

Comparison of six generalized linear models for occurrence of lightning-induced fires in northern Daxing'an Mountains, China

Futao Guo^{1,2} · Guangyu Wang² · John L. Innes² · Zhihai Ma³ · Aiqin Liu¹ · Yurui Lin⁴

Received: 18 September 2014 / Accepted: 4 November 2014 / Published online: 27 November 2015
© Northeast Forestry University and Springer-Verlag Berlin Heidelberg 2015

Abstract The occurrence of lightning-induced forest fires during a time period is count data featuring over-dispersion (i.e., variance is larger than mean) and a high frequency of zero counts. In this study, we used six generalized linear models to examine the relationship between the occurrence of lightning-induced forest fires and meteorological factors in the Northern Daxing'an Mountains of China. The six models included Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), Poisson hurdle (PH), and negative binomial hurdle (NBH) models. Goodness-of-fit was compared and tested among the six models using Akaike information criterion (AIC), sum of squared errors, likelihood ratio test, and Vuong test. The predictive performance of the models was

assessed and compared using independent validation data by the data-splitting method. Based on the model AIC, the ZINB model best fitted the fire occurrence data, followed by (in order of smaller AIC) NBH, ZIP, NB, PH, and Poisson models. The ZINB model was also best for predicting either zero counts or positive counts (≥ 1). The two Hurdle models (PH and NBH) were better than ZIP, Poisson, and NB models for predicting positive counts, but worse than these three models for predicting zero counts. Thus, the ZINB model was the first choice for modeling the occurrence of lightning-induced forest fires in this study, which implied that the excessive zero counts of lightning-induced fires came from both structure and sampling zeros.

Keywords Poisson · Negative binomial (NB) · Zero-inflated Poisson (ZIP) · Zero-inflated negative binomial (ZINB) · Poisson hurdle (PH) · Negative binomial hurdle (NBH) · Likelihood ratio test (LRT) · Vuong test

Project funding: This work was funded by Asia-Pacific Forests Net (APFNET/2010/FPF/001); National Natural Science Foundation of China (Grant No. 31400552).

The online version is available at <http://www.springerlink.com>

Corresponding editor: Chai Ruihai

✉ Guangyu Wang
guangyu.wang@ubc.ca

✉ Aiqin Liu
fjlaq@126.com

¹ College of Forestry, Fujian Agriculture and Forestry University, Fuzhou 350002, China

² Sustainable Forest Management Laboratory, Faculty of Forestry, University of British Columbia, 2424 Main Mall, Vancouver, BC V6T 1Z4, Canada

³ Department of Medicine, University of Calgary, 3280 Hospital Drive NW, Calgary, AB T2N 4Z6, Canada

⁴ College of Computer and Information Science, Fujian Agriculture and Forestry University, Fuzhou 350002, China

Introduction

Over-dispersed count data are common in ecological research. Similarly, the occurrence of forest fires is characterized by over-dispersion and a high frequency of zeros. These features of fire occurrence data present challenges for better understanding the ecological processes of forest fires and effectively modeling the future scenarios of forest fires using climate variables. Therefore, selecting an appropriate model or models to address both over-dispersion and excessive zeros is crucial for developing realistic prediction systems of forest fires in order to provide reliable information for fire prevention, land-use planning, and decision-making in natural resources management in China (Guo et al. 2010a; Xu 2014).

In past decades, forest researchers devoted considerable time and effort to model the ignition and occurrence of forest fires. In the literature, logistic regression models have been used to estimate the ignition probability of forest fires, while Poisson regression models have been applied to predict numbers of fire occurrences (Martell et al. 1987; Chou et al. 1993; Poulin-Costello 1993; Vega-Garcia et al. 1995; Mandallaz and Ye 1997; García Diez et al. 1999; Preisler et al. 2004; Griffith and Haining 2006; Liu and Cela 2008; Podur et al. 2009). However, the Poisson model is criticized for its restrictive assumption of equality between the sample mean and variance. It has been noted in various applications that the observed dispersion of count data is commonly underestimated by Poisson models. As an alternative, negative binomial (NB) models have been adopted for count data when the sample variance exceeds the sample mean (i.e., over-dispersion) (Cameron and Trivedi 1998).

In reality, forest fire occurrence data not only exhibit over-dispersion, but also include excessive numbers of zero counts. Zero-inflated models (Lambert 1992) and hurdle models (Mullahy 1986) have been utilized to address these situations. Both zero-inflated and hurdle models assume that count data are a mixture of two separate data generation processes: one generates only zeros, and the other is either a Poisson or an NB data-generating process. However, these two types of models are distinct in their interpretation and analysis of zero counts. Zero-inflated models allow for two separate processes. Conceptually, the first step is to model the structural zeros using a logistic regression and the second step is to model the Poisson distribution conditional on the structural zeros; i.e., a Poisson or NB model is used for the sampling zeros and positive counts. In contrast, hurdle models are interpreted as two-part models, in which a logistic regression model governs the binary outcome of whether a count variable has a zero or a positive realization. If the realization is positive, “the hurdle” is crossed, and the conditional distribution of the positive counts is then determined by a truncated-at-zero Poisson or NB model (Cameron and Trivedi 1998; Rose et al. 2006). In summary, Zero-inflated models assume that zero counts have two different origins-structure and sampling, whereas Hurdle models assume that all zero counts are from one structural source (Erdman et al. 2008; Hu et al. 2011).

The applications of Zero-inflated and hurdle models can be found in different study fields (e.g., Yau and Lee 2001; Ridout et al. 2001; Affleck 2006; Lee et al. 2011), but only few have been related to the prediction of forest fire occurrence (e.g., Krawchuk et al. 2009). Most studies that assessed the relationship between forest fire occurrence and meteorological conditions either emphasized the prediction function of the models or focused on the selection of independent variables in order to improve model performance. Only a few

studies undertook comprehensive analysis of the process of model selection based on the principles of statistics (Mandallaz and Ye 1997; Guo et al. 2010b).

In this study, we used six generalized linear models, viz. Poisson, NB, zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), Poisson hurdle (PH), and negative binomial hurdle (NBH) models, to fit the occurrence of lightning-induced fires (count data) to examine the relationship between forest fires and corresponding meteorological factors in the northern Daxing'an Mountains, China. The objective of this study was to provide comparative assessment for researchers to deal with the challenges of analyzing and modeling forest fire occurrence (count data) with over-dispersion and excessive zeros.

Materials and methods

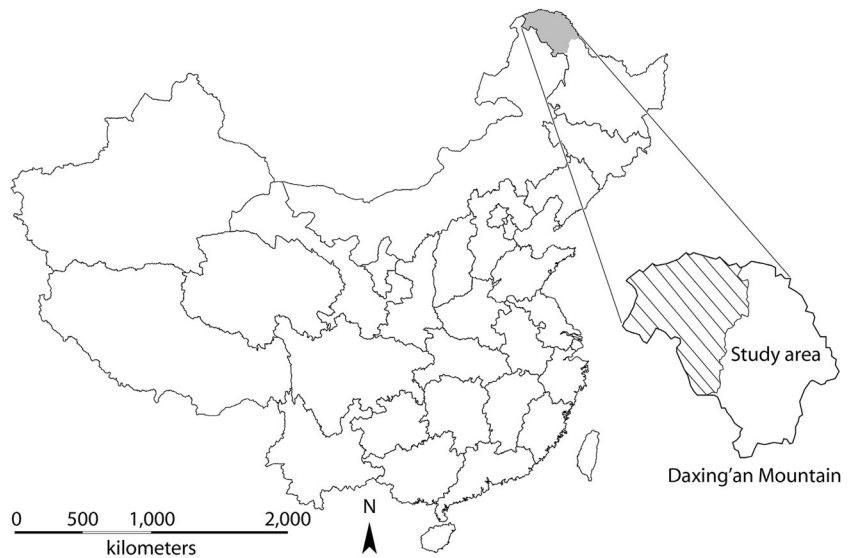
Study site description

The study site is located in high-latitude boreal forest regions of the Daxing'an Mountains (50°10'–53°33'N, 121°12'–127°00'E) with a total area of 8.46×10^6 ha in northeast China (Fig. 1). The Daxing'an Mountains support the largest natural forests in China. Dominant tree species include Daurian larch (*Larix gmelinii* Rupr.), White birch (*Betula platyphylla* Suk.), Mongolian pine (*Pinus sylvestris* L. var. *mongolica* Litv.), and Mongolian oak (*Quercus mongolica* Fischer ex Ledebour) (Xu 1998). Mean annual temperature is -2 to 4 °C, with extremes ranging from -52.3 to 39.0 °C. Total annual precipitation is 350–500 mm and most is received in winter and early spring as snow. Elevation ranges from 300 to 1400 m. The Daxing'an Mountains consist of seven sub-administrative regions (Xu 1998). Our study area was located in the northern Daxing'an Mountains, and included three sub-administrative regions (namely Mohe, Huzhong, and Tahe), covering an area of about 42×10^5 ha (Fig. 1).

Fire frequency data

The Daxing'an Mountains have an extremely high fire risk and the highest average area burned annually in China. The fire occurrence data used in this study were collected from 1980 to 2005. According to the records, there were over 1000 forest fires and nearly 1.3×10^5 ha burned area during this 26-year time period. Our fire data, including location, ignition dates and total burned area, were provided by the Fire Prevention Office of Heilongjiang Province (FPOHP). We chose to focus on the Mohe, Huzhong, and Tahe sub-administrative regions in the Northern Daxing'an Mountains because the records of fire occurrence were relatively complete compared to the other sub-

Fig. 1 Map of the study area (*shaded*) within the Daxing'an Mountains in northeast China



administrative regions. We addressed only lightning-induced fires in this study due to the completeness of the data records and the significant relationship between lightning fires and meteorological factors (Yu et al. 2007; Guo et al. 2010a, b; Chang et al. 2013). The fire frequency or count was calculated on a monthly scale from January to December of each year. Hence, the dependent variable was the monthly occurrence (count) of lightning-induced fires over the 26-year time period (1980–2005).

Meteorological variables

We focused on five meteorological variables, viz. average monthly wind speed (AMWS), average monthly temperature (AMT), average monthly precipitation (AMP), average monthly relative humidity (AMRH), and average monthly evaporation (AME). These variables significantly impact forest fire occurrence in the Daxing'an Mountains (Yu et al. 2007; Guo et al. 2010a, b; Chang et al. 2013). The meteorological data were provided by the China Meteorological Data and Sharing Network (<http://cdc.cma.gov.cn/>), which included more than seven hundred national meteorological stations across China. Three national meteorological stations were located in our study area, one in each sub-administrative region (Mohe, Huzhong, and Tahe).

The descriptive statistics for the fire occurrence counts and meteorological variables are listed in Table 1. The average monthly fire occurrence was 0.23, while the variance was 0.80. The ratio of the variance to the mean was 3.45, showing over-dispersion in the fire occurrence data. Figure 2 shows the frequency distribution of the observed counts of the lightning-induced fires, illustrating a large proportion of zero counts. The zero records contain some

Table 1 Descriptive statistics for the occurrence of lightning-induced forest fires (dependent variable) and meteorological factors (independent or predictor variables)

Variable	Mean	Variance	Minimum	Maximum
Fire count	0.23	0.80	0.00	9.00
AMWS (m/s)	2.31	0.50	0.50	4.50
AMP (mm)	38.24	2134.52	0.00	266.20
AMT (°C)	-2.16	259.82	-34.00	21.20
AMRH (%)	67.89	98.24	7.00	89.00
AME (mm)	82.64	4887.24	1.70	306.60

AMWS is average monthly wind speed; AMP is average monthly precipitation; AMT is average monthly temperature; AMRH is average monthly relative humidity; AME is average monthly evaporation

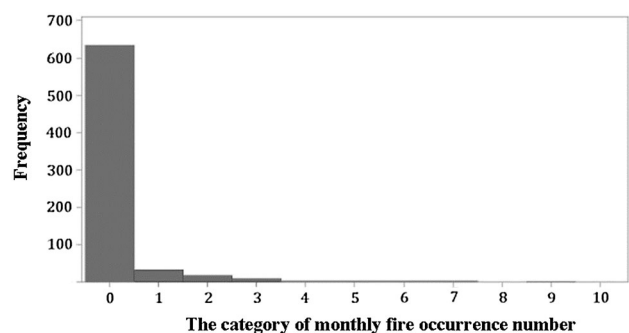


Fig. 2 Frequency distribution of the monthly occurrence of lightning-induced forest fires over the study period (1980–2005). X-axis represents the category of monthly fire occurrence number. Y-axis represents frequency of the category over the study period. The total number of counts (frequency) used was 704 for model fitting

“structural (or true) zeros” due to absence of lightning strikes during non-fire seasons and some “sampling zeros” recorded during fire seasons when fire was not recorded due to the combined effects of meteorological factors but lightning strikes actually occurred.

Statistical models

Poisson model

The Poisson model is used to model counts of events during a time period as a function of predictor variables and is based on the assumption that the conditional mean equals the conditional variance. The probability density function (*pdf*) of the Poisson model is:

$$P(Y) = \frac{e^{-\mu} \cdot \mu^Y}{Y!} = \frac{e^{-\mu} \cdot \mu^Y}{\Gamma(Y + 1)} \tag{1}$$

where $P(Y)$ is the probability that the number of events (Y) occurs during a time period, and μ is the parameter representing the expected value of Y ; i.e., $E(Y) = \mu$ and $\text{Var}(Y) = \mu$, and $\Gamma()$ is gamma function. The set of predictor variables X impacts the mean of the response variable μ via a link function such that $\eta = g(\mu) = \ln(X\beta)$, and the inverse link function (mean function) is

$$\mu = g^{-1}(\eta) = e^{X\beta} \text{ or } \ln(\mu) = X\beta \tag{2}$$

where β is the model coefficient to be estimated from data. Thus, Eq. 2 is a regression model relating the natural logarithm of the response mean or expected number of events to the explanatory or predictor variables (Cameron and Trivedi 1998; Osgood 2000).

Negative binomial (NB) model

The NB distribution can be used for count data with over-dispersion, i.e., when the sample variance exceeds the sample mean. The NB model addresses over-dispersion by including a dispersion parameter to accommodate unobserved heterogeneity in count data. The NB model used in this study has the following *pdf*:

$$P(Y) = \frac{\Gamma(Y + \frac{1}{\kappa})}{\Gamma(Y + 1)\Gamma(\frac{1}{\kappa})} \left(\frac{1}{1 + \kappa\mu}\right)^{1/\kappa} \left(\frac{\kappa\mu}{1 + \kappa\mu}\right)^Y \tag{3}$$

The mean of Y is $E(Y) = \mu$ and the variance of Y is $V(Y) = \mu + \kappa\mu^2$, where $\kappa \geq 0$ which is usually referred to as the dispersion parameter. Equation 3 allows the variance to exceed the mean. Consequently, the Poisson model can be regarded as a limiting model of NB model as the dispersion parameter κ approaches 0 (Miaou 1994). Given a set of predictor variables X , the link function of the NB model is also $\eta = g(\mu) = \ln(X\beta)$

Zero-inflated models: ZIP and ZINB

Observed count data are frequently characterized by over-dispersion and many zero counts. Zero-inflated models are powerful in these situations. Zero-inflated models generate two models as follows: a logistic model is first generated for the “certain zero” in order to predict whether a case would happen. Then, a Poisson or NB model is generated to predict the counts for the case (≥ 0). In other words, Zero-inflated models consider two sources of zero observations: “structural or true zeros” which cannot score anything other than “0”, and “sampling zeros” which are part of the underlying sampling distribution (either a Poisson model (ZIP) or an NB model (ZINB)). Zero-inflated Poisson model can be expressed as (Numna 2009):

$$P(Y) = \begin{cases} \omega + (1 - \omega)e^{-\mu} & Y = 0 \\ (1 - \omega) \frac{e^{-\mu}\mu^Y}{Y!} & Y \geq 1 \end{cases} \quad 0 \leq \omega \leq 1 \tag{4}$$

The mean and variance of the ZIP model are, respectively, $E(Y) = (1 - \omega)\mu$ and $V(Y) = (1 - \omega)(\mu + \omega\mu^2)$, where ω denotes the probability of being an individual having zero count and μ denotes the mean of the underlying distribution. Equation 4 shows that the marginal distribution of Y exhibits over-dispersion if $\omega < 0$, and it reduces to the standard Poisson model when $\omega = 0$

The alternative is that Y has the Zero-inflated NB distribution, specifically:

$$P(Y) = \begin{cases} \omega + (1 - \omega) \left(\frac{1}{1 + \kappa\mu}\right)^{1/\kappa} & \text{if } Y = 0 \\ (1 - \omega) \frac{\Gamma\left(Y + \frac{1}{\kappa}\right)}{\Gamma(Y + 1)\Gamma\left(\frac{1}{\kappa}\right)} \left(\frac{1}{1 + \kappa\mu}\right)^{1/\kappa} \left(\frac{\kappa\mu}{1 + \kappa\mu}\right)^Y & \text{if } Y \geq 1 \end{cases} \tag{5}$$

where $\kappa \geq 0$ is a dispersion parameter that is assumed independent of covariates. The mean and the variance of the distribution are $E(Y) = (1 - \omega) \mu$ and $V(Y) = (1 - \omega)[\mu(1 + \mu\kappa) + \omega\mu^2]$, respectively. The ZINB model reduces to the ZIP model in the limit $\kappa \rightarrow 0$

Hurdle models: PH and NBH

Hurdle models, first discussed by Mullahy (1986), are popular for modeling count data with many zeros. In contrast to Zero-inflated models, Hurdle models can be interpreted as two-part models: a logistic model is used to predict the binary outcome whether a count variable has a zero or a positive realization. If the realization is positive (i.e., the hurdle is crossed), a truncated-at-zero Poisson or NB model is used to predict the conditional distribution of the positive counts (≥ 1) (Cameron and Trivedi 1998). The Hurdle model can be expressed as:

$$P(Y) = \begin{cases} \omega & Y = 0 \\ (1 - \omega) \frac{f(Y=y)}{1 - f(Y=0)} & Y \geq 1 \end{cases} \quad 0 \leq \omega \leq 1 \tag{6}$$

where ω is the probability of a zero count and $(1 - \omega)$ is the probability of overcoming the hurdle. We can define two Hurdle models by specifying $f(Y)$ as a Poisson or an NB distribution. If we substitute Eq. 1 into Eq. 6 we obtain the Poisson Hurdle model (PH) as follows:

$$P(Y) = \begin{cases} \omega & \text{if } Y = 0 \\ [1 - \omega] \frac{e^{-\mu} \mu^Y}{(1 - e^{-\mu}) \cdot \Gamma(Y + 1)} & \text{if } Y \geq 1 \end{cases} \tag{7}$$

Alternatively, if we substitute Eq. 3 into Eq. 6 for $f(Y)$ we generate a NBH model as follows:

$$P(Y) = \begin{cases} \omega & \text{if } Y = 0 \\ [1 - \omega] \frac{\Gamma\left(Y + \frac{1}{\kappa}\right)}{\left[1 - \left(\frac{1}{1 + \kappa\mu}\right)^{1/\kappa}\right] \Gamma(Y + 1) \Gamma\left(\frac{1}{\kappa}\right)} \left(\frac{1}{1 + \kappa\mu}\right)^{1/\kappa} \left(\frac{\kappa\mu}{1 + \kappa\mu}\right)^Y & \text{if } Y \geq 1 \end{cases} \tag{8}$$

Model fitting and selection

In this study, the total number of dependent variable observations was expected to be 936 (12 months \times 3 regions \times 26 years). However, there were some missing fire records, resulting in 834 observations. A random sample of 704 observations (84.4 % of the fire occurrence data) was selected for model development (model

calibration), and the remaining 130 observations (15.6 %) were reserved for independently testing the model’s predictive capability (model validation). Five weather variables were used as the predictor variables, viz. AMWS, AMT, AMP, AMRH, and AME. The statistical software R (R Development Core Team, 2005) was used for data analyses and modeling.

The multicollinearity among the 5 predictor variables was diagnosed by a variance inflation factor (VIF), with $VIF > 10$ as the threshold or red-flag for multicollinearity (O’Brien 2007). In addition, we used a stepwise approach to select significant meteorological factors at the significance level of $\alpha = 0.05$ through the Poisson model. The theory underlying this approach was that nested models can be obtained by restricting a parameter to zero in a more complex model. Because the other five models were all based on the Poisson model, we were able to use this model to select significant meteorological factors.

Model assessment and evaluation

(1) The Akaike information criterion was used to evaluate the goodness of fit of the six models and is defined as follows:

$$AIC = -2 \log L + 2p \tag{9}$$

where $\log L$ is the maximum of the likelihood function for a fitted model and p is the number of parameters in the fitted model. The preferred model is the one with the minimum AIC value (Burnham and Anderson 2004).

(2) While AIC enables comparison of models for goodness-of-fit, it does not reveal anything about how well a model fits the data in an absolute sense (Burnham and Anderson 2004). Thus, we also computed the sum of

squared errors (SSE) to assess the general goodness-of-fit of each model as follows:

$$SSE = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \tag{10}$$

where Y_i is the observed count and \hat{Y}_i is the predicted count from the models.

(3) A likelihood ratio test (LRT) was used to compare nested models (i.e., NB vs. Poisson, ZINB vs. ZIP, and NBH vs. PH) in order to test whether the over-dispersion parameter would be necessary. In LRT, the null hypothesis is for the restricted or constrained model (null model) with the log-likelihood $\log L_R$ and degrees of freedom df_R , and the alternative hypothesis is for the unrestricted or unconstrained model (alternative model) with log-likelihood $\log L_U$ and degrees of freedom df_U . Then, LRT follows a χ^2 distribution such that

$$D = -2 \log \left(\frac{L_R}{L_U} \right) = -2(\log L_R - \log L_U) \sim \chi^2, \text{ with } df = df_U - df_R \tag{11}$$

(4) The Vuong test is a likelihood ratio based test for model selection using the Kullback–Leibler information criterion (Vuong 1989). This statistic makes probabilistic statements about two models. It tests the null hypothesis that two models equally approximate the actual model against the alternative hypothesis that one model more accurately represents the actual model (i.e., is preferred). It cannot make the decision that the “more accurate” model is the true model. Suppose we attempt to test between a model $f(Y|X; \hat{\theta})$ (e.g., ZIP) against a model $g(Y|Z; \hat{\gamma})$ (e.g., Poisson) Under the null hypothesis that these two models are indistinguishable, and the test statistic is asymptotically distributed standard normal, the formula is:

$$V = \frac{1}{\sqrt{n}} \frac{\log Lf(Y|X; \hat{\theta}) - \log Lg(Y|Z; \hat{\gamma})}{\varpi} \sim N(0, 1) \tag{12}$$

where $\varpi^2 = \frac{1}{n} \sum_{i=1}^n \left[\log \frac{f(Y|X; \hat{\theta})}{g(Y|Z; \hat{\gamma})} \right]^2 - \left[\frac{1}{n} \sum_{i=1}^n \log \frac{f(Y|X; \hat{\theta})}{g(Y|Z; \hat{\gamma})} \right]^2$;

If $V > 1.648$, reject the null hypothesis and conclude that $f(Y|X; \hat{\theta})$ is better than $g(Y|Z; \hat{\gamma})$; if $V < -1.648$, reject

the null hypothesis and conclude that $g(Y|Z; \hat{\gamma})$ is better than $f(Y|X; \hat{\theta})$; and if $|V| \leq 1.648$, we cannot reject the null hypothesis and conclude that the two models are the same (Vuong 1989). Thus, the Vuong test can be used to test between pairs of non-nested models (i.e., ZIP vs. Poisson, ZINB vs. NB, PH vs. Poisson, NBH vs. NB, ZIP vs. PH, and ZINB vs. NBH). Using the Vuong test for ZIP versus Poisson and ZINB versus NB pairings also enables testing whether the over-dispersion in count data is attributable to high frequencies of zero counts.

Results

Using the recorded fire occurrence data, the Poisson model was used as a benchmark model for screening the predictor variables, with the result that AMWS was not significant at $\alpha = 0.05$, while four other meteorological factors (AMT, AMP, AMRH, and AME) were statistically significant. In addition, the VIF values of the four predictor factors were all less than 10, indicating that there was no serious multicollinearity among these predictor variables. Thus, we fitted the other five models (i.e., NB, ZIP, ZINB, PH and NBH) using these four meteorological factors. The model fitting results are listed in Tables 2, 3, 4. According to the AIC and SSE of the six models, the zero-inflated models fitted the fire occurrence data better than other models. The ZINB model had the smallest AIC, and the Poisson model had the largest AIC. The rank order of the model AICs was ZINB < NBH < ZIP < NB < PH < Poisson (Tables 2, 3, 4)

The LRT was used to compare nested models (i.e., NB vs. Poisson, ZINB vs. ZIP, and NBH vs. PH) and to test if the over-dispersion parameter in the NB-type models was necessary. All LRT tests were highly significant ($p < 0.01$) for differences in the three pairs (Table 5). It was evident

Table 2 Parameter estimates, standard errors (S.E.), and model goodness of fit statistics for Poisson and negative binomial (NB) models

Predictor	Poisson model		NB model	
	Parameter estimate	S.E.	Parameter estimate	S.E.
Intercept	-1.9960**	0.73991	-0.58104	1.19793
AMP (mm)	-0.01228***	0.00304	-0.01200**	0.00463
AMT (°C)	0.13095***	0.02255	0.14523***	0.03400
AMRH (%)	-0.02222*	0.01087	-0.04577**	0.01727
AME (mm)	0.00973***	0.00198	0.009296**	0.00353
Model goodness of fit statistics				
SSE	437.65	508.18		
AIC	639	538		

AMP is average monthly precipitation; AMT is average monthly temperature; AMRH is average monthly relative humidity; AME is average monthly evaporation; SSE is sum of squares of error; AIC is Akaike information criterion (the smaller, the better)

* denotes $p < 0.05$; ** denotes $p < 0.01$; and *** denotes $p < 0.001$

Table 3 Parameter estimates, standard errors (S.E.) and model goodness of fit statistics for the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models

Predictor	ZIP model				ZINB model			
	Zero-inflated portion		Count portion		Zero-inflated portion		Count portion	
	Parameter estimate	S.E.	Parameter estimate	S.E.	Parameter estimate	S.E.	Parameter estimate	S.E.
Intercept	-9.2782	5.28,745	-5.6118***	1.60074	-3391.17	5023.86	-4.46012**	1.49516
AMP (mm)	-0.00097	0.00701	-0.01408**	0.00449	-0.6231	0.9278	-0.00803	0.00496
AMT (°C)	-0.21864**	0.08344	0.00073	0.03435	-16.703	25.164	0.11247***	0.03320
AMRH (%)	0.17915*	0.07169	0.07324**	0.02226	45.1346	66.893	0.01791	0.02190
AME(mm)	0.00579	0.53292	0.01209***	0.00303	2.1863	3.2645	0.01174**	0.00359
Model goodness of fit statistics								
SSE	411.06				389.30			
AIC	536				487			

AMP is average monthly precipitation; AMT is average monthly temperature; AMRH is average monthly relative humidity; AME is average monthly evaporation; SSE is sum of squares of error; AIC is Akaike information criterion (the smaller, the better)

* denotes $p < 0.05$; ** denotes $p < 0.01$; and *** denotes $p < 0.001$

Table 4 Parameter estimates, standard errors (S.E.) and model goodness of fit statistics for the Poisson hurdle (PH) and negative binomial hurdle (NBH) models

Predictor	PH model				NBH model			
	Hurdle portion		Count portion		Hurdle portion		Count portion	
	Parameter estimate	S.E.	Parameter estimate	S.E.	Parameter estimate	S.E.	Parameter estimate	S.E.
Intercept	-1.18491	1.31917	-2.9002*	1.47209	-1.18491	1.31917	-3.55228	2.31646
AMP (mm)	-0.00785	0.00483	-0.01375**	0.00461	-0.007850	0.00483	-0.01482*	0.00678
AMT (°C)	0.10417**	0.03534	0.03799	0.03859	0.10417**	0.03534	0.05341	0.05831
AMRH (%)	-0.04389*	0.01873	0.03676	0.02285	-0.04389*	0.01873	0.03732	0.03559
AME (mm)	0.01178**	0.00399	0.00746*	0.00314	0.01178**	0.00399	0.00853	0.00500
Model goodness of fit statistics								
SSE	425.90				425.53			
AIC	542				535			

AMP is average monthly precipitation; AMT is average monthly temperature; AMRH is average monthly relative humidity; AME is average monthly evaporation; SSE is sum of squares of error; AIC is Akaike information criterion (the smaller, the better)

* denotes $p < 0.05$; ** denotes $p < 0.01$; and *** denotes $p < 0.001$

Table 5 The likelihood ratio test (LRT) and Vuong test among the six models, Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), Poisson hurdle (PH) and negative binomial hurdle (NBH) models

LRT (χ^2)	NB vs. Poisson 103.4 ($p < 0.0001$)		ZINB vs. ZIP 50.585 ($p < 0.0001$)		NBH vs. PH 9.061 ($p = 0.0026$)	
Vuong	Poisson versus ZIP	NB versus ZINB	ZIP versus PH	ZINB versus NBH	Poisson versus PH	NB versus NBH
	-3.876 ($p = 0.00005$)	-5.519 ($p < 0.0001$)	0.995 ($p = 0.1598$)	4.249 ($p < 0.0001$)	-3.708 ($p < 0.0001$)	-1.781 ($p = 0.037$)

For the Vuong tests, in each pair of model comparison (model 1 vs. model 2), the positive testing statistic indicated that the model 1 was preferred over the model 2, while the negative testing statistic indicated that the model 2 was preferred over the model 1

that the NB-type models (i.e., NB, ZINB, and NBH) were more suitable than Poisson-type models (i.e., Poisson, ZIP, and PH) to handle the over-dispersion of the fire occurrence data in this study (Table 5).

The Vuong test was used to test between the pairs of non-nested models. In this study, we compared ZIP versus Poisson, ZINB versus NB, PH versus Poisson, and NBH versus NB to test if the over-dispersion in the fire occurrence data was attributable to high frequencies of zero counts (excessive zeros). We also compared ZIP versus PH and ZINB versus NBH to investigate if the excessive zeros were due to two sources (structure and sampling) or only one source (structure). The ZIP model was preferred over the Poisson model, and the ZINB model was preferred over the NB model, indicating that the Zero-inflated models were effective to handle the excessive zero counts. Similarly, the PH model was preferred over the Poisson model, and the NBH model was preferred over the NB model, meaning that the Hurdle models were also better than the Poisson and NB models at handling excessive zero counts. There was no difference between ZIP and PH models, indicating both Zero-inflated Poisson and hurdle Poisson handled the excessive zeros equally well, without accounting for over-dispersion. In contrast, the ZINB model was definitely preferred over the NBH model, meaning that when the over-dispersion was accounted for by the NB-type models, the ZINB model was a better choice than the NBH model (Table 5).

Furthermore, the four meteorological factors in the six models showed some differences (Tables 2, 3, 4). For the Poisson and NB models, the estimated parameters of four meteorological factors (AMP, AMT, AMRH, and AME) were all statistically significant ($p < 0.05$) (Table 2). In these two models, AMP (precipitation) and AMRH (relative humidity) were negatively related to lightning-induced fire occurrence, while AMT (temperature) and AME (evaporation) were positively related to fire occurrence. In contrast, the four meteorological factors behaved differently between Zero-inflated and hurdle models, as well as between the two components (i.e., logistic models and count models) of these models. For example, AMP, AMRH and AME were statistically significant ($p < 0.05$) for the count portion of the ZIP model, while only AMT and AMRH were significant (p value < 0.05) for the logistic model (Table 3).

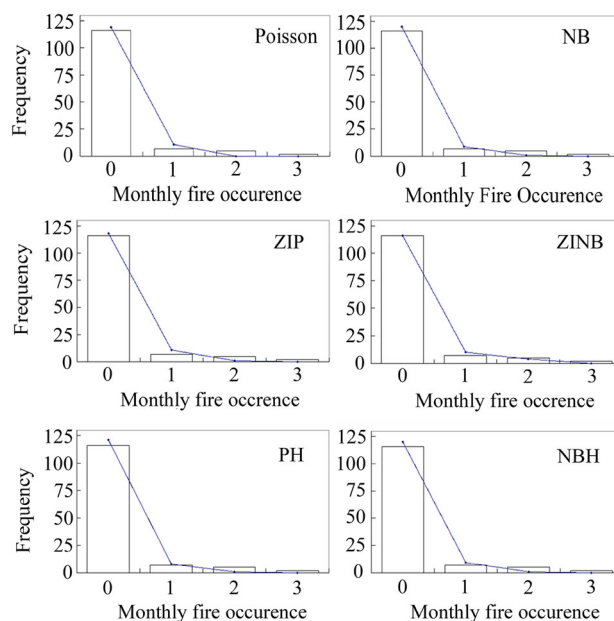


Fig. 3 The observed and predicted frequencies of fire occurrence for Poisson, Negative Binomial (NB), Zero-inflated Poisson (ZIP), Zero-inflated NB (ZINB), Poisson hurdle (PH), NB hurdle (NBH) models using the 130 validation data

In order to assess the predictive capacity of the six models, the independent validation data (130 observations) were used to compare the observed fire counts against the predictions from the six models. The prediction error was defined as the difference between observed count and predicted count. We computed the mean prediction errors (MPE) for predicting the zero counts and for predicting the positive counts (≥ 1) for each of the six models (Table 6). We found that: (1) all models over-predicted ($MPE < 0$) zero counts and the ZINB model was the best (smallest MPE), followed by the Poisson, ZIP, NB, NBH models, and the PH model was the worst (largest MPE); and (2) all models under-predicted positive counts ($MPE > 0$) and the ZINB model was still the best (smallest MPE), followed by the NBH, PH, ZIP, Poisson models, and the NB model was the worst (largest MPE). Figure 3 illustrates the observed frequency of fire occurrence in the 130 validation data points (bar chart) and predicted frequencies of fire occurrence from each of the six models. It was clear that the

Table 6 The mean prediction error (MPE) of the six models using the model validation data (130 observations)

	Models					
	Poisson	NB	ZIP	ZINB	PH	NBH
MPE for zero counts	-0.118	-0.122	-0.119	-0.102	-0.188	-0.182
MPE for count ≥ 1	0.904	0.921	0.758	0.340	0.646	0.593

The mean prediction error is defined as the difference between observed count–predicted count

ZINB model yielded better predictions for both zero counts and positive counts than did the other five models.

Discussion

The fire season of Daxing'an Mountains usually runs from April to October every year, and can be extended or shortened due to the specific meteorological conditions of the year, as well as the specific geographical areas. In contrast, forest fire is rare during winter months (e.g., November to March), resulting in many zero counts when we calculated the number of lightning-induced fires for each month in every annual fire cycle. To avoid dealing with the zero counts during non-fire seasons, some studies limited their fire data to active fire seasons (e.g., Mandallaz and Ye 1997; Martell et al. 1987). In this study, we analyzed fire occurrence over the full year rather than during the fire season only, because the fire seasons in the study area had various lengths from year to year. We anticipated that analysis of fire occurrence data for the entire year would be beneficial to capturing the impacts of meteorological factors on forest fire occurrence.

As described in Methods, the zero-inflated models treat the zero counts from two sources: the structural zeros that cannot score anything other than zero, and the sampling zeros that are a part of the underlying sampling distribution (Poisson or NB). In this study, we considered that all zero records contained some structural zeros due to no lightning strikes during non-fire seasons, and some sampling zeros because zero fire was recorded due to the combined effects of meteorological factors when lightning strikes actually occurred during active fire seasons. The zero-inflated models assume that some lightning strikes may not cause a forest fire due to unfavorable weather conditions, while the Hurdle models presume that each lightning strike would cause a measurable forest fire. Consequently, the zero-inflated models performed better than the Hurdle models in this study.

The ZINB model proved to be the most suitable model for fitting the monthly occurrence of lightning-induced forest fire in the Northern Daxing'an Mountains. However, rather than propose a new approach to forest fire prediction, we attempted to provide comparative assessment and evaluation so that researchers can effectively deal with the challenges of analyzing and modeling count data characterized by over-dispersion and excessive zero counts. Generally speaking, to some extent the data structure decides the model application. Hence, the most suitable model can differ if the data structure of fire occurrence changes. In this study, for example, if we collected fire occurrence data based on a daily scale instead of monthly. As a result, the fire occurrence data were more over-dispersed and zero-inflated. Had we increased the time scale to a yearly basis, the fire occurrence data would likely

be less over-dispersed and have fewer zero counts. In addition, more appropriate explanation variables may also affect the modeling of fire occurrence data. Beside the meteorological factors used in this study, other factors such as topography and fuel types and conditions might also be important. Thus, other useful explanation variables and various time scales should be taken into account when forest fire managers and researchers use our results to predict forest fire occurrence in the future.

Conclusions

Our results showed that, based on the model AIC, the ZINB model best fitted the fire occurrence data, followed by (in an order of declining AIC) NBH, ZIP, NB, PH, and Poisson models. It was possible that excessive zeros impacted model fitting more than over-dispersion, because the improvement of model fitting for Poisson vs. ZIP was more than that for Poisson vs. NB (Tables 2, 3). The ZINB model proved best for fitting the fire occurrence data and for predicting either zero counts or positive counts (≥ 1). The two Hurdle models (PH and NBH) were better than ZIP, Poisson and NB models for predicting positive counts, but worse than these three models for predicting zero counts. The performance of the ZINB model in this study implied that the excessive zero counts arose from both structural and sampling sources, i.e., some lightning strikes occurred, but other environment factors prevented the fire ignition from developing into a measurable forest fire.

Acknowledgments This work was funded by Asia-Pacific Forests Net (APFNET/2010/FPF/001); National Natural Science Foundation of China (Grant No. 31400552).

References

- Affleck DLR (2006) Poisson mixture models for regression analysis of stand-level mortality. *Can J For Res* 36(11):2994–3006
- Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model selection. *Sociol Methods Res* 33:261–304
- Cameron AC, Trivedi PK (1998) *Regression analysis of count data*. Cambridge University Press, Cambridge
- Chang Y, Zhu ZL, Bu RC et al (2013) Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China. *Landsc Ecol* 28(10):1989–2004
- Chou YH, Minnich RA, Chase RA (1993) Mapping probability of fire occurrence in San Jacinto Mountains California, USA. *Environ Manag* 17:129–140
- Erdman D, Jackson L, Sinko A et al (2008) Zero-inflated Poisson and zero-inflated negative binomial models using the countreg procedure. SAS Global Forum, SAS Institute Inc, Cary
- García Diez EL, Rivas SL, de Pablo F et al (1999) Prediction of the daily number of forest fires. *Int J Wildland Fire* 9(3):207–211
- Griffith DA, Haining R (2006) Beyond mule kicks: the Poisson distribution in geographical analysis. *Geogr Anal* 38(2):123–139

- Guo FT, Hu HQ, Jin S, Ma ZH, Zhang Y (2010a) Relationship between forest lightning fire occurrence and weather factors in Daxing'an Mountains based on negative binomial model and zero-inflated negative binomial models. *Chin J Plant Ecol* 34(5):571–577
- Guo FT, Hu HQ, Ma ZH et al (2010b) Applicability of different models in simulating the relationships between forest fire occurrence and weather factors in Daxing'an Mountains. *Chin J Appl Ecol* 21(1):159–164
- Hu M, Pavlicova M, Nunes EV (2011) Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-Risk reduction intervention trial. *Am J Drug Alcohol Abuse* 37(5):367–375
- Krawchuk MA, Cumming SG, Flannigan MD (2009) Predicted changes in fire weather suggest increases in lightning fire initiation and future area burned in the mixed wood boreal forest. *Clim Change* 92:82–97
- Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34:1–14
- Lee JH, Han G, Fulp WJ et al (2011) Analysis of overdispersed count data: application to the human papillomavirus infection in men (HIM) study. *Epidemiol Infect* 30:1–8
- Liu W, Cela J (2008) Count data models in SAS. SAS Institute Inc, SAS Global Forum, Cary, NC
- Mandallaz D, Ye R (1997) Prediction of forest fires with Poisson model. *Can J For Res* 27(10):1685–1694
- Martell DL, Otukol S, Stocks BJ (1987) A logistic model for predicting daily people-caused forest fire occurrence in Ontario. *Can J For Res* 17:394–401
- Miaou SP (1994) The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accid Anal Prev* 26(4):471–482
- Mullahy J (1986) Specification and testing of some modified count data models. *J Econom* 33:341–365
- Numna S (2009) Analysis of extra zero counts using zero-inflated Poisson models. M.Sc. Thesis, Department of Mathematics, Prince of Songkla University, Hat Yai, Songkhla
- O'Brien RM (2007) A caution regarding rules of thumb for variance inflation factors. *Qual Quant* 41:673–690
- Osgood W (2000) Poisson-based regression analysis of aggregate crime rates. *J Quant Criminol* 16(1):21–43
- Podur JJ, Martell DL, Stanford D (2009) A compound Poisson model for the annual area burned by forest fires in the province of Ontario. *Environmetrics* 21(5):457–469
- Poulin-Costello M (1993) People-caused forest fire prediction using Poisson and logistic regression. M.Sc. Thesis, Department of Mathematics and Statistics, University of Victoria, Victoria, B.C
- Preisler HK, Brillinger DR, Burgan RE et al (2004) Probability based models for estimation of wildfire risk. *Intermt Fire Sci Lab* 13:133–142
- Ridout M, Hinde J, Demetrio CG (2001) A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* 57(1):219–223
- Rose CE, Martin SW, Wannemuegler KA et al (2006) On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *J Biopharm Stat* 16(4):463–481
- R Development Core Team, 2005. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, www.R-project.org
- Vega-Garcia C, Woodard PM, Titus SJ et al (1995) A logistic model for predicting the daily occurrence of human caused forest fires. *Intermt Fire Sci Lab* 5:101–111
- Vuong QH (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57(2):307–333
- Xu H (1998) Forests in Daxing'anling mountains. Science Press, Beijing
- Xu Z (2014) Predicting wildfires and measuring their impacts: case studies in British Columbia. Ph.D. Dissertation, University of Victoria
- Yau KKW, Lee AH (2001) Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention program. *Stat Med* 20(19):2907–2920
- Yu C, Hu H, Wei R (2007) Dynamic analysis of meteorological conditions of forest fire in Tahe forestry bureau of Daxing'an mountains. *J Northeast For Univ* 35(8):23–25