables were insignificant when we control for exposure through average daily traffic; however, network variables still remain significant. The lack of significance of land use variables may be an artifact of our dataset. We also note that in addition to the key network design variables we found to be significant, social policy variables may serve a key role as well, if one corrects for self-selection. Community intervention programs and their effect on child pedestrian safety have been researched but the dataset at hand in our study lacked sufficient observations to explore that effect.

4.2. Recommendations

The dataset used in this study is somewhat limited in its coverage of geographic and land use effects. National databases that account for geographic variability and scale effects, as well as disaggregate land use measures of density and area, demographic variables such as corridor age, income and education profiles, would likely provide greater insights into the interactions between land use and traffic factors in the pedestrian crash context. Such interactions were found to be insignificant in this study; however, it is important that these effects be researched further, for they could provide key insights into the behavioral dimension of pedestrian crash involvement. In addition to the above-mentioned factors, a database that includes network variables such as presence and width of sidewalks and shoulders, illumination, length of unsignalized blocks or signal spacing, presence of center-turn lanes, detailed land use variables in the form of multi-family, single-family, commercial, and institutional characteristics, measures of pedestrian accessibility, and sociological variables such as community intervention and educational programs would provide for a comprehensive basis for assessing pedestrian safety.

Methodologically, some issues remain as well. It is inevitable that overdispersion will occur in pedestrian crash contexts; it remains to be seen if the overdispersion is of a different nature at intersections versus corridors as a whole. For example, it is likely that pedestrian crash overdispersion at intersections could follow the negative binomial, while mid-block corridor locations could follow a mixture distribution, or vice versa. For simplicity, this study combined mid-blocks and intersections into 1-mile sections and concluded that a mixture model such as the zero-inflated Poisson is plausible.

References

Ben-Akiva, M., Lerman, S.L., 1985. Discrete choice analysis: theory and application to travel demand. The MIT Press, Cambridge, Massachusetts.

Greene, W., 1994. Accounting for excess zeros and sample selection in poisson and negative binomial regression models. Working Paper EC-94–10, Stern School of Business, New York University.

Hess, P.M., Moudon, A.V., Snyder, M.C., Stanilov, K., 1999. Site design and pedestrian travel. Transportation Research Record 1674, 9–19.

Huang, H., Zegeer, C., Nassi, R., 2000a. Innovative treatments at unsignalized pedestrian crossing locations. ITE 2000 Annual Meeting and Exhibit, Nashville, Tennessee, 6–9 August, Institute of Transportation Engineers, Washington, DC.

Huang, H., Zegeer, C., Nassi, R., 2000b. Effects of innovative pedestrian signs at unsignalized locations: three treatments. Transportation Research Record 1705, 43–52.

Jarrett, K.L., Saul, R.A., 1998. Pedestrian injury-analysis of the PCDS field collision data. Proceedings of the 16th International Technical Conference on the Enhanced Safety of Vehicles, 31 May–4 June, Windsor, Ontario, Canada, vol. 2, 1204–1211.

Johnston, J., 1990. Econometric Methods, 3rd ed. McGraw-Hill Companies Inc.

Lascala, E.A., Gerber, D., Gruenewald, P.J., 2000. Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. Accident Analysis and Prevention 32 (5), 651–658.

Manski, C.F., Lerman, S.L., 1977. The estimation of choice probabilities from choice-based samples. Econometrica 45, 1977–1988.

McMahon, P.J., Duncan, C., Stewart, J.R., Zegeer, C.V., Khattak, A.J., 1999. Analysis of factors contributing to "walking along roadway" crashes. Transportation Research Record 1674, 41–48.

Moudon, A.V., Hess, P.M., Snyder, M.C., Stanilov, K., 1997. Effect of site design on pedestrian travel in mixed-use, medium-density environments. Transportation Research Record 1578, 48–55.

Shankar, V.N., Mannering, F.L., 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.N., Milton, J.C., Mannering, F.L., 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accident Analysis and Prevention 29 (6), 829–837.

Stevenson, M., Iredell, H., Howat, P., Cross, D., Hall, M., 1999. Measuring community/environmental interventions: The Child Pedestrian Injury Prevention Project. Injury Prevention 5 (1), 26–30.

Vuong, Q., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–334.

Zegeer, C.V., Seiderman, C., Lagerwey, P., Cynecki, M., Ronkin, M., Schneider, B., 2000. Pedestrian facilities users guide. Providing safety and mobility. University of North Carolina, Chapel Hill, Highway Safety Research Center, and Federal Highway Administration.