

小班马尾松毛虫虫灾发生类型的辨识模型研究^{*}

黄英姿¹, 古德祥¹, 张文庆¹, 梁承丰², 陈占芳³

(1. 中山大学生物防治国家重点实验室, 广东 广州 510275;

2. 广东省林业局森林病虫害防治站, 广东 广州 510173;

3. 广东省德庆县林业局, 广东 德庆 526000)

摘要: 应用判别函数分析及线性 Logistic 回归分析两种方法, 对影响小班马尾松毛虫虫灾发生类型的各种因素的重要性进行分析, 并筛选出相对重要的因子建立起小班发生类型的 4 个辨识模型。

关键词: 马尾松毛虫; 小班; 发生类型; 判别函数分析; 线性 Logistic 回归分析; 辨识模型

中图分类号: Q968.1 **文献标识码:** A **文章编号:** 0529-6579 (2001) 02-0093-05

森林害虫防治比起农作物害虫防治有着更大的困难。这是由于森林分布区内地形、地势复杂, 使得在不同的分布区内, 甚至同一分布区不同的小区域内的虫情都不一样。为了便于管理, 人们把管理的单位一般落实到小班 (subcompartment)。并视成灾的频率在管理上将小班划为安全类、偶灾类及常灾类。然而要确定一个小班所属的类型 (安全类、偶灾类还是常灾类), 并不是一件易事, 这需要对该小班作多年连续的观测, 所需耗费的人力物力是可想而知的。然而实际防治工作又迫切要知道每个小班的发生类型, 以便做到防治有的放矢, 事半功倍。

判别函数分析是研究个体归属问题的一个常用统计方法, 它要求识别变量是定量变量或数值型变量, 且服从有相等方差的正态分布。在小班二类调查资料中, 含有大量的定性指标, 比如小班的地形、坡位等, 且不一定满足正态性的条件。虽然也可以通过赋值的方式将定性指标转成定量指标参与判别函数分析, 但赋值的大小及结构都会对模型产生影响, 增加了建模的人为干扰。如果不选用定性指标, 无疑会丢掉大量信息。线性 Logistic 回归分析通过确定个体归属某类别的概率的大小对个体进行分类。模型对识别变量的类型及分布均未作任何要求。本文分别应用判别函数分析及线性 Logistic 回归分析的方法, 对影响小班马尾松毛虫灾发生类型的因素的重要性进行分析, 并筛选出重要因子建立各种小班发生类型的辨识模型, 为提高森林虫害管理水平和效率奠定必要的基础。

1 数据来源及分析方法

1.1 数据来源

分析数据来自广东省德庆县境内 10 973 个小班的二类调查资料, 其中包括安全类小班 5 508 个, 偶灾类小班 2 616 个, 常灾类小班 2849 个。每个小班的特征通过小班的优势树种、公顷株数、林地起源、郁闭度、平均树龄、平均树高、平均胸径、地类、地形、坡位、坡向、坡度及海拔 13 个定性定量指标所描述。

从 10 973 个小班中随机选取 5 447 个小班 (约占总小班数的 50%, 其中安全类小班 2 720 个, 偶灾类小班 1 288 个, 常灾类小班 1 439 个) 的数据用于建模, 剩余 5 526 个小班 (其中安全类 2 788 个, 偶灾类小班 1 328 个, 常灾类小班 1 410 个) 的数据留作预测, 以检验所建模型的可信度。

1.2 线性判别分析的基本思想及应用方法

设有来自属于 N 个类别的 n 个个体, 每个个体均可以由 m 个数值变量 $x_1, x_2, \dots, x_m$ 所描述。线性判别分析的基本思想就是根据处于这 N 个类别中的 n 个个体的 m 个变量的取值特征, 总结客观分类的规律性, 拟合出如下称为判别函数的方程^[1~4]:

$$y_1 = b_{10} + b_{11}x_1 + b_{12}x_2 + \dots + b_{1m}x_m$$

$$y_2 = b_{20} + b_{21}x_1 + b_{22}x_2 + \dots + b_{2m}x_m$$

⋮

$$y_N = b_{N0} + b_{N1}x_1 + b_{N2}x_2 + \dots + b_{Nm}x_m$$

^{*} 收稿日期: 2000-05-17; 作者简介: 黄英姿 (1959—), 女, 博士, 副教授; 通讯联系人: 古德祥。

这里, y_k 给出了一个个体与第 k 个类别的接近程度. 对于一个未知类别的个体, 只要将其变量值 $x_1, x_2, \cdots, x_m$ 代入上述方程, 就可求出 N 个 y_k 值. 如果 $y_{k_0} = \max_{1 \leq k \leq N} \{y_k\}$, 则表明该个体与第 k_0 类的接近程度最高.

1.3 Logistic 回归分析的基本思想及应用方法

设有来自属于 N 个类别的 n 个个体, 每个个体均可以由 m 个变量 $x_1, x_2, \cdots, x_m$ 所描述, 其中, $x_i (i = 1, 2, \cdots, m)$ 既可以是定性指标, 也可以是定量指标. 令 $p^{(k)}$, ($k = 1, 2, \cdots, N$) 表示取值为 $x_1, x_2, \cdots, x_m$ 的个体属于第 k 个类别的概率, 线性 Logistic 回归分析就是要拟合如下方程^[2, 4, 5]:

$$\ln \frac{p^{(k)}}{1 - p^{(k)}} = \beta_0 + \sum_{i=1}^m \beta_i x_i$$

式中, β_0 为截距, β_i 为待估系数 ($i = 1, 2, \cdots, m$). 一旦模型确定之后, 对每一个体, 就可求出相应的 $p^{(k)} (k = 1, 2, \cdots, N)$ 值. 若 $p^{(k_0)} = \max_{1 \leq k \leq N} \{p^{(k)}\}$, 则个体属于第 k_0 类别的可能性最大. 在实际应用中, 往往只求出其中的 $N-1$ 个 $p^{(k)}$ 值, 剩下的一个为 1 减去这些 $p^{(k)}$ 值之和的差值.

2 结果及分析

2.1 判别分析的结果及分析

2.1.1 只有定量指标参与分析的结果 选择公顷株数 (株/hm²)、郁闭度、平均树龄、平均树高、平均胸径、坡度及海拔高度 7 个定量指标进行逐步回归分析. 类别均值向量的检验采用 Wilks 的 λ 统计量, 按引入变量的 F 概率值取 0.01, 剔出变量的 F 概率值取 0.05 的标准对变量进行逐步筛选, 先验概率根据实际观测值所确定. 经过 14 次反复引入与剔除的过程, 最后筛选出海拔高度、坡度及平均树高 3 个指标作为识别变量构建判别函数模型, 模型系数及相关检验值列于表 1.

模型检验的 χ^2 值为 2 438.748, 自由度为 6, 而 $\chi^2_{0.005}(6) = 18.5$, 因此有 $\chi^2 > \chi^2_{0.005}(6)$. 说明平均树高、坡度及海拔高度 3 个变量具有区分小班灾情发生类型的能力. 由这 3 个变量参与构建的判别模型对安全类小班的回判及预测正确率均为 80.8%, 对偶灾类小班的回判及预测正确率分别为 22% 及 20.3%, 对常灾类小班的回判及预测正确率分别为 69.2% 及 71.9%, 总的回判

及预测正确率分别为 63.8% 及 64%. 总而言之, 模型对安全类及常灾类小班的识别效果都较好, 但对偶灾类小班的识别效果却不太理想.

表 1 由平均树高、坡度及海拔高度参与建立的判别函数的系数 (广东, 德庆)

Tah 1 Coefficients of discrimination function made up average height of tree, slope and elevation (Guangdong, Deqing)

变量	安全类	偶灾类	常灾类
平均树高	0.959	0.898	0.855
坡度	0.570	0.531	0.468
海拔高度	6.621×10^{-3}	-1.795×10^{-3}	-3.311×10^{-3}
常数	-15.684	-12.893	-10.554

2.1.2 有量化的定性指标参与分析的结果 将 6 个定性指标进行量化处理. 选择量化后的 6 个定性指标连同前面 7 个定量指标共 13 个指标施行逐步判别分析, 变量检验统计量及引入剔除标准也如 2.1.1 中所述.

经过 24 步引入、剔除的交替过程, 先后选取海拔高度、地形、坡度、平均树高、优势树种、地类 6 个指标作为识别变量参与建模. 模型系数由表 2 列出.

表 2 由优势树种、平均树高、地类、地形、坡度及海拔高度参与建立的判别函数的系数

Tah 2 Coefficients of discrimination function made up dominant species, average height of tree, topography, type, slope and elevation

变量	安全类	偶灾类	常灾类
优势树种	0.694	0.706	0.728
平均树高	-0.453	-0.530	-0.608
地类	67.160	67.408	68.129
地形	7.693	6.693	5.988
坡度	0.473	0.443	0.385
海拔高度	-2.054×10^{-2}	-2.663×10^{-2}	-2.661×10^{-2}
截距	-122.977	-118.663	-117.112

模型检验的 χ^2 值为 3 060.26, 自由度为 12, 而 $\chi^2_{0.005}(12) = 28.3$, 因此有 $\chi^2 > \chi^2_{0.005}(12)$. 说明所建立的模型区分 3 种小班虫灾发生类型的效果是显著的. 所建模型对安全类小班的回判及预测正确率分别为 83.2% 及 83.7%, 对偶灾类小班的回判及预测正确率分别为 18.6% 及 15.9%, 对常灾类小班的回判及预测正确率分别为

69.3%及 68.2%，总的回判及预测正确率分别为 64.2%及 63.4%。由此可见，加入一些定量化的定性指标之后，所建模型对安全类的判别效果有所提高，但对偶灾类小班的判别效果却有所下降。进一步的分析结果表明，将定性指标的赋值方式加以调整之后，不仅模型参数会随之改变，而且模型的判别结果也会改变。

2.2 线性 Logistic 回归分析的结果及分析

2.2.1 将全部指标作为定量变量 将描述小班

特征 6 个定性指标量化连同 7 个定量指标作为变量参与线性 Logistic 回归分析。选用主效应模型，不考虑变量之间的交互作用。根据各变量的似然比检验结果，选取显著性水平低于 0.01 的变量优势树种、平均树高、地类、地形、坡度及海拔高度 6 个变量参与构建线性 Logistic 回归模型。模型的回归系数及有关检验值列入表 3。

表 3 由优势树种、平均树高、地类、地形、坡度及海拔高度参与建立的 Logistic 回归模型的系数及有关检验结果（广东，德庆）

Tah 3 Parameters of Logistic regression model made up dominant species, average height of tree, type, topography, slope and elevation (Guangdong, Deqing)

变量	安全类				常灾类			
	系数	SE	Wald 检验值	显著水平	系数	SE	Wald 检验值	显著水平
截距	− 4 127	0 552	55.965	0.000	2 284	0 614	13.821	0.000
优势种	0 521	0 080	42.636	0.000	− 0 020 83	0 091	0 053	0 818
平均高度	0 036 34	0 014	6 966	0.008	− 0 012 95	0 015	0 700	0 403
地类	− 0 259	0 164	2 486	0.115	0 498	0 187	7.058	0.008
地形	0 641	0 070	84.740	0.000	− 0 377	0 069	29.473	0.000
坡度	2 6 79× 10 ^{−2}	0 006	21.016	0.000	− 0 053 07	0 006	78.204	0.000
海拔高度	7.175× 10 ^{−3}	0 001	159.11	0.000	− 9.78× 10 ^{−3}	0 001	103.52	0.000

模型显著性检验的 χ^2 值为 3 210.355，远远大于自由度为 12 的 χ^2 临界值，说明模型效果是显著的。所建模型对安全类小班的回判及预测正确率分别为 85.2%及 85%，对偶灾类小班的回判及预测正确率分别为 13.8%及 18.8%，对常灾类小班的回判及预测正确率分别为 75.9%及 73.6%，总的回判及预测正确率分别为 65.9%及 66.2%。从模型系数的检验值大小可以看出，对小班是否成为安全类影响最大的前 3 个因素依次为海拔高度、地形、坡度，而对小班是否能成为常灾类影响最大的前 3 个重要因子则依次为海拔高度、坡度、地形。

从判别结果来看，模型对安全类及常灾类小班的识别效果都更好，而对偶灾类小班的识别效果却较差。进一步的分析同样表明，将定性指标的赋值方式加以调整之后，只含有变量的 Logistic 回归模型的参数及判别结果也会随之改变。

2.2.2 将定性指标作为因子，定量指标作为协变量的分析结果 将描述小班特征的 13 个指标中的 7 个定量指标作为协变量，6 个定性指标作为因子进行建模分析。仍选用主效应模型，不考

虑变量之间、因子之间及变量与因子之间的交互作用，模型系数的显著性检验采用 Wald 检验法。根据各因子及变量的似然比检验结果，选取显著性水平低于 0.01 的变量平均树高、坡度、海拔高度及因子优势树种、起源、地类、地形及坡向参与构建线性 Logistic 回归模型，模型的系数及有关检验值列入表 4。

模型显著性检验的 χ^2 值为 3 327.128，远远大于自由度为 44 的 χ^2 临界值，说明模型效果是显著的。模型对安全类小班的回判及预测正确率分别为 84%及 84.5%，对偶灾类小班的回判及预测正确率分别为 19.7%及 22.6%，对常灾类小班的回判及预测正确率分别为 75.5%及 73.3%，总的回判及预测正确率分别为 66.6%及 66.8%。总而言之，模型对安全类及常灾类的识别能力都比较强，但对于偶灾类的误判率仍较高。这里不需要将定性指标量化，也就不存在因对定性指标的量化方式不同而对模型造成的影响。因此建模过程中受人为因素的干扰就相对较小。

表 4 由变量参与建立的 Logistic 回归模型的回归系数及有关检验结果(广东, 德庆)

Tah 4 Parameters of Logistic regression made up coevaniables average height of tree, slope and elevation and factors dominant species, origin, type, topography and direction of slope (Guangdong, Deqing)

变量	安全类				常灾类			
	系数	SE	Wald 检验值	显著水平	系数	SE	Wald 检验值	显著水平
截距	25. 838	2 184	139. 943	0. 000	31. 829	0. 364	7 649. 211	0. 000
平均树高	0. 043 4	0. 015	8. 838	0. 003	$4. 492\times 10^{-4}$	0. 017	0. 001	0. 979
坡度	0. 0300 9	0. 006	24. 652	0. 000	-0. 052 51	0. 006	70. 344	0. 000
海拔高度	$6. 607\times 10^{-3}$	0. 001	128. 489	0. 000	-0. 011 02	0. 001	119. 877	0. 000
优势种	马尾松	-0. 529	0. 205	6. 677	0. 010	0. 169	0. 210	0. 645
	湿地松	-1. 292	0. 323	16. 046	0. 000	-0. 041 39	0. 237	0. 030
	针阔混	0. 405	0. 266	2. 322	0. 128	-0. 050 49	0. 337	0. 022
	针叶混	0	0	—	—	0	0	—
起源	飞播	-1. 332	0. 349	14. 569	0. 000	-1. 351	0. 346	15. 277
	人工	0. 038 67	0. 174	0. 050	0. 824	0. 153	0. 162	0. 889
	天然	0	0	—	—	0	0	—
地类	林分	-13. 899	1. 597	75. 769	0. 000	-15. 684	0. 193	6 626. 872
	疏林	-13. 646	1. 597	73. 021	0. 000	-16. 270	0. 000	—
	未成	0	0	—	—	0	0	—
地形	低丘	-13. 421	1. 590	71. 245	0. 000	-11. 009	0. 470	548. 039
	低山	-12. 488	1. 445	74. 660	0. 000	-13. 633	0. 210	4 232. 828
	高丘	-13. 374	1. 441	86. 159	0. 000	-13. 309	0. 091	21 333. 004
	中丘	-13. 905	1. 440	93. 219	0. 000	-13. 039	0. 000	—
	中山	0	0	—	—	0	0	—
坡向	北	0. 142	0. 192	0. 553	0. 457	0. 084 11	0. 195	0. 186
	东北	-0. 336	0. 174	3. 740	0. 053	-0. 572	0. 177	10. 421
	东南	-0. 109	0. 168	0. 424	0. 515	-0. 410	0. 173	5. 604
	东	-0. 167	0. 174	0. 921	0. 337	-0. 330	0. 174	3. 589
	南	0. 162	0. 168	0. 934	0. 334	-0. 201	0. 172	1. 366
	无	-0. 227	0. 180	1. 598	0. 206	-0. 757	0. 175	18. 602
	西北	-0. 358	0. 177	4. 073	0. 044	-0. 460	0. 184	6. 248
	西南	$7. 596\times 10^{-3}$	0. 169	0. 002	0. 964	-0. 468	0. 174	7. 225
	西	0	0	—	—	0	0	—

2. 3 小班发生类型辨识模型的比较

以上借助判别函数分析及线性 Logistic 回归分析的方法, 针对变量的不同组合形式建立了 4 个小班发生类型的辨识模型.

4 种模型将来自安全类判为安全类的概率分别 0. 808、0. 837、0. 85、0. 845, 将来自常灾类判为常灾类的概率分别 0. 719、0. 682、0. 736、0. 733, 这表明它们识别安全类及常灾类小班的能力均较强; 但 4 种模型将来自偶灾类判为偶灾类的概率分别 0. 203、0. 159、0. 188、0. 226, 说明这些模型对偶灾类小班的识别不够准确. 无论是对哪一灾情发生类型小班而言, Logistic 回归模型的识别能力均强于判别分析模型.

在实际应用中, 人们往往更关心的是判别的

后验概率, 即当把一个未知类别的小班判为某一类别时, 它确实属于该类别的可能性有多大, 也即对新个体正确归类的概率有多大. 由前面的结果可知, 经 4 种模型判为安全类实际也属于安全类小班的比例分别为 0. 772、0. 752、0. 754 及 0. 765, 判为偶灾类实际也属于偶灾类小班的比例分别为 0. 378、0. 334、0. 427 及 0. 44, 判为常灾类实际也属于常灾类小班的比例分别为 0. 535、0. 536、0. 578 及 0. 586. 说明 4 种模型在对安全类的判别方面, 效果相差不大, 而在对偶灾类及常灾类, 尤其是偶灾类的判别方面, Logistic 回归模型要优于判别分析模型.

4 种模型总的识别正确率分别为 0. 64、0. 634、0. 662 及 0. 668. 由此可见, 从总体识别效果

上来讲, Logistic 回归模型也优于判别分析模型, 而含有定性指标因子的 Logistic 回归模型则是 4 个模型中最好的一个。

3 讨论及结论

通过给定性指标赋值使之转化为定量指标, 固然可使定性指标能够参与一些分析, 但由于模型本身及效果均随赋值方式的不同而变化, 从而增加了建模过程中的人为干扰及模型的不稳定性。因此, Logistic 回归分析应当是处理含有定性指标问题的理想方法。

本文就变量的两种不同处理方式, 分别用判别分析及 Logistic 回归分析的方法建立了四个模型。四个模型的预测及回测结果都较一致, 说明模型具有较高的可信度。比较而言, 运用 Logistic 回归所建模型的效果均优于判别函数模型。两类模型对安全类的识别及判别效果均较好, 对常灾类的识别效果要好于判别效果, 而对于偶灾类小班, 识别及判别效果均不理想。偶灾类小班是介于安全类与常灾类之间的一类小班, 与安全类小班及常灾类小班之间的界限并不明确, 识别起来有一

定的难度, 更何况人们在确定这类小班时, 本身就存在一定的不准确性。

在被 4 个模型判为安全类的小班中, 本属于常灾类小班的比例不超过 7%。说明当利用所建模型把一个未知发生类型的小班确定为安全类并忽略对它的监控和管理, 而结果又因该小班经常遭受虫灾而导致损失的可能性是很小的。因此, 本研究所建模型对于提高森林害虫管理水平及管理效益都具有较高的应用价值。

参考文献:

- [1] 罗积玉, 邢英. 经济统计分析及预测[M]. 北京: 清华大学出版社, 1987.
- [2] 陈家鼎, 戴中维, 译. 生存数据的统计方法[M]. 北京: 中国统计出版社, 1998.
- [3] 任若恩, 王惠文. 多元统计数据分析—理论、方法、实例[M]. 北京: 国防工业出版社, 1997.
- [4] 柯惠新, 黄京华. 调查研究中的统计分析方法[M]. 北京: 北京广播学院出版社, 1992.
- [5] 王静龙, 梁小筠, 李宝慧, 译. 应用线性回归[M]. 北京: 中国统计出版社, 1998.

Studies on Distinguishing Models of Occurrence Type of *Dendrolimus punctatus* in a Subcompartment

HUANG Ying-zi¹, GU De-xiang¹, ZHANG Wen-qing¹, LIANG Cheng-feng², CHEN Zhan-fang³

(1. State Key Laboratory for Biocontrol, Zhongshan University, Guangzhou 510275, China;

2. Forest Pest & Disease Control Station, Forestry Department of Guangdong Prov, Guangzhou 510173;

3. Forestry Bureau of Deqing Country, Guangdong, Deqing 526000)

Abstract: The importance of factors influencing occurrence type of *Dendrolimus punctatus* in a subcompartment was analysed. Four distinguishing models of occurrence type of *D. punctatus* in a subcompartment were developed based on discriminant analysis and logistic regression analysis.

Keywords: *Dendrolimus punctatus*; subcompartment; occurrence type; discriminant function analysis; Logistic regression analysis