Accident Analysis and Prevention 128 (2019) 148–158

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Adequacy of negative binomial models for managing safety on rural local roads

Thomas Hall[a,*], Andrew P. Tarko[b]

[a] Lyles School of Civil Engineering, Purdue University, 550 Stadium Mall Drive, West Lafayette, IN, 47907, United States
[b] Center for Road Safety, Lyles School of Civil Engineering, Purdue University, 3000 Kent Avenue, Suite C2-103, West Lafayette, IN, 47906, United States

ARTICLE INFO

Keywords:
Rural local roads
Rural local intersections
Low-volume county roads
Low sample mean
Bivariate negative binomial model
Bivariate ordered probit model

ABSTRACT

Count models, such as negative binomial regression, are well-established statistical methods for analyzing road safety. Although count models are widely used for arterial roads, their application to rural local roads is sparse, partly due to the concern of possible estimation bias caused by low crash counts. This paper revisits the matter to further evaluate the suitability of negative binomial models for rural local roads with low crash frequencies, comparing the performance of the model to probabilistic regression (ordered probit) proposed in the past.

The negative binomial model was estimated to predict crashes for rural local intersections and compared to predictions obtained from the ordered probit model. Bivariate versions of both models were applied to improve model efficiency by incorporating correlation between two severity outcomes, fatal/injury (FI) and property damage only (PDO) crashes. The estimated models included several significant variables with intuitive signs. These results are discussed in the paper to support the claim that both models are adequate. Furthermore, the cumulative sums of the model-predicted and observed crashes conditioned on the estimated effects were compared to detect any systematic bias in the results. Although both models showed similar performance and no obvious biases could be detected, the negative binomial model seemed to behave slightly better than the ordered probit model, demonstrating the model's suitability in the analyzed case. The results point to the possibility of applying the Highway Safety Manual methodology to lower-volume county roads with focus shifted from individual high-crash locations to safety-deficient road features present at multiple locations.

1. Introduction

Most studies on rural road safety focus on state-administered arterial roads, whereas the majority of rural roads are local roads maintained by counties and townships. In heavily agricultural states, more than 80% of all rural road miles are local (Federal Highway Administration, 2015). Rural local roads often have outdated geometrical designs, poor visibility, and roadside obstructions that make them particularly hazardous to roadway users. Despite the fact that they have less total crashes than their rural state road counterparts, rural local roads tend to have greater crash rates when adjusted for vehicle miles travelled (VMT) (Souleyrette et al., 2010).

Past studies aimed at investigating safety and identifying potential issues on rural local roads have utilized a variety of methods, including both non-statistical and statistical techniques. Non-statistical techniques have included road inspections by human observers (Cafiso et al., 2011, 2015), as well as crash location mapping, pattern identification, and field studies (Hall et al., 2003). Hall et al. (2003) found that local agencies commonly used field studies and road user complaints to determine which road locations were most in need of safety improvements. However, techniques such as field studies and road user complaints may introduce some degree of subjectivity. On the other hand, relying on the crash history for the inspected roads may be difficult due to the typically low crash frequencies that hinder the ability to draw confident conclusions.

Applying statistical models to acquire and utilize transferable safety knowledge may be a good approach to overcome the mentioned hurdles for rural local roads. Negative binomial count models, abundantly used for examining safety on arterial roads, have seen less usage on rural local roads. The concerns about potential estimation issues stemming from a low sample mean and from frequently poor and incomplete road data are likely deterrents. Only a few negative binomial models for rural local road segments were found in the literature (Labi, 2006; Avelar et al., 2015; Stapleton et al., 2018). These studies examined the impact of various segment features on total and injury crashes. Alternative statistical investigations included analysis of covariance (Zegeer

[*] Corresponding author.
E-mail addresses: hall198@purdue.edu (T. Hall), tarko@purdue.edu (A.P. Tarko).

https://doi.org/10.1016/j.aap.2019.03.001
Received 24 September 2018; Received in revised form 24 February 2019; Accepted 5 March 2019
Available online 22 April 2019

et al., 1994), proportion tests (Souleyrette et al., 2010), correlation analysis (Ewan et al., 2016), multivariate linear regression (Ewan et al., 2016), and the ordered probit model (Souleyrette et al., 2010). Tarko et al. (2012) and Hall (2017) applied multivariate ordered probit models for identifying the features affecting traffic safety on rural roads.

Previous studies indicated the risk of erroneous parameter estimates in count models estimated with a low sample mean (Lord, 2006; Lord and Mannering, 2010). Model-based treatments proposed to mitigate the potential problem included zero-inflated count models (Shankar et al., 1997) and ordered probit models with crash counts as alternative outcomes of the data-generating process (Tarko et al., 2012; Hall, 2017). This study builds upon past research by shedding more light on the suitability of negative binomial models for county roads (representative of rural local roads) with a low frequency of crashes, evaluating the model for potential biases that may affect its predictions. This is accomplished by comparing the cumulative crash counts with predictions produced with fitted regression models and by contrasting the performance of a bivariate negative binomial model with the performance of a bivariate ordered probit model. Both the bivariate models are estimated with respect to fatal/injury (FI) and property damage only (PDO) crashes reported at rural local intersections in Tippecanoe County, Indiana. The results, findings, and implications of the study are presented and discussed.

2. Methodology

Apart from the aforementioned studies by Labi (2006), Avelar et al. (2015) and Stapleton et al. (2018), negative binomial models have seen relatively little usage in analyzing the effect of road features on traffic safety for rural local roads. This is partially due to the concerns over the model's estimation with a low-mean sample. The study by Lord and Mannering (2010) suggested that low sample means with crash counts skewed towards zero may lead to improperly estimated parameters and incorrect inferences. Additionally, Lord (2006) studied the impact of a low sample mean and small sample size on the fixed dispersion parameter. Using simulation-based methods under various combinations of the sample mean, sample size, and dispersion parameter, Lord recommended sample sizes sufficiently large to reduce the risk of unreliable estimates of the dispersion parameter. Further studies by Shirazi et al. (2017) and Shirazi and Lord (2018) used machine learning techniques to investigate heuristics (including percentage-of-zeros, skewness, and kurtosis) used in the selection of competing distributions for negative binomial and Poisson models. Finally, Khazraee et al. (2018) utilized simulation in evaluating Bayesian Poisson-hierarchical models from the perspective of crash predictions at specific locations.

This paper investigates the application of negative binomial models to study traffic safety on rural local roads, facilities that tend to have the lowest crash sample means among all types of public roads. From this point of view, the focus is on a practical aspect of safety modeling rather than on specific statistical conditions defining what sample mean and sample size are too low to guarantee acceptable results already studied in the past. The practical importance of the study is defined by the set of negative binomial models already developed and implemented for other public roads in the Highway Safety Manual (American Association of State Highway and Transportation Officials (AASHTO), 2010). County roads are a notable missing element.

The bivariate negative binomial model, introduced by Maher (1990) for crash data modeling, is proposed in the current context to model the joint distribution of crash counts at two severity levels. The model structure accounts for possible overdispersion (sample variance greater than the sample mean) that is common in crash data (Washington et al., 2011). The first equation applies to fatal/injury (FI) crashes, which are crashes categorized as fatal (K), disabling injury (A), evident injury (B), and possible injury (C) on the KABCO scale (National Safety Council, 2016). The second equation applies to property damage only (PDO) crashes, or those categorized as O on the KABCO scale. It is expected that the crash types within each group share similarities in terms of the crash-causing mechanism.

The derivation of the bivariate negative binomial model is discussed extensively in the work of Famoye (2010) and summarized by Xu and Hardin (2016). Let Y_{ig} represent a count dependent variable, where $i = 1,2, \ldots, n$ represent each observation and $g = 1,2$ represents the two groups (FI and PDO crash types). Furthermore, the vector of explanatory variables (road and roadside features) is represented as: $\boldsymbol{x}_{ig} = (x_{ig0} = 1,\ x_{ig1}, x_{ig2}, \ldots, x_{igk})$. The bivariate negative binomial model represents the joint distribution of the count dependent variables (Y_{i1}, Y_{i2}) given the explanatory variables (x_{i1}, x_{i2}). The probability function of the bivariate negative binomial distribution may be written as shown in Eq. (1):

$$P(y_{i1}, y_{i2}) = \prod_{g=1}^{2} \binom{m_g^{-1} + y_{ig} - 1}{y_{ig}} \left(\frac{\mu_{ig}}{(m_g^{-1} + \mu_{ig})}\right)^{y_{ig}} \left(\frac{m_g^{-1}}{m_g^{-1} + \mu_{ig}}\right)^{m_g^{-1}} \times [1 + \lambda(e^{-y_{i1}} - c_1)(e^{-y_{i2}} - c_2)] \tag{1}$$

where: μ_{ig} = marginal means of negative binomial distribution,

m_g = overdispersion parameter,

λ = the multiplicative factor parameter,

$c_g = \left[\frac{1-\theta_g}{1-\theta_g e^{-1}}\right]^{m_g^{-1}}$

θ_g = mean of the Poisson distribution.

If $\lambda = 0$, then the count dependent variables Y_{i1} and Y_{i2} are independent and the FI and PDO crash types can be modeled separately using two univariate negative binomial models. However, if λ is found to be significantly greater than 0, then the crash types are positively correlated and may be modeled using the bivariate negative binomial model. The bivariate negative binomial model is estimated using maximum likelihood estimation to obtain the parameters $\boldsymbol{\beta}_{gj}$ ($g = 1,2$ and $j = 0,1,\ 2,\ \ldots, k$), m_g, and λ.

The bivariate ordered probit model was applied by Hall (2017) for rural local intersections to simultaneously estimate the correlated distributions of grouped FI and PDO crash counts. More specifically, the bivariate ordered probit model estimated the risk of having varying numbers of FI and PDO crashes based on the intersection's features (for instance, the risk of having zero, one, or two or more FI crashes). The application of the model in this way varies from its usual application in estimating the probability of various severity outcomes from individual crashes. Eq. (2) presents the basic form of the bivariate ordered probit model (Greene and Hensher, 2009):

$$\begin{aligned} z_{i1} &= \boldsymbol{\beta}_1 \boldsymbol{x}_{i1} + \varepsilon_{i1} \quad o_{i1} = l \ \ \textit{if} \ \ \mu_{l-1} < z_{i1} < \mu_l \quad l = 0, \ldots, L_1 \\ z_{i2} &= \boldsymbol{\beta}_2 \boldsymbol{x}_{i2} + \varepsilon_{i2} \quad o_{i2} = l \ \ \textit{if} \ \ \delta_{l-1} < z_{i2} < \delta_l \quad l = 0, \ldots, L_2 \end{aligned} \tag{2}$$

The error terms are related by Eq. (3):

$$\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right] \tag{3}$$

where: z_{ig} = latent variable used as basis for modeling ordinal ranking for $g = 1,\ 2$,

$\boldsymbol{\beta}_g$ = vector of regression parameters,

$\boldsymbol{x}_{ig}$ = vector of explanatory variables,

ε_{ig} = random error term (normally distributed),

o_{ig} = observed ordinal data,

l = integer ordered outcome,

μ and δ = estimable parameters (thresholds) defining z, and estimated jointly with β parameters,

λ = correlation coefficient for error terms.

Statistical significance of the λ term justifies the bivariate model form. If this term is found to be insignificant, then two separate ordered probit models could be estimated for the two crash types without any loss of the estimation efficiency. The bivariate ordered probit model is estimated using the maximum likelihood principle.

The features of the studied intersections are included in the bivariate negative binomial and bivariate ordered probit models for FI and PDO crashes. Although the zero-inflated count models discussed in Shankar et al. (1997) were considered for the analysis, these models were not estimated due to the reservation expressed in the published literature towards this model if inherently safe roads are difficult to claim (Lord et al., 2005, 2007). The following section discusses the available crash data and the distribution of the observed crashes among the intersections. It also discusses other data used to estimate the models.

3. Data

This analysis focuses on Tippecanoe County, which is located in northwest Indiana. Tippecanoe County's road network consists of 840 miles of county roads with 1 million VMT daily (Indiana Department of Transportation, 2015).

Two types of intersections are evaluated in this study: intersections of two county roads (218 observations) and intersections of one state road and one county road (61 observations). Crash data was obtained for the period 2012–2015 from the Automated Reporting Information Exchange System (ARIES), a database of crashes reported by Indiana law enforcement agencies. Since the focus of this analysis is on identifying intersection features that affect safety, crashes with deer (constituting 21.3% of the total crashes at the studied intersections) were removed from the sample. Table 1 shows the breakdown of crashes by severity level that were used in the statistical modeling.

Intersection-related crashes fell within the impact zone, which included the area within a 150-foot radius surrounding the midpoint of each intersection. The intersection locations were identified during the field data collection (discussed later in this section). Figs. 1 and 2 show the distribution of the crash frequency per intersection for FI and PDO crashes, respectively. These figures also display the ordinal grouping of crashes into bins that were used in estimating the bivariate ordered probit model, determined based on the breaks observed in the crash frequency distributions.

Annual Average Daily Traffic (AADT) for each intersection approach was obtained from the Tippecanoe County Highway Department and the Indiana Department of Transportation (INDOT) for the period from 2012 to 2016. When applicable, the values were adjusted to the middle of the crash period (2013) using yearly adjustment factors available from INDOT (Indiana Department of Transportation, 2016). Separate adjustment factors were used for the local road AADTs and state road AADTs (at applicable state-local intersections).

In the context of the studied intersections, the major road was considered to be the road without traffic controls, while the minor road was the road with traffic controls. In the case of an all-way stop or uncontrolled intersection, the road with higher AADT was considered to be the major road. For the major road, the mean AADT was 1912 vehicles per day (vpd), with a minimum of 8 vpd and a maximum of 12,489 vpd. AADT on the minor road had a mean of 295 vpd, with a minimum of 8 vpd and a maximum of 4645 vpd. Furthermore, the AADT was subdivided into different categories based on typical volume cutoffs used in past research (Souleyrette et al., 2010; Ewan et al., 2016). This categorization of roads by volume may facilitate model estimation in the absence of reliable AADT, data that is often unavailable at the local level. For the major road, the categorization was as follows:

- Low AADT (AADT ≤ 400 vpd)
- Medium AADT (400 vpd < AADT ≤ 1000 vpd) (base condition)
- High AADT (AADT > 1000 vpd)

For the minor road, the categories included:

- Low AADT (AADT ≤ 100 vpd)
- Medium AADT (100 vpd < AADT ≤ 400 vpd) (base condition)
- High AADT (AADT > 400 vpd)

The road surface type, obtained using data from the Tippecanoe County Highway Department and supplemented using Google Earth aerial imagery, was also split into three categories each for the major and minor roads. These included major and minor roads with all approaches unpaved, one unpaved and one paved approach, and all approaches paved (base condition).

Data collection of road and roadside features was conducted in August 2016 over the entire Tippecanoe County road network. A minivan equipped with portable GPS, cameras, and laptops facilitated the data collection (see Fig. 3). The data collection quality was ensured by keeping the workload at the manageable level. This was achieved by splitting the work between two observers, by equipping them with a convenient data recording tool, and by controlling the data flow level by adjusting the van's speed accordingly. The observers noted the features of segments and intersections using a special Data Collector software application developed in the Purdue Center for Road Safety (see Fig. 4). For example, each time the vehicle passed by an intersection, an observer would press keys noting the type of traffic controls (stop, yield, etc.) present on each intersection approach. Road features were assigned evenly to reduce the potential for overload among the observers; furthermore, the vehicle's driver maintained a conservative speed that was adjusted according to the preference of the observers and the number of road features present. During the post-processing phase, the road features were associated with the GPS coordinate data in order to determine the exact intersection locations. All field-collected

Table 1
Crash Data Summary by Severity Level at 279 Sites (2012–2015).

Crash severity	Total number of crashes	Number of intersections with zero crashes	Average crashes per intersection
FI	116	214	0.416
PDO	247	173	0.885

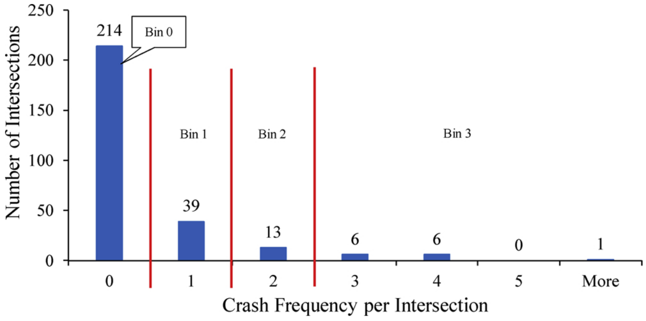


Fig. 1. Distribution of the Frequency of FI Crashes per Intersection.

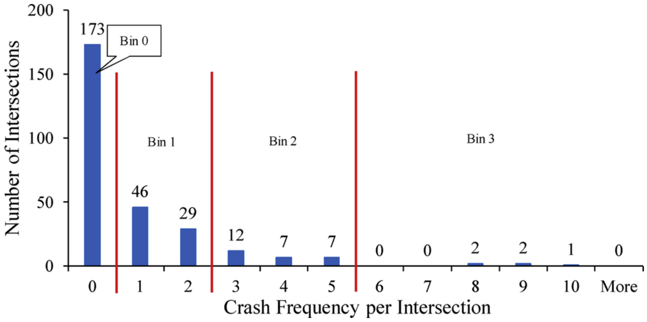


Fig. 2. Distribution of the Frequency of PDO Crashes per Intersection.

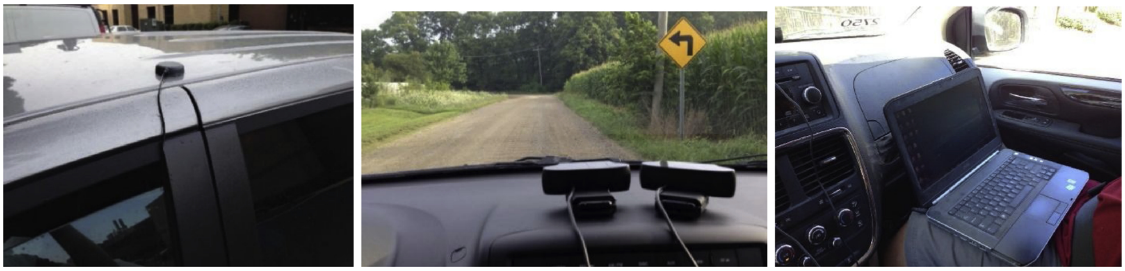


Fig. 3. Data Collection Setup.

data were reviewed in the lab for accuracy using the recorded video and current aerial imagery.

Most of the intersection features (for example, the number of legs or intersection angle) were obtained or measured using color aerial imagery in Google Earth. The historical imagery was reviewed to ensure that there were no changes in geometry, roadside conditions, or road surface type over the entire crash period at each of the studied intersections. Intersections with such changes that occurred during the crash period were not considered in the analysis. Table 2 summarizes the data that was used to create variables. It should be noted that there were no posted speed limits on the approach roads to most of the studied county intersections; hence, speed limit was not considered as a variable in the analysis. However, the unposted maximum speed limit on these roads was 55 mph.

The crashes, AADT, and intersection features were compiled in an Excel spreadsheet and utilized to create variables. Statistical modeling was conducted using packages developed by Sajaia (2008) and Xu and Hardin (2016) in the Stata statistical software.

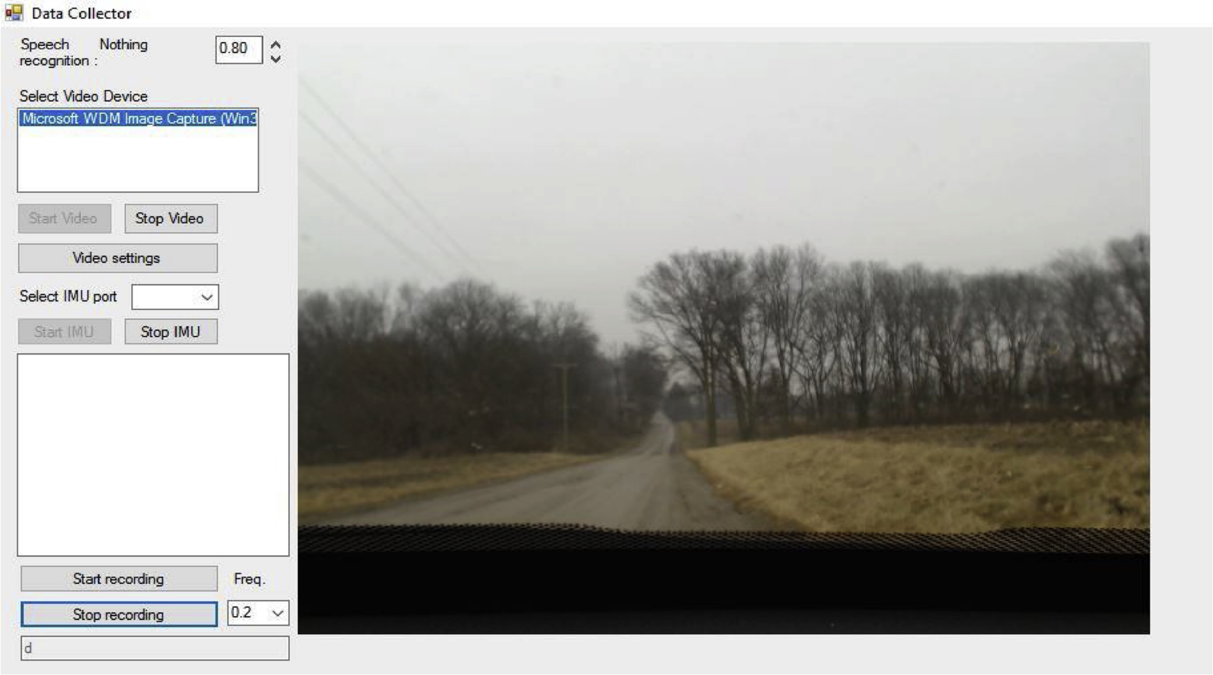
4. Results and discussion

4.1. Model estimation

Model specifications for the bivariate negative binomial were determined by considering all the variables derived from the AADT and intersection features in Table 2 in the initial model. A sequence of trials involved removing insignificant variables (at a confidence level of 0.80) and keeping the significant variables. Each of the variables was independently considered in each of the two equations for FI and PDO crashes. The model was re-estimated each time with the remaining variables until the final model was obtained.

The relatively low confidence level was selected to retain in the final models as many variables as possible with a particular attention given to variables that might prompt practical safety improvements. Retaining as many variables in the model is justified as a good practice if the primary purpose of the analysis is unbiased estimation of individual safety effects rather than the most parsimonious prediction model. A higher confidence could be achieved by expanding the data collection to additional counties. The limited resources did not allow for this treatment.

Two bivariate negative binomial models were estimated: the first included the AADT on the major and minor roads represented as categorical variables (Low, Medium, and High volume) (Table 3), and the second included the log of AADTs as continuous variables (Table 4). The first model is justified as a practical option for rural local roads where AADT values are unknown, but categorization of the traffic at the proposed three levels is possible to local agencies. On the other hand, an accurate traffic representation using the direct AADT improves the model's performance. This claim was confirmed by determining each model's Akaike Information Criterion (AIC) (see Eq. (4)). *LL* represents the log-likelihood for each model, while ν is the number of estimated model parameters (Famoye, 2010):



Data Collector

Speech recognition : Nothing

0.80

Select Video Device

Microsoft WDM Image Capture (Win3

Select IMU port

Freq. 0.2

d

Fig. 4. Data Collector Software.

Table 2
Summary of Intersection Features.

Data type	Details and number of intersections (in parentheses)	Data source(s)
Traffic	Major road: Low AADT (101) Medium AADT (61) High AADT (117) Minor road: Low AADT (135) Medium AADT (87) High AADT (57)	Tippecanoe County Highway Department and INDOT
Road surface type	Major road: All approaches unpaved (35) One unpaved and one paved approach (7) All approaches paved (237) Minor road: All approaches unpaved (80) One unpaved and one paved approach (25) All approaches paved (174)	Tippecanoe County Highway Department and Google Earth aerial imagery
Number of legs	3-leg intersections (178) 4-leg intersections (101)	Google Earth aerial imagery
Intersection offset	Offset (12) Not offset (267)	Ruler for measuring offset (Google Earth)
Intersection angle	90 degrees (210) 75-89 degrees (36) 60-74 degrees (20) <60 degrees (13)	Ruler for measuring headings (directions) of intersecting roads (Google Earth)
Intersection corners with agricultural land use	One or more corners (214) No corners (65)	Google Earth aerial imagery
Intersection corners with trees/bushes	One or more corners (95) No corners (184)	Google Earth aerial imagery
Classification of major road	State (61) Local (218)	Maps in Google Earth
Facility type of major road	Undivided highway (268) Divided highway (11)	Google Earth aerial imagery
Type of traffic control	Two-way stop (259) All-way stop (9) Yield-control (10) Uncontrolled (1)	Field data collection
Number of lanes on major road	2 lanes (267) 4 lanes (12)	Google Earth aerial imagery
Turn lane(s) on major road	Right turn lane (4) Left turn lane (8)	Google Earth aerial imagery
Intersection passing lane	With intersection passing lane (4) Without intersection passing lane (275)	Google Earth aerial imagery
Driveway(s) within impact zone	On major road (83) On minor road (44)	Google Earth aerial imagery and ruler for measuring distance from intersection midpoint
Access point(s) within impact zone*	On major road (9) On minor road (1)	Google Earth aerial imagery and ruler for measuring distance from intersection midpoint
Roadway curve(s) within impact zone	On major road (26) On minor road (12)	Google Earth aerial imagery and ruler for measuring distance from intersection midpoint

* For the purposes of this study, access points represent major driveways for groups of three or more houses as well as entrances to housing subdivisions.

$$AIC = -2LL + 2\nu \tag{4}$$

The model with categorical AADT produces an AIC value of 1001.816, while the model with continuous AADT has an AIC value of 980.342, implying that the model with continuous AADT outperforms the model with categorical AADT. Since volumes were available in this study, the model with continuous AADTs provides the main focus of this text, while the implementation implications for both the models are discussed at the end of this section.

Table 5 shows the estimation results for the bivariate ordered probit model with continuous AADTs. The predictive performance of this model is compared with its bivariate negative binomial counterpart.

4.2. Predictive performance

To assess the predictive performance of each model, cumulative sums (partial sums) of the observed and predicted crashes were compared. First, the predicted crash frequencies were arranged in ascending order and paired with their corresponding observed crash frequencies. Then, the curves representing the cumulative sums of the predicted and observed crash frequencies were plotted. Fig. 5 presents the obtained cumulative sum curves for the bivariate negative binomial and bivariate ordered probit models. A vertical distance between the two curves at a certain point indicates an overall prediction bias accumulated across all predictions lower than the prediction represented by the point. On the other hand, any difference in the slopes of the two curves indicates a bias contributed locally by the corresponding prediction. Obviously, limited local discrepancies are expected due to the random fluctuation of crash counts. Overestimation (or underestimation) bias consistently present within a sequence of predictions (sites) makes the predicted and observed curves gradually divert from one another. Although the resulting gap may then persist for a consecutive set of predictions (sites), an approximately fixed gap magnitude indicate that the predictions and observations are consistent again within the sequence. Overall good prediction in the entire range is manifested with only local curve deviations, while in general the two curves coincide with one another.

Although the cumulative predictions produced in the bivariate

Table 3
Bivariate Negative Binomial Model with Categorical AADT.

Variable name	Estimated parameter	Standard error	*t*-statistic	*p*-value
FI crashes				
Intercept	−1.658	0.437	−3.79	<0.01
Low or Medium AADT on the major road (AADT ≤ 1000)	0			
High AADT on the major road (AADT > 1000)	0.585	0.314	1.86	0.06
Low AADT on the minor road (AADT ≤ 100)	−0.538	0.345	−1.56	0.12
Medium AADT on the minor road (100 vpd < AADT ≤ 400 vpd)	0			
High AADT on the minor road (AADT > 400)	1.081	0.292	3.71	<0.01
All approaches unpaved on the major road (1 if yes, 0 if no)	−1.581	1.048	−1.51	0.13
Intersection has three legs (1 if yes, 0 if no)	−0.598	0.267	−2.24	0.03
Agricultural land use in one or more intersection corners (1 if yes, 0 if no)	0.524	0.310	1.69	0.09
Major road is a state highway (1 if yes, 0 if no)	0.359	0.283	1.27	0.20
Overdispersion parameter	0.669	0.299	2.24	0.03
PDO crashes				
Intercept	−0.811	0.245	−3.31	<0.01
Low or Medium AADT on the major road (AADT ≤ 1000)	0			
High AADT on the major road (AADT > 1000)	1.023	0.208	4.93	<0.01
Low AADT on the minor road (AADT ≤ 100)	−0.670	0.247	−2.71	0.01
Medium AADT on the minor road (100 vpd < AADT ≤ 400 vpd)	0			
High AADT on the minor road (AADT > 400)	0.833	0.216	3.85	<0.01
All approaches unpaved on the major road (1 if yes, 0 if no)	−2.015	1.031	−1.95	0.05
Intersection has three legs (1 if yes, 0 if no)	−0.397	0.193	−2.05	0.04
Presence of driveways and/or access points on the major road approach (1 if yes, 0 if no)	0.335	0.195	1.71	0.09
Presence of trees and/or bushes obstructing sightlines in one or more intersection corners (1 if yes, 0 if no)	0.323	0.183	1.77	0.08
Major road is a four-lane divided highway (1 if yes, 0 if no)	−0.636	0.389	−1.63	0.10
Overdispersion parameter	0.575	0.166	3.45	<0.01
λ	1.177	0.660	1.78	0.08
Log-likelihood	−480.908	–	–	–
Number of observations	279	–	–	–
Likelihood ratio test of independence: $\chi^2 = 3.2155$, Probability $> \chi^2 = 0.0729$				

negative binomial models locally deviate above or below the cumulative observations, the two cumulative quantities tend to follow each other rather closely. Interestingly, the cumulative predictions for FI crashes from the bivariate negative binomial model with categorical AADT appear to follow the cumulative observations slightly better than the bivariate negative binomial model with continuous AADT. The cumulative predictions from the bivariate ordered probit model also tend to follow the cumulative observations. However, the cumulative predictions from the bivariate ordered probit model appear to depart from the cumulative observations to a larger extent than for the bivariate negative binomial models. This may be an undesirable side effect of aggregating crash counts in ranges that represent discrete outcomes in the ordered probit model. The root mean square error displayed in Table 6 shows a similar finding when comparing the bivariate negative binomial and bivariate ordered probit models with continuous AADT, albeit a relatively small magnitude of difference. The results prompt that the bivariate negative binomial model with continuous AADT may produce slightly more accurate predictions than the corresponding bivariate ordered probit model.

The following discussion focuses on the results of the bivariate negative binomial model with continuous AADT. The model predictions conditioned on a given variable were compared with the corresponding

Table 4
Bivariate Negative Binomial Model with Continuous AADT.

Variable name	Estimated parameter	Standard error	*t*-statistic	*p*-value
FI crashes				
Intercept	−6.677	0.942	−7.09	<0.01
Log of AADT on the major road	0.389	0.092	4.24	<0.01
Log of AADT on the minor road	0.505	0.110	4.59	<0.01
Intersection has three legs (1 if yes, 0 if no)	−0.558	0.265	−2.11	0.04
Agricultural land use in one or more intersection corners (1 if yes, 0 if no)	0.696	0.316	2.20	0.03
Overdispersion parameter	0.591	0.280	2.11	0.04
PDO crashes				
Intercept	−5.978	0.585	−10.22	<0.01
Log of AADT on the major road	0.407	0.075	5.42	<0.01
Log of AADT on the minor road	0.544	0.075	7.25	<0.01
Intersection has three legs (1 if yes, 0 if no)	−0.337	0.181	−1.87	0.06
Presence of driveways and/or access points on the major road approach (1 if yes, 0 if no)	0.302	0.179	1.69	0.09
Presence of trees and/or bushes obstructing sightlines in one or more intersection corners (1 if yes, 0 if no)	0.224	0.172	1.30	0.19
Major road is a four-lane divided highway (1 if yes, 0 if no)	−0.943	0.382	−2.47	0.01
Overdispersion parameter	0.406	0.146	2.77	0.01
λ	0.928	0.684	1.36	0.18
Log-likelihood	−475.171	–	–	–
Number of observations	279	–	–	–
Likelihood ratio test of independence: $\chi^2 = 1.8624$, Probability $> \chi^2 = 0.1723$				

Table 5
Bivariate Ordered Probit Model with Continuous AADT.

Variable name	Estimated parameter	Standard error	*t*-statistic	*p*-value
FI crashes				
Intercept	−4.075	0.665	−6.13	< 0.01
Log of AADT on the major road	0.189	0.091	2.09	0.04
Log of AADT on the minor road	0.358	0.085	4.24	< 0.01
Intersection has three legs (1 if yes, 0 if no)	−0.308	0.190	−1.62	0.11
Agricultural land use in one or more intersection corners (1 if yes, 0 if no)	0.486	0.229	2.12	0.03
Major road is a state highway (1 if yes, 0 if no)	0.335	0.262	1.28	0.20
Threshold 1	0.786	0.117	6.74	< 0.01
Threshold 2	1.317	0.169	7.79	< 0.01
PDO crashes				
Intercept	−4.670	0.518	−9.02	< 0.01
Log of AADT on the major road	0.352	0.068	5.15	< 0.01
Log of AADT on the minor road	0.421	0.075	5.58	< 0.01
Intersection has three legs (1 if yes, 0 if no)	−0.418	0.172	−2.43	0.02
Presence of driveways and/or access points on the major road approach (1 if yes, 0 if no)	0.319	0.169	1.89	0.06
Presence of trees and/or bushes obstructing sightlines in one or more intersection corners (1 if yes, 0 if no)	0.372	0.162	2.29	0.02
Major road is a four-lane divided highway (1 if yes, 0 if no)	−1.007	0.402	−2.51	0.01
Threshold 1	1.348	0.138	9.75	< 0.01
Threshold 2	2.630	0.258	10.19	< 0.01
ρ	0.221	0.105	2.21	0.03
Log-likelihood	−369.572	–	–	–
Number of observations	279	–	–	–

cumulative observations to detect any prediction bias that could be attributed to incorrectly estimated model parameters (discussed in Lord and Mannering, 2010). Fig. 6 shows the cumulative sum curves for three indicator variables: (1) presence of driveways and/or access points on the major road approach, (2) presence of trees/bushes obstructing sightlines in one or more intersection corners, and (3) agricultural land use in one or more intersection corners. These three indicator variables were selected because they represent non-exposure features that prompt potential countermeasures aimed at improving safety at county intersections. The dataset was split into two subsets for each of these indicator variables: intersections without the particular characteristic of interest (indicator = 0) and intersections with the characteristic (indicator = 1). Within each subset, the predicted crash frequencies were arranged in ascending order and paired with their corresponding observed crash frequencies. An estimation issue of a model parameter may be claimed if the two cumulative sum curves diverge from one another for the particular condition represented by the same value of the indicator variable (0 or 1).

Although the slopes of the two cumulative sum curves vary locally from one another, these discrepancies may be attributed to the randomness associated with crash counts. In general, the curves tend to remain parallel. The only puzzling result is a group of 18 intersections with surrounding non-agricultural land-use shown in Fig. 6 that experienced no FI crashes, represented by the horizontal line of cumulative observations. Except in this limited range, the cumulative sum curves do not seem to diverge from each other and there is no obvious indication of systematic prediction bias among the intersections without and with the studied three road features.

4.3. Significant variables

The final bivariate negative binomial model includes the AADT on the major and minor roads as well as a variety of other explanatory variables related to the intersection features. All variables in the FI crash equation and most in the PDO crash equation are statistically significant at the 0.90 confidence level (or greater), with one variable in the PDO equation significant at the 0.80 confidence level.

In each of the equations, the significant overdisperison parameters indicate that the variance is greater than the mean, justifying the negative binomial model specification over the Poisson. The *p*-value of λ shows that the dependence between the FI and PDO outcomes is rather weak. For further investigation, separate univariate negative binomial models were estimated for FI and PDO crashes utilizing the same significant variables. As shown in Table 7, the difference in estimated parameters resulting from the bivariate negative binomial model and univariate negative binomial models is negligible from the practical point of view. While both models are suitable, the bivariate negative binomial model provides a small improvement in efficiency in comparison to having separate univariate negative binomial models, which is confirmed from the results of the likelihood ratio test of independence.

The log of AADT is positively correlated with both FI and PDO crashes, a finding that is observed for both the major and minor roads. The effect of AADT on crashes has been widely studied by other researchers for rural state-administered intersections (for instance, Tarko et al., 2012). AADT represents a measure of exposure.

The presence of driveways and/or access points on the major road approach (no stop sign in almost all cases) to the intersection is associated with an increase in the PDO crash frequency. Similar findings have been observed in past studies at rural intersections for various crash types (Kim et al., 2006). Drivers on the major road who pay attention to driveways may divert their attention from the intersection, thus increasing the risk of a crash with drivers who violated the traffic controls on the crossing road. While this variable is significant at the PDO severity level, it is not found to be an influential factor at the FI severity level. This might be simply the result of there being fewer observations of FI crashes. Another possibility is that drivers approaching an intersection on the major road with driveways tend to slow down, thereby reducing the severity of a crash.

The presence of trees and/or bushes obstructing sightlines in one or more intersection corners is associated with an increased frequency of PDO crashes. In addition, intersections with agricultural land use in one or more corners have an increased frequency of FI crashes. The former factor represents a sight obstruction that is present year-round at the intersection, while the latter may only be relevant during certain times of year (late summer) if agricultural crops become tall enough to block driver sightlines. Limited sight distance may prevent drivers from seeing other potentially conflicting vehicles approaching the intersection, which may be particularly consequential at intersections with frequent traffic control violations. At intersections with trees/bushes

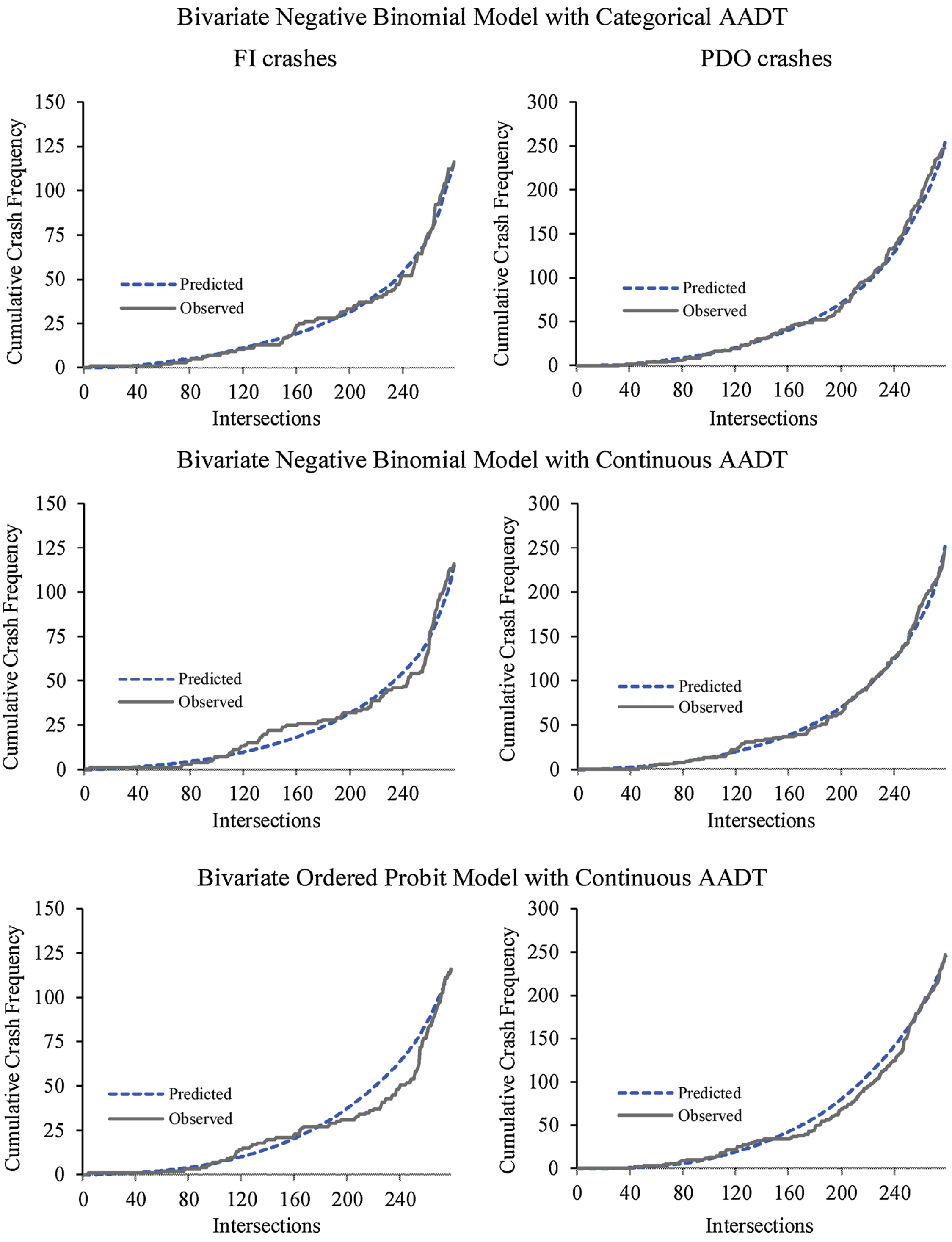


Fig. 5. Cumulative Sum Curves for Study Intersections.

Table 6
Prediction Accuracy Comparison using Root Mean Square Error.

Crash type	Bivariate negative binomial model with categorical AADT	Bivariate negative binomial model with continuous AADT	Bivariate ordered probit model with continuous AADT
FI	0.828	0.825	0.831
PDO	1.364	1.219	1.250

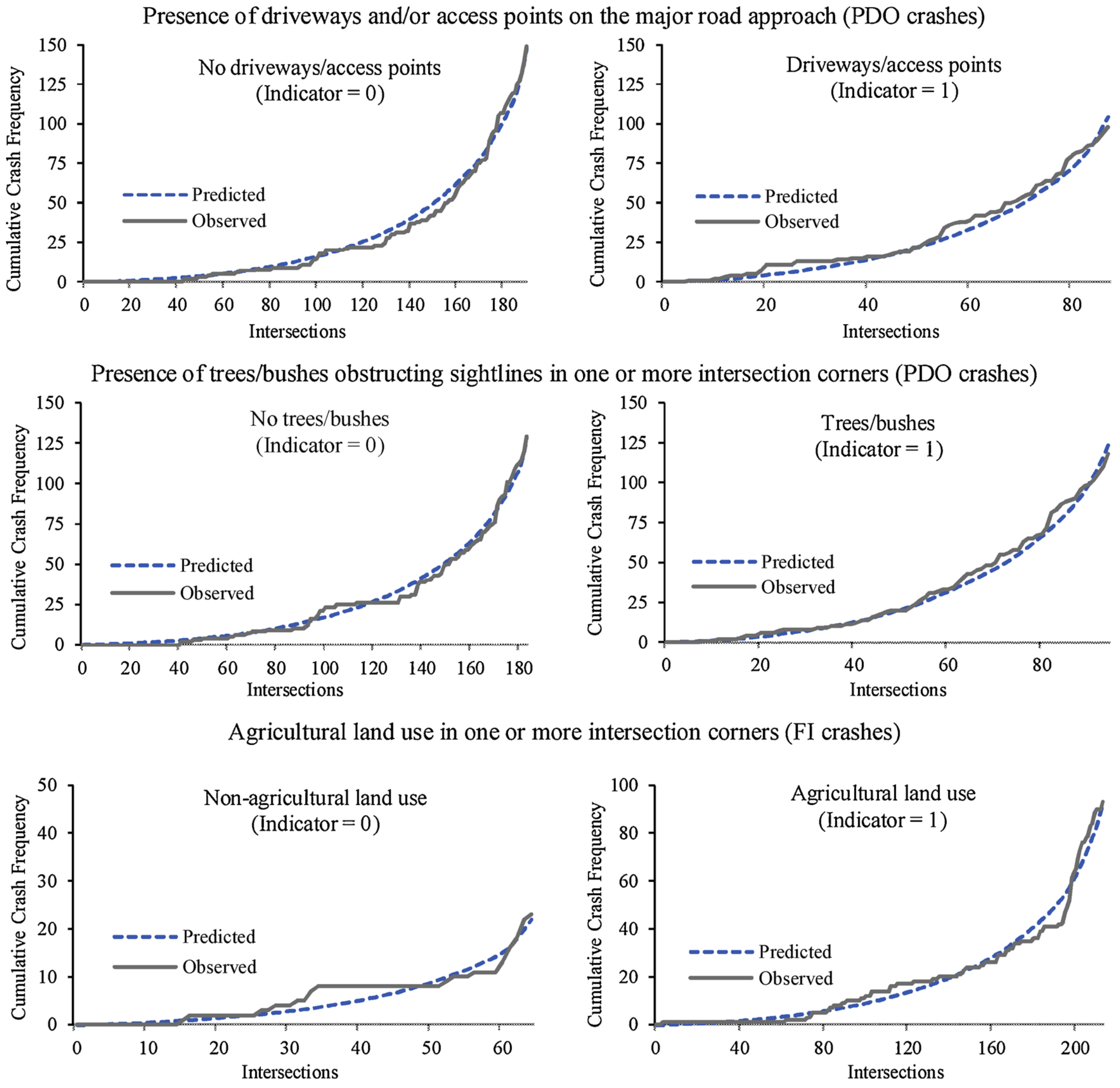


Fig. 6. Bivariate Negative Binomial Model Cumulative Sum Curves for Selected Indicator Variables.

Table 7
Comparison of Estimated Parameters from Bivariate Negative Binomial Model and Univariate Negative Binomial Models.

Variable name	Estimated parameter (standard error in parentheses)	
	Bivariate Negative Binomial Model	Univariate Negative Binomial Models
FI crashes		
Intercept	−6.677 (0.942)	−6.669 (0.945)
Log of AADT on the major road	0.389 (0.092)	0.396 (0.091)
Log of AADT on the minor road	0.505 (0.110)	0.502 (0.110)
Intersection has three legs (1 if yes, 0 if no)	−0.558 (0.265)	−0.586 (0.262)
Agricultural land use in one or more intersection corners (1 if yes, 0 if no)	0.696 (0.316)	0.656 (0.314)
Overdispersion parameter	0.591 (0.280)	0.572 (0.273)
PDO crashes		
Intercept	−5.978 (0.585)	−5.973 (0.586)
Log of AADT on the major road	0.407 (0.075)	0.406 (0.075)
Log of AADT on the minor road	0.544 (0.075)	0.542 (0.075)
Intersection has three legs (1 if yes, 0 if no)	−0.337 (0.181)	−0.331 (0.181)
Presence of driveways and/or access points on the major road approach (1 if yes, 0 if no)	0.302 (0.179)	0.311 (0.180)
Presence of trees and/or bushes obstructing sightlines in one or more intersection corners (1 if yes, 0 if no)	0.224 (0.172)	0.221 (0.173)
Major road is a four-lane divided highway (1 if yes, 0 if no)	−0.943 (0.382)	−0.898 (0.379)
Overdispersion parameter	0.406 (0.146)	0.408 (0.147)

continually present, drivers are typically aware of the sight obstruction and may compensate for the heightened risk by lowering their speed, consequently reducing the crash severity. However, the sight obstruction by agriculture crops varies by the type of crop grown and the time of the year. Drivers might not be as conscious of potential risks and take fewer precautions (such as lowering their speed) to avoid conflicts, hence leading to higher severity crashes. Consistent with these findings and notwithstanding other intersection features, the percentage of FI crashes occurring during the timeframe when crops tend to be their tallest (July-November) was greater at intersections with agricultural land use (54.8%) than at intersections with non-agricultural land uses (43.5%).

The variable representing three leg intersections is negatively associated with crash frequency at both the FI and PDO severity levels (in comparison to four leg intersections). At three leg intersections, there is typically less traffic entering the intersection, and hence less potential interactions between vehicles on the major and minor roads. Moreover, the simpler traffic pattern of six turning movements at three leg intersections is less conducive to driving errors than at four leg intersections with twelve turning movements. Finally, the geometry of most three leg intersections is such that vehicles on one of the intersection legs, usually the minor road, are forced to slow down or stop since the road terminates at the intersection. This provides more time to react to the presence of other vehicles and to safely proceed through the intersection.

Finally, intersections where the major road is a four-lane divided highway are associated with a decrease in PDO crash frequency. The medians considered in this study were sufficiently wide to allow vehicles crossing the major road in two phases, thus simplifying the crossing maneuver and lowering the risk of an error. The effect of medians on decreasing FI crashes was not confirmed. The main factor of crash severity at two-way stop-controlled intersections, speed on the major road, was not affected or could even be slightly higher on divided roads than on undivided roads.

4.4. Implementation

The bivariate negative binomial with continuous AADT provided an improvement in model fitness over the bivariate negative binomial with categorical AADT. Other significant conditions represented in the models include:

- intersection has three legs (FI and PDO crashes),
- agricultural land use in one or more intersection corners (FI crashes),
- presence of driveways and/or access points on the major road approach (PDO crashes),
- presence of trees and/or bushes obstructing sightlines in one or more intersection corners (PDO crashes), and
- major road is a four-lane divided highway (PDO crashes).

The estimated parameters of these common variables are shown in Table 8.

Estimated parameters between corresponding variables are typically similar, with the greatest difference in variables that may jointly represent themselves and the AADT in the model with less precise representation of the exposure. This applies to the four-lane divided highway variable in the PDO crash model. Moreover, two other variables appear in the categorical AADT model: (1) All approaches unpaved on the major road, and (2) Major road is a state highway. These variables may also be simultaneously representing themselves and the AADT in this model.

Given the above findings, the categorical AADT model should not be the basis for evaluating non-exposure effects that are correlated with the exposure, as they may be biased. On the other hand, the results prompt that either model is suitable when evaluating the specific impacts of non-AADT-correlated variables. Furthermore, the model with categorical AADT is suitable for predicting the number of crashes and it may be used as a practical alternative if the precise AADT is not known but can be assessed as a range.

5. Summary and conclusions

Rural local roads have among the greatest crash rates across all road facility types. Such road facilities comprise more than 80% of the total rural road mileage in numerous states, including Indiana. The established statistical methods for evaluating safety problems on higher-volume rural arterial roads, which typically involve the use of negative binomial regression, have not seen as widespread usage in studying and managing safety of lower-volume rural local roads. This is due in part to concerns raised over the model's estimation under a low sample mean and the lack of crash concentration at particular spots.

This paper sought to attain more insight into the adequacy of negative binomial models for studying and managing safety on the roads with the lowest crash frequencies, rural local roads. Bivariate negative binomial models were estimated for rural local intersections including various road and roadside features and the AADT as either a categorical or continuous variable, the former of which is useful in the absence of reliable traffic volumes. Model results were compared to the results of the bivariate ordered probit model used in past research for studying safety on roads with low numbers of crashes. For the models, cumulative sum curves of the predicted and observed crashes were compared for both FI and PDO crashes to detect any possible prediction bias. Although both models performed well, the bivariate negative binomial model seemed to produce slightly more accurate predictions than the bivariate ordered probit model. Additionally, the cumulative sum curves conditioned on individual safety effects indicated no obvious systematic bias in the estimated parameters. When resources warrant

Table 8
Comparison of Estimated Parameters of Common Variables from Bivariate Negative Binomial Models with Categorical and Continuous AADTs.

Variable name	Estimated parameter (standard error in parentheses)	
	Model with Categorical AADT	Model with Continuous AADT
FI crashes		
Intersection has three legs (1 if yes, 0 if no)	−0.598 (0.267)	−0.558 (0.265)
Agricultural land use in one or more intersection corners (1 if yes, 0 if no)	0.524 (0.310)	0.696 (0.316)
PDO crashes		
Intersection has three legs (1 if yes, 0 if no)	−0.397 (0.193)	−0.337 (0.181)
Presence of driveways and/or access points on the major road approach (1 if yes, 0 if no)	0.335 (0.195)	0.302 (0.179)
Presence of trees and/or bushes obstructing sightlines in one or more intersection corners (1 if yes, 0 if no)	0.323 (0.183)	0.224 (0.172)
Major road is a four-lane divided highway (1 if yes, 0 if no)	−0.636 (0.389)	−0.943 (0.382)

the researcher or practitioner, increasing the sample size by adding data from additional counties can help to further minimize the risk of bias in the dispersion parameter, discussed by Lord (2006).

The considerable number of significant variables and their intuitive signs further demonstrated the suitability of the estimated models. It was found that the AADTs on the major and minor roads strongly affected the frequency of FI and PDO crashes at the intersections. Driveways and access points on the major road approach and trees/bushes at the intersection corners increased the frequency of PDO crashes, while intersections with agricultural land use in the corners had an increased frequency of FI crashes. Three leg intersections had a decreased frequency of both crash types. Finally, intersections where the major road was a four-lane, divided highway were associated with decreased PDO crashes.

The results suggest an opportunity for applying the Highway Safety Manual methodology for rural local roads with low-volumes, with a focus on identifying safety-deficient road features at multiple locations rather than identifying individual high-crash locations. Based on the methodology presented in this study, agencies can determine locations with features that increase the frequency of crashes. Safety-deficient intersections can be determined from those intersections with features found to increase crash frequency. These intersections may be appropriate candidates for low-cost safety improvements (for example, clearing sight-obstructed intersection corners of trees and bushes). For rural roads with low volumes, past studies have indicated that safety remediation with multiple low-cost improvements applied at the road network level can offer greater benefits than costly improvements applied only at a few dangerous locations (Ivey and Griffin, 1992). The improvements can be grouped together into safety projects and implemented widely over the road network to achieve maximum benefits. The safety and monetary benefits obtained from applying low-cost improvements to multiple locations will be evaluated in future research.

Future research should compare safety effects estimated on rural local roads with those effects on well-studied rural arterial roads. If the similarity of effects on both road types are confirmed, then this could allow transferring at least part of the knowledge of highway safety factors (for example, crash modification factors) to rural local roads to facilitate the development of road screening and safety improvement measures for these roads.

While the current study focused on rural local intersections, future research should extend further to cover rural local segments. A similar statistical methodology may be used in examining the effect of segment-level variables related to the road curvature, driveways and access points, and the presence of roadside obstructions, among other factors. Segments may also offer further opportunities for safety improvements through the alignment, cross-sectional, and roadside components.

Acknowledgements

This research was conducted at the Purdue University Center for Road Safety (CRS). The research was facilitated in part through the funding of the Dwight David Eisenhower Transportation Fellowship Program. Additionally, the authors greatly appreciated the contributions of Dr. Mario Romero and Mr. Jose Thomaz from CRS for their help in data processing and software development that assisted in streamlining the more time-consuming tasks of this research. Mr. Jeremy Brown and Mr. Chibuzor Ezeamaku helped during the data collection and extraction effort. Mr. Cristhian Lizarazo and Mr. Tariq U. Saeed provided invaluable guidance in enhancing the paper, including in the statistical modeling and other research phases. Finally, Ms. Opal Kuhl and Mr. Mike Parks from the Tippecanoe County Highway Department provided valuable information on road features throughout the county.

References

American Association of State Highway and Transportation Officials (AASHTO), 2010. Highway Safety Manual.

Avelar, R., Dixon, K.K., Schertz, G., 2015. Identifying low-volume road segments with high frequencies of severe crashes. Transp. Res. Rec. (2472), 162–171.

Cafiso, S., La Cava, G., Montella, A., 2011. Safety inspections as supporting tool for safety management of low-volume roads. Transp. Res. Rec. (2203), 116–125.

Cafiso, S., Di Graziano, A., Pappalardo, G., 2015. Safety inspection and management tools for low-volume road network. Transp. Res. Rec. (2472), 134–141.

Ewan, L., Al-Kaisy, A., Hossain, F., 2016. Safety effects of road geometry and roadside features on low-volume roads in Oregon. Transp. Res. Rec. (2580), 47–55.

Famoye, F., 2010. On the bivariate negative binomial regression model. J. Appl. Stat. 37 (6), 969–981.

Federal Highway Administration, 2015. Highway Statistics. United States Department of Transportation, Washington, D.C.

Greene, W.H., Hensher, D.A., 2009. Modeling Ordered Choices. New York University Stern School of Business and University of Sydney Institute of Transport and Logistics Studies.

Hall, T., 2017. Project-Oriented Safety Management of Rural Local Roads. (Doctoral dissertation). Purdue University.

Hall, J.W., Rutman, E.W., Brogan, J.D., 2003. Highway Safety Challenges on Low-Volume Rural Roads. Institute of Transportation Engineers Annual Meeting, Seattle.

Indiana Department of Transportation, 2015. Mileage and DVMT by Year, County, and System. Retrieved September 9, 2018, from. http://www.in.gov/indot/files/TrafficStastics_HistoricINVMTByCntyAndSys(2006-2014).pdf.

Indiana Department of Transportation, 2016. Latest INDOT Traffic Adjustment Factors. Retrieved September 10, 2018, from. http://www.in.gov/indot/files/2016%20INDOT%20Adjustment%20Factors.pdf.

Ivey, D.L., Griffin III, L.I., 1992. Safety Improvements for Low Volume Rural Roads (Publication FHWA/TX-90/1130-2F). Texas Department of Transportation and Texas Transportation Institute.

Khazraee, S.H., Johnson, V., Lord, D., 2018. Bayesian Poisson hierarchical models for crash data analysis: investigating the impact of model choice on site-specific predictions. Accid. Anal. Prev. 117, 181–195.

Kim, D.G., Washington, S., Oh, J., 2006. Modeling crash types: new insights into the effects of covariates on crashes at rural intersections. J. Transp. Eng. 132 (4), 282–292.

Labi, S., 2006. Effects of Geometric Characteristics of Rural Two-Lane Roads on Safety (Publication FHWA/IN/JTRP-2005/2). Joint Transportation Research Program, Indiana Department of Transportation and Purdue University, West Lafayette, Indiana.

Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accid. Anal. Prev. 38 (4), 751–766.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. A Policy Pract. 44 (5), 291–305.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. 37 (1), 35–46.

Lord, D., Washington, S., Ivan, J.N., 2007. Further notes on the application of zero-inflated models in highway safety. Accid. Anal. Prev. 39 (1), 53–57.

Maher, M.J., 1990. A bivariate negative binomial model to explain traffic accident migration. Accid. Anal. Prev. 22 (5), 487–498.

National Safety Council, 2016. Estimating the Costs of Unintentional Injuries, 2014.

Sajaia, Z., 2008. Maximum likelihood estimation of a bivariate ordered probit model: implementation and Monte Carlo simulations. Stata J. 4 (2), 1–18.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accid. Anal. Prev. 29 (6), 829–837.

Shirazi, M., Lord, D., 2018. Characteristics based heuristics to select a logical distribution between the Poisson-gamma and the Poisson-lognormal for crash data modeling. Paper Presented at the 97th Annual Meeting of the Transportation Research Board.

Shirazi, M., Dhavala, S.S., Lord, D., Geedipally, S.R., 2017. A methodology to design heuristics for model selection based on the characteristics of data: application to investigate when the negative binomial Lindley (NB-L) is preferred over the negative binomial (NB). Accid. Anal. Prev. 107, 186–194.

Souleyrette, R.R., Caputeu, M., McDonald, T.J., Sperry, R.B., Hans, Z.N., Cook, D., 2010. Safety Analysis of Low-Volume Rural Roads in Iowa (No. InTrans Project 07-309). Institute for Transportation, Iowa State University.

Stapleton, S.Y., Ingle, A.J., Chakraborty, M., Gates, T.J., Savolainen, P.T., 2018. Safety performance functions for rural two-lane county road segments. Transp. Res. Rec., 0361198118799035.

Tarko, A.P., Leckrone, S., Anastasopoulos, P., 2012. Analysis and Methods of Improvement of Safety at High-Speed Rural Intersections (Publication FHWA/IN/JTRP-2012/01). Joint Transportation Research Program, Indiana Department of Transportation and Purdue University, West Lafayette, Indiana.

Washington, S.P., Karlaftis, M.G., Mannering, F., 2011. Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed. CRC press., Boca Raton, Florida.

Xu, X., Hardin, J., 2016. Regression models for bivariate count outcomes. Stata J. 16 (2), 301–315.

Zegeer, C.V., Stewart, R., Council, F., Neuman, T.R., 1994. Accident relationships of roadway width on low-volume roads. Transp. Res. Rec. (1445), 160–168.