以豚草为例利用 GIS 和信息理论的方法预测外来入侵物种在中国的潜在分布

陈 浩[①②] 陈利军[②*] Thomas P. Albright[③④]

(① 武汉大学遥感信息工程学院, 武汉 430079; ② 国家基础地理信息中心, 北京 100080; ③ SAIC, USGS Center for EROS, Sioux Falls, SD 57198, USA; ④ Department of Zoology, University of Wisconsin-Madison, WI, USA. * 联系人, E-mail: chenlj@nsdi.gov.cn)

摘要 外来物种入侵已经对世界各国的经济、公共健康、农业生产力和生态完整性造成了越来越大的威胁, 描述和预测外来入侵物种的空间分布对物种入侵的防治和早期预警起着重要的作用. 生物多样性数据的不对称性以及生物建模过程中模型选择的不确定性给入侵物种空间建模带来了困难和局限. 本研究利用统计学和信息理论的方法, 从地学空间制图和生物建模的角度研究了外来入侵物种(以豚草 *Ambrosia Artemisiifolia* L.为例)的潜在分布以及环境影响因子, 提出了一种改进的 logistic 回归模型. logistic 回归模型的选择基于 Akaike 信息标准(AIC), 针对物种数据的不对称性, 本研究提出了一种新的频率统计的方法去划分物种源生地的适应性生存环境. 最后, 我们把在源生地建立的模型和分类标准投影到入侵地绘制了该物种在入侵地的相对适应性分布图.

关键词 外来入侵物种 潜在分布 **Akaike** 信息标准**(AIC)** **logistic** 回归 频率统计 **GIS**

外来物种入侵已经成为人类历史上最重大的生态事件之一, 它影响到了国民经济、公众健康和农业生产力, 同时对生物多样性、生态系统的稳定性以及所有物种都赖以生存的自然界的平衡造成了长期的威胁 [1]. 阐释外来入侵物种的自然特性、物种-生存环境关系以及预测该物种在入侵地的空间分布对开展入侵物种的监测预警和治理具有重要的意义 [2,3]. 预测物种潜在分布最常用的方法是利用生存环境理论(一定的环境条件下物种能保持一定的种群数量)对入侵物种的适应性生存环境进行重建: 利用物种标本的地理信息和一系列GIS环境数据(气候、地形、生态等)进行机器学习和统计分析得出规则, 通过对规则的优化选择, 在多维的生态空间中建立入侵物种最适宜生存的生存环境模型, 最后把模型投影到地理空间中, 利用GIS技术绘制物种在源生地和入侵地的生存环境适应性分布图 [4].

广义线性模型(generalized linear models, GLM)的特殊形式——logistic回归模型是生态生存环境建模常用到的方法之一, 这种方法通过在多维的生态空间中建立模拟物种适应性程度的生存环境适应性概率曲面, 从而达到预测物种的目的. 对于logistic回归而言, 传统的模型选择方法是基于零假设的显著性检验, 这种方法的有效性, 特别是在使用潜在的不相关的零假设、显著性水平设定的人为武断、基于单一模型的模型选择不确定性上已经受到了很多学者的质疑 [5~7]. 近年来, 基于信息理论的模型选择和推论被认为是显著性检验的一种替代方法 [6]. 信息理论的基本思想是: 不存在单一、真实的模型, 模型只能无穷的接近于现实, 模型选择的目标是确认哪个模型更接近真实, 即Kulback-Leibler信息量损失最小 [8]. 这种基于信息量的模型选择方法考虑了模型的拟合程度和模型的复杂程度两个方面的因素, 同时可以进行多模型平均和变量评估, 克服了传统显著性检验的一些不足 [9].

生存环境适应性建模中, 生物多样性采样(物种出现点/物种不出现点)是一个重要的组成部分 [10,11]. 在现实中, 准确的物种不出现信息是很难获得的, 在绝大多数博物馆的标本数据库或网络标本数据库中往往只记录了该物种标本的采集地点(物种出现点), 而没有该物种在什么地方没有出现(物种不出现点)的地理空间信息. 即使在野外调查中, “物种不出现”的概念也是不确定的, 在采样单元(地理栅格)内任何

2006-10-25 收稿, 2007-02-07 接受

国家自然科学基金(批准号: 40371084)、美国地质调查局基金(批准号: 03CRCN0001)、美国威斯康星大学麦迪逊校区基金(批准号: 03CRAG0016)资助项目

一个物种出现事件，我们就可以确定该采样单元属于物种出现点，但是采样单元中的任何一个区域都没有该物种出现，才可以确认该采样单元属于物种不出现点，这种生物多样性数据的不对称性和不确定性就给物种调查和生存环境建模带来了很大的困难和局限[11]. 近些年来，把地理空间中的随机采样点作为“伪-物种不出现点”或“背景像素点”代替真实的物种不出现点是解决这一问题常用的折中方法[12]. 这种方法不可避免的会带来较大的采样误差，在此基础上建立的模拟物种适应性程度的绝对生存环境适应性概率是不准确的. 因此在进行生存环境建模的过程中需要建立一种更好的方法去解决在缺少真实物种不出现信息情况下进行合理预测的问题.

本研究利用地理空间技术和信息理论建立了一种改进的 logistic 回归方法，研究了外来入侵物种(以豚草 *Ambrosia Artemisiifolia* L.为例)的潜在分布及其环境影响因子. 利用信息理论(AIC 标准)及其推论来计算和评估模型的不确定性，在此基础上建立了多模型框架下的 logistic 加权平均模型. 由于获取的物种标本数据只包括物种出现点的信息，为了减少这种采样误差，本研究提出了一种利用 logit 阈值和频率统计进行相对适应性生存环境划分的方法，减小多样性数据的不对称性带来的影响.

1 入侵物种生存环境建模的原理和方法

1.1 物种数据的准备和前期处理

物种出现点数据来源于美国/加拿大的自然历史博物馆和中国科学院植物研究所标本馆，真实的物种不出现点由研究区域的随机采样获得的“伪不出现点”代替，并获得了每个样点的地理定位信息(经度/纬度). 所需的前期处理包括：(ⅰ) 物种样点的栅格化，使物种样点与环境 GIS 图层具有相同的空间分辨率；(ⅱ) 利用陆地掩模排除落入水中的随机采样点；(ⅲ) 利用缓冲区分析排除落在物种出现点附近 8 个邻域内的随机采样点以减少空间自相关带来的影响.

很多环境GIS图层来源于相同生态学/气候学特征的不同测量或者来源于相关图层之间的计算，运用这些具有相关性的环境GIS 图层进行建模可能会带来共线性的问题[13]. 鉴于此，在进行生态适应性建模之前进行环境GIS 图层的两两相关分析，相关系数小于设定阈值的环境变量被选择.

1.2 logistic 回归模型

利用物种出现点和“伪-物种不出现点”建立GLM模型去预测入侵物种的潜在分布. 由于响应变量是二值的(0 或 1)，因此GLM合适的形式是二值logistic回归[10]. Logistic回归模型中，物种出现的几率计算如下(B_0, …, B_K 是系数, X_1, …, X_k是预测变量):

$$\text{几率} = \frac{e^{B_0+B_1X_1+B_2X_2+\cdots+B_KX_K}}{1+e^{B_0+B_1X_1+B_2X_2+\cdots+B_KX_K}}.$$

1.3 Akaike 信息标准(Akaike's information criterion, AIC)

利用Akaike信息标准进行logistic回归模型的评估和选择. AIC代表了最大释然和Kullback-Leibler信息量的关系，在简约性的原则下加入了参数个数作为惩罚项去选择最接近真实(最适合)的模型[9,14]. AIC是模型灵活性和模型拟合的折衷评分标准，其中AIC值越低，模型越好. AIC定义如下:

$$\text{AIC} = -2\log_e\left(L\left(\hat{\theta}\,|\,\text{data}\right)\right)+2k.$$

其中 $\log_e\left(L\left(\hat{\theta}\,|\,data\right)\right)$ 是给定数据 data、候选模型、未知参数($\hat{\theta}$)的最大对数似然(log-likelihood)，k为模型中参数个数(包括变量个数和截距).

1.4 模型选择方法

一般来说，有两种常见的模型选择方法. 第 1 种方法是所谓的保守模型选择策略，即在进行模型选择之前，根据生物学和生态学知识建立少量的先验模型，根据 AIC 选择出最佳的先验模型. 这种方法可以降低数据挖掘中出现的“over-fitting”和其他缺陷的风险，但是，如果对于物种生态学和生态地理学知识、理解有局限，那么这种方法的有效性也将受到限制. 第 2 种方法是基于数据驱动的模型选择策略，即选择一系列分布具有相关性的变量，建立包含这些变量所有可能组合的候选模型，根据 AIC 选择最佳的候选模型. 这种方法可以降低出现遗漏重要信息的风险，但是由于没有考虑生态学和生态地理学的知识，最后选择的模型往往具有统计学意义但从生态学上很难解释.

综合上述两种模型选择方法，本研究提出了一种“核心变量加边缘变量”的模型选择策略：根据物种生长、扩散的先验知识和生态学家的建议选择了几种重要的生态学变量作为“核心变量”，这些核心变量组成的模型称为“核心模型”，然后将剩下的变量

("边缘变量")加入到核心模型中组成更为复杂的候选logistic模型, 最后根据AIC选择最佳的模型. 这种方法同时考虑到了生态学和统计学的意义, 是一种较为合理的折中策略.

1.5 logistic 加权平均模型

ΔAIC (Δ_i)是一种评价候选模型相对适应性的指标. 它描述了候选模型 AIC 值与所有模型中最小的 AIC 值之间的差, 根据Δ_i从小到大的顺序, 我们可以得到最佳模型到最差模型的排列. ΔAIC 定义如下:

$$\Delta\mathrm{AIC} = \Delta_i = \mathrm{AIC}_i - \mathrm{AIC}_{\min}.$$

其中AIC_i是模型i的AIC值, $\mathrm{AIC}_{\min}$是模型中最小的AIC值(相对最佳模型), 随着Δ_i增加, 候选模型最接近真实模型的可能性减小 [6].

Akaike权重(W_i), 是另一种候选模型评价指标, 它代表了该候选模型是最佳模型的概率 [5~8], 定义如下:

$$W_i = \frac{\exp(-\Delta_i/2)}{\sum_{r=1}^{R}\exp(-\Delta_i/2)}.$$

其中 R 是所有候选模型的数量. 单一的模型(具有最小的AIC或最大W_i的模型)往往不能达到完全反映真实模型的目标, 可以利用信息理论的推论建立基于多模型框架下的 logistic 加权平均模型. 利用下面的公式计算平均模型各预测变量的系数$\hat{\theta}$ (其中$\hat{\theta}_i$表示候选模型 i 的系数):

$$\hat{\theta} = \sum_{i=1}^{R} w_i\hat{\theta}_i.$$

logistic加权平均模型包括了"核心变量加边缘变量"策略中所有的变量, 只是每个变量具有不同的权重. 对包含某一给定预测变量的所有的候选模型的 Akaike 权重求和可以在多模型框架下对预测变量的相对重要性进行评估 [9,15].

1.6 频率统计方法

把 logistic 加权平均模型运用到源生地得到每个像素的 logit 值(logistic 模型的线性部分). 由于缺少物种真实的不出现点信息, 把每个未知像素的 logit 值转换到代表物种在该像素范围中可能出现的概率(0%~100%)的传统方法是不准确的. 本研究运用物种样点的 logit 阈值建立了一种简单频率统计算法确定了研究区域的相对适应性水平.

首先, 我们选择源生地真实的物种出现点的最小logit值作为相对适应性阈值, 将所有源生地的样点(包括物种出现点和"伪-物种不出现"点)重新划分为两族: 利用R统计软件 [16]分别对这两族进行四分位数统计, 将源生地划分为 8 个相对适生类别, 最后这 8 个logit值区间成为了相对适应性生存环境的划分标准(表 1). 最后利用预测外推(extrapolative predictions)的思想, 将特定环境数据集(源生地)建立的模型运用到其他不同的环境数据集(入侵地)中来预测入侵物种在入侵地的适应性分布.

表 1 基于频率统计的相对适应性划分方法

相对适应性类别	划分标准 a)
不适生区域:	小于 logit[1)]
1 最不适应	小于 25%分位数 [1)]
2	25%分位数 [1)] ~ 50%分位数 [1)]
3	50%分位数 [1)] ~ 75%分位数 [1)]
4 轻度不适应	75%分位数 [1)] ~ logit[1)]
适生区区域:	大于 logit[1)]
5 轻度适应	logit[1)] ~ 25%分位数 [2)]
6	25%分位数 [2)] ~ 50%分位数 [2)]
7	50%分位数 [2)] ~ 75%分位数 [2)]
8 最适应	大于 75%分位数 [2)]

a) logit[1)], 物种出现点的最小 logit 值; 分位数 [1)], logit 值小于 logit[1)]的"伪-不出现点"的基于 logit 值的分位数统计; 分位数 [2)], 包括 logit 值大于 logit[1)]的"伪-不出现点"和所有物种出现点的基于 logit 值的分位数统计

1.7 豚草(*Ambrosia artemisiifolia* L.)的例子

豚草(*Ambrosia artemisiifolia* L.), 属菊科一年生草本植物, 别名艾叶破布草, 原产美国和加拿大, 其花粉是过敏性鼻炎和季节性哮喘的主要病源, 是世界公认的公害性杂草之一 [17,18]. 自从 1935 年在我国东北发现以来, 豚草就在我国迅速扩张, 并在东北、华北、华中和华东等地约 15 个省\直辖市有分布, 而且有继续扩张的势头. 预测入侵物种豚草在中国的潜在分布能帮助政府的管理和决策者进行豚草入侵的监测和防治.

豚草样本记录包括 243 个源生地(美国和加拿大)标本记录和 83 个入侵地(中国)标本记录. 在源生地随机采样获得 1000 个样点, 通过前期处理剩下 852 个作为"伪-不出现点". 物种生存环境建模包括 32 个潜在生存环境GIS图层, 它们包括地形学图层(高程、坡度、地形指数等), 气候学图层(温度、降水、太阳辐射和蒸发等), 人为干扰图层(农业和城镇密度)以及森林覆盖密度图层. 地形学图层来源于美国地质调查局HYDRO-1K数据集(http://edc.usgs.gov/products/

elevation/gtopo30/gtopo30.html); 气候学图层来源于Worldclim 1.4 数据集(http://biogeo.berkeley.edu/worldclim/worldclim.htm). 人类活动的干扰是影响入侵物种扩散和生存的重要因素, 本研究用农业和城镇密度代表了人类干扰的程度, 它由美国地质调查局全球土地利用数据集(USGS Global Landcover Characterization dataset)[19]计算得来, 其中每个像素值代表了在该像素内农业和城镇所占得比例(取值范围从 0%~100%). 森林覆盖密度是基于高分辨率辐射计(AVHRR)数据和TM影像数据进行回归分析计算得出的, 该图层来源于联合国森林资源评估数据集(UN/FAO Forest Resources Assessment 2000 dataset), 其中每个像素值代表了在该像素内森林所占得比例(取值范围从 0%~100%).

从以上 32 个图层中选择了 6 个图层作为建模的环境变量, 它们分别是高程(elev)、年平均降水(precip)、最冷月份的最低温(mintmp)、太阳辐射(sun)、单位像素内农业和城镇密度(agurb)、单位像素内森林覆盖密度(fordens). 根据"核心变量加边缘变量"的模型变量选择策略, 在豚草生长和扩散的生物学先验知识的基础上, 我们选择了 3 个变量(precip, mintmp 和 sun)及其各自的平方项作为模型的核心预测变量, 剩下的 3 个变量(elev, agurb 和 fordens)及其各自的平方项加上两个交互项(sun*mintmp 和 precip*mintmp)成为模型的边缘变量. 把边缘变量的不同组合加入到由核心变量建立的核心模型中组成更为复杂的候选 logistic 模型.

2 结果与分析

2.1 信息理论的方法

(ⅰ) logistic回归模型的信息理论指标. 通过R统计工具建立入侵物种豚草的logistic回归候选模型集, 并计算出每个候选模型相应的信息理论指标, 它们包括: Akaike 信息标准(AIC)、Δ_i指标、Akaike 权重(W_i)以及累积Akaike 权重(W_s). 根据AIC从小到大的顺序选择了 9 个较优模型, 如表 2 所示. 具有最小AIC值(最大的Akaike 权重)的模型包括 3 个核心变量以及高程(elev)、单位像素内农业和城镇密度(agurb)、单位像素内森林覆盖密度(fordens)和一个交互变量sun*mintmp. 该模型在整个候选模型集中是相对最优的, 但是模型的Akaike 权重却相对较低(0.2460), 说明把这个单一的模型作为最终模型缺乏足够的说服力. 选取的这 9 个模型的ΔAIC指标都小于或接近 4, 可以认为它们都有助于反映真实模型的信息[6], 同时这 9 个模型的累积Akaike 权重为 0.9006, 用多模型结构去近似描述现实世界的真实模型比其他的单一模型具有更高的置信概率(90.06%). 所有的 9 个模型都包含了 3 个核心变量以及高程(elev)和单位像素内农业和城镇密度(agurb), 则表明这些变量对于豚草的生长和扩散都是重要的因子.

(ⅱ) 环境变量的重要性评估. logistic 加权平均模型中环境变量对入侵物种豚草的重要性评估的结果如表 3 所示. 最终模型是所选 logistic 候选模型基于 Akaike 权重的加权平均, 它包括了上述 9 个候选

表 2 豚草的最佳候选 logistic 回归模型的信息理论统计结果

模型描述	AIC	Δ_i	W_i	W_s
precip + precip^2 + mintmp + mintmp^2 + sun + sun^2 + elev + agurb + agurb^2 + fordens + fordens^2 + sun*mintmp	938.42	0.00	0.2460	0.2460
precip + precip^2 + mintmp + mintmp^2 + sun + sun^2 + elev + agurb + agurb^2 + fordens + fordens^2 + precip*mintmp	938.98	0.56	0.1859	0.4319
precip + precip^2 + mintmp + mintmp^2 + sun + sun^2 + elev + fordens + fordens^2 + precip*mintmp+ sun*mintmp	940.17	1.75	0.1026	0.5345
precip + precip^2 + mintmp + mintmp^2 + sun + sun^2 + elev + elev^2 + agurb + agurb^2 + fordens + fordens^2 + sun*mintmp	940.42	2.00	0.0905	0.6250
precip + precip^2 + mintmp + mintmp^2 + sun + sun^2 + elev + fordens + fordens^2 + sun*mintmp	940.64	2.22	0.0811	0.7061
precip + precip^2 + mintmp + mintmp^2 + sun + sun^2 + elev + elev^2 + fordens + fordens^2 + agurb + agurb^2 + sun*mintmp + precip*mintmp	940.94	2.52	0.0698	0.7759
precip + precip^2 + mintmp + mintmp^2 + sun + sun^2 + elev + fordens + fordens^2 + agurb + sun*mintmp	941.51	3.09	0.0524	0.8283
precip + precip^2 + mintmp + mintmp^2 + sun + sun^2 + elev + elev^2 + fordens + fordens^2 + sun*mintmp + precip*mintmp	942.00	3.58	0.0411	0.8694
precip + precip^2 + mintmp + mintmp^2 + sun + sun^2 + elev + elev^2 + fordens + fordens^2 + sun*mintmp	942.55	4.13	0.0312	0.9006
precip + precip^2 + mintmp + mintmp^2 + sun + sun^2 (核心模型)	989.16	50.7	0.0000	0.0000

表 3　豚草的环境变量重要性评估

模型变量	变量系数	权重 c)
precip	−0.00645 a) −1.34339 b)	1.0000
mintmp	0.13393 a) −0.00002 b)	1.0000
sun	0.01544 a) −4.84667 b)	1.0000
elev	−0.00158 a) −0.00158 b)	0.9996
fordens	0.03382 a) −0.00032 b)	0.9806
agurb	0.00145 a) −1.28464 b)	0.7533
sun*mintmp	−5.91351	0.9501
precip*mintmp	3.04714	0.4441
常数项	−219.680	–

a) 一次项系数; b) 二次项系数; c) 重要性权重计算时包括所有的一次项和二次项

模型中包含的所有变量. 重要性权重是评估气候环境变量对豚草入侵影响的标准, 是包含某一给定预测变量的所有的候选模型的 Akaike 权重的和. 除 3 个核心变量(权重为 1)外, 高程(0.999)和单位像素内森林覆盖密度(0.9806)都具有很高的权重, 并出现在所有 9 个候选 logistic 回归模型中, 说明这是两个对豚草入侵非常重要的因子. 同时温度和太阳辐射的交互作用(0.950)也是影响物种入侵特别是短日照植物豚草引种生长的重要因子, 而降水和温度的交互作用(0.4441)是重要性相对较弱的因子.

2.2　基于频率统计的相对适应性划分结果

结合适应性生存环境划分标准, 源生地建立的 logistic 加权平均模型运用到 GIS 图层中, 把豚草的源生地(美国和加拿大南部)划分为 8 个相对适生区类别, 并绘制了豚草在北美的分布图(图 1(a)). 考虑到物种生存环境在不同地理空间的变化, 本研究建立了两种相对适应性生存环境的划分标准: (ⅰ) 利用源生地物种出现点的 logit 阈值进行划分得出的相对适应性生存环境的划分标准(表 4, 标准Ⅰ); (ⅱ) 利用入侵地已知物种出现记录点的 logit 阈值进行划分得出的相对适应性生存环境的划分标准(表 4, 标准Ⅱ). 利用这两种划分标准把模型投影到物种入侵地中国, 绘制了豚草在中国的相对适应性分布图(图 1(b)和(c)). 豚草在中国的相对适应性分布结果与适应性生存环境划分标准有着很强的依赖关系, 但是从图

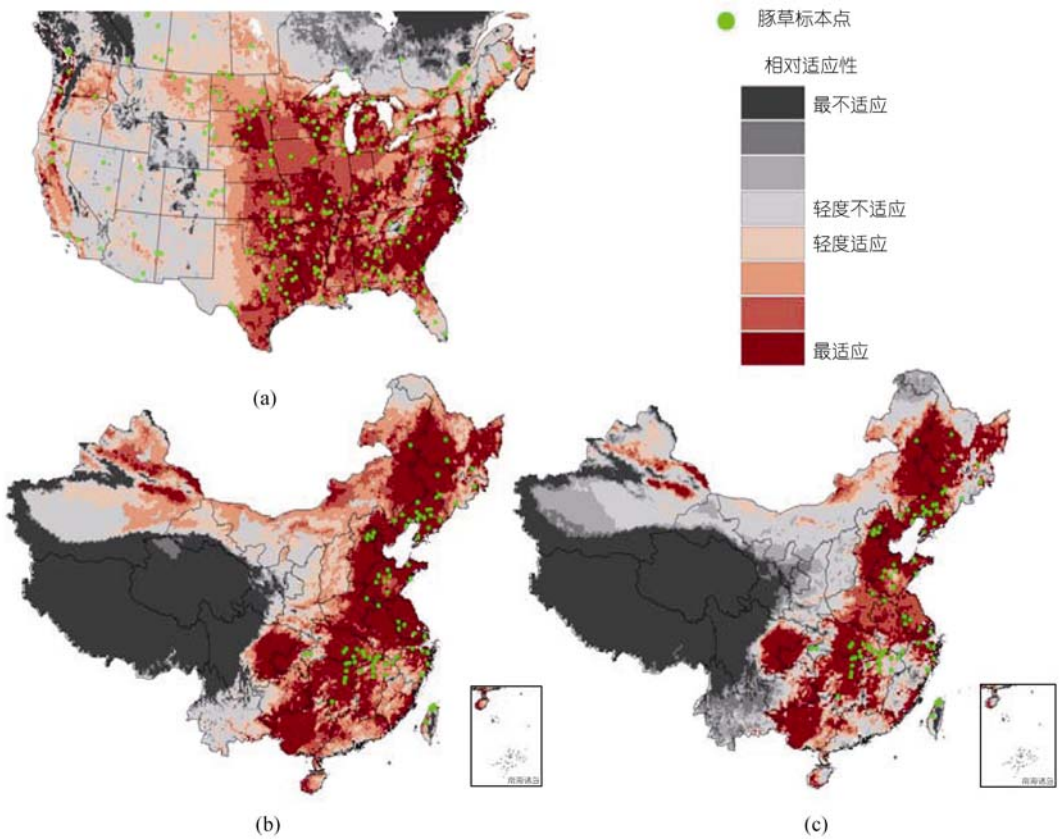

图 1　豚草在源生地(美国和加拿大南部)和入侵地(中国)的相对适应区分布示意图

随着灰色的逐步加深豚草在该地区不出现的可能性加大; 随着栗色的逐步加深豚草在该地区出现的可能性加大. (a)和(b) 相对适应性生存环境的划分标准源于源生地样点的 logit 阈值; (c) 相对适应性生存环境的划分标准源于入侵地样点的 logit 阈值

表 4 豚草的相对适应性划分标准 a)

相对适应性类别	标准 I	标准 II
不适生区域	< –2.09197	< –0.93756
1 最不适应	< –3.47050	< –3.17542
2	–3.47050 ~ –2.96187	–3.17542 ~ –2.46708
3	–2.96187 ~ –2.53550	–2.46708 ~ –1.80273
4 轻度不适应	–2.53550 ~ –2.09197	–1.80273 ~ –0.93756
适生区域:	> –2.09197	> –0.93756
5 轻度适应	–2.09197 ~ –1.40881	–0.93756 ~ –0.70046
6	–1.40881 ~ –0.78746	–0.70046 ~ –0.50788
7	–0.78746 ~ –0.42575	–0.50788 ~ –0.27703
8 最适应	> –0.42575	> –0.27703

a) 标准 I，基于源生地(北美)物种出现点的 logit 阈值得到的相对适应性生存环境划分标准; 标准 II: 基于入侵地(中国)物种出现点的 logit 阈值得到的相对适应性生存环境划分标准

1(b)和(c)都可以看出，豚草在中国具有强大的扩散潜力. 源于北美样点的logit阈值计算的8个分类区间是一种较为自由和宽松的划分方法，图 1(b)中豚草的适生区分布表现出了一种“过度预测”(over-prediction)的趋势，除了西藏和青海等区域外，豚草几乎在全中国都能生长，而源于中国已知样点的 logit 阈值计算的 8 个分类区间是一种相对保守的划分方法，图 1(c)中豚草的适生区分布与已知豚草在中国的分布更为接近，也更为合理. 同时，除了已知的发现豚草记录的区域以外，四川盆地、新疆的部分地区、中国南方的一些省份，如贵州、广西、广东和海南都是豚草最适应生长的地区，如果没有采取足够的预防措施，在不久的将来豚草这种危害性巨大的恶性杂草将可能蔓延到这些区域.

3 讨论

3.1 信息理论的方法 vs.显著性检验的方法

一般来说，传统的显著性检验对于可操控的实验是适用的，但是由于生态学上的重要性和统计学的显著性有很大的不同，在生存环境适应性建模中利用基于零假设的显著性检验方法进行模型选择往往是不适用的. 此外，在建模过程中，为了提高模型拟合程度而选择更为复杂的模型意味着增加了模型预测中出现“over-fitting”的风险，同时也增加模型的复杂度的. 信息理论的方法考虑到了模型拟合和模型复杂度两个方面的因素，同时，避免了传统的显著性检验中设定人为武断的概率阈值(如置信度设为 0.01 或 0.05)来评估单一模型的显著性. 信息理论的方法提供了一种定量的手段去比较候选假设，在有多个似是而非的候选模型或假设时，信息理论的方法可以在基于多模型框架下进行比较.

模型选择的不确定性是评价模型准确性的一个重要的部分 [20]，对于logistic回归而言，在模型选择上基于显著性检验的传统的阶梯式逐步回归会误导研究者选择一个看似最佳的模型，但是往往该模型具有很大的不确定性. 在豚草的例子中，所有候选模型中相对最佳的模型的Akaike权重仅为 0.2460，说明在模型选择上有很高的不确定性. 信息理论方法的一个优点是允许多模型平均，而不是选择具有很高不确定性的单一模型，即使该模型是相对最佳的. 这种多模型结构可以减小模型选择不确定带来的误差. 在本研究中，9 个相对最佳模型的累积Akaike权重为 0.9006，说明了这个模型集比任何其他的模型有更高的置信度. 同时，信息理论的方法的另一个优点就是它能在多模型框架下评估变量的重要性，而不是基于单一的模型.

3.2 利用改进的 **logisitic** 模型在缺少真实物种不出现信息情况下进行合理预测

传统的 logistic 回归需要准确的物种出现/不出现样点去建立生存环境适应性的概率曲面. 然而在很多情况下，准确的物种不出现数据是很难得到的，利用“伪-不出现点”代替真实的物种不出现点必然会给模型的准确性带来影响. 在 logistic 建模过程中有两个阶段的结果值: 一是方程的线性部分，本研究中称 logit 值; 另一个是代表物种出现的可能性的几率. 由于缺少真实的物种不出现的地理定位信息，我们不能按照传统的方法把 logit 值转换到代表物种在每个像素范围中可能出现概率的传统的空间尺度(0%~100%)上去. 而是把 logit 值作为衡量物种相对适应性程度的指标，使用这种相对可能性指标的分位数统计代替传统的概率进行相对适生区划分.

在本研究中, logit 值的使用是在生物多样性数据不对称的情况下进行生态生存环境建模的一种重要的策略. 首先，logit 值是基于物种源生地数据集的模型计算而来，它反映了物种在源生地的适应性; 此外，把物种出现点的最小 logit 值作为相对适应性阈值，将所有源生地的样点，特别是“伪-不出现点”重新划分为“适应性族点”和“非适应性族点”两类，这种类似于采样点重分类的方法能减小生物多样性数据不对称性带来的采样误差. 其次，把从源生地建立的模型向外投影到入侵地时，这种基于 logit 阈值的分位数统计方法具有更强的鲁棒性(robust). 一般来说，不同地理空间中生存环境的变化，特别是在入侵初期当地生物因素对入侵物种种群建立的制约(如竞争、

捕食等)是模型出现“过度预测”的重要原因. 目前, 对生物学因素的科学、直接的数学描述是非常困难的, 然而入侵地已知记录的物种个体本身可以看作这种复杂的生物学因素(如竞争、捕食等)和非生物学因素共同作用的结果. 因此我们可以用入侵地已知标本点的最小 logit 值代替源生地标本点的最小 logit 阈值, 间接地把生存环境在不同地理空间中的变化加入到生存环境适宜性模型中. 利用这种方法得出的物种的适生区结果与已知物种的分布记录更为一致, 因此预测的分布更接近于现实.

3.3 气候学预测变量 vs.地形学预测变量

根据Austin[21]的观念, 气候学因素是影响生态生存环境建模的直接因素, 而地形学因素是间接因素. 一般来说, 大尺度的预测(如国家或洲)模型常常只需要生物气候学参数, 而不需要考虑地形学因素, 因为人们认为这种粗糙分辨率下的地形因素已经失去了它的预测得能力. 在本研究中, 模型既使用了生物气候学因素如降水、温度、太阳辐射, 同时也使用了地形学因素如高程. 在建模过程中, 所有的 9 个较优的候选模型都包含了高程这一个变量, 同时它还具有很高的Akaike重要性高权重(0.9996), 这些说明在大尺度预测模型中, 高程也是影响豚草分布的一个非常重要的因素, 也具有一定的预测能力. 因为间接变量通常可以表现为不同的资源变量(resource variables)和直接变量的组合 [10,22], 在本研究中, 地形因素可能是除了气候学因素以外的其他重要的环境因素(如土壤等)或者这些环境因素复合作用的间接的反映. 因此在一些预测物种分布的特定实例中, 地形学因素可能和其他一些生物气候学因素一样有效甚至有更强的预测能力.

4 结论

本研究利用地理空间技术和统计学的方法进行了外来入侵物种的适生区分析, 预测了豚草在中国的潜在分布, 在信息理论的框架下建立了一种改进加权平均 logistic 回归模型, 该模型能根据生物多样性数据的不对称的特点, 在只有物种出现数据的情况下, 预测入侵物种的潜在分布. 同时根据 logit 阈值和频率统计的方法生成了两种相对适应性生存环境的划分标准, 改进了一般生态建模中由于忽略生物因素产生的不足之处, 得出的结果更接近现实.

致谢 感谢美国地质调查局(USGS)的朱志良博士和郭勤峰博士的帮助和合作, 国家基础地理信息中心(NGCC)的陈军教授、赵有松高工提出的宝贵建议.

参 考 文 献

1 Guo Q. Perspectives on trans-pacific biological invasion. Acta Phytoecol Sin, 2002, 26(6): 724—730
2 Waage T K, Reaser J K. A global strategy to defeat invasive species. Science, 2001, 292(5521): 1468—1486[DOI]
3 Wang R, Wang Y. Invasion dynamics and potential spread of the invasive alien plant species *Ageratina adenophora* (Asteraceae) in China. Diversity Distrib, 2006, 12(4): 397—408[DOI]
4 Peterson A T. Predicting the geography of species' invasions via ecological niche modeling. Q Rev Boil, 2003, 78(4): 419—433[DOI]
5 Anderson D R, Burnham K P. Avoiding pitfalls when using information-theoretic methods. J Wildl Manage, 2002, 66(33): 912—918 [DOI]
6 Anderson D R, Burnham K P, Thompson W L. Null hypothesis testing: Problems, prevalence, and an alternative. J Wildl Manage, 2000, 64(4): 912—923[DOI]
7 Anderson D R, Burnham K P, White G C. Kullback–Leibler information in resolving natural resource conflicts when definitive data exist. Wildl Soc Bull, 2001, 29(4): 1260—1270
8 Greaves R K, Sanderson R A, Rushton S P. Predicting species occurrence using information-theoretic approaches and significance testing: An example of dormouse distribution in Cumbria, UK. Biol Conserv, 2006, 130(2): 239—250[DOI]
9 Johnson J B, Omland K S. Model selection in ecology and Evolution. Trends Ecol Evol, 2004, 19(2): 101—108[DOI]
10 Guisan A, Zimmermann N E. Predictive habitat distribution models in ecology. Ecol Model, 2000, 135(4): 147—186[DOI]
11 Anderson R P, Lew D, Peterson A T. Evaluating predictive models of species' distributions: Criteria for selecting optimal models. Ecol Model, 2003, 162(3): 211—232[DOI]
12 Phillips S J, Anderson R P, Schapire R E. Maximum entropy modeling of species geographic distribution. Ecol Model, 2006, 190(3/4): 231—259[DOI]
13 Legendre P, Legendre L. Numerical ecology developments in environmental modelling. 2nd ed. Amsterdam: Elsevier, 1998
14 Kullback S, Leibler R A. On information and sufficiency. Ann Math Statist, 1951, 22(1): 79—86
15 Burnham K P, Anderson D R. Kullback–Leibler information as a basis for strong inference in ecological studies. Wildl Res, 2001, 28(2): 111—119[DOI]
16 A Language and Environment for Statistical Computing. Version 2.0.1. R Development Core Team. 2004
17 Boulet L P, Turcotte H, Laprise C, et al. Comparative degree and type of sensitization to common indoor and outdoor allergens in subjects with allergic rhinitis and/or asthma. Clin Exp Allergy, 1997, 27(1): 52—59[DOI]
18 Creticos P S, Reed C E, Norman P S et al. Ragweed Immunotherapy in adult asthma. J Allergy Clin Immunol, 1996, 334(8): 501—506
19 Loveland T R, Reed B C, Brown J F et al. Development of a global land cover characteristics database and IGBP Discover from 1-km AVHRR data. Int J Remote Sensing, 2000, 21(6/7): 1303—1330[DOI]
20 Breiman L. The little booststrap and other methods for dimensionality selection in regression: X-fixed prediction error. J Amer Statist, 1992, 87(419): 738—754[DOI]
21 Austin M P. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. Ecol Model, 2002, 157(2): 101—118[DOI]
22 Guisan A, Hofer U. Predicting reptile distributions at the mesoscale: Relation to climate and topography. J Biogeogr, 2003, 30(8): 1233—1243[DOI]