林业科学研究 2013,26(5):554～561
Forest Research

文章编号:1001-1498(2013)05-0554-08

长白落叶松林分进界模型的研究

雷渊才[1], 张雄清[1,2]

(1. 中国林业科学研究院资源信息研究所,北京 100091; 2. 中国林业科学研究院林业研究所, 北京 100091)

摘要:利用吉林省汪清林业局金沟岭林场落叶松林分连续观测数据,以计数类模型为基础,分别利用 Poisson 回归模型、负二项模型、零膨胀模型和 Hurdle 模型拟合林木进界株数,并通过 AIC 值,Pearson 残差图以及 Vuong 检验对这些模型进行了详细分析比较。结果表明:Poisson 回归模型不适用于模拟林木枯损株数;负二项回归模型相对于 Poisson 回归模型比较适用,但是对于零枯损过多的数据,这两类模型拟合效果较差;零膨胀模型和 Hurdle 模型对这类数据有很好的解决办法,而且,零膨胀负二项模型拟合效果最好。

关键词:进界;Poisson 模型;负二项模型;零膨胀模型;Hurdle 模型;长白落叶松

中图分类号:S791.22 **文献标识码**:A

Tree Recruitment Model of *Larix olgensis*

LEI Yuan-cai[1], *ZHANG Xiong-qing*[1,2]

(1. Research Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China;
2. Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China)

Abstract: Tree recruitment model play an important role in simulating stand dynamic processes. Considering the fact that in permanent sample plots some of the plots have no occurrences of recruitment even over periods of several years, it means that data are bounded and characteristically exhibit varying degrees of dispersion and skewness in relation to the mean. Additionally, the data often contain an excess number of zero counts. If least squares method is still used to deal with the data with large proportion of zero counts, the estimated results will be biased. Based on the data from permanent plots of *Larix olgensis* in Wangqing Forest Farm, Poisson model, negative binomial model, zero-inflated models and Hurdle models were used to analyze tree recruitment. The best model was chosen according to the AIC value, Pearson residual plot and Vuong test. The results showed that Poisson model was not suitable for recruitment, and negative binomial was superior to the Poisson model. But both of them were not competent for the over-dispersion data. Zero-inflated model and hurdle model were fitted into the data. Additionally, zero-inflated negative binomial model (ZINB) outperformed than other models. The result provided a feasible method for analyzing tree recruitment.

Key words: recruitment; Poisson model; negative binomial model; zero-inflated models; Hurdle models; *Larix olgensis*

进界是指调查间隔期内林木达到一定的测量水平(如:1.3 m 树高,5 cm 胸径),是林分动态变化发展的一个重要过程之一,进界模型与生长模型、枯损模型构成了林分动态变化模型。对于生长模型和枯损模型,研究学者已经做了大量的研究。由于林分进界数据结构离散的特点,人们在分析林分动态变

收稿日期: 2012-10-11
基金项目: "十二五"科技支撑项目(2012BAD22B0201)和"863"科技计划项目(2012AA12A306)
作者简介: 雷渊才,男,研究员,博士生导师,研究方向:森林模型模拟和抽样. E-mail:yclei@caf.ac.cn.

化模型时,往往忽略了进界模型。进界过程在复层林的林分材积研究中发挥了重要的作用,如果不考虑进界过程,就无法正确预测将来的林分生长与收获[1]。许多学者利用线性或者非线性回归方程研究了进界模型[2-3]。随着对进界过程的认识,发现在森林调查间隔期内,可能存在没有林木进界发生的现象,这就意味着进界数据中存在着大量的零数据,此类数据不服从正态分布,若继续采用单个简单的线性或非线性方程,预测会产生较大的偏差;而如果只考虑那些有出现进界的林分,将会高估进界株数[4]。为解决此问题,Ferguson 等[5]利用两阶段法分析 PROGNOSIS 模型中的进界模型。第 1 步先利用 Logistic 模型判断进界的概率,第 2 步在发生进界的基础上利用差分方程估算进界株数。之后,Lexerød[6]利用两阶段法分析了挪威不同树种的进界情况;然而,两阶段法在判断进界与否时,没有一个公认的确定概率阈值的方法。

对于离散数据,可以采用计数类模型分析。Poisson 模型作为计数模型分析的一种基本方法,已经被广泛应用于医药学、灾害等领域[7-8];然而由于 Poisson 回归均值和方差相等的假设条件过于严格,很多数据结构达不到这个要求,一些研究学者提出利用负二项回归模型或者零膨胀模型来分析[9-10]。虽然计数类模型在林业上的应用也有一些:如林分枯损[11-13],林木更新[14-15],但是在进界模型上的应用研究,国内外报道的比较少。Fortin 等[16]在硬木混交林中提出利用零膨胀模型来研究各树种的进界情况,文中用到的零膨胀 Poisson 模型由于 Poisson 模型严格的条件而在拟合计数部分时拟合不够准确。Zhang 等[17]利用零膨胀负二项和 Hurdle 负二项分析了北京油松林分的进界状况。落叶松是东北地区主要三大针叶用材林树种之一,资源十分丰富,蓄积量大,占东北林区针叶材总蓄积量的 40% 以上。据全国第七次森林资源清查结果,经过几十年的培育,落叶松已经成为全国继杉木、杨树和马尾松之后蓄积量最多的树种。研究落叶松林分的进界状况,对分析落叶松林分的动态变化和指导可持续森林经营有着重要的作用。同时,考虑到至今国内对落叶松林分进界模型的研究很少,尤其是计数模型方法在落叶松林分进界的研究未曾有报道。因此,本文以东北汪清长白落叶松(*Larix olgensis* Henry)固定调查样地数据基础,应用 Poisson 模型、负二项回归模型、零膨胀模型和 Hurdle 模型研究分析落叶松林分的进界情况,并以 AIC 和 Vuong 模型评价检验准则选出进界最优预测模型,为定量研究落叶松林分进界提供一种新的研究思路和方法。

1　研究区域概况与数据收集

研究区位于吉林省汪清林业局金沟岭林场,123°56′～131°04′E,43°05′～43°40′N,海拔 550～1 100 m,属长白山系的中低山丘陵区,母岩为玄武岩,土壤为暗棕色森林土。温带大陆性季风气候,降水量 670 mm,年平均气温 1.5 ℃,1 月份平均气温 -18.3 ℃,年积温 2 114 ℃。该区植物为长白山植物区系的一部分,从优势树种看,此地区人工林以长白落叶松林为主。

2　研究方法

2.1　数据整理

长白落叶松人工林于 1962 年造林,共有 20 块固定样地,样地面积为 0.077 5～0.250 0 hm^2。样地的主要调查因子有:胸径、方位角、林分年龄、平均树高、郁闭度、坡向、坡位、坡度、海拔、土层厚度等因子。每隔 2～3 年复测 1 次样地的主要调查因子。本研究采用 1988—2006 年的调查数据。按照区组试验设计的方法,样地共进行了 4 个处理 5 个重复,每个处理包括 3 种间伐强度和 1 个对照,其中,3 种间伐强度为弱度、中度和强度。去掉间伐时期的样本和间隔 3 年的样本,本研究数据组成间隔 2 年的样本 114 个。根据 Affleck[11]的建模分析方法,直接利用落叶松的 114 个样本进行建模分析。长白落叶松林分进界株树频数表及主要变量因子见表 1、2。

表 1　样地进界株数频数表

样地进界株数	频数	百分比/%
0	64	56.14
1	19	16.67
2	9	7.89
3	6	5.26
4	6	5.26
5	3	2.63
7	2	1.75
8	2	1.75
13	1	0.88
32	1	0.88
34	1	0.88

由表1可以看出:未进界的林分超过50%,也就是说数据结构中含有大量的0,数据呈较离散状态。

表2 长白落叶松林分的主要生长指标

变量	最小值	最大值	均值	标准差
年龄/a	24	45	34.08	6.00
优势木平均高/m	15	24.6	20.14	2.09
林分密度/(株·hm^{-2})	380	1 805	1 005.20	279.21
林分断面积/(m^2·hm^{-2})	13.91	37.77	26.71	5.47
林分平方平均直径/cm	13.24	22.34	17.61	2.06

林木进界主要与林分特征变量、优势木平均高(H)有关,其中,林分特征变量包括林分年龄(A)、林分株数密度(N)、林分平均直径(Dg)、林分断面积(B)、林分相对植距($Rs=(10\ 000/N)^{0.5}/H$)[6]。进界过程的发生与立地质量好坏关系很大。林分断面积反映了林木大小和株数密度,是一个常用的林分密度指标,也影响了进界的过程。因此,本论文以上述这些变量为自变量分析落叶松进界情况,而且在变量选择中,首先通过方差膨胀因子(VIF)选择不具有多重共线性的变量。一般认为当 $VIF>5$ 时,就存在较高的共线性。

2.2 Poisson 模型

Poisson 模型是分析计数型数据的一种最简单的方法,其概率质量函数如下[7]:

$$F(y_i)=P(Y_i=y_i)=\frac{e^{-\lambda_i}\lambda^{y_i}}{y_i!} \tag{1}$$

则 Poisson 回归模型为:

$$\lambda_i=\mathrm{Exp}(X_i\beta) \tag{2}$$

式(1)(2)中:Exp( )是以自然对数为底的指数函数,y_i 为随机变量,λ_i 为 Poisson 分布的期望,X_i 为自变量(年龄、密度、优势木平均高、相对植距等),β 为参数向量。

2.3 负二项模型(NB)

负二项分布又称复合 Poisson 分布。在 Poisson 分布中,参数 θ 为一常数,在负二项分布中,参数 θ 是服从 Γ 分布时的随机变量,即负二项分布是当 Poisson 分布强度参数 θ 服从 Γ 分布时所得的复合分布。负二项模型是 Poisson 模型的广义形式[18],不同之处是多了个离散参数 θ,它能够解释数据的异质性,因此,它比 Poisson 分布更具有适用性[19]。在 Poisson 分布中,事件数的方差等于 λ_i,但在负二项分布中,事件数的方差等于 $\lambda_i(1+\theta\lambda_i)$,当 $\theta\to0$ 时,说明事件发生是随机的,此时负二项分布退化为 Poisson 分布;反之,说明事件的发生不独立,因而存在着聚集性。负二项分布中的参数 λ_i 是不定的、变化的,且其变化是有规律的。也就是说,负二项分布个体出现的概率是不相等的,一部分个体出现的概率要大一些,另一部分则要小一些,从而使方差偏大[20]。负二项模型的概率质量函数(PMF)为[11]:

$$F(y_i)=P(Y_i=y_i)=\frac{\Gamma(y_i+\theta^{-1})}{\Gamma(y_i+1)\Gamma(\theta^{-1})}\left(\frac{\theta^{-1}}{\theta^{-1}+\lambda_i}\right)^{\theta^{-1}}\left(\frac{\lambda_i}{\theta^{-1}+\lambda_i}\right)^{y_i} \tag{3}$$

$$\lambda_i=\mathrm{Exp}(X_i\beta+e_i)=\mathrm{Exp}(X_i\beta)\mathrm{Exp}(e_i)$$

式(3)中:Γ 为伽玛函数,θ 为离散参数。

2.4 零膨胀(Zero-inflated)模型

在现实生活中,会有很多的过离散数据,零膨胀模型就是为了拟合零过多数据而发展起来的[21-22],其基本思想是把事件发生数的发生看成2种可能的情形:第1种对应零事件的发生假定服从贝努里分布,第2种对应事件假定服从 Poisson 分布或负二项分布。在零膨胀模型中,零数据有2个主要来源:一是那些从未可能发生的零部分;二是在 Poisson 或负二项理论分布下没有发生的离散部分[13]。实际上,Logit 模型常用来拟合零部分,离散部分可以用 Poisson 模型或负二项模型来模拟。设有一个服从零膨胀分布的离散随机变量 y(样地中林木进界数),p_i 为零部分的概率,它的概率质量函数为[17]:

$$P(Y_i=y_i)=\begin{cases}p_i+(1-p_i)f(0)\\(1-p_i)f(y_i)\end{cases} \tag{4}$$

在公式(4)中,一般认为 $0<p_i<1$,它是对模型的多零部分的解释。在零部分中常用 logit 模型来拟合,即:$\mathrm{logit}(p_i)=\mathrm{Log}\left(\frac{p_i}{1-p_i}\right)=X_i\delta$,Log( )是以自然对数为底的对数函数,$\delta$ 为参数向量。

2.4.1 零膨胀 Poisson 模型(ZIP) 在公式(4)中,如果 y_i 服从一个参数为 λ_i 的 Poisson 分布,就可以得到 ZIP 模型。当 $P_i=0$ 时,ZIP 模型将变成一个普通的 Poisson 模型。ZIP 模型的概率质量函数为[11]:

$$P(Y_i=y_i)=\begin{cases}p_i+(1-p_i)\mathrm{Exp}(-\lambda_i) & y_i=0\\ \dfrac{(1-p_i)\mathrm{Exp}(-\lambda_i)\lambda_i^{y_i}}{y_i!} & y_i>0\end{cases} \tag{5}$$

对于 ZIP 分布,其期望($E(Y_i)$)和方差 $Var(Y_i)$ 分别为:

$$E(Y_i) = (1-p)\lambda_i \tag{6}$$

$$Var(Y_i) = E(Y_i)(1+\lambda_i - E(Y_i)) \tag{7}$$

2.4.2　零膨胀负二项模型(ZINB)　在计数模型中,负二项分布是泊松分布的广义形式。同样,也可以将 ZIP 模型推广到 ZINB 模型,即:在离散部分,用负二项模型来模拟。ZINB 的概率质量函数(PMF)有[12]:

$$P(Y_i = y_i) = \begin{cases} p_i + (1-p_i)\left(\dfrac{\theta^{-1}}{\theta^{-1}+\lambda_i}\right)^{1/\theta} & y_i = 0 \\ (1-p_i)\dfrac{\Gamma(y_i+\theta^{-1})\lambda_i^{y_i}}{\Gamma(y_i+1)\Gamma(\theta)}\left(\dfrac{\theta^{-1}}{\theta^{-1}+\lambda_i}\right)^{1/\theta}\left(\dfrac{\lambda_i}{\theta^{-1}+\lambda_i}\right)^{y_i} & y_i > 0 \end{cases} \tag{8}$$

式(8)中:当 $\theta \to 0$ 时,ZINB 模型就退化为 ZIP 模型。对于 ZINB 分布,其期望和方差分别为:

$$E(Y_i) = (1-p)\lambda_i \tag{9}$$

$$Var(Y_i) = E(Y_i)(1+\lambda_i(1+\theta) - E(Y_i)) \tag{10}$$

从(10)式中可以看出:观测数据中的过离散现象可以通过方差项中 $E(Y_i)\lambda_i\theta$ 来描述。参数 θ 的引入并没有改变 ZINB 模型的均值函数,方差总是大于 ZIP 模型的方差。当在负二项部分引入解释变量后,可以得到 ZINB 回归模型。

2.5　Hurdle 模型

Hurdle 模型最早由 Mullahy[23] 提出,Hurdle 模型又叫两部分(Two-part)模型[24]:第 1 部分模拟零个数,如二分类模型(如 logit 模型);第 2 部分是模拟正数计数,如 Poisson 模型、负二项模型等。Hurdle 模型跟零膨胀模型相似,都可以看作是 2 个统计过程的混合,但是 Hurdle 模型与零膨胀模型的区别是 Hurdle 模型假设零数据来源于 1 个统计过程,而零膨胀模型有 2 个来源[9]。Hurdle 模型的概率质量函数(PMF)为[17]:

$$P(Y_i = y_i) = \begin{cases} p & y_i = 0 \\ (1-p)\dfrac{f(y_i)}{1-f(0)} & y_i > 0 \end{cases} \tag{11}$$

式(11)所对应的均值和方差为:

$$E(Y_i) = \frac{1-p}{1-f(0)}\lambda_i \tag{12}$$

$$Var(Y_i) = P(Y_i>0)Var(Y_i \mid Y_i>0) + P(Y_i=0)E(Y_i \mid Y_i>0) \tag{13}$$

在本研究中,第 1 部分(Hurdle 部分)利用常用的 logit 模型(与零膨胀模型零部分一样),第 2 部分分别利用 Poisson 模型和负二项模型进行比较研究,记为 Hurdle-Poisson 模型和 Hurdle-NB 模型。

2.5.1　Hurdle-Poisson 模型(HP)　Hurdle-Poisson 模型是 Hurdle 模型中较为常用的一种,其概率质量函数(PMF)为:

$$P(Y_i = y_i) = \begin{cases} p & y_i = 0 \\ \dfrac{(1-p)\mathrm{Exp}(-\lambda_i)\lambda_i^{y_i}}{(1-\mathrm{Exp}(-\lambda_i))y_i!} & y_i > 0 \end{cases} \tag{14}$$

2.5.2　Hurdle-NB 模型(HNB)　Hurdle-NB 模型是 Hurdle 模型的另一种形式,在截尾非零计数部分用负二项模型来拟合,其概率质量函数(PMF)为[17]:

$$P(Y_i = y_i) = \begin{cases} p & y_i = 0 \\ (1-p)\dfrac{\Gamma(y_i+\theta^{-1})}{\Gamma(\theta^{-1})\Gamma(y_i+1)}\left[(1+\theta\lambda_i)^{\theta^{-1}} - 1\right]^{-1}\left(\dfrac{\lambda_i}{\lambda_i+\theta^{-1}}\right)^{y_i} & y_i > 0 \end{cases} \tag{15}$$

对上述几类模型,利用 R 软件中 pscl 软件包通过极大似然估计方法估计参数[25]。

3　模型评价

为了比较 Poisson 模型、负二项模型、零膨胀模型和 Hurdle 模型的拟合情况,用 -2 似然对数(LogL)和 AIC 值统计量进行比较。LogL 和 AIC 值越小,说明模型越好。AIC 的公式为:

$$\mathrm{AIC} = \mathrm{LogL} + 2j \tag{16}$$

式(16)中:j 为模型中参数个数。

Vuong 检验在复合模型(ZIP 、ZINB)和一般计数模型(Poisson/NB 模型)的比较分析中有相当高的检验效能[26]。因此,本文也选用了 Vuong 检验择优方法。该方法在 2 个模型解释能力相等的假设下

得到Z统计量并进行似然比检验,根据Vuong的检验值和p值判断选优,极大地提高了模型择优的效果[27]。如果Vuong检验值大于1.96,则说明在显著水平0.05条件下模型A优于模型B;如果Vuong检验值小于-1.96,则相反。

对于模型的拟合优度检验,利用残差值进行诊断。常用的残差值为:$r = y - \hat{y}$。在传统的线性模型中,残差服从同方差正态分布$(y - \hat{y}) \sim N[0, \sigma^2]$。对于计数类的数据结构,残差具有异质性,而且是不对称的。常用的描述异质性残差为Pearson残差[28]:

$$pr = \frac{y_i - \hat{y}_i}{\sqrt{\hat{v}_i}} \tag{17}$$

式(17)中:$\hat{y}_i$为第i个样本的估计值,$\hat{v}_i$为方差。为了更详细的比较几类模型估计的准确性,一般通过χ^2检验完成。

4 结果与分析

由表1可以发现:进界株数结构离散,而且零数据比较多。根据方差膨胀因子(*VIF*)检验,发现林分年龄、林分优势高、相对植距以及林分株数密度、林分平均直径、林分胸高断面积中任意2个变量均不存在多重共线性(*VIF*<5)。

由表3、4可知:Poisson模型、负二项模型、零膨胀模型和Hurdle模型的各参数估计均在0.05水平上显著。在离散部分,林木进界与林龄、林分断面积、相对植距和平均直径显著相关。林龄、林分断面积和相对植距变量的参数估计为负值,说明随着林龄、林分断面积和相对植距的增加,进界株数减少。随着林龄的增加进界株数减少,这是由于本研究中的长白落叶松林处于中龄林时期,随着林龄增加使得林分更趋于稳定,进界株数减少。相对植距是株数密度和林分优势高的反函数,进界株数随相对植距的增加(林分优势高减小)而减少。这意味着进界株数随着林分优势高的增加而增加,林分优势高是反映立地质量的一个重要指标,Vanclay[29]发现,在立地质量好的林地更容易出现进界过程。林分断面积的增加,会导致进界株数的增加,这与前人的研究结果一致[6, 30]。在零部分,各参数估计也都在0.05水平上显著。

表3 Poisson模型和负二项模型的参数估计及评价统计量

参数	Poisson模型		NB模型	
	估计值	*p*值	估计值	*p*值
截距	9.27±1.34	<0.05	7.67±2.869 8	<0.05
A	-0.15±0.024 6	<0.05	-0.222 3±0.059 5	<0.05
B	-0.15±0.028 3	<0.05	-0.12±0.059 6	<0.05
Dg	0.40±0.063 2	<0.05	0.54±0.160 8	<0.05
Rs	-46.31±6.677 5	<0.05	-42.69±13.268 8	<0.05
LogL	633.165 6		353.760 9	
AIC	643.165 6		365.760 9	

注:*A*-林龄;*B*-林分断面积;*Dg*-林分平均直径;*Rs*-林分相对植距;logL--2似然对数;下同。

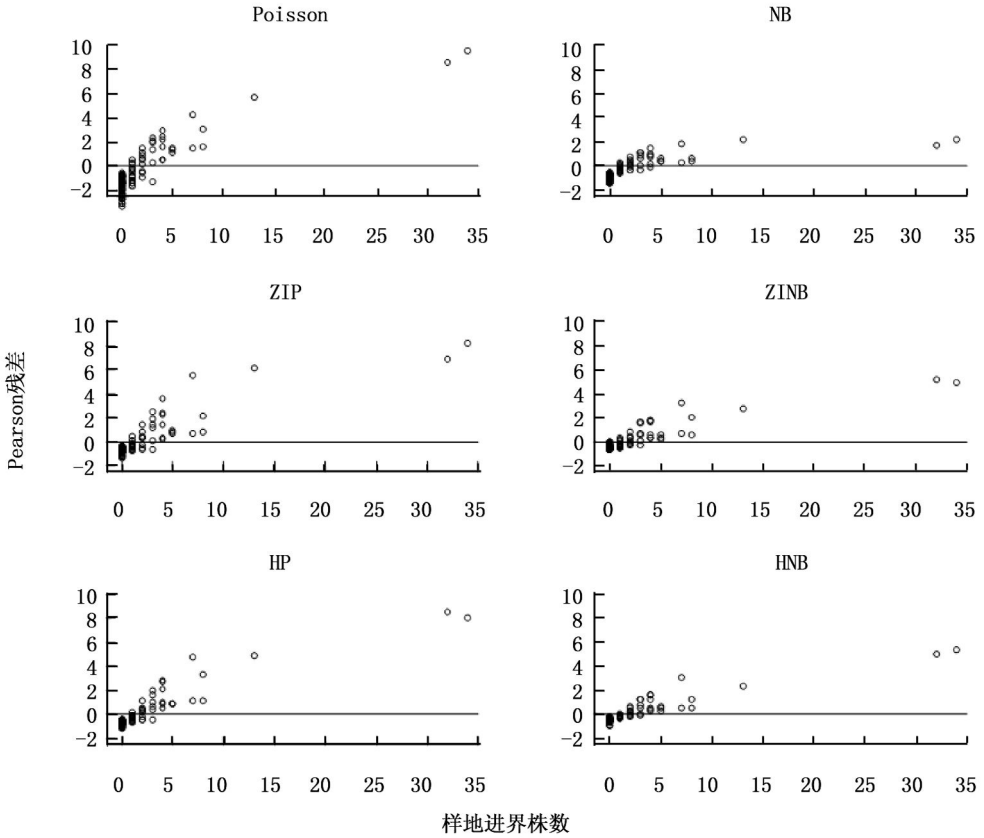
表4 零膨胀模型和Hurdle模型参数估计及评价统计量

参数	ZIP模型		ZINB模型		Hurdle-Poisson模型		Hurdle-NB模型	
	估计值	*p*值	估计值	*p*值	估计值	*p*值	估计值	*p*值
离散部分								
截距	9.47±1.591 1	<0.05	10.19±2.877 9	<0.05	10.17±1.632 7	<0.05	7.383 4±1.182 7	<0.05
A	-0.11±0.047 8	<0.05	-	-	-	-	-	-
B	-0.158 8±0.041 1	<0.05	-0.20±0.068 5	<0.05	-0.20±0.037 1	<0.05	-0.22±0.112 4	<0.05
Dg	0.36±0.087 8	<0.05	0.18±0.105 6	<0.05	0.16±0.046 6	<0.05	0.27±0.100 6	<0.05
Rs	-44.12±7.196 5	<0.05	-47.50±13.568 9	<0.05	-42.51±7.807 4	<0.05	-55.78±24.621 9	<0.05
零部分								
A	0.14±0.066 2	<0.05	1.11±0.535 3	<0.05	0.15±0.058 2	<0.05	0.15±0.058 1	<0.05
Dg	-0.25±0.118 0	<0.05	-2.23±1.082 5	<0.05	-0.25±0.104 2	<0.05	-0.26±0.104 2	<0.05
LogL	498.403 4		346.379 2		509.746 9		356.524 1	
AIC	512.403 4		360.379 2		521.746 9		368.524 1	

由表3、4可知:ZINB模型的LogL值(346.379 2)、AIC值(360.379 2)依次小于HNB模型(LogL=356.524 1,AIC=368.524 1)、NB模型(LogL=353.760 9,AIC=365.760 9)、ZIP模型(LogL=498.403 4,AIC=512.403 4)、HP模型(LogL=509.746 9,AIC=521.746 9)、Poisson模型(LogL=633.165 6,AIC=643.165 6),发现ZINB模型相对优于其它5个模型。

另外,根据 Vuong 检验,比较 HNB 模型和 HP 模型,其 Vuong 检验值为 2.164 1,p 值 < 0.05;比较 ZINB 和 ZIP 模型,其 Vuong 检验值为 2.306 2,p 值 <0.05。这说明 ZINB 模型或者 HNB 模型比 ZIP 模型或 HP 模型好。同样,也比较了 ZINB 模型和 NB 模型,发现 Vuong 检验值为 1.591 8,p 值为 0.055 7,说明 ZINB 模型在 0.1 水平上优于 NB 模型。对比 ZINB 模型和 HNB 模型,其检验值为 1.126 0,p 值为 0.130 0。这说明 ZINB 模型和 HNB 模型在 0.05 水平上差异不显著,但是从 LogL 值、AIC 值比较,ZINB 模型略优于 HNB 模型。

由图 1 可以发现:6 个模型都在进界数少的地方高估,而在进界数多的地方低估。Poison 模型、ZIP 模型和 HP 模型的残差波动范围比较大,拟合效果相对较差,而 NB 模型、ZINB 模型和 HNB 模型的波动范围则比较小,大部分落在 -2 到 2 之间。由此可以发现,NB、ZINB、HNB 这 3 个模型比较稳定。通过 χ^2 检验,发现 NB 模型、HNB 模型、ZINB 模型的 p 值均大于 0.05,而以 Poisson 模型为基础的这 3 种模型却小于 0.000 1。说明这 3 种模型的拟合分布与实际分布在 0.05 水平上差异显著。ZIP 模型和 HP 模型很好的反应了零数据部分,但是对于非零部分则拟合效果较差(图 2)。这是由于 Poisson 模型中期望与方差相等的条件比较苛刻,对于过离散的数据不能满足该条件;而对于负二项模型,离散参数的存在,使得负二项模型能够解释数据的异质性[19]。因此,以负二项模型为基础的 3 个模型(NB、ZINB、HNB)比 Poisson 模型为基础的 3 个模型(Poisson、ZIP、HP)更具有适用性。



Poisson:泊松;NB:负二项模型;ZIP:零膨胀泊松;ZINB:零膨胀负二项模型;HP:Hurdle - 泊松模型;HNB:Hurdle - 负二项模型

图 1　6 个进界模型的 Pearson 残差图

零膨胀模型对于处理离散数据有独特的优势,实际上它由 2 个模型组成:一是判断是否发生枯损的模型,二是模拟计数个数的模型。另外,相对 Poisson 模型和负二项模型,零膨胀模型的另一个优势是在模拟非零数据的同时对零数据进行研究[31]。零膨胀负二项模型优于 Poisson 模型、负二项模型和零膨胀 Poisson 模型。当数据由于零数据相对较多而离散时,零膨胀 Poisson 模型能够较为精确地反映

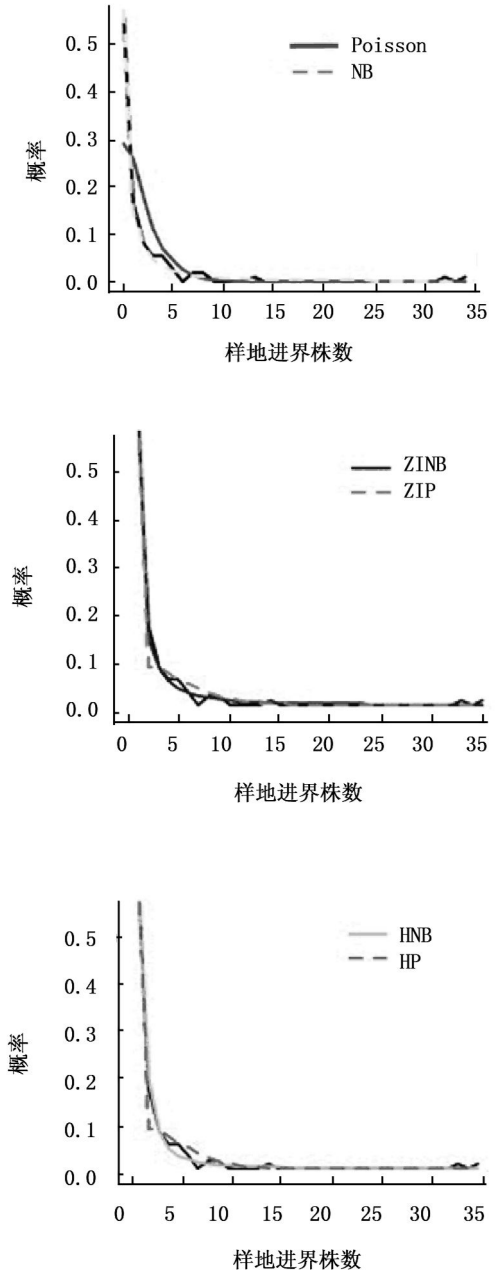


图2 进界实际概率分布和预测概率分布的比较

出数据的规律;然而当零数据膨胀时,零膨胀负二项模型模拟效果比零膨胀 Poisson 模型的好。零膨胀 Poisson 模型中,Poisson 部分只有 1 个变量——期望,当该分布的方差大于期望时,零膨胀 Poisson 分布不能够充分地拟合进界数据。

Hurdle 模型是处理离散数据的另外一种方法。对于 Hurdle 模型和零膨胀模型,两类模型差异很小。Hurdles 模型的零数据只来源于 Logit 部分,而零膨胀模型的零部分即来源于 Logit 部分,也来源于截尾的 Poisson 部分或者 NB 部分。当选择这类模型时,要根据零数据的生态含义选择[32]。比如,当多零数据主要是由实验设计、测量误差或者其它原因引起的,称为"伪"零数据[33],Hurdle 模型会是一个不错的选择;然而当多零数据是由"伪"零数据和确实存在的零数据组合而成,那么可以选择零膨胀模型。当然,HNB 模型和 ZINB 模型预测精度相当,要根据实际的数据情况来选出最优模型。

综合以上分析, ZINB 模型、HNB 模型和 NB 模型精度远优于 ZIP 模型和 HP 模型,并且 ZINB 模型和 HNB 模型精度相当,略优于 NB 模型;然而,知道进界出现的多零数据并不是唯一的进界数据结构。只有在分析罕见树种或在分析混交林林分中各树种的进界株数、小样地进界株数和在短间隔期分析慢生树种才会出现多零数据。随着间隔期的增加,零数据发生的概率逐渐减少[34]。进界是林分动态变化发展的一个重要过程之一,在下阶段的研究中可以结合进界的株树和枯损的株树估计林分的株树密度,为林分株树密度的研究提供一种新的研究思路,也能够更详细的说明株树变化的动态过程。

5 结论

本文以长白落叶松林为对象,研究分析林分进界模型的构建方法。当数据过离散时,负二项模型是个很好的选择。这是因为负二项模型的方差大于期望,能够更好地解释数据的异质性[35]。在本研究中,进界数据存在着多零现象。对于此类数据,零膨胀负二项模型和 Hurdle 负二项模型具有更好的适用性。通过各种评价指标的对比发现,零膨胀负二项模型拟合效果比其他几个模型都好,略好于 Hurdle 负二项模型。因此,利用 HNB 或者 ZINB 模型分析长白落叶松林分进界都具有一定的可行性。当然,进界是个相对比较复杂的过程,影响因素比较多,这些因素之间可能存在着相互作用,很难比较准确描述其它因素导致的进界发生;因此,模型中引入随机效应分析长白落叶松林分的进界问题有待于进一步研究。

参考文献:

[1] Andreassen K. Development and yield in selection forest [J]. Meddelelser fra Skogforsk, 1994, 47 (5): 1 - 37

[2] Adams D M, Ek A R. Optimizing the management of uneven-aged forest stands[J]. Canadian Journal of Forest Research, 1974, 4: 274 - 287

[3] Shifley S R, Ek A R, Burk T E. A generalized methodology for esti-

mating forest ingrowth at multiple threshold diameters[J]. Forest Science, 1993, 39(4): 776 - 798

[4] Li R, Weiskittel A R, Kersha J A J. Modeling annualized occurrence frequency, and composition of ingrowth using mixed-effects zero-inflated models and permanent plots in the Acadian Forest Region of North America. Canadian Journal of Forest Research, 2006, 41: 2077 - 2089

[5] Ferguson D E, Stage A R, Boyd R J. Predicting regeneration in the grand fir-cedar-hemlock ecosystem of the northern rocky mountains [J]. Forest Science, Supplement 26, 1986, 32 (1): a0001 - z0001

[6] Lexerød N L. Recruitment models for different tree species in Norway [J]. Forest Ecology and Management, 2005, 206: 91 - 108

[7] Perdeck A C. Poisson regression as a flexible alternative in the analysis of ring-recovery data[J]. Eyring Newsletter, 1998, 2: 30 - 36

[8] Jesper R, Igor R. A note on estimation of intensities of fire ignitions with incomplete data[J]. Fire Safety Journal, 2006, 41 (5): 399 - 405

[9] Liu W, Cela J. Count data models in SAS[R]. Statistics and Data Analysis in SAS Global Forum, 2008

[10] 许　飞. 负二项回归模型在过离散型索赔次数中的应用研究[J]. 统计教育, 2009, 115(4): 53 - 55

[11] Affleck D L R. Poisson mixture models for regression analysis of stand-level mortality[J]. Canadian Journal of Forest Research, 2006, 36(11): 2994 - 3006

[12] 张雄清, 雷渊才, 雷相东, 等. 基于计数模型方法的林分枯损研究[J]. 林业科学, 2012, 48(8): 54 - 61

[13] Eskelson B N I, Temesgen H, Barrett T M. Estimating cavity tree and snag abundance using negative binomial regression models and nearest neighbor imputation methods[J]. Canadian Journal of Forest Research, 2009, 39(9): 1749 - 1765

[14] Rathbun S L, Fei S. A spatial zero-inflated Poisson regression model for oak regeneration[J]. Environmental and Ecological Statistics, 2006 (13): 409 - 426

[15] Keefe R F. Two-stage and zero-inflated modelling of forest regeneration on the Pacific Northwest Coast[R]. Moscow: Univ of Idaho, 2004: 1 - 79

[16] Fortin M, DeBlois J. Modeling tree recruitment with zero-inflated models: the example of hardwood stands in southern Québec, Canada[J]. Forest Science, 2007, 53 (4): 529 - 539

[17] Zhang X, Lei Y, Cai D, *et al*. Predicting tree recruitment with negative binomial mixture models[J]. Forest Ecology and Management, 2012, 270: 209 - 215

[18] Evans M, Hastings N, Peacock B. Statistical Distributions[M]. New York, USA: John Wiley, 2000: 1 - 221

[19] MacNeil M A, Carlson J K, Beerkircher L R. Shark depredation rates in pelagic longline fisheries: a case study from the Northwest Atlantic[J]. ICES Journal of Marine Science, 2009, 66(4): 708 - 719

[20] 陈　平, 程晓明. 住院次数的负二项分布模型[J]. 卫生经济研究, 1998, 12 (12): 23 - 25

[21] Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing[J]. Technometrics, 1992, 34 (1): 1 - 14

[22] Welsh A H, Cunningham R B, Donnelly C F, *et al*. Modelling the abundance of rare species: statistical models for counts with extra zeros[J]. Ecological Modeling, 1996, 88(10): 297 - 308

[23] Mullahy J. Specification and testing of some modified count data models[J]. Journal of Econometrics, 1986, 33(3): 341 - 365

[24] Heilbron, D. Zero-altered and other regression models for count data with added zeros[J]. Biometrical Journal, 1994, 36(5): 531 - 547

[25] Zeileis A, Kleiber C, Jackman S. Regression models for count data in R[J]. Journal of statistical software, 2008, 27(8): 1 - 25

[26] 郭福涛. 基于负二项和零膨胀负二项回归模型的大兴安岭地区雷击火与气象因素的关系[J]. 植物生态学报, 2010, 34 (5): 571 - 577

[27] Vuong Q H. Likelihood ratio tests for model selection and non-nested hypotheses[J]. Econonmetrica, 1989, 57 (2): 307 - 333

[28] Cameron A C, Trivedi P K. Regression analysis of count data[M]. Cambridge: Cambridge University Press, 1998: 140 - 143

[29] Vanclay J K. A growth model for north Queensland rainforest[J]. Forest Ecology and Management, 1989, 27 (2): 245 - 271

[30] Vanclay J K. Modelling regeneration and recruitment in a tropical rain forest[J]. Canadian Journal of Forest Research, 1992, 22 (9): 1235 - 1248

[31] Karazsia B T, van Dulmen M H. Regression models for count data: illustrations using longitudinal predictors of childhood injury[J]. Journal of Pediatric Psychology, 2008, 33(10): 1076 - 1084

[32] Zuur A F, Ieno E N, Walker N J, *et al*. Mixed effects models and extensions in ecology with R [M]. New York: Springer, 2009

[33] Martin T G, Wintle B A, Rhodes J R, *et al*. Zero tolerance ecology: improving ecological inference by modeling the source of zero observation[J]. Ecology Letters, 2005, 8 (11): 1235 - 1246

[34] Yao X, Titus S J, Macdonald S E. A generalized logistic model of individual tree mortality for aspen, white spruce, and lodegepole pine in Alberta mixedwood forests[J]. Canadian Journal of Forest Research, 2001, 31 (2): 283 - 291

[35] Yaacob W F W, Lazim M A, Wah Y B. A practical approach in modeling count data[R]. Proceedings of Regional Conference on Statistical Sciences, 2010: 176 - 183