Accident Analysis and Prevention 137 (2020) 105456

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Application of the Poisson-Tweedie distribution in analyzing crash frequency data

Dibakar Saha[a,*], Priyanka Alluri[b], Eric Dumbaugh[a], Albert Gan[b]

[a] School of Urban and Regional Planning, Florida Atlantic University, Boca Raton, FL, 33431, United States
[b] Department of Civil and Environmental Engineering, Florida International University, Miami, FL, 33174, United States

ARTICLE INFO

Keywords:
Traffic safety
Poisson-Tweedie distribution
Negative binomial model
Geometric Poisson model
Poisson inverse Gaussian model
Dispersion parameter

ABSTRACT

This paper describes a study that applies the Poisson-Tweedie distribution in developing crash frequency models. The Poisson-Tweedie distribution offers a unified framework to model overdispersed, underdispersed, zero-inflated, spatial, and longitudinal count data, as well as multiple response variables of similar or mixed types. The form of its variance function is simple, and can be specified as the mean added to the product of dispersion and mean raised to the power *P*. The flexibility of the Poisson-Tweedie distribution lies in the domain of P, which includes positive real number values. Special cases of the Poisson-Tweedie distribution models include the linear form of the negative binomial (NB1) model with *P* equal to 1.0, the geometric Poisson (GeoP) model with *P* equal to 1.5, the quadratic form of the negative binomial (NB2) model with *P* equal to 2.0, and the Poisson Inverse Gaussian (PIG) model with *P* equal to 3.0. A series of models were developed in this study using the Poisson-Tweedie distribution without any restrictions on the value of the power parameter as well as with specific values of the power parameter representing NB1, GeoP, NB2, and PIG models. The effects of fixed and varying dispersion parameters (i.e., dispersion as a function of covariates) on the variance and expected crash frequency estimates were also examined. Three years (2012–2014) of crash data from urban three-leg stop-controlled intersections and urban four-leg signalized intersections in the state of Florida were used to develop the models. The Poisson-Tweedie models or the GeoP models were found to perform better when the dispersion parameter was constant or fixed. With the varying dispersion parameter, the NB2 and PIG models were found to perform better, with both performing equally well. Also, the fixed dispersion parameter values were found to be smaller in the models with a higher value of the power parameter. The variation across the models in their estimates of weight factor, expected crash frequency, and potential for safety improvement of hazardous sites based on the empirical Bayes method was also discussed.

1. Introduction

Identifying high crash locations is the first step in the roadway safety management process. The Highway Safety Manual (HSM), published by the American Association of State Highway and Transportation Officials (AASHTO), describes various methods for identifying high crash locations (AASHTO, 2010). Of them, the most reliable methods are based on empirical Bayes (EB) procedure or the adjusted EB procedure (Kolody et al., 2014). The EB method provides an estimate of the expected crash frequency by combining the observed crash frequency with the predicted crash frequency from a regression model (Hauer, 1997). A reliable estimate of the predicted crash frequency is thus key to enhancing the reliability of the expected crash frequency estimate for a location needing improvements (Gan et al., 2016).

Fitting a suitable crash prediction model depends on the nature of crash frequency data. Crashes are independent, random, discrete, and can only be described by zero or positive integers. These properties lead to the assumption that crashes are Poisson distributed (Miaou, 1994). However, the Poisson distribution cannot accommodate the effect of overdispersion, which is another common phenomenon of crash frequency data. Overdispersion may be caused by unobserved variables, inappropriate sampling, possible correlations between individual observations, and other unforeseen issues. If overdispersion is not accounted for in a model, it may lead to biased estimates of regression parameters and incorrectly rejecting the null hypothesis of zero associations between model covariates and crash frequency (Mannering et al., 2016). To incorporate the effect of overdispersion, a random error term is included in the Poisson model. A suitable distribution is

* Corresponding author.
E-mail addresses: sahad@fau.edu (D. Saha), palluri@fiu.edu (P. Alluri), edumbaug@fau.edu (E. Dumbaugh), gana@fiu.edu (A. Gan).

https://doi.org/10.1016/j.aap.2020.105456
Received 20 October 2019; Received in revised form 26 January 2020; Accepted 27 January 2020
Available online 06 February 2020
0001-4575/

then assumed for the random error term, and the resulting distribution is called the Poisson mixture distribution (Cameron and Trivedi, 1998). The Poisson-Gamma mixture distribution (i.e., incorporating a Gamma-distributed random error term into the Poisson process) is the most widely used technique for analyzing the overdispersed crash data (Lord and Mannering, 2010). The Poisson-Gamma model is commonly referred to as the negative binomial (NB) model, as the resulting probability density function of the Poisson-Gamma distribution resembles that of the NB distribution (Hilbe, 2011).

The NB regression models implemented in the HSM and in the majority of other studies were based on the quadratic form of the variance parameter, formulated as $variance = mean + dispersion \times mean^2$. It has recently been shown that the quadratic form of the NB variance function and the associated probability density function may not always provide the best estimates of the model parameters (Green, 2008). For example, Wang et al. (2019) developed crash frequency models for various crash types in rural two-lane intersections using three functional forms of the NB regression models varied by the power of the second "mean" term in the aforementioned formula for variance function. The functional forms that were examined include NB1 (i.e., linear variance function with 1 being the power of the second mean term), NB2 (i.e., quadratic variance function as aforementioned), and NBP (i.e., variance function with no fixed value assigned to the power of the second mean term). In most cases, the NBP models outperformed the NB1 and NB2 models, with the exponent of the variance function being estimated to be significantly different from 2. These findings, along with similar results reported by Green (2008) in the analysis of health care data, suggest a need for a flexible approach in estimating the variance of NB models.

Recently, several studies have employed various other forms of Poisson mixture distributions in crash analysis. For example, Lord et al. (2008) chose to apply the Conway-Maxwell-Poisson distribution for its flexibility in handling both overdispersion and underdispersion in the crash data. Geedipally et al. (2012) presented the NB-Lindley (NB-L) distribution to model crash data with a large number of zeros and a long tail. Cheng et al. (2013) demonstrated the suitability of the Poisson-Weibull distribution in analyzing crash data characterized by small sample sizes and low sample means. To understand the contributions of various sources of heterogeneity to the observed variability in crash data, Peng et al. (2014) applied the generalized Waring distribution, also referred to as beta NB distribution, using three parameters. In addition, Zou et al. (2013) and Wu et al. (2014) exemplified the application of the Poisson-generalized inverse Gaussian distribution, also known as the Sichel distribution, in the EB estimates for hot spot identification. A special case of the Sichel distribution is the Poisson-inverse Gaussian distribution, which was applied in a study by Zha et al. (2016). Other applications of the Poisson mixture models in crash analysis have included the Poisson-lognormal model (Miranda-Moreno et al., 2005), the NB-generalized exponential model (Vangala et al., 2015), the random parameter NB model (Venkataraman et al., 2013), the random parameter tobit model (Anastasopoulos et al., 2012), the zero-inflated NB model (Raihan et al., 2019), and the Poisson-Tweedie model (Debrabant et al., 2018).

Given the plethora of methods for analyzing specific crash frequency data, it is often difficult for the analysts to select the most suitable approach. In two recent studies, Shirazi et al. (2017) and Shirazi and Lord (2019) proposed a heuristic technique based on characteristics of data, such as kurtosis, skewness, and percentage of zeros, to select a distribution between two competing distributions. The results, however, cannot be generalized to all types of data set. Also, a simulated data set must be generated in order to run the heuristics technique that employs machine-learning-based classification algorithms (e.g., random forests). Another alternative for choosing the best fitting model is to use likelihood-based measures. However, the analysts may still face issues, which may include identifying the appropriate set of models for the data set and correctly specifying different parameters for different fitting algorithms, leading to misjudgment of distribution parameters and poor fitting of likelihood function (Bonat et al., 2018). The unified modeling approach based on the Poisson-Tweedie family of distributions, introduced by Jørgensen and Kokonendji (2016), is a suitable alternative to address these issues. The class of Poisson-Tweedie models fits not only overdispersed data, but it can also automatically fit underdispersed count data similar to the aforementioned Conway-Maxwell-Poisson distribution model. In addition, the Poisson-Tweedie distribution can model zero-inflated count data without the need to introducing two parts as are typically done in the zero-inflated NB or hurdle models. Several other data structures, such as spatial data, longitudinal data, multiple response variables, and response variables of mixed type, can also be fitted using the Poisson-Tweedie distribution (Bonat and Jørgensen, 2016; Bonat et al., 2018).

Table 1
Summary of covariates.

Variable by Intersection Type	Mean	Min	Max	Std. Dev.
Urban three-leg stop-controlled intersection (sample size = 316)				
Total crash frequency	7.00	0.00	80.00	9.40
FI crash frequency	3.10	0.00	25.00	3.73
AADT on the major AADT (major AADT)	18,987	900	62,667	12,732
AADT on the minor AADT (minor AADT)	3,167	123	21,500	2,939
Urban four-leg signalized intersection (sample size = 369)				
Total crash frequency	65.18	0.00	365.00	56.06
FI crash frequency	24.03	0.00	98.00	17.24
AADT on the major AADT (major AADT)	31,983	5,700	73,666	14,082
AADT on the minor AADT (minor AADT)	18,694	1,016	59,833	12,006

Note: FI = Fatal and Injury. AADT = Annual Average Daily Traffic. Statistics shown for total and FI crash frequency indicate three-year estimates (i.e. crashes in three years per intersection). The unit for AADT is vehicles per day.

The variance function in the Poisson-Tweedie distribution model is analogous to that in the NB model, but with a simple-yet-powerful tweak. The variance function is given by $mean + dispersion \times (mean)^P$, where P is the power parameter and can assume any real number. This makes the Poisson-Tweedie models very flexible (Kokonendji et al., 2004). It includes, as special cases, the Hermite distribution ($P = 0$), the Neyman Type A or NB1 distribution ($P = 1.0$), the Pólya-Aeppli or geometric Poisson (GeoP) distribution ($P = 1.5$), the NB2 distribution ($P = 2.0$), and the Poisson Inverse Gaussian (PIG) distribution ($P = 3.0$). El-Shaarawi et al. (2011) has shown that the Poisson Generalized Inverse Gaussian model is also a special case of the class of Poisson-Tweedie distribution models. Debrabant et al. (2018) used the Poisson-Tweedie distribution to fit an autoregressive model for identifying hot spots. The number of crashes in a neighborhood of 3911 grids of one square kilometer was modeled as a function of the number of intersections and street length at a location (i.e., grid) and the average crash count at the neighboring grids of the subject grid. The value of the power parameter was estimated to be 1.60, indicating that the Pólya-Aeppli or GeoP distribution rather than the NB2 distribution would better fit the data.

Another advantage offered by the Poisson-Tweedie models is that the dispersion parameter can be estimated as a function of observed covariates (Petterle et al., 2019). There are several studies that showed the importance of fitting separate regression models to estimate the overdispersion parameter (El-Basyouny and Sayed, 2006; Lord and Park, 2008; Miaou and Lord, 2003; Ukkusuri et al., 2012). Also, Geedipally et al. (2009) examined various parameterizations in modeling the dispersion parameter for segment crashes as a function of traffic volume and segment length. In the NB setting, such models are referred to as generalized NB models (Lord and Park, 2008; Ukkusuri et al., 2012).

Table 2
Model results for total crashes at urban three-leg stop-controlled intersections.

Parameter	Fixed Dispersion					Varying Dispersion				
	PTw	NB1	GeoP	NB2	PIG	PTw	NB1	GeoP	NB2	PIG
Regression Coefficient for Mean										
Intercept	−11.781	−12.112	−11.835	−11.625	−11.395	−11.897	−11.820	−11.846	−11.872	−11.915
	(0.787)	(0.797)	(0.788)	(0.788)	(0.804)	(0.772)	(0.792)	(0.788)	(0.781)	(0.762)
ln(major AADT)	1.005	1.039	1.011	0.986	0.950	1.010	1.014	1.013	1.011	1.010
	(0.077)	(0.078)	(0.077)	(0.077)	(0.076)	(0.077)	(0.077)	(0.077)	(0.077)	(0.076)
ln(minor AADT)	0.341	0.340	0.340	0.346	0.360	0.350	0.334	0.340	0.346	0.353
	(0.057)	(0.056)	(0.057)	(0.056)	(0.052)	(0.054)	(0.058)	(0.057)	(0.055)	(0.053)
Regression Coefficient for Dispersion										
Intercept	0.083[d] (0.587)	1.247	0.303 (0.159)	−0.663	−2.764	10.218[c]	−4.405[c]	0.541[d]	5.380 (2.692)	14.643
		(0.170)		(0.154)	(0.146)	(6.248)	(2.890)	(2.785)		(2.508)
ln(major AADT)	n/a	n/a	n/a	n/a	n/a	−1.210[c]	0.577[b]	−0.024[d]	−0.615	−1.757
						(0.793)	(0.297)	(0.285)	(0.274)	(0.254)
ϕ (constant)[a]	1.087	3.481	1.354	0.515	0.063	n/a	n/a	n/a	n/a	n/a
Power Parameter										
P	1.615	1.000 (fixed)	1.500 (fixed)	2.000 (fixed)	3.000	2.514 (0.840)	1.000 (fixed)	1.500 (fixed)	2.000 (fixed)	3.000
	(0.304)				(fixed)					(fixed)
Goodness-of-fit Measures										
pLL	−927.470	−940.46	−927.980	−930.500	−954.760	−923.680	−933.510	−927.970	−924.690	−924.590
pAIC	1864.940	1888.920	1863.960	1869.000	1917.520	1859.360	1877.020	1865.940	1859.380	1859.180
pBIC	1883.719	1903.943	1878.983	1884.023	1932.543	1881.894	1895.799	1884.719	1878.159	1877.959
MAD	4.332	4.326	4.331	4.334	4.337	4.337	4.329	4.331	4.333	4.339
Modified R^2	0.482	0.484	0.483	0.481	0.476	0.484	0.483	0.483	0.483	0.484

Note: PTw = Poisson-Tweedie model. NB1 = Negative binomial model with linear variance function. NB2 = Negative binomial model with quadratic variance function. GeoP = Geometric Poisson model. PIG = Poisson Inverse Gaussian model. Standard errors of the estimated coefficients are shown in parentheses. "n/a" indicates that the associated parameters were not estimated (or not applicable) for the models.

[a] ϕ (constant) = exp(Intercept).

[b] Statistically significant at 90% confidence level (i.e., p-value < 0.10).

[c] Statistically significant at 85% confidence level (i.e., p-value < 0.15).

[d] Statistically non-significant at 85% confidence level (i.e., p-value > 0.15).

In highway safety literature, researchers have shown the application of a variety of distributions to model crash frequency data. A specific distribution or model has been suggested to account for the effect of one or more of the many inherent issues present in data (Lord and Mannering, 2010). However, it is difficult for analysts to pick a single distributional model as being the most appropriate one. Also, fitting different distribution models oftentimes require different programs or algorithms and their different parameters. For instance, zero-inflated NB models, Sichel distribution models, and Conway-Maxwell-Poisson distribution models each require a different set of parameters (Lord et al., 2008; Raihan et al., 2019; Wu et al., 2014). A unified approach like the family of Poisson-Tweedie distributions can facilitate both researchers and practitioners to examine various models under a specific set of parameters (the number of parameters may vary depending on the model) and identify the model that is best suited for the data. Again, to the best of the Authors' knowledge, there have been no such studies that examined generalized NB models with different power parameters of the variance function. The Poisson-Tweedie model, on the other hand, allows to fit such models, i.e., generalized NB1, generalized NB2, generalized NB without a fixed value of the power parameter *P*, and similarly others. Also, while the NBP model allows the value of the power parameter to vary between 1.0 and 2.0 (Greene, 2008), a value higher than 2.0 is possible (Holla, 1967), which can be modeled under the Poisson-Tweedie family of distributions.

The study by Debrabant et al. (2018) shows a good application of the Poisson-Tweedie distribution in modeling crash counts data; however, it did not explore the many possibilities that the Poisson-Tweedie distribution model offers. For example, the variation of dispersion parameter across study units (i.e., rectangular grids) was not examined. Also, it would be more appealing if the degree of variation of model parameters between the fitted model with *P* equal to 1.60 and the typical NB2 model was explored.

This study further explored the application of the Poisson-Tweedie family of distributions in analyzing crash frequency data. As part of the evaluation, a number of models were fitted including without first fixing the power parameter value and then fixing it at 1.0, 1.5, 2.0, and 3.0. In addition, the dispersion parameter was modeled both as a constant and as a function of covariates. All of these were conducted using two sets of intersection data from the state of Florida, one involving urban three-leg stop controlled intersections and another involving urban four-leg signalized intersections.

2. Methodology

Let y_i denote the number of crashes that occurs at an intersection i during a specific period. Typically, y_i is assumed to follow a Poisson distribution that is defined by a single parameter λ_i, i.e.,

$$y_i \sim Poisson(\lambda_i) \quad (1)$$

Under the Poisson-Tweedie class of models, the Poisson mean parameter λ_i is assumed to follow a Tweedie distribution as

$$\lambda_i \sim Tw_P(\mu_i, \phi_i) \quad (2)$$

where $\mu_i > 0$ is the mean parameter, $\phi_i > 0$ is the dispersion parameter, and P indicates the Tweedie power parameter. The parameters μ_i and ϕ_i are modeled as a function of covariates using a link function expressed as

$$\mu_i = g^{-1}(x_i^T \beta) \quad (3)$$

and

$$\phi_i = h^{-1}(z_i^T \gamma) \quad (4)$$

Table 3
Model results for fatal and injury (FI) crashes at urban three-leg stop-controlled intersections.

Parameter	Fixed Dispersion					Varying Dispersion				
	PTw	NB1	GeoP	NB2	PIG	PTw	NB1	GeoP	NB2	PIG
Regression Coefficient for Mean										
Intercept	−11.512	−11.542	−11.500	−11.469	−11.502	−11.599	−11.505	−11.539	−11.573	−11.636
	(0.832)	(0.824)	(0.834)	(0.838)	(0.854)	(0.828)	(0.828)	(0.831)	(0.830)	(0.823)
ln(major AADT)	0.952	0.959 (0.081)	0.949	0.939	0.933	0.959	0.953	0.955	0.957	0.962
	(0.082)		(0.082)	(0.082)	(0.081)	(0.082)	(0.081)	(0.082)	(0.082)	(0.082)
ln(minor AADT)	0.274	0.268 (0.058)	0.276	0.285	0.296	0.276	0.271	0.273	0.275	0.277
	(0.060)		(0.060)	(0.059)	(0.055)	(0.058)	(0.060)	(0.059)	(0.059)	(0.058)
Regression Coefficient for Dispersion										
Intercept	−0.148[c]	0.194[c]	−0.321[b]	−0.975	−2.566	10.647[c]	−2.318[c]	2.368[c]	6.987 (3.160)	16.013
	(0.491)	(0.193)	(0.192)	(0.184)	(0.192)	(10.502)	(3.262)	(3.214)		(3.023)
ln(major AADT)	n/a	n/a	n/a	n/a	n/a	−1.219[c]	0.259[c]	−0.274[c]	−0.800	−1.837
						(1.218)	(0.328)	(0.323)	(0.318)	(0.304)
ϕ (constant) [a]	0.863	1.214	0.725	0.377	0.077	n/a	n/a	n/a	n/a	n/a
Power Parameter										
P	1.347 (0.373)	1.000	1.500	2.000	3.000	2.402	1.000	1.500	2.000	3.000
		(fixed)	(fixed)	(fixed)	(fixed)	(1.295)	(fixed)	(fixed)	(fixed)	(fixed)
Goodness-of-fit Measures										
pLL	−708.920	−710.710	−709.110	−712.570	−723.650	−707.610	−709.870	−708.480	−707.750	−707.970
pAIC	1427.840	1429.420	1426.220	1433.140	1455.300	1427.220	1429.740	1429.960	1425.500	1425.940
pBIC	1446.619	1444.443	1441.243	1448.163	1470.323	1449.754	1448.519	1445.739	1444.279	1444.719
MAD	2.036	2.034	2.036	2.038	2.039	2.035	2.035	2.035	2.035	2.035
Modified R^2	0.536	0.536	0.536	0.535	0.535	0.536	0.536	0.536	0.536	0.536

Note: PTw = Poisson-Tweedie model. NB1 = Negative binomial model with linear variance function. NB2 = Negative binomial model with quadratic variance function. GeoP = Geometric Poisson model. PIG = Poisson Inverse Gaussian model. Standard errors of the estimated coefficients are shown in parentheses. "n/a" indicates that the associated parameters were not estimated (or not applicable) for the models.

[a] ϕ (constant) = exp(Intercept).

[b] Statistically significant at 90% confidence level (i.e., p-value < 0.10).

[c] Statistically non-significant at 85% confidence level (i.e., p-value > 0.15).

where x_i and z_i are vectors of observed covariates, β and γ are vectors of unknown regression coefficients including the intercept, and $g(.)$ and $h(.)$ are inverse link functions (e.g., log). Note that x_i and z_i may consist of either a similar or a different set of covariates. The expectation and variance of y_i are then given by (Jørgensen and Kokonendji, 2016)

$$E(y_i) = \mu_i \tag{5}$$

and

$$Var(y_i) = \mu_i + \phi_i \mu_i^P \tag{6}$$

The parameters, including β, γ,and P, can be estimated using the estimating function approach proposed by Bonat et al. (2018). The estimating function approach combines the quasi-likelihood score function and Pearson estimating function to estimate model parameters. The quasi-score function shown in Eq. (7) was used to estimate $\beta$and the Pearson estimating function shown in Eq. (8) was used to estimate the set of parameters $\theta = (\gamma, P)$ (Petterle et al., 2019), as follows:

$$\psi_\beta(\beta, \theta) = \left(\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_1} \sigma_i^{-1} (y_i - \mu_i), \ ..., \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_k} \sigma_i^{-1} (y_i - \mu_i) \right)^T \tag{7}$$

and

$$\psi_\theta(\theta, \beta) = \left(\sum_{i=1}^{n} \omega_{i\theta_1} [(y_i - \mu_i)^2 - \sigma_i], \ ..., \sum_{i=1}^{n} \omega_{i\theta_{q+1}} [(y_i - \mu_i)^2 - \sigma_i] \right)^T \tag{8}$$

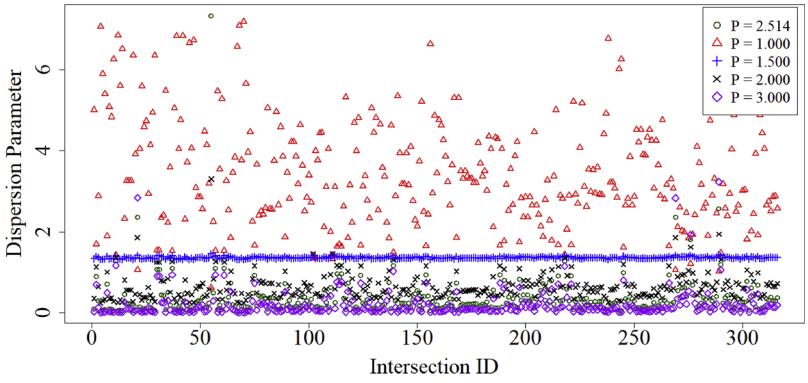
where $\sigma_i = \mu_i + \phi \mu_i^P$with $i = 1, ..., n$ (i.e., number of observations); $\partial \mu_i / \partial \beta_k = \mu_i x_{ik}$ with k denoting the number of covariates for mean; and $\omega_{i\theta_l} = \partial \sigma_i^{-1} / \partial \theta_l$ with l denoting the number of covariates (i.e., $q + 1$) for dispersion and power. The details of the estimation procedure can be found in Bonat and Jørgensen (2016), Bonat et al. (2018), and Petterle et al. (2019).

A series of models were developed in this study with different settings of dispersion and power parameters. The dispersion parameter was estimated using an intercept-only model as well as a covariate-based model. The intercept-only model provides a constant or mean value of the dispersion parameter. The dispersion parameter in the covariate-based model is estimated as a function of one or more covariates in addition to the intercept, allowing the values of the dispersion parameter to vary across the set of observations (Geedipally et al., 2009). Henceforth, the dispersion parameter obtained from the intercept-only model is referred to as fixed dispersion parameter and that obtained from covariate-based model is referred to as varying dispersion parameter. With each setting of dispersion parameter, the following models were fitted with different values of Tweedie power parameters under the Poisson-Tweedie distribution:

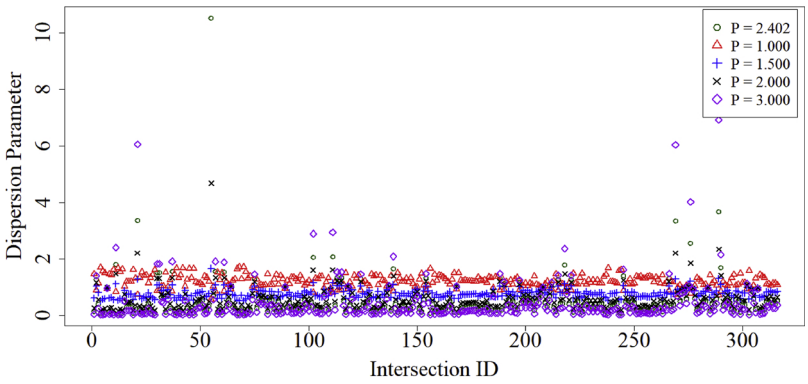
- Poisson-Tweedie model with P not fixed to any specific value
- NB1 model with $P = 1.0$ (Green, 2008)
- GeoP model with $P = 1.5$ (Özel and İnal, 2010)
- NB2 model with $P = 2.0$ (Miaou, 1994)
- PIG model with $P = 3.0$ (Holla, 1967)

Several goodness-of-fit measures were computed, including the pseudo Akaike information criterion (pAIC), the pseudo Bayesian information criterion (pBIC), the mean absolute deviation (MAD), and the modified R^2. Because the Poisson-Tweedie family of models are based on the quasi log-likelihood, Bonat (2018) suggested the pAIC and pBIC measures based on the Gaussian pseudo log-likelihood proposed by Carey and Wang (2011). The pseudo log-likelihood (pLL) is given by

$$\text{pLL} = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - (\boldsymbol{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \tag{9}$$



(a) Total crash model



(b) Fatal and Injury (FI) crash model

Fig. 1. Variation in dispersion parameter for urban three-leg stop-controlled intersections.

where n is the number of observations (e.g., number of intersections), Σ is the estimated covariance matrix, and $\boldsymbol{y} = (y_1, ..., y_n)^T$ and $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$ are the vectors of the observed and the predicted crash frequency, respectively.

The pAIC and pBIC values are then estimated as

$$\text{pAIC} = 2(k + l) - 2 \times \text{pLL} \tag{10}$$

$$\text{pBIC} = (k + l)\log(n) - 2 \times \text{pLL} \tag{11}$$

where k and l are the number of regression parameters for the mean and dispersion (including power parameter, if estimated), respectively. The lower the values of pAIC and pBIC, the better is the model fit.

MAD measures the magnitude of variability in model prediction. It is the sum of the absolute value of the difference between predicted crashes and observed crashes over the sample size. Models with smaller MAD values are preferred to models with larger MAD values. MAD is given by

$$\text{MAD} = \frac{\sum_{i=1}^{n} |\mu_i - y_i|}{n} \tag{12}$$

Modified R^2 measures the amount of systematic variation explained by the fitted model. A larger value indicates a better fit; however, R^2 values greater than 1.0 indicates that the model incorrectly explains some of the random variation as systematic (Lyon et al., 2016). The modified R^2 is given by

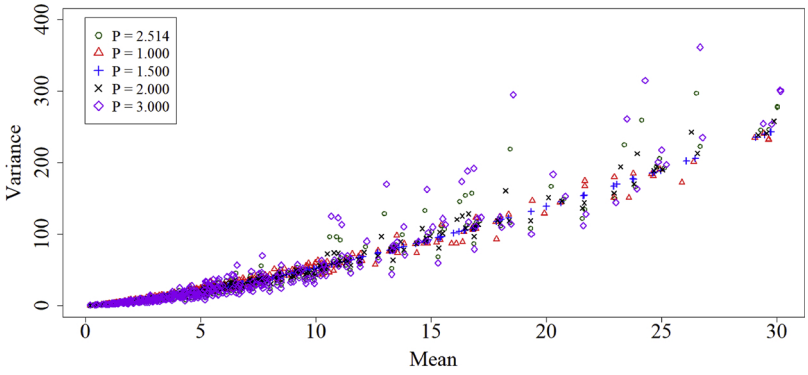
$$\text{Modified } R^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (y_i - \mu_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} \mu_i} \tag{13}$$

The open-source software application ***R*** (R Core Team, 2019) was used for data processing and model fitting and evaluation in this study. In particular, the "dplyr" package (Wickham et al., 2019) was used to prepare the study data set and the "mcglm" package (Bonat, 2018) was used to fit the Poisson-Tweedie class of models.

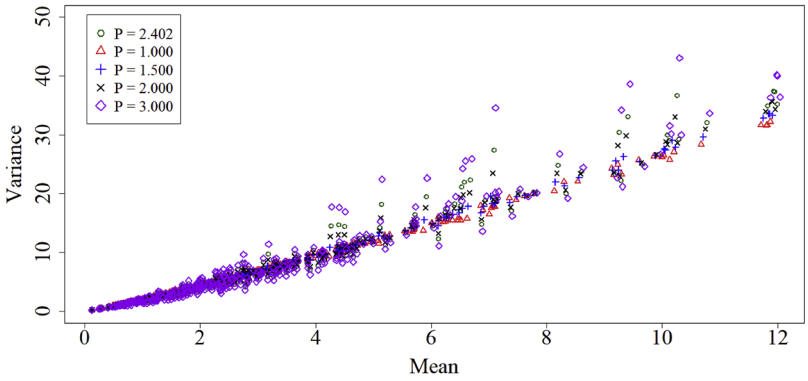
3. Data

The data set used in this study was prepared as part of a Florida Department of Transportation (FDOT) project to enhance the calibration procedures of the HSM for Florida. The data collection process was facilitated by identifying intersections on state roads and for which the annual average daily traffic (AADT) volumes were available. Data were initially collected for a total of 1038 signalized intersections and 1555 unsignalized intersections using a Google Maps application (Alluri et al., 2014). The data were then filtered to exclude the intersections that exhibited the following characteristics:

- five or more approaches;
- any of the approaches carrying one-way traffic;
- yield-controlled, roundabouts or traffic circles, and presence of flasher;
- undergoing construction work; or
- unreliable (i.e., zero or extremely high) AADT values.



(a) Total crash model



(b) Fatal and Injury (FI) crash model

Fig. 2. Estimated mean vs. variance for urban three-leg stop-controlled intersections.

Further segregation of intersections by area type (rural, urban) and number of approaches (three, four) considerably reduced the sample size for each intersection type. Only urban three-leg stop-controlled intersections and urban four-leg signalized intersections were found to have a reasonable sample size and were thus analyzed in this study. In total, the study sample consisted of 316 urban three-leg stop-controlled intersections and 369 urban four-leg signalized intersections. Note that the sample for urban three-leg stop-controlled intersections included only those intersections, where the stop control applied to the minor road approach only (i.e., no control for traffic on the major road approach).

Three years (2012–2014) of crash data were used in this study. The data were retrieved from the FDOT's Unified Basemap Repository portal. A 250-ft buffer around the center of each intersection was generated to identify intersection-related crashes. Crashes that were located in the overlapping portion of two or more buffers were assigned to the nearest intersection. The number of crashes in each intersection was then counted by crash severity, i.e., fatal (K), incapacitating injury (A), non-incapacitating injury (B), possible injury (C), and property damage only (O). Separate models were developed for total crashes including all crash severities and fatal and injury (FI) crashes comprising K, A, B, and C crashes.

In addition, the AADT data were extracted for the same three years (i.e., 2012–2014) from the FDOT's Roadway Characteristics Inventory database. A three-year average of AADT on the major road approach, henceforth referred to as major AADT, and a three-year average of AADT on the minor road approach, henceforth referred to as minor AADT, were used as the covariates. Table 1 provides the summary of the crash statistics and covariates used in developing the models.

It is worth mentioning here that several other covariates, such as number of approaches with exclusive left-turn and right-turn lanes, skew angle, presence of lighting, and presence of school and bus stops within 250 ft were initially explored in this study. The covariates, however, did not show any meaningful correlations with crash frequencies. As found in the literature, crash frequency models were developed based only on AADT or covariates related to AADT, when other covariate were not available (Harwood et al., 2010; Lord et al., 2008; Lord and Park, 2008; Wang et al., 2019; Wu et al., 2014).

4. Results

4.1. Urban three-leg stop-controlled intersections

Tables 2 and 3 summarize the model parameters and the goodness-of-fit measures for total and FI crashes, respectively, at urban three-leg stop-controlled intersections. The estimates are shown for 10 models fitted with two settings of dispersion parameters and five power parameters under each crash category. As can be seen in the Tables, both ln (major AADT) and ln(minor AADT) were found to have statistically significant associations at the 95% confidence interval with the mean function for crash frequency in all the models for total and FI crashes. The varying dispersion parameter models were found to have the best estimates in terms of convergence and model fit with ln(major AADT) only.

The value of the power parameter (i.e., when the parameter is not fixed) in the total crash frequency models was estimated to be 1.615 with the fixed dispersion parameter and 2.514 when the dispersion parameter was estimated as a function of major AADT (Table 2). In the FI crash frequency models, the estimated values of the power parameter under two types of dispersion parameter were smaller than the

Table 4
Model results for total crashes at urban four-leg signalized intersections.

Parameter	Fixed Dispersion					Varying Dispersion				
	PTw	NB1	GeoP	NB2	PIG	PTw	NB1	GeoP	NB2	PIG
Regression Coefficient for Mean										
Intercept	−9.998	−10.370	−9.681	−9.121	−8.471	−10.177	−10.070	−10.114	−10.163	−10.273
	(0.770)	(0.794)	(0.751)	(0.722)	(0.713)	(0.777)	(0.776)	(0.776)	(0.777)	(0.779)
ln(major AADT)	1.010	1.051	0.972	0.900	0.796	1.021	1.019	1.020	1.021	1.021
	(0.093)	(0.094)	(0.091)	(0.089)	(0.084)	(0.091)	(0.093)	(0.092)	(0.092)	(0.090)
ln(minor AADT)	0.264	0.258	0.271	0.290	0.334	0.270	0.261	0.264	0.268	0.280
	(0.057)	(0.057)	(0.057)	(0.056)	(0.053)	(0.053)	(0.058)	(0.056)	(0.053)	(0.049)
Regression Coefficient for Dispersion										
Intercept	1.972	3.036	1.014	−0.950	−4.714	11.101[b]	0.411[d]	5.050	9.765	19.466 (2.138)
	(0.771)	(0.097)	(0.099)	(0.108)	(0.135)	(5.882)	(2.593)	(2.473)	(2.355)	
ln(major AADT)	n/a	n/a	n/a	n/a	n/a	−1.232[c]	0.254[d]	−0.393[c]	−1.047	−2.380
						(0.794)	(0.251)	(0.239)	(0.228)	(0.206)
ϕ (constant) [a]	7.185	20.822	2.755	0.387	0.009	n/a	n/a	n/a	n/a	n/a
Power Parameter										
P	1.261	1.000	1.500	2.000	3.000	2.140	1.000	1.500	2.000	3.000
	(0.187)	(fixed)	(fixed)	(fixed)	(fixed)	(0.603)	(fixed)	(fixed)	(fixed)	(fixed)
Goodness-of-fit Measures										
pLL	−1827.890	−1830.000	−1829.080	−1840.250	−1896.150	−1826.160	−1828.910	−1827.050	−1826.200	−1828.360
pAIC	3665.780	3668.000	3666.160	3688.500	3800.300	3664.320	3667.820	3664.100	3662.400	3666.720
pBIC	3685.334	3683.643	3681.803	3704.143	3815.943	3687.785	3687.374	3683.654	3681.954	3686.274
MAD	29.294	29.166	29.402	29.625	29.922	29.245	29.268	29.257	29.247	29.237
Modified R^2	0.502	0.504	0.499	0.493	0.480	0.503	0.503	0.503	0.503	0.504

Note: PTw = Poisson-Tweedie model. NB1 = Negative binomial model with linear variance function. NB2 = Negative binomial model with quadratic variance function. GeoP = Geometric Poisson model. PIG = Poisson Inverse Gaussian model. Standard errors of the estimated coefficients are shown in parentheses. "n/a" indicates that the associated parameters were not estimated (or not applicable) for the models.

[a] ϕ (constant) = exp(Intercept).

[b] Statistically significant at 90% confidence level (i.e., p-value < 0.10).

[c] Statistically significant at 85% confidence level (i.e., p-value < 0.15).

[d] Statistically non-significant at 85% confidence level (i.e., p-value > 0.15).

corresponding values in the total crash frequency model. The value of the power parameter was 1.347 with a fixed dispersion parameter and 2.402 with a varying dispersion parameter (Table 3).

The estimated value of the fixed dispersion parameter was found to be statistically not significant at the estimated value of the power parameter (i.e., when not fixed) in both total and FI crash frequency models. As can be seen in Tables 2 and 3, the estimated value of the fixed dispersion parameter was the highest in the NB1 model and the lowest value in the PIG models. In general, the value of the fixed dispersion parameter was found to decrease as the value of the power parameter increased. The regression coefficient of major AADT associated with the varying dispersion parameter was significant at the 95% confidence level for both total and FI crash frequencies under the NB2 and PIG models only.

Fig. 1(a) and 1(b) show the variation in the varying dispersion parameter estimates across the study set of intersections for total and FI crash frequency models, respectively. The plots in Fig. 1 also show the distribution of the varying dispersion parameter values under each of five power parameters. In Fig. 1(a), the values of the dispersion parameter were found to greatly vary in the model fitted with the power parameter equal to 1.0 (i.e., NB1 model) and found to have the least variation in the model fitted with the power parameter equal to 1.5 (i.e., GeoP model). In Fig. 1(b), the variation in the values of dispersion parameter between the models was much less compared to what was observed in Fig. 1(a). There were only a few observations in the NB2 and PIG models, where the values of dispersion parameters were estimated to be greater than 2.0.

To understand the effects of the dispersion parameter on the variance function, additional plots were prepared. Fig. 2 shows the crash variance plotted against the mean crash rate for models fitted with different power parameters, where the mean was estimated using regression coefficients for mean and the variance was estimated using the coefficients of the varying dispersion parameter and the given power parameter. The crash variance across the models was found to be quite similar in both Fig. 2(a) and 2(b). The variance function under the Poisson-Tweedie and PIG (i.e., $P = 3.0$) models tended to increase at a slightly higher rate than that under the NB1 (i.e., $P = 1.0$), GeoP (i.e., $P = 1.5$), and NB2 (i.e., $P = 2.0$) models, which were probably due to higher values of the power parameter associated with the Poisson-Tweedie and PIG models. The Spearman's rank correlation coefficients were found to be a minimum of 0.89 and a maximum of 0.99 between the estimated variance in the total crash frequency models. Similarly, the correlation coefficients were found to vary between 0.95 and 0.99 for the estimated variance in the FI crash frequency models. This indicates that although the dispersion parameter varies widely, the variance function across the models has shown quite similar performance, which could be attributed to the higher value of the power parameter.

4.2. Urban four-leg signalized intersections

Tables 4 and 5 show the estimated model parameters and the goodness-of-fit measures of the total and FI crash frequency models, respectively, at urban four-leg signalized intersections. Similar to the findings for urban three-leg stop-controlled intersections, both AADT-related covariates, i.e., ln(major AADT) and ln(minor AADT), had positive and statistically significant associations with total and FI crashes at urban four-leg signalized intersections.

The value of the power parameter was estimated to be 1.261 in the total crash frequency model and 1.016 in the FI crash frequency model, when fitted with fixed dispersion parameter. On the other hand, in the models estimated with varying dispersion, the value of the power parameter was found to varied greatly between total and FI crashes. For

Table 5
Model results for fatal and injury (FI) crashes at urban four-leg signalized intersections.

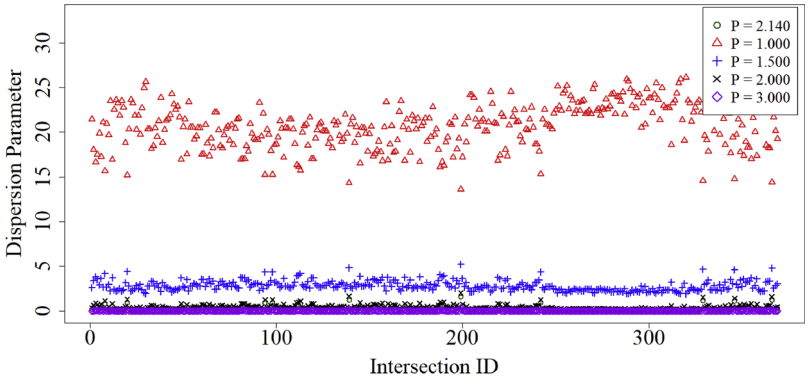
Parameter	Fixed Dispersion					Varying Dispersion				
	PTw	NB1	GeoP	NB2	PIG	PTw	NB1	GeoP	NB2	PIG
Regression Coefficient for Mean										
Intercept	−8.435 (0.696)	−8.439 (0.697)	−8.312 (0.678)	−8.199 (0.665)	−8.015 (0.653)	−8.433 (0.695)	−8.433 (0.696)	−8.431 (0.698)	−8.430 (0.700)	−8.426 (0.704)
ln(major AADT)	0.812 (0.084)	0.812 (0.084)	0.799 (0.083)	0.785 (0.081)	0.759 (0.078)	0.812 (0.084)	0.812 (0.084)	0.811 (0.083)	0.810 (0.083)	0.808 (0.082)
ln(minor AADT)	0.215 (0.051)	0.215 (0.051)	0.216 (0.051)	0.220 (0.051)	0.228 (0.049)	0.215 (0.052)	0.215 (0.051)	0.215 (0.050)	0.216 (0.049)	0.218 (0.046)
Regression Coefficient for Dispersion										
Intercept	1.714 (0.832)	1.760 (0.104)	0.238 (0.113)	−1.265 (0.125)	−4.256 (0.141)	0.668[c] (7.888)	1.510[c] (3.024)	5.211[b] (2.925)	8.918 (2.834)	16.345 (2.677)
ln(major AADT)	n/a	n/a	n/a	n/a	n/a	0.140[c] (1.019)	0.024[c] (0.291)	−0.485[b] (0.282)	−0.995 (0.273)	−2.015 (0.258)
ϕ (constant) [a]	5.549	5.816	1.268	0.282	0.014	n/a	n/a	n/a	n/a	n/a
Power Parameter										
P	1.016 (0.259)	1.000 (fixed)	1.500 (fixed)	2.000 (fixed)	3.000 (fixed)	0.886[b] (0.891)	1.000 (fixed)	1.500 (fixed)	2.000 (fixed)	3.000 (fixed)
Goodness-of-fit Measures										
pLL	−1443.800	−1443.810	−1447.880	−1460.210	−1504.510	−1443.780	−1443.800	−1444.230	−1445.240	−1449.120
pAIC	2897.600	2895.620	2903.76	2928.420	3017.020	2899.560	2897.600	2898.460	2900.480	2908.240
pBIC	2917.154	2911.263	2919.403	2944.063	3032.663	2923.025	2917.154	2918.014	2920.034	2927.794
MAD	10.038	10.037	10.043	10.044	10.045	10.038	10.038	10.038	10.037	10.037
Modified R^2	0.472	0.472	0.471	0.471	0.470	0.472	0.472	0.472	0.472	0.472

Note: PTw = Poisson-Tweedie model. NB1 = Negative binomial model with linear variance function. NB2 = Negative binomial model with quadratic variance function. GeoP = Geometric Poisson model. PIG = Poisson Inverse Gaussian model. Standard errors of the estimated coefficients are shown in parentheses. "n/a" indicates that the associated parameters were not estimated (or not applicable) for the models.

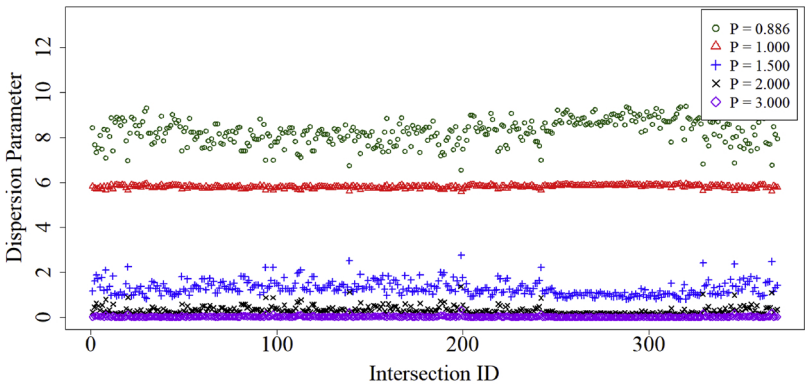
[a] ϕ (constant) = exp(Intercept).

[b] Statistically significant at 90% confidence level (i.e., p-value < 0.10).

[c] Statistically non-significant at 85% confidence level (i.e., p-value > 0.15).



(a) Total crash model



(b) Fatal and Injury (FI) crash model

Fig. 3. Variation in dispersion parameter for urban four-leg signalized intersections.

instance, the value of the power parameter was greater than 2.0 in the total crash frequency model and less than 1.0 in the FI crash frequency model. However, the estimated value of 0.886 associated with the power parameter of the FI crash model (Table 5) was not significant. It is worth mentioning that the value of a power parameter below 1.0 is not reliable for overdispersed count data (Bonat et al., 2018).

An examination of the dispersion parameters from Tables 4 and 5 show that the estimated values of the fixed dispersion parameter were significant at the 95% confidence level in all the models. Also, the values of the fixed dispersion parameter were higher in the models fitted with smaller values of the power parameter and vice versa, which is consistent with the findings from models for urban three-leg stop-controlled intersections. Nonetheless, the estimates of the fixed dispersion parameter were found to greatly vary between the models with power parameter equal to 1.0 and the models with power parameter equal to 3.0. The regression coefficients for the covariate major AADT in the varying dispersion parameter models were found to be significant in the Poisson-Tweedie models for total crashes and in the GeoP, NB2, and PIG models for both total and FI crash frequencies.

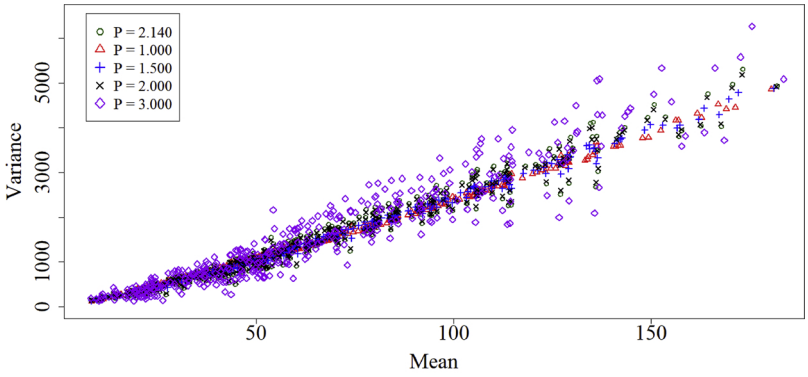
Fig. 3(a) and 3(b) show the variation in the dispersion parameter estimates across the study intersections for the total and FI crash frequency models, respectively. In the total crash model, the values of the varying dispersion parameter were found to vary in the range from 0 to 5 under the Poisson-Tweedie ($P = 2.14$), GeoP ($P = 1.5$), NB2 ($P = 2.0$), and PIG ($P = 3.0$) models. However, the range was from 13 to 26 for the varying dispersion parameter values in the NB1 model ($P = 1.0$). In the FI crash model, the values of the varying dispersion parameter were larger under the Poisson-Tweedie model than those under the NB1, GeoP, NB2, and PIG models. Contrary to the total crash model, the range of values of the dispersion parameter in the FI crash model was narrower. The least variation in the varying dispersion parameter values was observed under the PIG model.

Fig. 4 shows the plot of crash variance versus mean for models fitted with different power parameters. The estimates of crash variance under the PIG model were found to have a slightly wider interval and a higher slope. This is expected, as the variance function in the PIG model is estimated by raising the power of the mean to 3.0. The Spearman's rank correlation coefficients between the estimated variance in both the total and the FI crash frequency models were at a minimum of 0.93 and a maximum of 0.99. This indicates that the variance function across the models has shown quite similar performance.

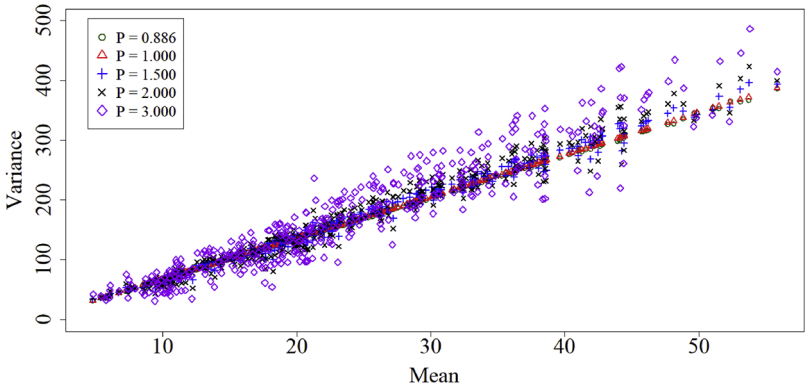
4.3. Comparative performance of the models

Of the models fitted with a fixed dispersion parameter, the GeoP models were found to provide a better fit than all the other models in terms of statistical significance of model parameters and the two measures of information criteria, pAIC and pBIC. In terms of the MAD and the modified R^2 measures, the NB1 models were found to predict crashes better than the other models.

Assessing the performance of the models with the varying dispersion parameter requires additional consideration. For urban three-leg stop-controlled intersections, the PIG model was found to have the lowest pAIC and pBIC values in the total crash frequency model and the NB2 model was found to have the lowest pAIC and pBIC values in the FI crash frequency model. However, both the NB2 and PIG models had approximately similar fit, with a very small difference (i.e., < 0.50) between the pAIC and the pBIC values. For urban four-leg signalized



(a) Total crash model



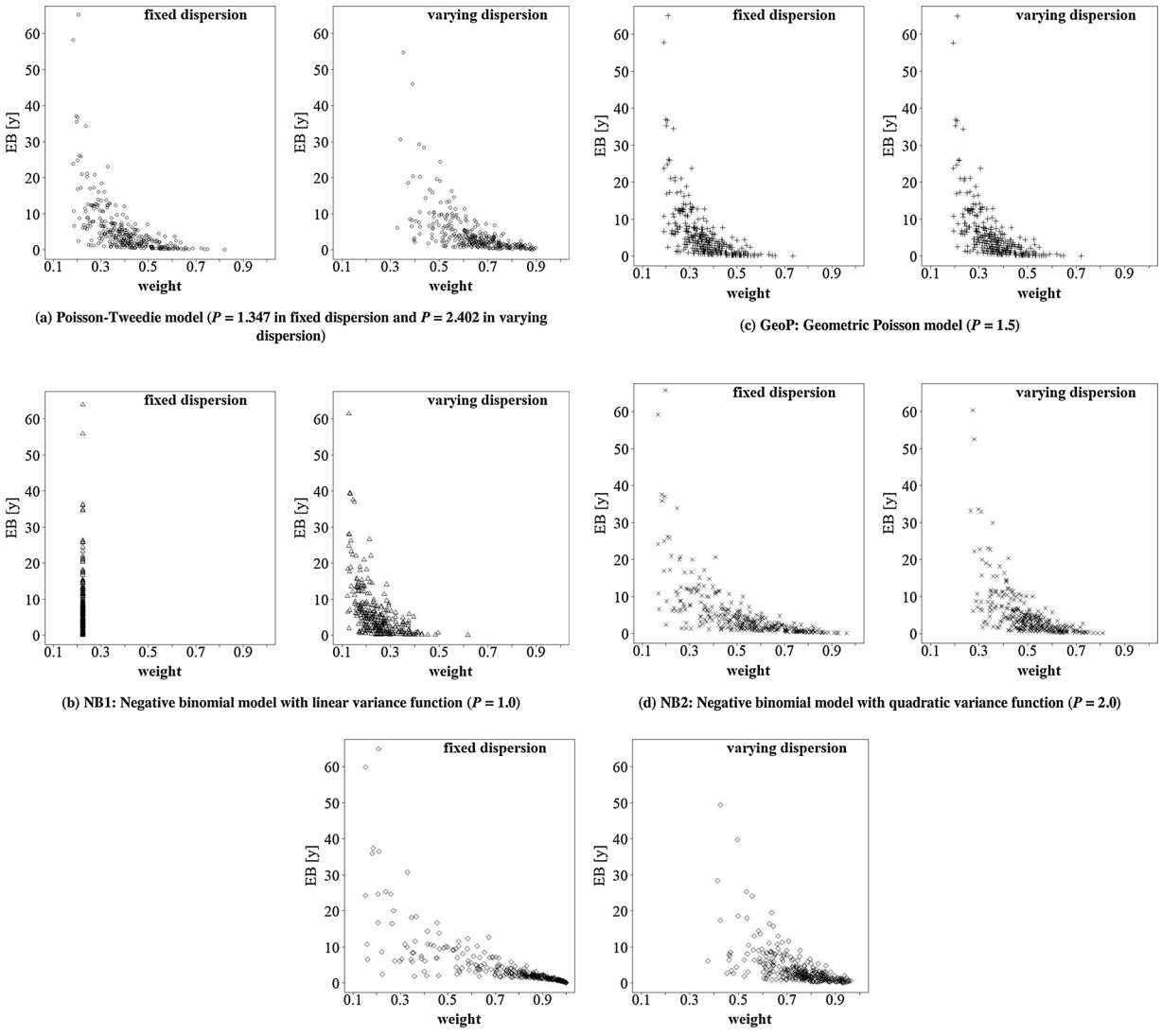
(b) Fatal and Injury (FI) crash model

Fig. 4. Estimated mean vs. variance for urban four-leg signalized intersections.

Table 6
Likelihood ratio test results.

Competing Models	χ^2 test statistic	p-value
Total crashes at urban three-leg stop-controlled intersections		
PTw with fixed dispersion vs. PTw with varying dispersion	7.58	0.0059
NB1 with fixed dispersion vs. NB1 with varying dispersion	13.9	0.0002
GeoP with fixed dispersion vs. GeoP with varying dispersion	0.02	0.8875
NB2 with fixed dispersion vs vs. NB2 with varying dispersion	11.62	0.0007
PIG with fixed dispersion vs vs. PIG with varying dispersion	60.34	< 0.0001
Fatal and Injury (FI) crashes at urban three-leg stop-controlled intersections		
PTw with fixed dispersion vs. PTw with varying dispersion	2.62	0.1055
NB1 with fixed dispersion vs. NB1 with varying dispersion	1.68	0.1949
GeoP with fixed dispersion vs. GeoP with varying dispersion	1.26	0.2617
NB2 with fixed dispersion vs vs. NB2 with varying dispersion	9.64	0.0019
PIG with fixed dispersion vs vs. PIG with varying dispersion	31.36	< 0.0001
Total crashes at urban four-leg signalized intersections		
PTw with fixed dispersion vs. PTw with varying dispersion	3.46	0.0629
NB1 with fixed dispersion vs. NB1 with varying dispersion	2.18	0.1398
GeoP with fixed dispersion vs. GeoP with varying dispersion	4.06	0.0439
NB2 with fixed dispersion vs vs. NB2 with varying dispersion	28.10	< 0.0001
PIG with fixed dispersion vs vs. PIG with varying dispersion	135.58	< 0.0001
Fatal and Injury (FI) crashes at urban four-leg signalized intersections		
PTw with fixed dispersion vs. PTw with varying dispersion	0.04	0.8415
NB1 with fixed dispersion vs. NB1 with varying dispersion	0.02	0.8875
GeoP with fixed dispersion vs. GeoP with varying dispersion	7.30	0.0069
NB2 with fixed dispersion vs vs. NB2 with varying dispersion	29.94	< 0.0001
PIG with fixed dispersion vs vs. PIG with varying dispersion	110.78	< 0.0001

Note: PTw = Poisson-Tweedie model. NB1 = Negative binomial model with linear variance function. NB2 = Negative binomial model with quadratic variance function. GeoP = Geometric Poisson model. PIG = Poisson Inverse Gaussian model.



(a) Poisson-Tweedie model (P = 1.347 in fixed dispersion and P = 2.402 in varying dispersion)

(c) GeoP: Geometric Poisson model (P = 1.5)

(b) NB1: Negative binomial model with linear variance function (P = 1.0)

(d) NB2: Negative binomial model with quadratic variance function (P = 2.0)

(e) PIG: Poisson Inverse Gaussian model (P = 3.0)

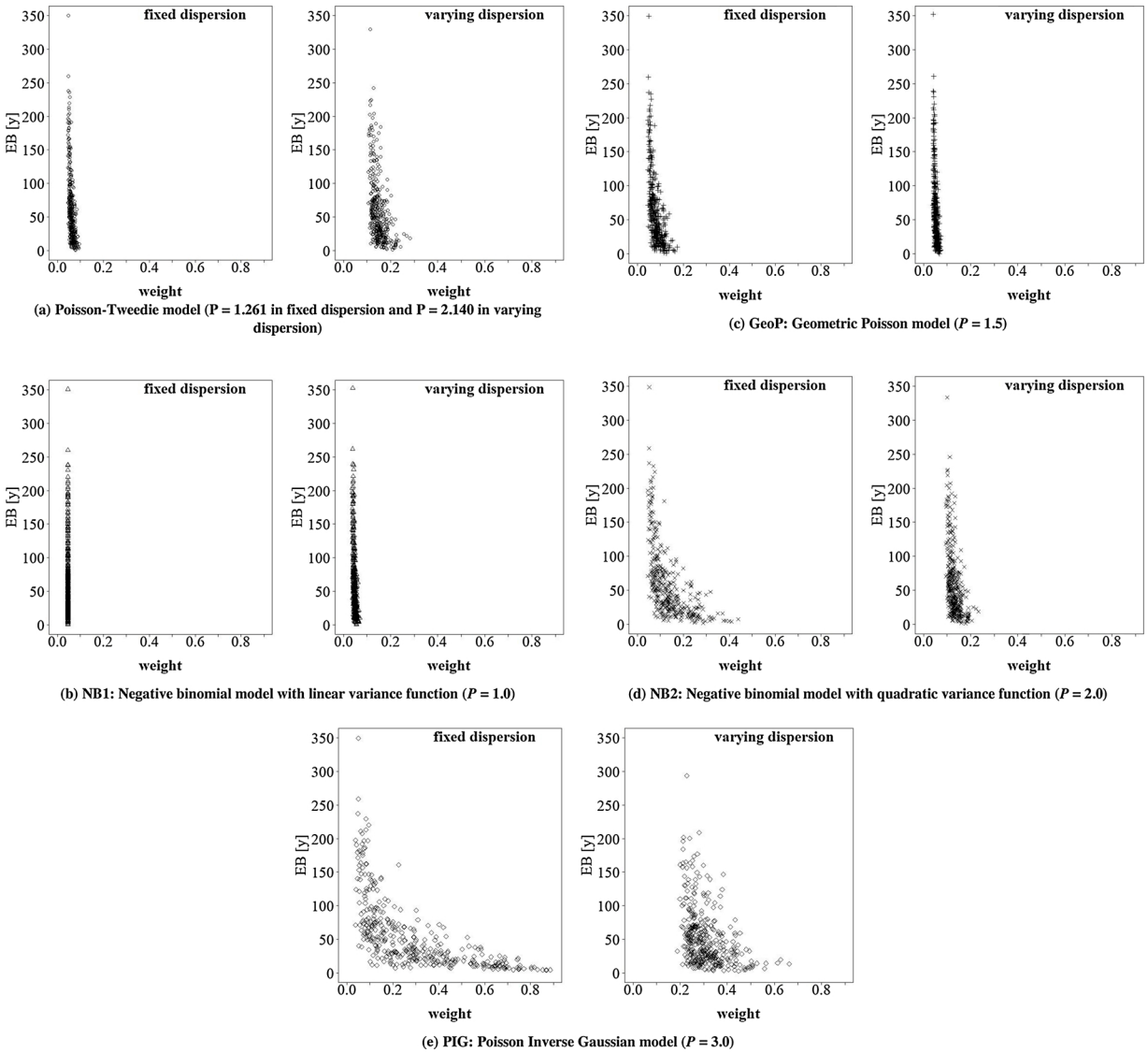
Fig. 5. Relationship between expected total crash frequency and weight for urban three-leg stop-controlled intersections.

intersections, the NB2 model had the lowest pAIC and pBIC values in the total crash frequency model and the NB1 model had the lowest pAIC and pBIC values in the FI crash frequency model. However, the regression coefficients associated with the varying dispersion parameter were not significant in the NB1 model for FI crash frequency. Considering the significance of the covariates, the GeoP model was found to have the best fit and the NB2 model the second best in the FI crash frequency model. The modified R^2 values were similar (the highest difference being 0.001) across the models with the varying dispersion parameter. The preferred models based on the MAD values may be different, but the difference in MAD values between the two models (i.e., models based on low pBIC and the models based on low MAD) was very small.

Between the models with fixed dispersion and the models with varying dispersion, the MAD and modified R^2 values were very close and, therefore, could not be relied upon to select the best models. Also, the dispersion parameter and the power parameter were not accounted for in the equations of MAD and modified R^2 measures. A likelihood ratio test between the corresponding (i.e., with respect to P) fixed and varying dispersion parameter models was therefore conducted to examine if there was any statistically significant difference between the models under the assumption of χ^2 distribution of log-likelihood values. Table 6 shows the results of the likelihood ratio test. The χ^2 test statistic between the GeoP models with the fixed and varying dispersion parameter was not statistically significant for both total and FI crashes at urban three-leg stop-controlled intersections. Also, the Poisson-Tweedie and NB1 models were found to display statistically non-significant difference between the fixed and varying dispersion parameters at the 0.05 level of significance for FI crashes at both urban three-leg stop-controlled intersections and urban four-leg signalized intersections. The results reflect that the dispersion parameter models that were not statistically different from the fixed dispersion parameter models had non-significant regression coefficients associated with the varying dispersion parameter estimates. In other words, the varying dispersion parameter models were statistically significantly different from the fixed dispersion parameter models only if its coefficients were statistically significant.

5. Practical applications

The EB method has been widely employed to identify locations for potential safety improvements (El-Basyouny and Sayed, 2006; Gan et al., 2016). It combines the observed crash frequency and the model-predicted crash frequency using a weight factor, and the output is known as the expected crash frequency. The EB estimate of the expected crash frequency is given by



(a) Poisson-Tweedie model (P = 1.261 in fixed dispersion and P = 2.140 in varying dispersion)

(c) GeoP: Geometric Poisson model ($P = 1.5$)

(b) NB1: Negative binomial model with linear variance function ($P = 1.0$)

(d) NB2: Negative binomial model with quadratic variance function ($P = 2.0$)

(e) PIG: Poisson Inverse Gaussian model ($P = 3.0$)

Fig. 6. Relationship between expected total crash frequency and weight for urban four-leg signalized intersections.

$$EB[y_i] = w_i\mu_i + (1 - w_i)y_i \tag{14}$$

where w_i is the weight factor varying between 0 and 1.

A reliable estimate of the expected crash frequency depends heavily on the weight factor. Hauer (1997) shows that the weight factor is a function of the mean and variance of the EB estimate. It has been shown that the weight factor for the NB2 model, which has a quadratic mean-variance relationship, is a function of the dispersion parameter and the estimated crash frequency, as follows:

$$w_i^{NB2} = \frac{1}{1 + \phi_i\mu_i} \tag{15}$$

Analogously, the weight factor in the Poisson-Tweedie model, where the variance is a power function of the mean, can be computed as

$$w_i^{PTw} = \frac{1}{1 + \phi_i\mu_i^{P-1}} \tag{16}$$

5.1. Relationship between expected total crash frequency and weight factor

Using Eqs. (14) and (16), the weight factor and the expected crash frequency were calculated considering both the fixed and varying dispersion parameters. Figs. 5 and 6 show the plots of expected crash frequency (i.e., EB[y]) versus weight factor for urban three-leg stop-controlled intersections and urban four-leg signalized intersections, respectively. Noted that all the results presented in this section, for brevity, are based on total crashes only.

In Fig. 5, the expected crash frequency at individual intersections was found to be lower under the varying dispersion parameter models compared to that under the fixed dispersion parameter models. However, the distribution of the values of the weight factors did not show any unique pattern. For instance, the values of the weight factors in the Poisson-Tweedie models (Fig. 5(a)) were mostly found to lie approximately in the range of 0.20-0.65 with fixed dispersion and in the range of 0.40-0.90 with varying dispersion. The weight factors in the NB1 models were found to vary within a very narrow range for fixed dispersion and within a wider range for varying dispersion (Fig. 5(b)). On the other hand, the weight factors in the NB2 and PIG models were found to have a slightly wider range for fixed dispersion compared to those for varying dispersion (Fig. 5(d) and (e)). There was a very small variation in the weight factors and the expected crash frequency between fixed dispersion and varying dispersion in the GeoP models (Fig. 5(c)).

Fig. 6 shows that the weigh factors for both the fixed and varying dispersion parameter models varied within a narrow range when the power parameter was less than 2.0. In the case of the NB2 and PIG

Table 7
Test results of potential for safety improvement (PSI) values based on total crashes.

Criterion	Urban Three-leg Stop-Controlled Intersection					Urban Four-leg Signalized Intersection				
	Paired *t*-Test			Spearman's Correlation		Paired *t*-Test			Spearman's Correlation	
	Mean diff.	*t*-statistic	p-value	ρ_s	p-value	Mean diff.	*t*-statistic	p-value	ρ_s	p-value
PSI between fixed dispersion parameter models										
PTw with fixed dispersion vs. NB1 with fixed dispersion	−0.651	−13.083	< 0.0001	0.989	< 0.0001	−0.536	−15.726	< 0.0001	1.000	< 0.0001
PTw with fixed dispersion vs. GeoP with fixed dispersion	−0.141	−14.416	< 0.0001	0.999	< 0.0001	0.337	7.603	< 0.0001	0.999	< 0.0001
PTw with fixed dispersion vs. NB2 with fixed dispersion	0.488	15.120	< 0.0001	0.991	< 0.0001	1.825	11.816	< 0.0001	0.997	< 0.0001
PTw with fixed dispersion vs. PIG with fixed dispersion	1.717	15.847	< 0.0001	0.910	< 0.0001	5.993	13.376	< 0.0001	0.970	< 0.0001
NB1 with fixed dispersion vs. GeoP with fixed dispersion	0.510	12.743	< 0.0001	0.993	< 0.0001	14.365	21.826	< 0.0001	0.994	< 0.0001
NB1 with fixed dispersion vs. NB2 with fixed dispersion	1.139	13.944	< 0.0001	0.964	< 0.0001	2.360	12.589	< 0.0001	0.996	< 0.0001
NB1 with fixed dispersion vs. PIG with fixed dispersion	2.368	15.504	< 0.0001	0.847	< 0.0001	6.529	13.611	< 0.0001	0.967	< 0.0001
NB2 with fixed dispersion vs. GeoP with fixed dispersion	−0.629	−14.980	< 0.0001	0.987	< 0.0001	12.004	19.316	< 0.0001	1.000	< 0.0001
NB2 with fixed dispersion vs. PIG with fixed dispersion	1.229	15.435	< 0.0001	0.951	< 0.0001	4.168	14.042	< 0.0001	0.985	< 0.0001
PIG with fixed dispersion vs. GeoP with fixed dispersion	−1.858	−15.890	< 0.0001	0.897	< 0.0001	7.836	11.734	< 0.0001	0.987	< 0.0001
PSI between varying dispersion parameter models										
PTw with varying dispersion vs. NB1 with varying dispersion	−2.471	−14.649	< 0.0001	0.977	< 0.0001	−4.763	−20.334	< 0.0001	1.000	< 0.0001
PTw with varying dispersion vs. GeoP with varying dispersion	−1.770	−14.200	< 0.0001	0.985	< 0.0001	−4.631	−20.693	< 0.0001	1.000	< 0.0001
PTw with varying dispersion vs. NB2 with varying dispersion	−0.926	−13.641	< 0.0001	0.995	< 0.0001	−0.965	−21.516	< 0.0001	1.000	< 0.0001
PTw with varying dispersion vs. PIG with varying dispersion	0.763	12.480	< 0.0001	0.995	< 0.0001	7.457	19.561	< 0.0001	0.998	< 0.0001
NB1 with varying dispersion vs. GeoP with varying dispersion	0.702	15.819	< 0.0001	0.999	< 0.0001	0.133	10.266	< 0.0001	1.000	< 0.0001
NB1 with varying dispersion vs. NB2 with varying dispersion	1.545	15.270	< 0.0001	0.993	< 0.0001	3.799	20.033	< 0.0001	1.000	< 0.0001
NB1 with varying dispersion vs. PIG with varying dispersion	3.234	14.135	< 0.0001	0.953	< 0.0001	12.220	19.885	< 0.0001	0.996	< 0.0001
NB2 with varying dispersion vs. GeoP with varying dispersion	−0.843	−14.813	< 0.0001	0.997	< 0.0001	−3.666	−20.452	< 0.0001	1.000	< 0.0001
NB2 with varying dispersion vs. PIG with varying dispersion	1.690	13.125	< 0.0001	0.980	< 0.0001	8.422	19.773	< 0.0001	0.998	< 0.0001
PIG with varying dispersion vs. GeoP with varying dispersion	−2.533	−13.691	< 0.0001	0.964	< 0.0001	−12.088	−20.027	< 0.0001	0.996	< 0.0001
PSI between fixed and varying dispersion parameter models										
PTw with fixed dispersion vs. PTw with varying dispersion	1.643	13.059	< 0.0001	0.986	< 0.0001	3.985	19.293	< 0.0001	1.000	< 0.0001
NB1 with fixed dispersion vs. NB1 with varying dispersion	−0.178	−3.336	0.0009	0.996	< 0.0001	−0.242	−6.645	< 0.0001	1.000	< 0.0001
GeoP with fixed dispersion vs. GeoP with varying dispersion	0.014	4.982	< 0.0001	1.000	< 0.0001	−0.982	−18.133	< 0.0001	1.000	< 0.0001
NB2 with fixed dispersion vs. NB2 with varying dispersion	0.229	3.488	0.0006	0.989	< 0.0001	1.196	5.643	< 0.0001	0.997	< 0.0001
PIG with fixed dispersion vs. PIG with varying dispersion	0.689	3.824	0.0001	0.917	< 0.0001	5.449	7.794	< 0.0001	0.973	< 0.0001

Note: PTw = Poisson-Tweedie model. NB1 = Negative binomial model with linear variance function. NB2 = Negative binomial model with quadratic variance function. GeoP = Geometric Poisson model. PIG = Poisson Inverse Gaussian model.

Table 8
Variation in ranking of urban three-leg stop-controlled intersections across models based on total crashes.

Obs. Crash Freq.	Fixed Dispersion										Varying Dispersion									
	PTw		NB1		GeoP		NB2		PIG		PTw		NB1		GeoP		NB2		PIG	
	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank
80	57.31	1	55.92	1	57.10	1	57.87	1	57.18	1	46.63	1	60.89	1	56.91	1	52.26	1	41.27	1
69	48.28	2	45.72	2	47.88	2	49.41	2	50.50	2	35.98	2	51.62	2	47.62	2	42.48	2	29.68	2
44	28.61	3	27.67	4	28.48	3	28.96	3	28.67	4	20.15	4	31.10	4	28.31	3	24.69	3	15.84	5
44	28.38	4	27.18	5	28.20	5	28.86	4	29.03	3	20.43	3	30.64	5	28.02	5	24.66	4	16.37	4
43	28.34	5	28.68	3	28.44	4	27.92	5	25.00	6	18.34	6	31.67	3	28.30	4	23.82	6	13.42	6
42	26.81	6	25.78	6	26.66	6	27.24	6	27.46	5	21.83	5	28.40	6	26.53	6	24.39	5	19.37	3
33	20.27	7	23.48	7	20.94	7	17.85	9	9.90	15	13.56	8	23.81	7	20.95	7	17.50	7	10.14	8
31	18.52	9	18.01	9	18.46	9	18.67	8	18.18	7	11.92	10	20.42	9	18.31	9	15.48	10	8.66	10
31	18.72	8	18.35	8	18.68	8	18.77	7	17.97	8	11.81	11	20.74	8	18.54	8	15.55	9	8.45	12
29	16.62	10	16.00	12	16.54	11	16.85	10	16.82	9	11.96	9	18.01	10	16.42	11	14.43	11	9.59	9
27	14.01	16	13.09	20	13.88	16	14.39	14	14.93	10	10.34	12	15.01	17	13.74	16	12.22	15	8.48	11
27	16.51	11	17.71	10	16.76	10	15.64	11	12.35	14	14.24	7	17.85	11	16.76	10	15.59	8	12.99	7
26	15.48	12	15.74	13	15.55	12	15.15	12	13.16	12	8.95	16	17.58	12	15.44	12	12.50	14	5.97	19
25	15.07	13	16.76	11	15.42	13	13.77	16	9.01	19	9.90	14	17.46	13	15.40	13	12.85	12	7.33	15
25	14.70	14	14.97	14	14.77	14	14.40	13	12.75	13	10.18	13	16.33	14	14.69	14	12.65	13	7.90	13
25	14.17	15	13.98	16	14.16	15	14.16	15	13.46	11	9.74	15	15.59	15	14.06	15	12.13	16	7.51	14
22	12.96	17	14.67	15	13.31	17	11.65	18	7.10	22	8.56	18	15.11	16	13.30	17	11.10	17	6.35	18
22	12.85	18	13.65	17	13.02	18	12.21	17	9.50	18	8.66	17	14.57	18	12.97	18	11.00	18	6.56	17
21	11.89	20	12.25	22	11.98	20	11.52	19	9.53	17	6.30	23	13.72	20	11.89	20	9.32	22	3.90	27
21	12.25	19	13.37	19	12.49	19	11.36	20	7.88	21	7.45	20	14.26	19	12.44	19	10.13	19	5.18	21
20	10.10	27	9.80	29	10.06	28	10.16	21	9.82	16	6.01	26	11.24	29	9.96	28	8.20	26	4.08	26
19	8.59	31	8.15	30	8.54	32	8.75	27	8.87	20	5.93	27	9.35	30	8.43	30	7.31	28	4.59	23
19	10.89	21	12.50	21	11.23	21	9.67	22	5.57	25	7.18	21	12.79	22	11.22	21	9.34	21	5.31	20
19	10.41	22	13.46	18	11.04	22	8.19	30	2.90	30	8.07	19	12.41	23	11.12	22	9.69	20	6.61	16
18	10.19	24	11.98	24	10.56	23	8.83	26	4.58	26	6.63	22	12.11	25	10.56	23	8.72	23	4.84	22
18	10.09	28	12.28	23	10.55	25	8.42	28	3.68	29	6.18	25	12.29	24	10.56	24	8.47	24	4.30	25
18	10.20	23	11.86	25	10.56	24	8.87	25	4.41	28	4.16	30	12.82	21	10.51	25	7.35	29	2.14	30
18	10.18	25	11.27	26	10.41	26	9.31	24	6.10	24	6.24	24	11.91	26	10.38	26	8.45	25	4.37	24
18	10.10	26	10.96	28	10.29	27	9.40	23	6.53	23	5.51	28	11.89	27	10.24	27	8.03	27	3.51	28
17	9.54	29	10.96	27	9.84	30	8.40	29	4.47	27	4.36	29	11.78	28	9.80	29	7.15	30	2.43	29

Note: PTw = Poisson-Tweedie model. NB1 = Negative binomial model with linear variance function. NB2 = Negative binomial model with quadratic variance function. GeoP = Geometric Poisson model. PIG = Poisson Inverse Gaussian model. PSI = Potential for Safety Improvement.

models, the weight factors had a wider distribution for fixed dispersion compared to the weight factors for varying dispersion. In particular, the range of weight factors in the PIG model with fixed dispersion parameter was in the range of 0.05-0.90. Similar to the plots in Fig. 5, the expected crash frequency estimates at urban four-leg signalized intersections were lower in the varying dispersion parameter models compared to that in the fixed dispersion parameter models.

Overall, the EB estimates of expected crash frequency were found to decrease with the increasing value of the power parameter in the varying dispersion models; however, the estimates are more or less consistent in the fixed dispersion parameter models. Also, with the increasing value of the power parameter, the weight factors in the fixed dispersion parameter models were found to have a wider range than those in the varying dispersion parameter models.

5.2. Potential for safety improvement

Based on the estimates of predicted and expected crash frequency, several studies have calculated the potential for safety improvement (PSI) to rank sites needing treatments, where PSI is equal to the expected crash frequency minus the predicted crash frequency (Lu et al., 2013). For each intersection in this study, the PSI was calculated based on the estimated model parameters with both the fixed and varying dispersion parameters.

Two tests, including the paired sample *t*-test and the Spearman's rank correlation test, were conducted based on the positive values of PSI, i.e., when the expected crash frequency was greater than the predicted crash frequency. The objective of the paired sample *t*-test was to examine if the mean difference between the pairs of PSI values estimated using different models was zero. The Spearman's rank correlation test was done to examine whether the order of the PSI values between two models was significantly different across intersections, under the null hypothesis of no correlation between the ordered pair of PSI values from two models. Note that the Shapiro-Wilk normality test of the resulting PSI values confirmed that a normally distribution could be assumed for the PSI values.

Positive PSI values were found in a total of 233 urban three-leg stop-controlled intersections and 329 urban four-leg signalized intersections. Based on these intersections, the aforementioned tests were performed between the PSI values for the following three sets: (i) all model combinations with the fixed dispersion parameter, (ii) all model combinations with the varying dispersion parameter, and (iii) corresponding models with the fixed and varying dispersion parameters. Table 7 shows the values of the Spearman's rank correlation coefficient (ρ_s) and the *t*-test statistics and their statistical significance.

Given the p-values in the paired *t*-test results were consistently less than 0.001, the mean difference between the PSI values of all relevant model combinations was statistically significant. On the other hand, the Spearman's rank correlation coefficients (ρ_s) between the PSI values were greater than 0.95 in most cases, indicating a strong correlation in the ordered pair of PSI values among all models. In summary, the test results indicate that, while the estimated mean of the PSI values was different in different models, the ordered pair of PSI values had strong associations. In other words, if the PSI value in one model was ranked high, the PSI value in the other model was also ranked high, even if the mean difference between the PSI values was statistically significantly different.

To further understand how the ranking of sites varies by different

Table 9
Variation in ranking of urban four-leg signalized intersections across models based on total crashes.

Obs. Crash Freq.	Fixed Dispersion										Varying Dispersion									
	PTw		NB1		GeoP		NB2		PIG		PTw		NB1		GeoP		NB2		PIG	
	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank	PSI	Rank
365	300.46	1	300.49	1	301.02	1	301.40	1	303.02	1	279.47	1	303.40	1	302.85	1	283.92	1	243.12	1
270	207.39	2	206.58	2	208.73	2	210.36	2	212.85	2	189.57	2	209.57	2	208.63	2	193.53	2	156.08	2
247	186.74	5	186.30	5	187.73	5	188.76	5	190.53	4	172.91	5	188.67	5	188.21	5	175.96	5	148.46	5
247	200.79	3	202.12	3	199.94	3	197.61	3	194.37	3	187.85	3	203.07	3	203.03	3	190.61	3	166.27	3
240	197.04	4	198.72	4	195.79	4	192.63	4	187.71	5	184.51	4	199.36	4	199.38	4	187.17	4	163.73	4
229	181.31	6	181.94	6	181.15	6	180.01	6	177.86	6	166.28	6	183.51	6	182.83	6	169.59	6	138.62	7
222	170.12	8	170.53	8	170.25	7	169.78	7	169.45	7	159.02	7	171.96	8	171.87	8	161.46	7	140.41	6
219	167.28	10	167.41	10	167.66	9	167.59	8	167.49	8	154.08	9	169.19	10	168.67	9	157.00	9	130.17	8
213	167.33	9	167.90	9	167.18	10	166.03	9	163.33	9	152.20	10	169.45	9	168.62	10	155.57	10	123.55	9
210	156.66	11	155.88	11	157.86	11	159.16	11	159.99	10	138.58	14	158.58	11	157.19	11	142.76	14	101.50	23
203	136.29	19	134.94	19	138.21	17	140.72	15	144.08	14	125.48	19	137.66	18	137.22	18	127.98	19	106.03	21
203	153.27	12	153.55	12	153.49	12	153.16	12	152.71	11	141.88	11	155.03	12	154.69	12	144.41	11	121.75	12
202	171.18	7	173.84	7	168.51	8	161.16	10	141.30	15	155.07	8	173.94	7	173.04	7	158.60	8	125.01	10
199	152.76	13	153.32	13	152.68	13	151.75	13	150.14	12	141.01	13	154.60	13	154.21	13	143.62	12	119.98	13
197	151.05	14	151.83	14	150.77	14	149.53	14	147.85	13	141.31	12	152.80	14	152.76	14	143.45	13	125.01	11
197	135.93	18	135.25	18	131.18	20	138.58	17	140.50	16	126.32	18	137.34	19	137.11	19	128.52	18	109.79	16
190	138.22	16	138.47	16	138.52	16	138.33	18	138.22	18	129.68	15	139.74	16	139.76	16	131.61	15	115.69	14
189	137.08	17	137.12	17	137.57	18	137.71	19	138.00	19	127.44	17	138.62	17	138.42	17	129.61	17	110.78	15
189	139.45	15	139.47	15	139.92	15	140.02	16	139.96	17	127.90	16	141.11	15	140.60	15	130.51	16	106.61	19
186	129.53	22	129.28	25	130.80	21	130.99	21	131.92	21	121.01	21	130.93	23	130.83	23	122.96	21	106.75	18
185	132.75	20	132.17	21	119.90	27	134.92	20	136.05	20	119.59	24	134.34	20	133.51	20	122.62	22	93.75	25
178	118.29	28	117.21	28	131.43	19	121.96	26	124.71	24	107.93	28	119.57	28	119.06	28	110.35	28	88.52	29
174	131.86	21	132.73	20	126.12	24	129.91	22	127.53	22	123.15	20	133.47	21	133.42	21	125.09	20	108.45	17
174	125.13	25	124.56	24	121.25	25	127.08	23	127.39	23	110.88	26	126.71	25	125.67	25	114.20	25	81.86	30
172	120.47	27	120.20	27	114.62	30	121.90	27	122.69	25	110.71	27	121.88	27	121.49	27	112.95	27	92.87	26
167	113.81	30	113.54	30	116.93	28	115.29	30	116.17	28	105.48	30	115.11	30	114.90	30	107.41	30	90.91	27
165	116.36	29	116.28	29	127.87	22	117.22	29	117.41	26	106.85	29	117.77	29	117.38	29	109.04	29	89.43	28
162	129.23	23	130.85	22	126.60	23	124.22	24	116.45	27	119.61	23	131.06	22	130.85	22	121.72	23	102.73	22
160	128.15	24	129.96	23	121.00	26	122.62	25	114.77	30	119.93	22	129.91	24	129.97	24	121.71	24	106.20	20
160	121.49	26	122.35	26	115.23	29	119.26	28	115.69	29	111.80	25	123.11	26	122.77	26	113.97	26	94.25	24

Note: PTw = Poisson-Tweedie model. NB1 = Negative binomial model with linear variance function. NB2 = Negative binomial model with quadratic variance function. GeoP = Geometric Poisson model. PIG = Poisson Inverse Gaussian model. PSI = Potential for Safety Improvement.

models, a sample of 30 intersections from the entire data set of each intersection type was selected based on the highest number of observed crashes during three years of the study period. Tables 8 and 9 show the PSI values and rank of the 30 selected urban three-leg stop-controlled intersections and urban four-leg signalized intersections, respectively. The PSI values for the varying dispersion parameter models were found to be lower as the value of the power parameter increased (e.g., PIG). This indicates that the predicted crash frequency estimates with the varying dispersion parameter were more reliable for models fitted with a higher value of the power parameter. However, there was no such pattern between the PSI values and the power parameter when the crash frequency was estimated using the fixed dispersion parameter. In terms of ranking, the intersections ranked between #1 and #6 in both the tables consisted of similar set of intersections, and among them, the intersections in rank #1 and #2 were consistent in all the models. However, there was no specific pattern in the ranking of other intersections, which sometimes varied within one position and sometimes varied by two or more positions.

Lord and Prak (2008) reported that the ranking based on PSI values between the fixed and varying dispersion parameter models may be quite different despite a strong correlation in the ordered pair of PSI values. This study extends the previous work by examining the rank of the sites across an array of models those varied not only by how dispersion parameter was modeled but also by the degree of the variance function (i.e., power parameter). In summary, a model fitted with a varying dispersion parameter and an increasing value of power parameter may provide more reliable ranking, as the difference between the expected crash frequency and the predicted crash frequency may gradually reduce.

6. Summary and conclusions

This study demonstrated the application of the flexible Poisson-Tweedie family of distributions in the regression analysis of crash frequency data. Having expressed the variance as a simple function of the power of the mean, the Poisson-Tweedie model can estimate the power parameter P without any restrictions to its value. The study analyses were based on two datasets of urban intersections, including urban three-leg stop-controlled intersections and urban four-leg signalized intersections, in Florida. For each intersection type, a total of 20 models were developed based on the combination of crash type (i.e., total and FI), dispersion parameter type (i.e., fixed and varying dispersion), and five power parameters (i.e., one with no fixed value and the others with 1.0, 1.5, 2.0, 3.0). Note that the models fitted with the power parameter 1.0, 1.5, 2.0, and 3.0 represent NB1, GeoP, NB2, and PIG models, respectively.

The study results show that values of the dispersion parameter were smaller in models fitted with a higher value of the power parameter. For example, the value of the fixed dispersion parameter was found to be the lowest in the PIG models and the highest in the NB1 models. The range of values of the varying dispersion parameter across the study intersections also showed similar patterns.

The coefficient of logarithm of major AADT (i.e., ln(major AADT)) associated with the varying dispersion parameter was negative, similar to the findings in Wang et al. (2019). Of them, the association was consistently found to be statistically significant at the 0.05 level in the NB2 and PIG models. On the other hand, the association between ln (major AADT) and the dispersion parameter was found to be not statistically significant in the GeoP models for both total and FI crashes at

urban three-leg stop-controlled intersections. Similar findings were also observed in three out of four cases of the NB1 model. It is obvious that when the covariate of the varying dispersion model is not significant, it may be better to fit models with the mean (i.e., fixed) dispersion parameter as this saves time and effort. In summary, the dispersion parameter of a model is very sensitive to how the model is parameterized in terms of the power parameter.

Of several goodness-of-fit measures, the MAD and modified R^2 values were found to be approximately similar. Based on other measures such as pLL, pAIC, and pBIC, and the significance of the model parameters, the Poisson-Tweedie models or the GeoP models were found to perform better than the NB2 models when the dispersion parameter was constant. On the other hand, when the dispersion parameter was modeled as a function of covariate(s), the NB2 and PIG models were found to perform better, with both performing equally equal. It is worth mentioning that Zha et al. (2016) also reported similar findings while comparing the performance of NB2 and PIG models.

The EB estimates of the expected crash frequency were lower and the values of the weight factors were higher in the varying dispersion parameter models compared to the fixed dispersion parameter models. This indicates that the predicted crash frequency was more reliable in the varying dispersion parameter model. The weight factors were also found to have a wider range when the value of the power parameter was 2.0 or greater. Also, with the increasing value of the power parameter, the PSI values decreased under the varying dispersion parameter models. Except for some high crash locations, the sites needing treatment based on PSI values had a discordant in ranking across different models.

This study showed that different count models based on the degree of variance parameter can be easily fitted under the Poisson-Tweedie modeling framework. Although the NB model appears to be quite efficient in fitting crash frequency data in several cases, it may not always be the most appropriate choice. As such, the Poisson-Tweedie class of models, which includes NB, PIG, and other distributions, is a suitable alternative to examine and select the appropriate function for crash analysis.

One limitation of this study is related to using only two covariates associated with intersection traffic flow characteristics (i.e., major road AADT and minor road AADT). Given several other covariates were initially examined but not found to have any significant associations, a rigorous examination with a comprehensive set of covariates is required to find the significant associations of additional variables with crash frequency. Additional covariates may contribute to lessening the effect of possible omitted-variables bias.

Further research based on different datasets including additional covariates as well as other intersection and roadway types can provide more insights on the performance of the Poisson-Tweedie class of models in crash analysis. Future research can also extend the Poisson-Tweedie models to fit crash data that are characterized by under-dispersion, zero-inflation, spatial-temporal correlation, repeated measures, or correlated multiple dependent variables.

CRediT authorship contribution statement

Dibakar Saha: Conceptualization, Software, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Priyanka Alluri:** Data curation, Writing - original draft, Writing - review & editing. **Eric Dumbaugh:** Writing - original draft, Writing - review & editing. **Albert Gan:** Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The intersections data used in this research was collected as part of the project BDK80-977-37, funded by the Research Center of the Florida Department of Transportation (FDOT).

References

Alluri, P., Saha, D., Liu, K., Gan, A., 2014. Improved Processes for Meeting the Data Requirements for Implementing the Highway Safety Manual (HSM) and Safety Analyst in Florida. Final Report. Florida Department of Transportation, Tallahassee, FL.

American Association of State Highway Transportation Officials, 2010. Highway Safety Manual. Federal Highway Administration, Washington, D.C.

Anastasopoulos, P.C., Mannering, F.L., Shankar, V.N., Haddock, J.E., 2012. A study of factors affecting highway accident rates using the random-parameters tobit model. Accid. Anal. Prev. 45, 628–633.

Bonat, W.H., 2018. Multiple response variables regression models in R: the mcglm package. J. Stat. Softw. 84 (1), 1–30.

Bonat, W.H., Jørgensen, B., 2016. Multivariate covariance generalized linear models. J. R. Stat. Soc. Ser. C Appl. Stat. 65 (5), 649–675.

Bonat, W.H., Jørgensen, B., Kokonendji, C.C., Hinde, J., Demétrio, C.G., 2018. Extended Poisson–Tweedie: properties and regression models for count data. Stat. Modelling 18 (1), 24–49.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge, UK.

Carey, V.J., Wang, Y., 2011. Working covariance model selection for generalized estimating equations. Stat. Med. 30 (26), 3117–3124.

Cheng, L., Geedipally, S.R., Lord, D., 2013. The Poisson–Weibull generalized linear model for analyzing motor vehicle crash data. Saf. Sci. 54, 38–42.

Debrabant, B., Halekoh, U., Bonat, W.H., Hansen, D.L., Hjelmborg, J., Lauritsen, J., 2018. Identifying traffic accident black spots with Poisson-Tweedie models. Accid. Anal. Prev. 111, 147–154.

El-Basyouny, K., Sayed, T., 2006. Comparison of two negative binomial regression techniques in developing accident prediction models. Transp. Res. Rec.: J. Transp. Res. Board 1950, 9–16.

El-Shaarawi, A.H., Zhu, R., Joe, H., 2011. Modelling species abundance using the Poisson–Tweedie family. Environmetrics 22 (2), 152–164.

Gan, A., Raihan, M.A., Alluri, P., Liu, K., Saha, D., 2016. Updating and Improving Methodology for Prioritizing Highway Project Locations on the Strategic Intermodal System (SIS). Final Report. Florida Department of Transportation, Tallahassee, FL.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-Lindley generalized linear model: characteristics and application using crash data. Accid. Anal. Prev. 45, 258–265.

Geedipally, S.R., Lord, D., Park, B.J., 2009. Analyzing different parameterizations of the varying dispersion parameter as a function of segment length. Transp. Res. Rec.: J. Transp. Res. Board 2103, 108–118.

Greene, W., 2008. Functional forms for the negative binomial model for count data. Econ. Lett. 99 (3), 585–590.

Harwood, D., Torbic, D., Richard, K., Meyer, M., 2010. SafetyAnalyst: Software Tools for Safety Management of Specific Highway Sites. Report FHWA-HRT-10-063. Federal Highway Administration, Office of Safety, McLean, VA.

Hauer, E., 1997. Observational Before-after Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Pergamon, Tarrytown, NY.

Hilbe, J.M., 2011. Negative Binomial Regression, 2nd edition. Cambridge University Press, Cambridge, UK.

Holla, M.S., 1967. On a Poisson-inverse Gaussian distribution. Metrika 11 (1), 115–121.

Jørgensen, B., Kokonendji, C.C., 2016. Discrete dispersion models and their Tweedie asymptotics. Asta Adv. Stat. Anal. 100 (1), 43–78.

Kokonendji, C.C., Dossou-Gbété, S., Demétrio, C.G., 2004. Some discrete exponential dispersion models: Poisson-Tweedie and Hinde-Demétrio classes. Stat. Oper. Res. Trans. 28 (2), 201–214.

Kolody, K., Perez-Bravo, D., Zhao, J., Neuman, T.R., 2014. Highway Safety Manual User Guide. CH2M HILL, Chicago, IL.

Lord, D., Guikema, S.D., Geedipally, S.R., 2008. Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. Accid. Anal. Prev. 40 (3), 1123–1134.

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

Lord, D., Park, P.Y.J., 2008. Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. Accid. Anal. Prev. 40 (4), 1441–1457.

Lu, J., Haleem, K., Alluri, P., Gan, A., 2013. Full versus simple safety performance functions: comparison based on urban four-lane freeway interchange influence areas in Florida. Transp. Res. Rec.: J. Transp. Res. Board 2398, 83–92.

Lyon, C., Persaud, B., Gross, F., 2016. The Calibrator: an SPF Calibration and Assessment Tool User Guide. Report No. FHWA-SA-17-016. Federal Highway Administration, Washington, D.C.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.

Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road

sections: Poisson versus negative binomial regressions. Accid. Anal. Prev. 26 (4), 471–482.
Miaou, S.P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transp. Res. Rec.: J. Transp. Res. Board 1840, 31–40.
Miranda-Moreno, L.F., Fu, L., Saccomanno, F.F., Labbe, A., 2005. Alternative risk models for ranking locations for safety improvement. Transp. Res. Rec.: J. Transp. Res. Board 1908, 1–8.
Özel, G., İnal, C., 2010. The probability function of a geometric Poisson distribution. J. Stat. Comput. Simul. 80 (5), 479–487.
Peng, Y., Lord, D., Zou, Y., 2014. Applying the Generalized Waring model for investigating sources of variance in motor vehicle crash analysis. Accid. Anal. Prev. 73, 20–26.
Petterle, R.R., Bonat, W.H., Kokonendji, C.C., Seganfredo, J.C., Moraes, A., da Silva, M.G., 2019. Double Poisson-Tweedie Regression Models. Int. J. Biostat. 15 (1). https://doi.org/10.1515/ijb-2018-0119.
R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.
Raihan, M.A., Alluri, P., Wu, W., Gan, A., 2019. Estimation of bicycle crash modification factors (CMFs) on urban facilities using zero inflated negative binomial models. Accid. Anal. Prev. 123, 303–313.
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.
Shirazi, M., Lord, D., 2019. Characteristics-based heuristics to select a logical distribution between the Poisson-gamma and the Poisson-lognormal for crash data modelling. Transp. A Transp. Sci. 15 (2), 1791–1803.
Ukkusuri, S., Miranda-Moreno, L.F., Ramadurai, G., Isa-Tavarez, J., 2012. The role of built environment on pedestrian crash frequency. Saf. Sci. 50, 1141–1151.
Vangala, P., Lord, D., Geedipally, S.R., 2015. Exploring the application of the negative binomial–generalized exponential model for analyzing traffic crash data with excess zeros. Anal. Methods Accid. Res. 7, 29–36.
Venkataraman, N., Ulfarsson, G.F., Shankar, V.N., 2013. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. Accid. Anal. Prev. 59, 309–318.
Wang, K., Zhao, S., Jackson, E., 2019. Functional forms of the negative binomial models in safety performance functions for rural two-lane intersections. Accid. Anal. Prev. 124, 193–201.
Wickham, H., François, R., Henry, L., Müller, K., 2019. dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr.
Wu, L., Zou, Y., Lord, D., 2014. Comparison of Sichel and negative binomial models in hot spot identification. Transp. Res. Rec.: J. Transp. Res. Board 2460, 107–116.
Zha, L., Lord, D., Zou, Y., 2016. The Poisson inverse Gaussian (PIG) generalized linear regression model for analyzing motor vehicle crash data. J. Transp. Saf. Secur. 8 (1), 18–35.
Zou, Y., Lord, D., Zhang, Y., Peng, Y., 2013. Comparison of Sichel and negative binomial models in estimating empirical Bayes estimates. Transp. Res. Rec.: J. Transp. Res. Board 2392, 11–21.