

DOI: 10.5846/stxb201304030600

许仲林, 彭焕华, 彭守璋. 物种分布模型的发展及评价方法. 生态学报, 2015, 35(2): 557-567.

Xu Z L, Peng H H, Peng S Z. The development and evaluation of species distribution models. Acta Ecologica Sinica, 2015, 35(2): 557-567.

物种分布模型的发展及评价方法

许仲林^{1, 2, *}, 彭焕华³, 彭守璋⁴

1 新疆大学资源与环境科学学院, 乌鲁木齐 830046

2 新疆大学智慧城市与环境建模重点实验室, 乌鲁木齐 830046

3 中国科学院武汉植物园, 武汉 430074

4 草地农业生态系统国家重点实验室, 兰州大学生命科学学院, 兰州 730000

摘要: 物种分布模型已被广泛地应用于以保护区规划、气候变化对物种分布的影响等为目的的研究。回顾了已经得到广泛应用的多种物种分布模型, 总结了评价模型性能的方法。基于物种分布模型的发展和应用以及性能评价中尚存在的问题, 本文认为: 在物种分布模型中集成样本选择模块能够避免模型预测过程中的过度拟合及欠拟合, 增加变量选择模块可评估和降低变量之间自相关性的影响, 增加生物因子以及将物种对环境的适应性机制(及扩散行为特征)和潜在分布模型进行结合, 是提高模型预测性能的可行方案; 在模型性能的评价方面, 采用赤池信息量可对模型的预测性能进行客观评价。相关建议可为物种分布建模提供参考。

关键词: 物种分布模型; 性能评价; 阈值相关; 阈值无关

The development and evaluation of species distribution models

XU Zhonglin^{1, 2, *}, PENG Huanhua³, PENG Shouzhang⁴

1 College of Resources and Environment Science, Xinjiang University, Urumqi 830046, China

2 Key Laboratory of City Intellectualizing and Environment Modelling, Xinjiang University, Urumqi 830046, China

3 Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China

4 State Key Laboratory of Grassland Agro-ecosystems, School of Life Science, Lanzhou University, Lanzhou 730000, China

Abstract: Species distribution models (SDMs) have been used in various applications, such as conservation planning, determining the impact of climate change on species distribution, and others. SDMs allow construction of the correlation relationship between occurrence of a target species and environmental conditions (including bioclimatic and anthropogenic conditions). The correlation relationship can then be applied to the entire environmental space to predict the potential distribution of a target species. In the present study, we first review widely-used SDMs and summarize their evaluation approaches. Generally, SDMs can be classified into two categories according to the data required for construction of the correlation relationship, i.e., SDMs that predict the potential distribution of species based on presence-only records (PO models), and SDMs that use presence-absence records (PA models). If reliable absence records are available, PA models are suggested. Additional classification of SDMs is based on output format, namely, SDMs that give prediction results in the format of continuous probabilities (the higher the probability, the more suitable for distribution), and those with results in the format of binary values (1 for suitable and 0 for unsuitable). According to the various SDM output formats, SDM performance can be evaluated by threshold-independent (for models with continuous probabilities) and threshold-dependent (for models with binary prediction) strategies. Threshold-independent strategies can be realized by calculating values of

基金项目: 国家自然科学基金资助(41361098)

收稿日期: 2013-04-03; 网络出版日期: 2014-03-25

* 通讯作者 Corresponding author. E-mail: galinwa@gmail.com

maximum overall accuracy, maximum kappa, maximum vertical distance, area under the receiver operating characteristic (ROC) Curve (AUC), Gini index, point biserial correlation coefficient, mean square error, root mean square error, coefficient of determination, mean absolute prediction error, and others. Threshold-dependent strategies can be realized by calculating values of sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio, negative likelihood ratio, true skill statistic, odds ratio, Yule's Y, Yule's Q, Phi coefficient, Kappa, normalized mutual information, extreme dependency score, and others. Based on existing problems related to the development, application and evaluation of SDMs, the present study suggests the following. (1) Because biased presence samples can influence the result of prediction, integrating a sample selection module in the SDM could improve the reliability of model prediction. (2) Bioclimatic variables (such as WorldClim) that are calculated from a Digital Elevation Model (DEM) may co-linearly correlate with each other, and such collinearity may result in overfitting when modeling the potential distribution of a species. As a result, selecting variables based on calculation of the Variance Inflation Factor (VIF) is a suitable means to avoid overfitting. (3) In addition to abiotic factors, biotic factors are also important determinants for species distribution. Thus, the use of biotic variables could improve the model results, although biotic factors are not easy to delineate within a geographic space. (4) The spatial and temporal extrapolation of SDMs, which deal with problems of species potential distribution at different geographic ranges and time points (past and/or future), respectively, are actually based on the assumption of an equilibrium relationship between the target species and environmental conditions. However, this assumption is challenged, because species have the abilities of adaption and dispersal. (5) The Partial AUC (PAUC) is suitable for evaluation of single model performance, and the Akaike Information Criterion (AIC) could provide an objective evaluation of the performance of several SDMs.

Key Words: species distribution models; performance evaluation; threshold-dependent; threshold-independent

物种分布模型(Species Distribution Models, SDMs),是将物种的分布样本信息和对应的环境变量信息进行关联得出物种的分布与环境变量之间的关系,并将这种关系应用于所研究的区域,对目标物种的分布进行估计的模型。物种分布模型的理论基础,是生态位的概念,生态位被定义为生态系统中的种群在时间和空间上所占据的位置及其与其他种群之间的关系与作用^[1]。Hutchinson 以数学方式描述了生态位的概念:在由多个环境变量定义的多维空间内,能够维持稳定种群的“超体积(Hyper-volume)”^[1]。围绕如何界定“超体积”,生态学家进行了各种尝试并依据不同的界定方法,发展了不同的物种分布模型。目前,物种分布模型已经成为基础生态学和生物地理学研究的重要工具,被广泛用于研究全球变化背景下物种的分布和气候之间的关系^[2-5]、区域气候变化对植物群落和功能的影响^[6-7]、生态系统功能群和关键种的监测和预测^[8]、生态系统不同尺度多样性的管理和保护^[9]、外来物种入侵区域的预测^[10]、面向生态系统恢复的关键物种的潜在分布预测和保护区规划^[11]等。物种分布模型的广泛应用,应考虑模型的性能评价。因此,对模型的性能进行评价的方法,也是模型的应用过程中值得重点关注的问题。本文首先对物种分布模型的发展过程和评价方法进行了回顾,其次总结了物种分布模型的发展和性能评价方面目前存在的问题并提出了可能的解决方案。

1 物种分布模型的发展

1.1 物种分布模型简介

物种分布模型的发展,始于 BIOCLIM 模型^[12]的开发和应用,随后的 20 多年内,涌现了 HABITAT、DOMAIN、生态位因子分析模型(Ecological Niche Factor Analysis, ENFA)、马氏距离(Mahalanobis Distance, MD)、边界函数方法(Border Function, BF)、最大熵模(Maximum Entropy, Maxent)、广义线性模型(Generalized Linear Model, GLM)、广义加法模型(Generalized Additive Model, GAM)、分类与回归树模型(Classification and Regression Tree, CART)、推动回归树模型(Boosted Regression Tree, BRT)、多元适应性回归样条(Multivariate

Adaptive Regression Splines, MARS)等基于统计的和基于规则集的遗传算法(Genetic Algorithm for Rule-set Prediction, GARP)、人工神经网络(Artificial Neural Network, ANN)等基于人工智能的模型。

BIOCLIM 模型将生态位定义为在环境变量空间中包含所有研究物种样本的超体积^[12]。按照此定义,该超体积是一个以各环境变量的极值(极大值和极小值)界定的超矩形,超矩形所界定的变量范围都被认为是适合于物种分布的。这种定义的缺陷在于,极限环境条件也被认为能够维持种群的稳定,这是不合理的,因为在极限环境条件下,物种虽然能够存活,但是不能维持种群的延续。因此,为减少极值对模型性能的影响,提高模型的预测能力,通常在界定多维超矩形的边界之前,对所有样本上的各变量值进行排序,在其中选择一定数量的极值样本(例如最高的 5%)并对这些环境变量的极值进行平均处理以得到超矩形的边界,由此估计物种的潜在分布区。

HABITAT 模型将生态位定义为物种在环境变量空间上的凸壳^[13],与 BIOCLIM 模型不同,在环境变量空间中 HABITAT 模型不再将环境变量的极值作为生态位的边界,即边界不再是刚性的,而以样本本身所对应环境变量的一个邻域作为适宜物种分布的环境条件,这样就排除了某些极限环境条件。遗憾的是,HABITAT 模型对边界的刻画仍然依赖于外围样本。

基于 Gower 距离算法的 DOMAIN 模型利用点-点的相似矩阵计算目标点上环境变量的适宜性,该适宜性表示了环境变量空间中(而非现实分布空间中),目标点与离它最近的分布样本点之间的相似程度^[14]。在确定物种的生境或者分布范围时,首先需要确定一个阈值以排除非适宜分布区。与之前 BIOCLIM 方法相比,DOMAIN 模型在环境变量空间中确定的环境超矩形并不一定是连续的。

生态位因子分析模型(ENFA)通过计算边际性和环境偏差来度量目标点的适宜性^[15]。在一维情况下,边际性表现为该环境变量的值域上,分布样本点所对应的环境变量的均值(样本均值)与研究区所有点对应的环境变量的均值(全局均值)之间的差;环境偏差其实就是该环境变量的样本方差与全局方差之间的差别。在多维情况下,边际性和环境偏差以多维向量的形式表示。在确定了边际性和环境偏差之后,应用阈值对环境条件进行筛选,可得到物种在环境变量空间中的适宜范围,将其映射到实际研究区,便可得到物种的分布区域^[15]。

MD 方法首先计算样本上各环境变量的平均值,并统计研究区各点上的环境变量到该平均值的 Mahalanobis 距离,依据一定的方法确定一个阈值用以确定生境空间的边界。Mahalanobis 距离算法计算得出的椭圆形超体积能够有效的表达环境变量之间的关系^[16]。

BF 模型以边界函数界定物种在环境变量空间上的边界。以二维环境变量(V_1 和 V_2)空间为例,边界函数的确定方法分为以下步骤:首先,收集物种分布样本及与之相关的环境变量 V_1 和 V_2 ,并作散点图;其次,对其中的一个变量(例如 V_1)进行分段并取各段的中值(或均值),对各段 V_1 值相对应的 V_2 进行排序,选择一定比例(如 5%)的极值并统计极值的平均;再次,在对每一段进行相应分析之后,可得一系列环境变量对偶值,对这些值进行拟合,便可得环境变量的边界函数^[17-18],对影响研究物种分布的各环境变量都进行相应的分析,则可得出物种在整个环境变量空间上的边界;最后,将所得边界映射至研究区的环境变量,便可估计物种的潜在分布区。该方法更加准确的刻画了环境变量空间上的生态位,其缺憾在于相关计算和处理较为繁琐。

Maxent 模型基于热力学第二定理。按照该定律,一个非均衡的生命系统通过与环境的物质和能量交换以保持其存在,也就是说,一个实测存在的系统具有“耗散”的特征,耗散使系统的熵不断增加,直至该生命系统与环境的熵最大,而使熵达到最大的状态,也是系统与环境之间的关系达到平衡的状态。在物种潜在分布的相关研究中,可将物种与其生长环境视为一个系统,通过计算系统具有最大熵时的状态参数确定物种和环境之间的稳定关系,并以此估计物种的分布^[19]。基于该原则,最大熵模型在已知样本点和对应环境变量的基础上,通过拟合具有熵值最大的概率分布对物种的潜在分布作出估计。最大熵模型自 2006 年被开发以来,得到了非常广泛的应用,现已被众多相关研究所采用^[20-24]。

GLM 模型被广泛的应用于物种的潜在分布建模。由于观测样本为布尔值(1 为存在分布,0 为不存在),

而不是连续变量,因此无法构建线性回归模型,于是转而预测物种出现(1)或不出现(0)的概率 μ 。此概率 μ 取值0—1,为连续变量。然而,若以环境变量为预测变量 μ 为应变量,仍无法建立合理的线性回归模型,这是因为物种分布概率与环境变量之间的关系往往不是线性相关的。因此,将问题转为求概率 μ 的转化形式 $\log(\mu/(1-\mu))$ 与环境变量的线性函数,在确定了此函数的参数后,再推求 μ ,便可得各环境条件下物种分布的概率^[25]。

GLM 模型是多元线性回归模型的推广,事实上,广义线性模型所得物种分布的概率是多维环境变量空间中的一个响应曲面,但该曲面对真实概率的拟合程度受样本数量的影响较大,即样本量越大,拟合程度越好,对潜在分布的估计越合理;反之,若样本量较少,则由广义线性模型估算的潜在分布概率不太可靠。另外,广义线性模型的响应变量需要服从高斯分布或者其他对称分布,但物种的响应曲面很可能不是对称的,因此即便增加拟合参数(如增加二次项),也不能很好的逼近真实响应曲面^[26]。在这种情况下,GAM 模型由于能够逼近更丰富的响应曲面而得到了广泛的应用^[27]。此外,广义相加模型受数据而不是模型驱动,即它能够按照数据的结构而不是预先设定的模型(例如高斯分布)对数据进行拟合,因此较广义线性模型更适用^[28]。

CART 模型通过二值递归分割产生二叉树,在每一个可能的节点根据变量的值进行判断并将变量的所有值分割为两个子类。每一次分割都只基于单个变量值,在此过程中,有些变量值可能会被采用多次,而另外一些变量值则可能不会被采用。在每一次分割之后,能够保证包含在两子类中的变量值是“有区别的最大化”。两个子类的变量值根据一定的准则被继续分割,直至达到分类的目的。在物种潜在分布模拟中,通常以物种分布样本上各变量的极值作为节点进行分割。

MARS 模型将线性回归、样条函数构建和二值递归分割相结合,产生一个线性的或者非线性的预测变量和响应变量之间的关系模型。为实现该过程,MARS 模型以一系列基函数对关系函数进行逼近。具体地,若令 $B_n = a_n + b_n x_n$ (式中 x_n 为预测变量,在物种空间分布建模中为环境变量)为基函数序列,则 MARS 模型的形式为: $\hat{f}(x) = \sum_{i=1}^n c_n B_n$ 。

GARP 模型利用物种的分布数据和环境数据运行产生不同规则的集合,判断物种的生态需求,然后预测物种的潜在分布区。GARP 模型是一个反复迭代,寻找最优种类分布规则的过程^[29],在模型中,遗传算法本身并不能刻画生物分布与环境因子的关系,被用来刻画这种关系的是“规则(Rule)”,如 Range 规则,Atomic 规则,Logit 规则等等,遗传算法的作用是为这些规则寻找最大的参数。GARP 模型具有以下优点:首先,它能快速有效的搜索多变量空间;其次,遗传算法是一种非参数方法,它对变量所属总体的统计分布形式没有严格要求;第三,GARP 模型中集成了多类规则,各类规则之间的互补提高了 GARP 的模拟能力。

1.2 各物种分布模型的性能表现

应用本文提及的物种分布模型,Elith 等对分布于全球 6 个不同地区的 226 种物种(澳大利亚湿润赤道地区的鸟类和植物种、加拿大安大略省的鸟类、澳大利亚新南威尔士州的植物种、哺乳动物和爬行动物种、新西兰的植物种、南美 5 个国家的植物种以及瑞士的植物种)的空间分布进行了模拟研究,并根据实测样本数据对结果进行了评价,相关结论认为,Maxent 模型具有较好的性能表现^[4]。自 2006 年被开发以来,应用该模型进行物种保护区规划、入侵物种的潜在分布预测以及物种的空间分布对气候变化的响应等研究的报道已超过 2000 次,这也从另一个侧面说明了相关研究领域对该模型的认可程度。GLM 模型、GAM 模型和 GARP 模型的性能位于 Maxent 模型之后,但优于 DOMAIN 模型和 BIOCLIM 模型^[4]。Tsor 等选择分布于以色列的 42 种蛇类、鸟类和蝙蝠物种进行了空间分布的研究和相关模型的性能评价,结论也同样认为,BIOCLIM 模型的性能较差^[30]。对其他模型性能的评价结果,却与 Elith 等的结果有所区别:GARP 模型和 MD 模型的性能优异,HABITAT 模型、DOMAIN 模型和 ENFA 模型的性能次之。关于物种分布模型性能评价的最新结果显示:选用不同的性能评价手段对模型的性能进行评价,会导致得出不同的结果;另外,若对适宜分布区范围较小的物种进行空间分布的模拟,则模型之间的差异较大,相反,若研究物种的分布较为广泛,则各模型的性能表现出较

高的一致性^[31]。

2 物种分布模型的评价方法

物种分布模型对潜在分布的预测,一般以布尔值(1 表示潜在分布,0 表示潜在不分布)或连续值(概率大小表示物种分布的可能性)给出模拟结果。对应的物种分布模型评价方法,也分为两类:阈值相关方法(Threshold-dependent)和阈值无关方法(Threshold-independent)。两类方法都需要首先定义误差矩阵(表 1),若令样本总数为 n ,则误差矩阵的基本元素分别为实测分布、预测也分布的元素 a ,实测未分布、预测也为未分布的元素 d ,实测未分布、预测分布的元素 b (Commission error)和实测分布、预测未分布的元素 c (Omission error)。

2.1 阈值相关法

阈值相关法指可被直接用以评价输出结果为布尔值(1 表示潜在分布,0 表示潜在不分布)的模型的评价方法。如表 2 所示,阈值相关的评价方法包括总体精度、敏感性、特异性、预测分布准确率、预测未分布准确率、预测分布似然比、预测未分布似然比、真实技巧统计值、分布-未分布比值比、尤拉 Y 系数(Yule's Y)、尤拉 Q 系数(Yule's Q)、 ϕ 系数(Phi coefficient)、卡帕系数(Kappa)、归一化互信息量和极端相关值。

表 1 误差矩阵

Table 1 Confusion matrix

预测结果 Predicted results	实测分布 Recorded presence	实测未分布 Recorded absence
预测分布 Predicted presence	a	b
预测未分布 Predicted absence	c	d

表 2 物种分布模型的阈值相关评价方法^[32]

Table 2 Threshold-dependent accuracy measures of species distribution models^[32]

系数 Coefficient	定义 Definition
总体精度 Overall accuracy	$(a + d)/(a + b + c + d)$
敏感性 Sensitivity	$a/(a + c)$
特异性 Specificity	$d/(b + d)$
预测分布准确率 Positive predictive value	$a/(a + b)$
预测未分布准确率 Negative predictive value	$d/(c + d)$
预测分布似然比 Positive likelihood ratio	$\frac{a/(a + c)}{b/(b + d)}$
预测未分布似然比 Negative likelihood ratio	$\frac{c/(a + c)}{d/(b + d)}$
卡帕系数 Kappa	$\frac{(a + d) - [(a + c)(a + b) + (b + d)(c + d)]/n}{n - [(a + c)(a + b) + (b + d)(c + d)]/n}$
分布-未分布比值比 Odds ratio	ad/bc
尤拉 Y 系数 Yule's Y	$(\sqrt{ad} - \sqrt{bc})/(\sqrt{ad} + \sqrt{bc})$
尤拉 Q 系数 Yule's Q	$(ad - bc)/(ad + bc)$
ϕ 系数 Phi coefficient	$(ad - bc)/n$
真实技巧统计值 True skill statistic	$\frac{a}{a + c} + \frac{d}{(b + d)} - 1$
归一化互信息量 Normalized mutual information	$\frac{-a \ln a - b \ln b - c \ln c - d \ln d + (a + b) \ln(a + b) + (c + d) \ln(c + d)}{n \ln n - [(a + c) \ln(a + c) + (b + d) \ln(b + d)]}$
极端相关值 Extreme dependency score	$2 \ln(\frac{a + c}{n}) \ln(\frac{a}{n}) - 1$

在这些指数中,总体精度粗略的反应了模型预测结果与观测样本相吻合的程度。敏感性和特异性是条件概率,前者指模型预测分布的样本个数占实测分布样本总数的比例,后者指预测未分布样本个数占实测未分布样本总数的比例。可以看到,敏感性和特异性是基于实测观测样本的准确性评价指数,对应的,可以构建基

于预测结果的准确性评价指数——预测分布准确率(与敏感性相对应,建议称为预测敏感性)和预测未分布准确率(与特异性相对应,建议称为预测特异性),前者指实测分布的样本个数占预测分布总数的比例,后者指实测未分布的样本个数占预测未分布样本总数的比例。由这些指数的定义可知,它们的值域为 $[0, 1]$,且值越高表明模型的性能越好。

预测分布似然比和预测未分布似然比最早被用以进行医学诊断,由定义(表2)可知,预测分布似然比指实测分布样本中被准确预测为分布的样本与实测未分布样本中被错误预测为分布样本的比例,对应的,预测未分布似然比指实测分布中被错误预测为未分布的样本与实测未分布样本中被准确预测为未分布样本之间的比例,在计算过程中,如果 b 为零,则预测分布似然比无意义,若 d 为零,则预测未分布似然比无意义。由于同时考虑了敏感性和特异性,因此相对于单独采用敏感性和特异性,采用预测分布似然比和预测未分布似然比对模型性能进行评价给出的结果更有效^[33]。由定义可知,预测分布似然比和预测未分布似然比取非负值,值域取决于样本个数,前者的值越大说明模型的预测性能越好,后者则相反,即值越小,模型的性能越好。

Kappa 系数在很多研究中得到了非常广泛的应用,该系数刻画了模拟值和观测值之间的接近程度是否大于随机分配的值。Kappa 系数的值域为 $[0, 1]$ 且该值越高表明模型的性能越好。由分布-未分布比值比的定义可以看出,该指数表示预测正确的样本个数与预测错误的样本个数之间的比值,Fielding 和 Bell 首次将其引入到物种分布建模中进行模型性能的评价^[33],分布-未分布比值比在 $b=0$ (模型预测未分布而实测记录到分布的样本个数为零)或 $c=0$ (模型预测分布而实测未分布的样本个数为零)的情况下是无意义的,但是在模型表现出色的情况下,有可能出现 $b=0$ 或 $c=0$,因此,在这种情况下,一般为误差矩阵(表1)中的各元素增加一个微量(如 0.1)来计算分布-未分布比值比。分布-未分布比值比为非负值,值域取决于样本个数,分布-未分布比值比越大,模型性能越优。尤拉 Y 系数和 Q 系数相对于分布-未分布比值比更具优势,因为这两个指数可被认为是在 $[-1, 1]$ 上的相关系数,另外,由计算公式可以看出,它们不存在无定义的情况且尤拉 Y 系数和 Q 系数越大,模型的预测性能越佳。

ϕ 系数被 Karl Pearson 引入并作为衡量两个二进制变量之间相关性的系数,事实上, ϕ 系数是皮尔逊相关系数(Pearson correlation coefficient)在计算二进制变量相关性时的简化形式。现有的文献资料表明,在物种分布建模中,尚未有研究涉及采用 ϕ 系数对模型的性能和模拟结果进行评价。不过由其定义可得,该系数在其非负的值域范围内,越大表明模型的性能越优。真实技巧统计值,为敏感性和特异性之和减 1,近似于敏感性和特异性的算术均值,表示实测样本(包括分布样本和未分布样本)上的净预测成功率。该指数在医学诊断中得到了广泛的应用,近年来,也被越来越多地应用于物种分布建模的模型性能评价。类似于敏感性特异性的性质,该值越大,表明模型的预测结果越好。归一化互信息量由 Fielding 和 Bell 引入生态学^[34],Manel 等将其应用于物种分布建模^[35],在表1所示的四个元素中,若任何一个为 0,则无法定义归一化互信息量,但是,若考虑 $\lim_{x \rightarrow 0} x \ln x \cong 0$,则可进行计算而不影响运算结果。归一化互信息量的值域为 $[0, 1]$,其值越高表明模型的性能越好。但是也有研究认为,归一化互信息量存在缺陷,即该指数只能估计两种模式的相似程度,而不能区别“劣于随机”和“优于随机”,因此不是一个精确的性能评价指标^[32]。

极端相关值利用误差矩阵中的 a 和 c 元素对模型的性能进行评价,其值域为 $[-1, 1]$, -1 表示预测性能极差, 1 表示极优, 0 为模型预测性能与随机分配相当。然而,极端相关值未考虑元素 b (实测未分布但预测分布的样本个数)和 d (实测未分布、预测也未分布的样本个数),因此利用极端相关值进行模型性能评价是不全面的。

2.2 阈值无关法

阈值无关法是与阈值相关法相对应的模型性能评价方法,阈值无关法被定义为能够评价输出结果为连续值(而非布尔值)的模型的评价方法,这类方法包括最大总体精度、最大卡帕系数(Maximum kappa)、最大真实技巧统计值、受试者操作特征曲线下面积(AUC)、基尼系数(Gini index)、点二列相关系数、均方差、均方根误差、决定系数和平均绝对预测误差。总体精度、Kappa 系数和真实技巧统计值已在阈值相关的评价方法中有

所介绍,阈值无关法中与它们相关的评价方法是它们各自的极值。均方差、均方根误差、决定系数和平均绝对预测误差是常见的误差统计值,但在物种分布模型的性能评价中,应用还较少,由于其定义简单易于理解,因此也不做具体的分析,以下主要分析受试者操作特征曲线下面积 AUC(按定义,基尼系数为 $2\text{AUC}-1$,因此也略过对基尼系数的具体分析)、点二列相关系数和平均交叉熵。

表 3 物种分布模型的阈值无关评价方法^[32]Table 3 Threshold-independent accuracy measures of species distribution models^[32]

系数 Coefficient	定义 Definition
最大总体精度 Maximum overall accuracy	$\max((a+b)/(a+c))$
最大 Kappa 系数 Maximum Kappa	$\max\left(\frac{(a+d) - [(a+c)(a+b) + (b+d)(c+d)]/n}{n - [(a+c)(a+b) + (b+d)(c+d)]/n}\right)$
最大真实技巧统计值 Maximum vertical distance	$\max\left(\frac{a}{a+c} + \frac{d}{(b+d)} - 1\right)$
受试者操作特征曲线下面积 Area Under receiver operating character Curve	$\frac{1}{(a+c)(b+d)} \sum_{i=1}^{b+d} \sum_{j=1}^{a+c} \varphi(X_i, Y_j)$, 其中,若 $Y > X$, 则 $\varphi(X, Y) = 1$, 若 $Y = X$ 则 $\varphi(X, Y) = \frac{1}{2}$, 否则 $\varphi(X, Y) = 0$. X_i 和 Y_j 分别为实测未分布样本 i 和实测分布样本 j 上的预测值。
基尼系数 Gini index	$2\text{AUC}-1$
点二列相关系数 Point biserial correlation coefficient	$\frac{\overline{P_1} - \overline{P_0}}{\delta} \sqrt{\frac{(a+c)(b+d)}{a+b+c+d}}$, 其中, $\overline{P_1} = \sum_{o_i=1} \frac{p_i}{a+c}$, $\overline{P_0} = \sum_{o_i=0} \frac{p_i}{b+d}$, $\delta = \sqrt{\sum_{i=1}^n \frac{(P_i - \bar{P})^2}{n}}$, $\bar{P} = \sum_{i=1}^n \frac{p_i}{n}$, p_i 和 o_i 为样本 i 上的预测值和观测值(1 为分布, 0 为未分布)
均方差 Mean square error	$\frac{1}{n} \sum_{i=1}^n (p_i - o_i)^2$
均方根误差 Root mean square error	$\sqrt{\frac{1}{n} \sum_{i=1}^n (p_i - o_i)^2}$
决定系数 Coefficient of determination	$1 - \frac{1}{n} \sum_{i=1}^n (p_i - o_i)^2 / [p(1-p)]$
平均绝对预测误差 Mean absolute prediction error	$\frac{1}{n} \sum_{i=1}^n (p_i - o_i)$

AUC 被广泛应用于物种分布建模的各种应用^[19,36]。事实上,AUC 是对模型在不同阈值条件下的性能进行评判的方法,AUC 曲线所处的坐标系中,横轴为 $\frac{b}{b+d}$ 的值(也即 1-specificity),纵轴为 $\frac{a}{a+c}$ 的值(也即 sensitivity),在模型预测的概率分布图中,若取 t_1 为阈值,可得预测的分布样本数和未分布样本数,对照调查的实测分布和实测未分布样本,可构建由 a_1 、 b_1 、 c_1 以及 d_1 组成的误差矩阵,并可通过计算得到 $\frac{b_1}{b_1+d_1}$ 和 $\frac{a_1}{a_1+c_1}$;类似的,若取 t_2 为阈值,可得预测的分布样本数和未分布样本数,对照调查的实测分布和实测未分布

样本,可构建由 a_2 、 b_2 、 c_2 以及 d_2 组成的误差矩阵,并可通过计算得到 $\frac{b_2}{b_2 + d_2}$ 和 $\frac{a_2}{a_2 + c_2}$;若重复以上步骤,则可得一系列点的坐标 $(\frac{b_1}{b_1 + d_1}, \frac{a_1}{a_1 + c_1})$ 、 $(\frac{b_2}{b_2 + d_2}, \frac{a_2}{a_2 + c_2}) \cdots (\frac{b_n}{b_n + d_n}, \frac{a_n}{a_n + c_n})$,用曲线拟合这些点,得到的便是受试者操作特征线,对该曲线下的面积进行积分,则可得 AUC。AUC 的值域为 $[0.5, 1]$, 0.5 代表完全随机的分类,1 表示完全正确的分类,因此, AUC 值越大,表明模型的预测性能越好。

在点二列相关系数的定义中^[37](表 3), $\overline{P_1}$ 为所有观测分布的样本上模型预测值(此处为连续值,值的大小表示物种可能分布的概率大小)的平均值, $\overline{P_0}$ 为所有观测未分布样本上模型预测值的平均值, δ 为标准差,表示所有样本上模型预测值的离散程度,事实上, $\frac{\overline{P_1} - \overline{P_0}}{\delta}$ 指示了模型是否能够有效的区分“观测分布”和“观测未分布”两个样本集合, $\sqrt{\frac{(a+c)(b+d)}{a+b+c+d}}$ 在点二列相关系数中,是一个跟模型的预测性能相关的调整参数。由定义可知,点二列相关系数的值域为 $[0, 1]$, 该值值越大,表明模型的预测性能越好

2.3 适用于不同模型的性能评价手段

按照模型输出结果的数据类型,可将前述的多种模型分为两类:布尔型结果模型和概率型结果模型。布尔型结果模型包括 BIOCLIM、HABITAT、DOMAIN、ENFA、MD、BF、CART、以及 GARP 等模型,因此阈值无关法所包含的多种评价方法可被用于评价这些模型的预测性能。概率型结果模型包括 Maxent、GLM、GAM、MARS 以及 ANN 等,因该类模型的输出结果为概率值,因此选择不同的概率值作为阈值,将对模型的预测结果产生影响。因此,应用阈值相关法包含的多种评价方法对该类模型的性能进行评价是合适的。

3 物种分布模型及评价手段展望

尽管物种分布模型得到了广泛的应用,但也存在值得深入研究的问题:

首先,分布样本的选取可能造成预测结果的偏差。由于在样本采集过程中存在诸如道路便利性、物种的空间分布受到自然或人为因素的干扰等问题,造成收集的样本集是“有偏的”——即样本集所代表的生态位只是实际生态位的一部分。因此,若不对样本集进行以无偏化为目的的选择和处理,则将造成模拟过程中模型算法过度利用部分区域的样本拟合物种和环境变量之间的关系,从而产生过度拟合问题。而与此同时,由于另外一些区域的样本数较少,造成欠拟合。与分布样点有关的另一个影响因素是:在建模过程中是否考虑不分布样点的信息。若对物种分布模型按其所需的样点类型进行分类,可将各种分布模型分为只依赖分布样点的(Presence-only)和共同依赖分布-不分布样点的(Presence-absence)模型。Presence-only 型模型只依据分布样点和环境变量之间的关系构建物种与环境条件之间的关系,而 Presence-absence 型模型除此之外,还考虑不利于物种分布的环境条件的影响。有研究认为,在具备可信的不分布样点(Absence)情况下,可以考虑采用 Presence-absence 型模型对物种的潜在分布进行预测^[38]。因此,在物种分布模型中集成样本选择模块,以避免模型预测过程中的过度拟合及欠拟合,或允许用户能够选择基于 Presence-only 或 Presence-absence 数据对物种的分布进行预测,是值得关注的问题。

其次,环境变量的选择也将影响模型的预测结果。目前的很多研究,都采用 Bioclim 数据集(该集合内的 19 个生物气候变量包括温度、降水量的极值以及变率等)以及坡向、坡度、海拔、归一化植被指数(NDVI)、植被覆盖度、土地利用等作为环境变量对陆地生态系统的物种做出潜在分布的估计^[39]。然而,这些变量之间存在的自相关在模型预测过程中引入了冗余信息,因此,在进行物种潜在分布建模时,应先考虑对环境变量进行选择。在诸多模型中,Maxent 模型以贡献率评价各环境变量对物种潜在分布的影响程度,但这种方法也存在一定的局限性^[19]。因此,在后续的模式开发过程中,增加变量选择模块(如计算方差膨胀因子),以排除冗余信息的影响,是物种分布模型发展中的关键问题。

再次,生物因子(如物种间的相互作用)的空间信息难以表达。除以上提及的诸多非生物因子外,生物因子对目标物种的潜在分布也会产生重要的影响。尽管有研究涉及采用生物因子(如收集的竞争物种的空间分布)对目标物种的潜在分布进行预测^[39],然而,目前开展的大部分研究都较少引入生物因子进行目标物种空间分布的预测。因而,对生物因子的可信表达和物种间相互作用的考量,也是物种空间分布模型的发展和应用中的前沿问题。需要注意的是,对不同生境类型的物种,需采用不同的方法描述生物因子。例如,影响植物种空间分布的生物因子,可从垂直方向(放牧水平、踩踏程度、植食动物的施肥功能等)和水平方向(以优势种盖度所表达的竞争、适应性等)进行描述。而影响动物种潜在分布的生物因子,则宜从其所处食物链中的级别进行考虑,例如上级捕食者、同级竞争者和下级捕食对象的空间分布和密度等。

最后,空间外推和时间外推和合理性问题。对物种历史时期的分布、入侵物种在入侵地的分布以及气候变化背景下物种分布的估计,是物种分布模型的重要应用方面。然而,相关估计的假设前提,是物种与环境条件之间的关系是均衡的,即物种在新的环境条件下,仍保持与原环境条件之间的关系。这是值得商榷的,因为物种具有对新的环境条件的适应性以及物种本身的空间扩散行为。因此,将物种对环境的适应性机制(及扩散行为特征)和潜在分布模型进行结合,是提高模型预测性能的可行方案。值得考虑的是:物种的适应性和扩散行为,需要根据目标物种的不同而区别对待,较合理的方案,是对适应性和扩散行为进行分级,例如,可将扩散行为分为时间和空间上的充分扩散、部分扩散以及无扩散^[40]。

模型性能的评价方面,也存在值得关注的问题:

目前应用较为广泛的 AUC 方法,于 1997 年被由 Fielding 和 Bell 首次引入物种分布模型的性能评价中^[34],由于 AUC 能够通过计算一个数值便能提供模型在所有可能阈值范围上的性能评价结果,因此在过去 10 多年间得到了广泛的应用。然而,有研究认为,对 AUC 的过度应用可能存在问题。相关的质疑是基于以下理由:(1) AUC 值在整个 ROC 曲线范围上计算得出,而极值阈值(上文所提及的 AUC 曲线构建方法中的 $t_1, t_2, \dots, t_n$ 取极大值或极小值)对 AUC 是有显著影响的;(2) AUC 值平等的对待 Omission Error 和 Commission Error,在数据样本较少或者物种分布-不分布(Presence-absence)样本个数相差较大的情况下,这是不合理的;(3) AUC 值不能给出模型在整个研究区上的误差分布^[41]。基于此,有研究者认为应该用部分 AUC(Partial AUC, 简记为 PAUC)作为 AUC 的替代方案对物种分布模型的性能进行评价^[39]。在评价单一模型对研究物种空间分布的预测性能时,PAUC 是值得采用的办法,然而,若要比多个模型的性能,则需要采用其他的评价手段,例如赤池信息量准则(Akaike Information Criterion, AIC)^[32]。该准则可被用来寻找能够最好地解释数据但包含最少自由参数的模型^[42]。在一般情况下,若模型的误差服从独立的正态分布,则 AIC 可以表示为: $AIC = 2k - k \ln(L)$, 式中, k 为环境变量的个数, L 为似然函数。在表 1 中,以 $p(i, j)$ 表示元素 (i, j) 发生的概率,则 $a/N, b/N, c/N, d/N$ 为元素的实现值, $N = \sum \sum n(i, j) = a + b + c + d$ 为样本总数,若以 $p(i, j)$ 为参数,则似然函数的对数可表示为: $l = \sum_{i=1}^2 \sum_{j=1}^2 n(i, j) \ln p(i, j)$ 。在条件 $\sum \sum p(i, j) = 1$ 下,4 个元素 $p(i, j)$ 中有 3 个是可自由变动的,且 l 的最大值可令 $p(i, j) = n(i, j)/N$ 得到。那么预测结果的 AIC 即为: $AIC = (-2) \sum_{i=1}^2 \sum_{j=1}^2 n(i, j) \ln \{n(i, j)/N\} + 2 \times 3$ ^[43]。

5 结论

本文回顾了应用较为广泛的物种分布模型、其性能表现以及现阶段对这些模型的指数评价方法。从文献回顾可知,Maxent、GLM、GAM、GARP 以及 MD 模型的性能相对优异,而 HABITAT、DOMAIN 和 ENFA 模型的性能相对较差。物种分布模型的进一步发展,应考虑增加避免过度拟合及欠拟合的模块,变量选择模块、生物变量的空间表达模块以及物种对新环境变量的适应性机制模块,以实现物种空间分布的可信估计。关于对物种性能的评价,现阶段被广泛应用的 AUC 方法存在缺陷,相关研究中可考虑用 PAUC 作为替代方案,也可

采用 AIC 指数进行评价。另外,在对各种评价指数的计算中,缺乏对它们的标准差以及置信区间的估计,在进一步的研究中,可通过 Z-检验(或 Monte-Carlo 检验)的办法获得各指数的显著性检验。然而,由于 Z-检验(或 Monte-Carlo 检验)的实行要求数据符合渐进分布,这就意味着在野外工作中要获取大量的调查数据。因此,能够依据有限的样本对模型的性能进行评价的指数和方法是有价值的。此外,本文列举的各种参数之间也存在相关性,因此,对它们之间变化关系的研究,也是值得考虑的问题。

参考文献(References):

- [1] Hutchinson G E. The Niche: an Abstractly Inhabited Hyper volume. The Ecological Theatre and the Evolutionary Play. New Haven: Yale University Press, 1995.
- [2] 蒋霞,倪健. 西北干旱区 10 种荒漠植物地理分布与大气候的关系及其可能潜在分布区的估测. 植物生态学报, 2005, 29(1): 98-107.
- [3] Anderson R P, Raza A. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. Journal of Biogeography, 2010, 37(7): 1378-1393.
- [4] Elith J, Graham C H, Anderson R P, Dudík M, Ferrier S, Guisan A, Hijmans R J, Huetmann F, Leathwick J R, Lehmann A, Li J, Lohmann L G, Loiselle B A, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton J M, Peterson A T, Phillips S J, Richardson K S, Scachetti-Pereira R, Schapire R E, Soberón J, Williams S, Wisz M S, Zimmermann N E. Novel methods improve prediction of species' distributions from occurrence data. Ecography, 2006, 29(2): 129-151.
- [5] 翟天庆,李欣海. 用组合模型综合比较的方法分析气候变化对朱鹮潜在生境的影响. 生态学报, 2012, 32(8): 2361-2370.
- [6] 冷文芳,贺红士,布仁仓,胡远满. 中国东北落叶松属 3 种植物潜在分布对气候变化的敏感性分析. 植物生态学报, 2007, 31(5): 825-833.
- [7] 张雷,刘世荣,孙鹏森,王同立. 气候变化对马尾松潜在分布影响预估的多模型比较. 植物生态学报, 2011, 35(11): 1091-1105.
- [8] 张志东,臧润国. 海南岛霸王岭热带天然林景观中主要木本植物关键种的潜在分布. 植物生态学报, 2007, 31(6): 1079-1091.
- [9] Svenning J C, Skov F. The relative roles of environment and history as controls of tree species composition and richness in Europe. Journal of Biogeography, 2005, 32(6): 1019-1033.
- [10] Larson E R, Olden J D. Using avatar species to model the potential distribution of emerging invaders. Global Ecology and Biogeography, 2012, 21(11): 1114-1125.
- [11] Xu Z L, Zhao C Y, Feng Z D. Species distribution models to estimate the deforested area of *Picea crassifolia* in arid region recently protected: Qilian Mts. national nature reserve (China). Polish Journal of Ecology, 2012, 60(3): 515-524.
- [12] Busby J R. BIOCLIM —A Bioclimate Analysis and Prediction System. Nature Conservation: Cost effective biological surveys and data analysis. Melbourne: CSIRO, 1991: 64-68.
- [13] Walker P A, Cocks K D. HABITAT: a procedure for modeling a disjoint environmental envelope for a plant or animal species. Global Ecology and Biogeography Letters, 1991, 1(4): 108-118.
- [14] Carpenter G, Gillison A N, Winter J. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity and Conservation, 1993, 2(6): 667-680.
- [15] Hirzel A H, Hausser J, Chessel D, Perrin N. Ecological niche factor analysis: How to compute habitat-suitability maps without absence data? Ecology, 2002, 83(7): 2027-2036.
- [16] Farber O, Kadmon R. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. Ecological Modelling, 2003, 160(1/2): 115-130.
- [17] 许仲林,赵传燕,冯兆东. 祁连山青海云杉林物种分布模型与变量相异指数. 兰州大学学报: 自然科学版, 2011, 47(4): 55-63.
- [18] Zhao C Y, Nan Z R, Cheng G D, Zhang J H, Feng Z D. GIS-assisted modelling of the spatial distribution of Qinghai spruce (*Picea crassifolia*) in the Qilian Mountains, northwestern China based on biophysical parameters. Ecological Modelling, 2006, 191(3/4): 487-500.
- [19] Phillips S J, Anderson R P, Schapire R E. Maximum entropy modeling of species geographic distribution. Ecological Modelling, 2006, 190(3): 231-259.
- [20] 马松梅,张明理,张宏祥,孟宏虎,陈曦. 利用最大熵模型和规则集遗传算法模型预测孑遗植物裸果木的潜在地理分布及格局. 植物生态学报, 2010, 34(11): 1327-1335.
- [21] 张颖,李君,林蔚,强胜. 基于最大熵生态位元模型的入侵杂草春飞蓬在中国潜在分布区的预测. 应用生态学报, 2012, 22(11): 2970-2976.
- [22] 段居琦,周广胜. 我国单季稻种植区的气候适宜性. 应用生态学报, 2012, 23(2): 426-432.

- [23] Xu Z L, Zhao C Y, Feng Z D, Peng H H, Wang C. The impact of climate change on potential distribution of species in semi-arid region: a case study of Qinghai spruce (*Picea crassifolia*) in Qilian Mountain, Gansu province, China // Proceeding of the 2009 IEEE International Geoscience and Remote Sensing Symposium. Cape Town: IEEE, 3: III412-III415.
- [24] Xu Z L, Zhao C Y, Feng Z D, Zhang F, Sher H, Wang C, Peng H H, Wang Y, Zhao Y, Wang Y, Peng S Z, Zheng X L. Estimating realized and potential carbon storage benefits from reforestation and afforestation under climate change: a case study of the Qinghai spruce forests in the Qilian Mountains, northwestern China. *Mitigation and Adaptation Strategies for Global Change*, 2013, 18(8): 1257-1268.
- [25] 戚鹏程. 基于 GIS 的陇西黄土高原落叶阔叶林潜在分布及潜在净初级生产力的模拟研究 [D]. 兰州: 兰州大学, 2009.
- [26] 朱源, 康慕谊. 排序和广义线性模型与广义可加模型在植物种与环境关系研究中的应用. *生态学杂志*, 2005, 24(7): 807-811.
- [27] 温仲明, 赫晓慧, 焦峰, 焦菊英. 延河流域本氏针茅(*Stipa bungeana*)分布预测——广义相加模型及其应用. *生态学报*, 2008, 28(1): 192-201.
- [28] Yee T W, Mitchell N D. Generalized additive models in plant ecology. *Journal of Vegetation Science*, 1991, 2(5): 587-602.
- [29] Li S C, Gao J B. Prediction of spatial distribution of *Eupatorium adenophorum* sprengel based on GARP model: A case study in longitudinal range-gorge region of Yunnan Province. *Chinese Journal of Ecology*, 2008, 27(9): 1531-1536.
- [30] Tsoar A, Allouche O, Steinitz O, Rotem D, Kadmon R. A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, 2007, 13(4): 397-405.
- [31] Aguirre-Gutiérrez J, Carvalheiro L G, Polce C, van Loon E E, Raes N, Reemer M, Biesmeijer J C. Fit-for-Purpose: Species Distribution Model Performance Depends on Evaluation Criteria - Dutch Hoverflies as a Case Study. *PLoS One*, 2013, 8(5): e63708.
- [32] Liu C, White M, Newell G. Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography*, 2011, 34(2): 232-243.
- [33] Riddle D L, Stratford P W. Interpreting validity indexes for diagnostic tests: an illustration using the Berg balance test. *Physical Therapy*, 1999, 79(10): 939-948.
- [34] Fielding A H, Bell J F. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 1997, 24(1): 38-49.
- [35] Manel S, Williams H C, Ormerod S J. Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 2001, 38(5): 921-931.
- [36] Raes N, ter Steege H. A null-model for significance testing of presence-only species distribution models. *Ecography*, 2007, 30(5): 727-736.
- [37] Tate R F. Correlation between a discrete and a continuous variable: point-biserial correlation. *The Annals of Mathematical Statistics*, 1954, 25(3): 603-607.
- [38] Li W K, Guo Q H. How to assess the prediction accuracy of species presence-absence models without absence data? *Ecography*, 2013, 36(7): 788-799.
- [39] Slater H, Michael E, Baylis M. Predicting the Current and Future Potential Distributions of Lymphatic Filariasis in Africa Using Maximum Entropy Ecological Niche Modelling. *PloS ONE*, 2012, 7(2), e32202.
- [40] Godsoe W, Harmon L J. How do species interactions affect species distribution models? *Ecography*, 2012, 35(9): 8111-820.
- [41] Lobo J M, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 2008, 17(2): 145-151.
- [42] Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, 19(6): 716-723.
- [43] Warren D L, Seifert S N. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Application*, 2011, 21(2): 335-342.