应用生态学报 2019年6月 第30卷 第6期 http://www.cjae.net
Chinese Journal of Applied Ecology, Jun. 2019, **30**(6): 2116-2128 DOI: 10.13287/j.1001-9332.201906.029

最大熵模型在物种分布预测中的优化

孔维尧[1,2] 李欣海[3] 邹红菲[1*]
([1]东北林业大学野生动物资源学院, 哈尔滨 150040; [2]吉林省林业科学研究院/长白山动物资源与生物多样性重点实验室, 长春 130033; [3]中国科学院动物研究所, 北京 100101)

摘 要 最大熵模型在物种分布的预测研究中得到广泛应用,但未经优化的模型的预测结果可能存在严重的拟合偏差.本文汇总了最大熵模型在取样偏差修正、模型复杂性调整、物种分布判定阈值选择以及模型检验过程中的若干优化方法.在取样偏差的修正中,空间筛除法的修正效果最好,而背景限制法表现不佳.模型复杂性受建模变量的数量、函数模式和调控系数的影响.在样本量小于建模变量的数量时需进行变量筛选,筛选标准应侧重其生态学意义,而非变量间的相关性;函数模式对模型表现影响不大,在预测结果相近情况下应选择简单模型;建模时需要调整调控系数以控制过度拟合,一般最优模型调控系数高于默认值.判定物种出现阈值应遵从客观性、等效性和判别力3个原则,敏感度和特异性加和最大是良好的阈值判定标准.模型检验可分为不依赖阈值的检验和依赖阈值的检验,在不依赖阈值的模型评估方法中,基于信息标准选择的模型表现优于基于AUC或相关系数(COR)选择的模型;在基于阈值的模型评估方法中,真实技能统计能够兼顾模型遗漏误差和错判误差,不受假设缺失影响,且受物种流行度的影响较小.

关键词 最大熵模型; 取样偏差; 模型复杂性; 出现-不出现阈值; 模型检验

Optimizing MaxEnt model in the prediction of species distribution. KONG Wei-yao[1,2], LI Xin-hai[3], ZOU Hong-fei[1*] ([1] *College of Wildlife Resource, Northeast Forestry University, Harbin* 150040, *China*; [2] *Jilin Provincial Academy of Forestry Science/Jilin Provincial Key Laboratory of Wildlife and Biodiversity in Changbai Mountain, Changchun* 130033, *China*; [3] *Institute of Zoology, Chinese Academy of Sciences, Beijing* 100101, *China*).

Abstract: Maximum Entropy (MaxEnt) model has been widely used in recent years. However, MaxEnt is highly inclined to produce misleading results if it is not well optimized. We summarized the researches about the model optimization for sampling bias correction, model complexity tuning, presence-absence threshold selection, and model evaluation. Spatial filtering performs best for sampling bias correction, while restricted background method shows the lowest efficacy. Model complexity is mainly determined by three factors: The number of environmental variables, model feature types, and regularization multiplier. Variables filtering is needed when sample size is less than the number of environment variables. The criterion of variables selection should focus on their ecological significance rather than the co-linearity between them. The choice of feature types has relatively limited effects on predictive performance of the model, therefore it is advised to choose simpler models. To control overfitting, it is necessary to conduct species-specific tuning on regularization multiplier, which was usually bigger than the default setting. There are three criteria called objectivity, equality and discriminability for selecting threshold to convert continuous predication (*e.g.* probability of presence) into binary results. Maximizing the sum of sensitivity and specificity is a sound method for threshold selection. Model evaluation methods could be classified into two main types: Threshold-independent and threshold-dependent. Among the threshold-independent evaluations, information criteria may offer significant advantages over AUC and COR. True Skill Statistics is a better index

本文由吉林省财政厅公益项目(GY-2017-08)、吉林省重点实验室项目(20170622017JC)资助 This work was supported by the Public Welfare Project of Jilin Provincial Finance Department (GY-2017-08) and the Key Laboratory Foundation of Jilin Province (20170622017JC).
2018-12-18 Received, 2019-03-18 Accepted.
* 通讯作者 Corresponding author. E-mail: hongfeizou@163.com

for threshold-dependent evaluations, because it takes both omission and commission errors into account, and is robust to pseudo-absence assumption and species prevalence.

Key words: MaxEnt model; sampling bias; model complexity; presence-absence threshold; model.

最大熵(MaxEnt)是根据部分信息评估分布概率的技术[1].最大熵模型在2004开始应用于物种分布模型(SDM).它假设物种的分布点是根据物种分布的概率p确定的,这个概率可以通过物种分布点的环境因子对物种的约束性计算出来[2-3].MaxEnt在所有满足约束条件的模型中,选择熵值最大(即分布最均匀)的结果预测物种分布[4].最大熵模型只需要获取物种出现点的位置信息和环境背景数据,从而将研究者从判断不出现位点可靠性的困扰中解脱出来[5-6].最大熵模型还可以有效处理变量间的复杂交互关系[7],对小样本、非规则取样、有少量位点偏差的数据耐受度更高,模型预测表现优异[8-10],MaxEnt软件的用户友好型界面也极大地简化了建模程序.由于最大熵模型的上述优点,再加上全球生物多样性信息数据(GBIF)等开放数据平台的增多,最大熵模型迅速在生态学、进化生物学、保护生物学和生物安全等众多研究领域得到广泛应用[7,11].中国知网检索结果表明,我国采用最大熵进行物种分布的研究自2012年显著增加,截止2018年11月已发表论文750篇(图1).

作为一种方便易用的模型,使用者常常对MaxEnt模型表现很满意,不加分析地采纳模型结果,而忽略模型优化的作用[12].未经优化模型的预测结果可能存在严重的拟合偏差,此类偏差不仅会导致对物种生态位的错误评估,还会对保护管理政策的制定造成误导.虽然MaxEnt软件中提供一些模型优化的选项,但相关参数的设置并没有一个清晰明确的标准[13].本文系统总结了近年来有关MaxEnt模型的取样偏差修正、模型复杂性调整、物种分布判定阈值选择、模型表现评估等方面的研究,汇总了模型优化以及相关参数选择的方法,以期为最大熵模型在物种分布研究中的应用提供借鉴.

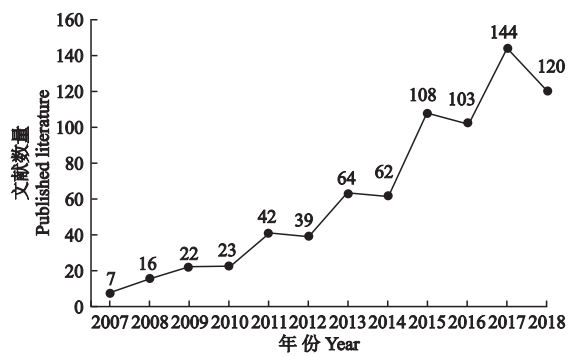

图1　中国知网检索最大熵模型研究文献数量(截至2018年11月15日)

Fig. 1　Statistics of published literatures of MaxEnt model research on CNKI (update to Nov. 15th 2018).

1　取样偏差的修正

1.1　取样偏差的影响

与所有的物种分布模型一样,MaxEnt模型建模的一个基本假设是研究区域都经过了系统或随机的无偏采样[4].但是,无偏取样由于成本高昂等问题往往很难实现[14],样本经常取自容易到达的区域,如城镇、道路、河流沿线,从而导致不同程度的取样偏差[15].Yackulic等[16]发现,以往87%的MaxEnt模型研究采用了容易导致取样偏差的数据.取样偏差会导致模型过度拟合,从而降低模型的预测能力[17-19],而相比于出现-不出现点模型,只采用物种出现位点和背景点建模的MaxEnt模型更容易受到取样偏差的影响[4,20].目前,对取样偏差的评估和修正已成为MaxEnt模型应用与研究中最受关注的热点领域.

1.2　基于物种出现点的取样偏差校正

取样偏差的修正主要包括对出现位点的处理和对背景位点的处理,处理的依据主要分为根据位点的地理空间属性以及位点对应的环境变量的属性两类.取样的空间偏差导致的取样偏差是建模者最关心的问题[4],而依据环境变量的属性进行偏差修正更符合最大熵的统计学原理[7].在对出现位点的偏差修正方法中,最常用且易于理解的方法是空间筛除法(spatial filtering,方法1),即在一定距离内只保留有限位点[13,21],相当于在出现位点中按照均匀、随机的原则再次进行系统取样.因此,一些学者也称其为系统重取样法(systematic sampling)[10].空间筛除法的一个问题是空间间距的设定很难确定标准[22-23],间距的设定应与环境变量在空间尺度上的变异程度相符.此外,由于环境变量在空间分布的异质性,地理位置上的均匀取样并不能保证其他环境变量的均匀分布.Varela等[24]采用聚类分组筛选法(cluster,方法2)修正取样偏差,根据出现点的不同变量特征进行分类,然后在每一类中随机选取一定

数量的位点作为建模数据.以上处理方式均在不同层面满足了系统随机取样原则,但都要去除一定数量的样本,由此可能导致建模数据不足[9],样本量的减少也会造成模型预测能力的降低[12].此外,物种密度较高的分布区会因出现位点的筛除而在建模过程中对该区域生态价值造成低估[10].分组建模法[25-26](split,方法 3)回避了数据筛选问题,对于由于取样强度不同造成出现概率产生显著差异的不同区域,将研究区域分组,使每个区域内物种出现概率相近,然后对每个区域内的数据分别建模.Dudík 等[14]在机械学习理论基础上,提出了补充样本数据去除取样偏差的去偏平均法(debiasing averages,方法 4),即根据取样点的分布概率获取样本点对数损失最小的无偏置信区间,在此区间内增加样本点使样本接近最优吉布森分布("best" Gibbs distribution).

1.3 基于背景点或致偏变量的取样偏差校正

去除取样偏差的另一个思路是对背景位点进行处理,使随机选取的背景位点与取样点具有同样的偏差[14,27],由此建模过程不再关注样本取样偏差,而是关注物种出现位点与随机选择背景位点的环境差异[4].根据上述原理,Phillips 和 Dudík[27]采用目标组物种背景法(target-group background,方法 5)定义背景位点.目标组物种是采用与建模物种相同的取样方法记录到的所有物种,以目标组物种记录点作为背景位点进行建模.由于该方法是以物种记录分布代替取样点分布,不需要获取取样点信息,避免了目前大多数物种信息数据库取样方法缺失对建模的限制.目标组背景法的一个严格假设是目标组物种是采用相同方法、相同设备进行记录的[4],不同物种的记录偏差也相同[28],且不同物种在所有位点被记录的几率相同[16].除了上述严格限制,目标组背景法的另一个问题是容易产生因物种丰富度差异引起的偏差,因此,该方法在物种丰富度空间分布均匀的区域效果较好,但在空间分布变异大的情况下会表现不佳[28-29].此外,目标组背景法也没有区分不适宜生境和未取样生境的差别[29].在目标组物种无法确定的情况下,Fourcade 等[10]采用背景限制法(restricted background,方法 6),即根据物种出现位点绘制缓冲区,在缓冲区内随机选取背景位点.此方法仅限制了背景位点选择的范围,并未考虑背景位点与取样点分布是否一致.针对这一问题,MaxEnt 软件提供了偏差栅格文件(bias file)的选项,软件根据使用者提供的体现取样点分布的栅格图层,生成与取样点布里格曼分歧(Bregman divergence)最小的背景位点进行建模[14].偏差栅格文件需要掌握取样点分布,在取样点未知的情况下,可以采用出现点的核密度分布概率估计取样点分布[30],通过核密度偏差栅格文件修正取样偏差(方法 7)[10].偏差栅格文件同样可以改善目标组背景法无法区分物种未出现点和未采样点的缺陷.Warren 等[29]根据目标组物种位点的密度(density bias estimate,方法 8)、栅格距离目标组物种位点的距离(distance bias estimate,方法 9)以及目标组物种位点对应引发取样偏差的环境变量(如道路、河流、坡度)的函数(behavioural bias estimate,方法 10)制作偏差栅格文件,估计取样偏差的分布.与方法 10 思路类似的一种校正取样偏差的方法是致偏变量修正法(bias covariates correction),但处理的对象从背景位点变为导致产生取样偏差的环境变量,将相关变量的影响调整到同一水平.最直接的修正方法是剔除致偏变量(方法 11)[31-32]或将导致产生取样偏差的变量统一设置为最小值(方法 12)[32].考虑到物种出现记录的实际分布,也可以建立出现位点频率对应致偏变量的函数模型,然后将预测值统一设为最大或最小的预测频率(方法 13)[32].

1.4 不同方法偏差校正效果的比较

在对取样偏差处理后,大多数情况下模型表现会更好,降低模型的过度拟合[4,28].但这一结果在减少遗漏误差的同时,增加了错判误差的风险.因此,对于未采样区域的预测结果一定要谨慎对待[32].从偏差修正效果来看,通过筛除物种出现数据的方法总体优于调整背景位点的方法,但通过增加取样点修正偏差的去偏平均法表现不及对背景位点的调整[12].Kramer-Schadt 等[33]比较了方法 1 和方法 7 的差异,结果表明,进行数据空间筛除的模型在降低遗漏误差和错判误差方面的表现均优于通过偏差栅格文件调整背景位点的模型.Fourcade 等[10]以 AUC(受试者工作特征曲线以下面积 Area under the receiver-operator-curve)和降低过度拟合而引起的预测分布范围增加程度为标准,比较了方法 1、2、3、6、7 对不同原因造成的取样偏差的修正效果,背景限制法(方法 6)修正偏差效果最差,而空间筛除法(方法 1)略优于方法 2、3、7.分组建模法(方法 3),对因不同区域采样强度差异引起的取样偏差修正效果更好,核密度偏差栅格文件法(方法 7)则在修正预测分布范围方面表现较好.Warren 等[29]研究显示,在对模型复杂性进行控制的前提下,方法 8、9、10 与未

表 1　13 种取样偏差校正方法的比较
Table 1　Comparison of 13 sampling bias correction methods

类别 Category	序号 Code	处理方法 Approach	去偏效果 Correction effect	局限性 Limitation
对物种出现点的处理	1	空间筛除	好	需要选择间距标准,损失原始数据,生态价值低估
Adjustment for	2	聚类分组筛选	较好	损失原始数据,生态价值低估
occurrence points	3	分组建模	较好	组内数据存在取样偏差
	4	去偏平均	较差	需要取样点数据
对背景点的处理	5	目标组背景	较好	未区分不出现点和未取样点,存在物种丰富度偏差
Adjustment for	6	背景限制	差	未考虑取样点分布
background	7	核密度偏差栅格文件	较好	对取样点分布的估计存在误差
	8	物种密度偏差栅格文件	与 9、10 等效	同 7
	9	距离偏差栅格文件	–	同 7
	10	变量函数偏差栅格文件	–	同 7
对致偏环境变量的处理	11	剔除致偏变量	–	低估致偏变量生态学意义
Adjustment for	12	致偏变量最小化	优于 11、13	同 11
environment variables	13	统一致偏变量函数	优于 11	同 11

经偏差调整的模型表现差异不大,但可以修正MaxEnt默认模型(未进行模型复杂性和取样偏差处理)的过度拟合.有关模型复杂性对取样偏差的影响详见本文第二节.El-Gabbas 等[32] 比较了方法 11、12、13 修正偏差的效果,将致偏变量设为最小值(方法 12)的效果最好,直接剔除致偏变量(方法 11)效果最差.上述 3 种方法在进行取样偏差修正时,要求取样偏差与其他环境变量无交互作用,但 MaxEnt 软件并没有具体控制交互作用的功能,特别是 Product feature 会考虑所有因子的交互作用.El-Gabbas 等[32] 建议在进行偏差修正时关闭 product feature,但采用此方法对模型的影响并不大.不同偏差校正方法的效果和局限性见表 1.

对取样偏差的修正并没有普适性的指导原则,需要根据导致取样偏差的原因以及校正偏差的表现选择适宜的修正方法.不同的修正方法并不是相互孤立的,实际应用中可尝试多种方法的结合[10,21].模型去偏修正的效果与物种的分布特征也有很大关系,狭域分布物种对取样偏差不敏感[10].这是因为狭域分布物种本身即存在空间分布上的偏差,另一方面,也是因为狭域分布格局对预测误差的耐受程度更高[5].有关物种分布格局对模型预测效果的影响详见本文第四节.

2　模型复杂性的调整

2.1　模型复杂性的选择

模型复杂性是物种分布模型研究关注的另一个重点领域,一个优良模型既应该很好地拟合数据,也要对不必要的复杂函数关系进行约束[12].作为一种算法模型,MaxEnt 倾向于建立非常复杂的函数关系以拟合观测数据.这也是最大熵模型预测表现较好的重要原因[5].但是,模型拟合复杂关系的能力本身并没有意义,为了使模型更容易在生态上解释,模型复杂性常常需要平衡[34].在建模时应比较复杂模型和简单模型的表现,如果效果类似,还应当选择在生态上更容易理解的简单模型,特别是对于濒危物种,简单模型的结果更容易提出有针对性的保护管理意见[35-36].复杂模型的另一个问题是容易因对建模数据的过度拟合造成模型外推时预测能力不佳[7],特别是建模数据存在取样偏差的情况下,过度拟合的问题会随着模型复杂性的增加而更加严重[37].而简单模型不仅更容易从生态学上解释,还能避免数据采集地特殊环境造成的过度拟合[13].

2.2　环境变量对模型表现的影响

影响 MaxEnt 模型复杂性的 3 个主要因素包括建模的环境变量数量、函数模式(即软件中的 feature)以及调控系数(regularization multiplier).环境变量的数量首先取决于研究者选择建模的原始变量数量,更重要的是受 MaxEnt 软件中 Product 函数模式中考虑变量交互关系的新增变量的影响.对于MaxEnt这种能处理复杂变量关系的模型,一般情况下过参数化比缺参数化影响小[8,12,38].常用的一些变量筛选方法,例如通过相关性分析或主成分分析去除部分变量并不能使 MaxEnt 模型表现更好,更有可能降低模型的拟合与预测能力[7,27].只有在一种情况下才有必要减少变量的数量,即变量个数大于出现点数量.此时过参数化会造成过度拟合,因此较少参数建立的模型表现更好[12,33],但变量取舍的标准是因子对物种的生态学意义,而不是变量间的相关性或其他因素[7].Kramer-Schadt 等[33] 比较了去除

共线性的变量和未筛选变量建模的效果,在进行样本去偏处理后,环境变量的共线性不影响模型表现,采用筛选后数据模型 AUC 会轻微降低,受变量减少影响,重要因子的相对增益贡献升高,但重要因子的相对贡献变动不大.但如果采用存在取样偏差的数据,未筛选变量的模型容易产生过度拟合.虽然变量的共线性对模型影响微乎其微,但在对物种分布与变量间的响应关系进行解释时,需要注意某个变量对模型的贡献是否受到与其高度相关的其他变量的影响.存在共线性变量间的相互影响可以通过MaxEnt运算结果中的刀切法检验进行评估,如果某个变量对模型贡献主要受其他相关变量的影响,该变量剔除后模型相对增益贡献变化会很小[39].此标准也可应用于样本量不足且对物种生态需求不明确的情况下对变量的筛选[29].

2.3 函数模式对模型表现的影响

MaxEnt 软件提供了 6 种函数模式(feature),分别是线性函数(linear,L)、二次项函数(抛物线 quadratic,Q)、考虑变量交互作用(interaction)的交互函数(product,P)、常数分段(阈值)函数(threshold,T)、线性分段函数(hinge,H)和处理分类变量的分类函数(category,C)[7].其中对模型复杂性影响最大的是会生成新变量的交互函数以及线性分段函数.MaxEnt 软件在建模过程中会尝试上述函数模式的组合,选择 AUC 最高的模型进行预测,因此默认设置预测效果在一般情况下优于其他函数模式组合[36].在样本量小的情况下,MaxEnt 软件的默认模式会限制复杂函数模式的使用,L 函数一直在运行,Q 函数至少在 10 个样本以上开始运行,H 函数至少需要 15 个样本,运行 P 和 T 函数则需要 80 个以上的样本[7].Syfert 等[13] 比较了 L、Q、H、T、LQ、HQ、LQP、LQT、QHP、QHT、QHPT 以及软件默认函数模式的预测表现,结果表明,模型函数模式对预测能力影响不大,复杂模型会轻微增加 AUC,但简单模型比自动选择的模型的响应曲线更平缓,在生态学上也更容易解释.简单模型的一个问题是受取样偏差影响较大,Anderson 等[37] 研究表明,只采用 L 函数模式建模时,随着样本量增加,会因取样偏差的增大导致模型 AUC 降低,而使用 LQ 模型时样本量多的模型 AUC 更高.Syfert 等[13] 研究也表明,LQ 模型去偏处理后预测能力显著提高.在模型组合中增加线性分段函数(H)可以显著提高模型预测能力[23,27,38],但使用分段函数时线性函数(L)和常数分段函数(T)会表现的过于复杂.由于 L、T 是 H 模型的特殊形式,一个使模型响应更平滑的方法是只采用 H 函数模式[7].在使用不同调控系数的情况下,H 模型与 LQH 模型的变动趋势类似(图 2,数据来源于笔者),模型的遗漏误差和检测数据 AUC 整体差别不大[23].控制模型复杂性的另一个选项是关闭交互函数模式,但会损失模型对变量交互影响的拟合能力[7].

2.4 调控系数对模型表现的影响

通过调控系数调整模型复杂性是最大熵模型研究中的一个热点领域.MaxEnt 为了平衡模型拟合度与外推能力,设置了调控系数(regularization)对变量权重进行约束,使模型不必精确地拟合建模数据,而是为建模数据设定一个误差边际(error bound),模型根据包含误差边际的观测值进行拟合.调控系数的调整实质上是对误差边际范围的调整.从这种意义来看,MaxEnt 是对模型复杂性进行约束的最大似然模型[6].因此,调控系数越小,引发模型过度拟合的可能性越大,增加遗漏误差;而过高的调控系数会导致模型平滑,增加错判误差,降低模型对不适宜地区的判别能力[21,37].虽然MaxEnt软件基于对全球

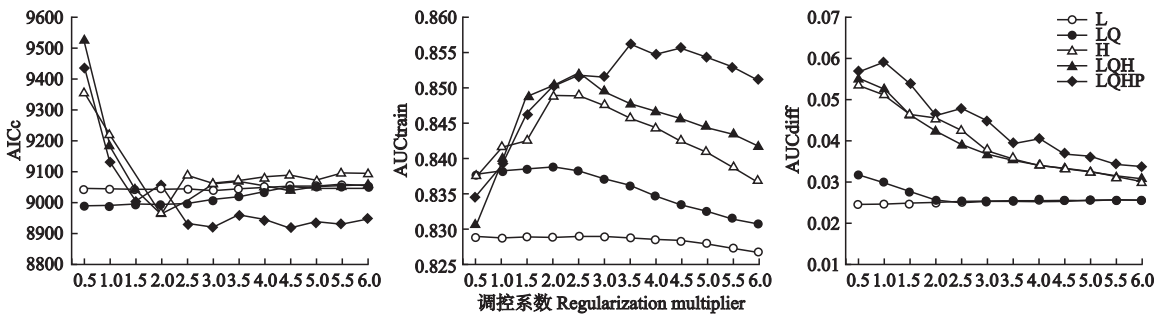

图 2 模型复杂性对模型表现的影响

Fig.2 Influence of model complexity on model performance.

RM: 调控系数 Regularization multiplier; AICc: 小样本赤池信息量准则 Akaike information criterion corrected for small sample size; AUCtrain: 建模数据 AUC AUC of training data; AUCdiff: 建模数据与检测数据 AUC 的差值 Differece between training and testing AUC.

6 个区域的 226 个物种的数据设置的默认调控系数在很大程度上降低了过度拟合的影响[27]，但很多研究表明，需要根据不同物种及其数据结构进行单独的调控系数选择[21,30,37-38,40]．过高或过低的调控系数都会导致建模数据和测试数据 AUC 降低[37]、二者 AUC 差距的增大以及遗漏误差的增加[21,40]（图 2）．相关研究表明，最优调控系数一般高于 MaxEnt 软件默认值 1，调控系数在 2～4 之间的模型综合表现最好，对于存在取样偏差的数据，调控系数应相应的增大[21,23,38,40]．调控系数对取样偏差校正的原理参照第一节中的去偏平均法（方法 4），调控系数增加时建模数据误差边际范围扩大，通过增加取样数据范围纠正偏差．应当注意的是，采用此原理由于对无偏平均值估计的误差，对取样偏差校正的效果不佳，仅在大样本模拟数据的情况下表现良好[14]，因此仅仅依靠调整调控系数并不能解决取样偏差问题[41]．适宜的调控系数也受函数模式影响，函数模式越多，数据吉布森分布越复杂，应相应地调高调控系数[27]．调控系数也隐含了对函数模式的选择，例如将一些函数模式的调控系数设为 0[7]．样本量也会影响调控系数的选择，数据量越小，估计区间越窄，因此需要调高调控系数以增加误差边际[2,27]．调控系数最终确定的误差边际与样本的标准误成正比[37]．这也是小样本数据需要调高调控系数的原因．Shcheglovitova 等[23]研究表明，对于小样本数据，复杂函数模式与较高的调控系数结合使用会使模型表现优于默认设置，但 Warren 等[29]的研究发现，样本量大小不影响最优调控系数的确定．

2.5　模型外推时模型复杂性的控制

在应用 MaxEnt 进行物种入侵、气候变化影响等需要进行模型外推的研究中，应当格外注重对模型复杂性的控制，以降低过度拟合[5,30]．在模型外推过程中，过参数化容易导致对栖息地适宜度的低估，而缺参数化则会导致高估[12,37]．Porfirio 等[42]研究气候变化对物种分布的影响时，采用全部变量、主成分分析变量、生态专家确定变量分别建模，结果表明，上述方法对现有数据拟合结果差别不大，但采用全部变量的模型 AUC 值最高．此外，不同物种对环境变化的耐受程度还受生活史的影响，对于生活史长的植物可采用较多变量以增加模型的拟合能力；对于生态幅较窄的动物则应减少变量以降低模型过度拟合．Warren 等[29]根据环境变量增益贡献（gain）以及变量间相关性进行变量筛除，发现减少环境变量会增加对栖息地适宜度变化的区别能力，消除具有相关性的变量会导致大部分物种生态幅变窄，并推断增益贡献高的限制性变量数量与物种生态幅预测结果呈负相关．Warren 等[29]的研究还发现，调高调控系数会导致预测结果生态幅增宽，从而避免对气候变化影响的极端评估．Moreno-Amat 等[38]在采用 MaxEnt 模型进行大跨度的时间尺度外推时也发现，简单函数模式、中等数量的环境变量和中等程度的调控系数的模型预测表现较好．

应当注意的是，不同物种对因预测变量、函数模式、调控系数的调整而产生的模型表现变化格局是不同的[23]，一些研究得出的模型复杂性影响很可能并不适用于目标物种，因此，研究者有必要针对目标物种进行独立的模型复杂性调整．

3　判定物种出现阈值的选择

3.1　阈值选择原则

MaxEnt 模型在应用于生物多样性评估、保护地设计、保护投入项目评估等内容时，需要将预测结果的连续变量转化为出现-不出现二元变量[42]，采用混淆矩阵对模型表现进行评估也需要确定判别阈值[2]．然而在研究中很难确定适宜的阈值选择标准[23,39]．Liu 等[44]提出了阈值选择的 3 个原则：1）客观性，即阈值根据客观数据求得；2）等效性，选择的阈值无论是基于出现-不出现数据还是仅用出现点的数据都具有相等的效力；3）判别力，阈值对出现-不出现点（而非出现-随机点）的判别能力达到最大化．很多研究直接选择一个估计概率（例如 0.5 或 0.8）作为判断物种出现阈值[45]．此方法武断而缺乏生态学依据，模型预测表现往往也很差[46]，本文不再进行讨论．

3.2　根据遗漏误差确定的阈值

MaxEnt 软件给出了 11 种判断阈值的方法，其中比较常用的 9 种方法可归纳为两类，即根据遗漏误差确定的阈值以及根据预测结果敏感度（sensitivity）和特异性（specificity）确定的阈值．第一类确定阈值的方法包括采用基于训练数据的最少存在率（minimum training presence，MTP），基于训练数据的 10%的存在率（10 percentile training presence，10P），固定累积结果值（fixed cumulative value）为 1、5 或 10（CV1、CV5、CV10）．MTP 阈值建模数据遗漏误差为 0，但对小样本非常敏感[37]；而在大样本情况下，受异常值影响则可能导致预测物种分布面积过大[2,33,47]．10P 阈值允许 10%的建模数据预测值小于阈值，避免了异常值的影响[21]，但同样面临小样本

表 2 9 种阈值选择方法比较

Table 2 Comparison of 9 presence-absence threshold selection methods

类别 Category	序号 Code	阈值判定标准 Threshold selection criteria	对小样本的稳健性 Robust to small sample	对异常值的稳健性 Robust to abnormal value	对流行度的稳健性 Robust to prevalence	客观性 Objectivity	等效性 Equality	判别力 Discriminability
根据遗漏误差确定的阈值 Threshold identified by ommision rate	1	基于训练数据的最少存在率 Minimum training presence	差	差	强	客观	是	较差
	2	基于训练数据的 10%的存在率 10 percentile training presence	差	强	强	较客观	是	较差
	3	固定累积结果值为 1 Fixed cumulative value was 1	强	强	强	客观	是	较差
	4	固定累积结果值为 5 Fixed cumulative value was 5	强	强	强	较客观	是	较差
	5	固定累积结果值为 10 Fixed cumulative value was 10	强	强	强	较客观	是	较差
根据敏感度和特异性确定的阈值 Threshold identified by sensitivity and specificity	6	建模数据敏感度与特异性相等 Equal sensitivity and specificity of training data	强	强	差	客观	否	较强
	7	检测数据敏感度与特异性相等 Equal sensitivity and specificity of testing data	强	强	差	客观	否	较强
	8	建模数据敏感度与特异性之和最大 Maximum sensitivity plus specificity of training data	强	强	较强	客观	是	强
	9	检测数据敏感度与特异性之和最大 Maximum sensitivity plus specificity of testing data	强	强	较强	客观	是	强

数据的局限,建模数据在低适宜度区域的不均衡分布也会影响阈值判定效果.固定累积结果值从遗漏误差的角度可解释为:如果选择某个物种分布位点所在栅格的固定累积结果值 c 作为判定物种出现的阈值,则该物种实际分布位点在模型中的遗漏误差为 c%[27].固定累积结果阈值回避了小样本以及建模数据不均衡分布的问题,但存在与 10P 阈值相同的问题,即设定的遗漏误差水平是否合理.根据遗漏误差确定阈值的方法一个共同的缺陷是没有考虑错判误差,不符合阈值选择的第三条原则,在模型评估时容易引起非对称误差[21].

3.3 根据敏感度和特异性确定的阈值

根据敏感度和特异性确定阈值综合考虑了模型的遗漏误差和错判误差[46].MaxEnt 软件给出了基于两种平衡策略的阈值判定方法,即使模型敏感度和特异性数值相等(equal sensitivity and specificity, ESS)或使二者之和最大(maximum sensitivity plus specificity, MSS).由于 MaxEnt 在计算预测结果的特异性时,将背景点作为假设缺失(pseudo absent)点而非使用真正的不出现点[44],放大了模型的错判误差,因此,此类模型选择标准受到一些学者的质疑[48].Liu 等[44]分析了包括 ESS、MSS 在内的 12 种确定阈值的方法,推导出 3 种不受假设缺失影响(即符合阈值选择第二条原则)的阈值标准:建模数据的流行度、大样本情况下一定量随机位点的平均预测值以及 MSS,但前两种阈值判定标准在物种流行度高时模型敏感度会降低,判别能力下降,不符合阈值选择第三条原则.MSS 同时满足阈值选择的 3 个原则,且受出现位点-背景点比例以及物种流行度影响较小,可降低流行度低的物种的遗漏误差和流行度高的物种的错判误差[47].MaxEnt 软件会根据建模数据和检测数据分别计算 MSS 和 ESS 阈值,在确定阈值时需要衡量建模数据和检测数据阈值的差距,如果二者之间差值较大,则意味着模型存在过度拟合或弥散.

3.4 阈值与栖息地适宜度的变化

很多进行环境变化影响的研究采用物种目前分布格局确定的阈值推测未来分布区的变化.这种方法值得商榷,因为栖息地适宜度的降低并不意味着物种的消失[29].这也是很多研究做出过于悲观预测的原因.在进行此类研究时,关注的重点应该是栖息地适宜度的变化,而不是分布区的扩大或消失.

4 模型预测能力的评估

4.1 检测数据的处理

4.1.1 检测数据的选择 模型表现的评估首先要明确的问题是检测数据的选择.对于所有的物种分布模型,最优方案是选择与建模数据相独立的通过系统采样法获得的出现-不出现数据[4,8,32].但实际上此类数据很难获得,除了成本问题,检测数据偏差的

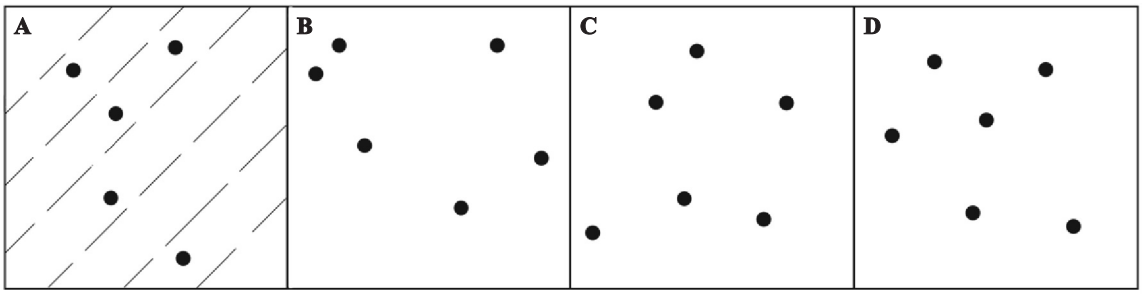

图 3　地理结构分区法与屏蔽地理结构分区法

Fig.3　Geographically structured and masked geographically structured methods.

A 区出现点留作检验，B、C、D 组出现点建模.二者的区别是前者背景位点在 A、B、C、D 中选取，后者仅在 B、C、D 中选取 They both use present points in B, C and D regions for model training and present points in A region for testing. The difference is that the former uses background points from all the 4 regions while the latter only use background points from B, C and D regions. (Imitated Radosavljevic, *et al.*[21]).

去除也非常困难[14,49].

4.1.2 检测数据的取样偏差校正　与出现位点的取样偏差相比，出现-不出现点数据的“空间分类偏差”往往被研究者忽略.模型检验的不出现点距离建模出现点越远，其预测值越低，此类偏差容易引起模型评估膨胀[8,43,50].Hijmans[51]建议采用配对等距法对模型检测的出现、不出现点进行选择，即评估模型的测试出现点、不出现点与建模出现点距离相近.“空间分类偏差”是否消除可采用“物种分布空白模型”(SDM null model，即模型只采用与建模出现点距离的倒数做变量，不考虑其他任何环境变量)的 AUC 值进行评估，消除分类偏差的测试点在空白模型中的 AUC 值应接近 0.5.

4.1.3 假设缺失点在模型检验中的应用　出现-不出现数据更难处理的问题是不出现点观测误差，在很多情况下预测物种，特别是动物的适宜栖息地往往因为观测遗漏而被错判[52-53].这也是很多研究者选用 MaxEnt 建模的原因.在缺乏可靠的不出现点检验数据时，可采用 MaxEnt 建模时的策略，在进行模型检验时同样只使用出现位点，将背景位点视为假设缺失点计算模型 AUC、COR 等指标.此时模型检验的实质从对出现-不出现位点的判别能力转化为出现-随机位点的判别[2].

4.1.4 交叉验证数据的分组　为保证检测数据的独立性，应选择与建模数据不同时间、不同方法(如不同的调查样线)获取的数据.在此类数据不可得的情况下，交叉验证是一种常用的替代手段[32].MaxEnt 软件可采用随机分组将出现点数据分为建模数据和检验数据，并通过迭代计算提高模型表现.但随机分组无法保证建模数据和检验数据的独立性，因此往往会高估模型预测能力[37].Radosavljevic 等[21]采用地理结构分区法(geographically structured)，将研究区分成 k 个区域以保证每个区域物种出现点相互独立(只有两个分区相邻区域出现点存在空间相关可能)，采用 k-1 组出现点建模，剩下 1 组检验.此方法用作检验的区域物种出现点未参与建模，但背景位点参与了建模过程，从而产生了取样偏差.Radosavljevic 等[21]针对此问题提出了屏蔽地理结构分区法(masked geographically structured，图 3)，即在建模过程中屏蔽检测数据区域，只有建模数据的背景位点进入模型.此方法可有效提高训练点和测试点 AUC 值，降低二者之间差值以及遗漏误差.与屏蔽地理结构分区法类似的分组方法还有棋盘格分组法(checkerboard grid)[32,40]，将研究区分为 k 组棋盘格状分布的区域(图 4)，将出现点和背景点分为 k 组进行交叉验证.棋盘格分组法的优点是不受研究区域面积限制，可对小范围区域内数据进行分组，组间数据的分布更均匀.应当注意的是，交叉验证只是在没有独立数据可用情况下的替代方案，特别是采用迭代运算时，交叉验证仍有可能低估模型过度拟

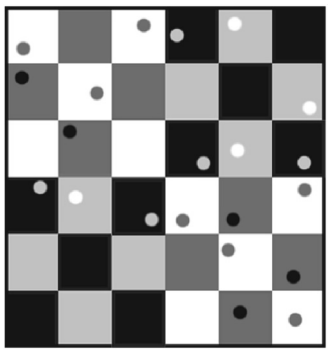

图 4　棋盘格分组法

Fig.4　Checkerboard grid method.

该方法将研究区划分到 k 种颜色棋盘格中，每种颜色棋盘格中的出现点(圆点)和背景点划分为 1 组独立数据 Checkerboard grid divides research area into k bins showing with different colors. The present point (circle points) and background point in each bin are independent to other bins (Imitated Muscarella, *et al.*[40]).

合程度.

在物种出现点样本量过小不足以支持数据分组的情况下,可采用刀切法交叉验证[23],将 1 个出现点列为测试数据,其余数据建模并进行迭代运算. Warren 和 Seifert[12]采用的先建模、再从结果中选取建模和检验数据的方法也是处理小样本数据的一种选择.在研究气候变化影响时,物种未来分布情况是不可得的,可采用历史记录数据检验模型在时间尺度上的外推能力[38].

4.2 模型判别能力的评估

4.2.1 不依赖阈值的模型表现评估 模型检验的方法可分为不依赖阈值评估和基于阈值评估 2 类.第一类评估方法较为常用的包括 AUC、COR、AIC 和 BIC.AUC 是对出现-不出现点或出现-背景点预测结果的秩检验,用于评估出现点分布概率高于不出现点或背景点(假设缺失点)的程度.虽然 AUC 是评估生态位模型表现最常用的方法[54],但其检验效力受到很多学者的质疑.AUC 检验应关注独立检测数据的 AUC 值(AUCtest),而不是建模数据(AUCtrain)[8].未经去偏处理的数据会因过度拟合而得出较高的 AUCtrain,但模型对测试数据判别能力很低;AUCtest 与 AUCtrain 的差值(AUCdiff)可以反映模型是否存在过度拟合,前提是确保建模数据与测试数据之间的相对独立,否则也会存在模型评估膨胀的情况[21].COR(correlation)用于评估预测结果与出现-不出现(或假设缺失)位点 1-0 数据的相关性.与秩检验相比,相关性检验对建模数据的采样强度更为敏感[4].COR 与 AUC 检验的另一个区别是 MaxEnt 给出的 3 种形式的预测值(raw,cumulative,logistic)采用 COR 评估效力是不同的,logistic 结果相关性最高,raw 结果最低.基于信息标准的模型检验与调控系数的原则类似[27],是对模型复杂性和拟合度的综合考评.此类方法常用于数据模型的检验,但 Warren 和 Seifert[12]研究表明,采用小样本赤池信息量准则(Akaike information criterion corrected for small sample size,AICc)和贝叶斯信息准则(bayesian information criteria,BIC)选择的最优模型在模型简单程度、对适宜栖息地判定、栖息地等级预测、环境变量对栖息地适宜度重要性的评估,以及模型在时间尺度上的外推方面普遍强于采用 AUC 标准选择模型.只有大样本的情况下,采用 AUCdiff 标准选择的最优模型在模型简洁程度方面高于信息标准确定的模型.这也是 BIC 标准表现强于 AICc 标准以及 AUCdiff 标准强于 AUCtest 标准的唯一情况.而采用 AUC 标准选择的模型中,AUC-test的表现普遍优于 AUCdiff,表现最差的是 AUCtrain,AUCtrain 仅在小样本情况下,在适宜栖息地判定和环境变量重要性评估方面强于表现最差的 AUCdiff 标准.Warren 和 Seifert[12]研究还发现,基于 AICc 选择的模型 AUCtest 会略低于 MaxEnt 默认参数模型,但这种判别能力的损失可能与研究采用的检测数据并非真正独立有关.

4.2.2 基于阈值的模型表现评估 基于阈值的模型评估首先需确定物种出现概率的阈值,将预测结果转化为 1-0 二元变量,阈值选择参见本文第 3 节. MaxEnt 根据 11 种判定阈值方法分别计算了检验出现位点相应的遗漏误差(模型错误预测出现点概率)以评估模型预测表现[2].模型敏感度是模型正确预测出现点的概率,遗漏误差与敏感度之和为 1.遗漏误差和敏感度只需要物种出现位点数据,但忽略了错判误差.评估错判误差的一个指标是模型特异性,即模型正确预测不出现位点的概率,但此方法没有考虑遗漏误差(表3).为了避免模型评估非对称

表 3 模型预测能力评估方法的比较

Table 3 Comparison of model evaluation methods

类别 Category	序号 Code	模型评估方法 Model evaluation approach	受物种流行度影响 Influenced by prevalence	受假设缺失影响 Influenced by pseudo-absence assumption	遗漏误差评估能力 Omission evaluate capacity	错判误差评估能力 Commission evaluate capacity
不依赖阈值的模型评估	1	赤池信息准则 AICc	较强	较强	强	较差
Threshold independent	2	贝叶斯信息准则 BIC	较强	较强	强	较差
model evaluation	3	建模数据 AUC AUCtrain	强	较小	差	差
	4	检测数据 AUC AUCtest	较强	较强	较强	较差
	5	建模、检测数据 AUC 差值 AUCdiff	较强	较强	较强	较差
	6	相关性 COR	强	较强	较强	较差
基于阈值的模型评估	7	模型敏感度 Sensitibity	很小	无	强	无
Threshold dependent	8	模型特异性 Specificity	小	大	无	强
model evaluation	9	Kappa 检验 Kappa test	大	大	较强	较差
	10	真实技能统计 True skill statistics	较小	无	强	强

误差，一些研究根据对目标物种生态学的了解，将预测结果中出现在该物种不可能出现生境中的位点视为错判误差[33]，但此方案容易引起错判误差的低估.Kappa 检验和真实技能统计（true skill statistics）是最为常用的 2 种综合考虑遗漏误差和错判误差的检验方法.Kappa 检验是参考模型预测精确度与随机预测精确度对模型表现进行评估，数值介于−1～1 之间，检测值为 1，表明模型完美拟合观测数据，低于 0，则代表预测结果不及随机预测.真实技能统计是模型敏感度与特异性之和减 1.遗漏误差、敏感度受物种流行度影响最小，特异性、真实技能统计与物种流行度负相关[55]，其中特异性对流行度耐受能力较强，Kappa 检验对物种流行度耐受能力最差[46,55].特异性、真实技能统计与物种流行度的负相关是因为广布种更容易因观测遗漏造成错判误差的高估[43,55]，而非模型真实预测能力的下降.特异性与 Kappa 检验的另一个缺陷是仅适用于出现-不出现数据，而真实技能统计的检验效力同样适用于出现-背景（假设缺失）数据[44].

4.2.3 模型表现的目视评估　由于 MaxEnt 软件提供了便捷的制图功能，目视判别也是一种简便有效的模型检验方式.目视判别既可基于阈值，也可以参考预测分布概率的高低，研究者根据对物种生态习性的了解对预测结果进行判断[21,37].由于模型评估数值无法显示错判位点的分布，不能确定误差是均匀分布还是集中在某个区域，而目视判别可以提供生态学、生物地理学方面的线索，对误差产生原因进行分析[43].在评估数值相近时，目视判别也更有利于确定更符合物种生态习性的预测结果，特别是在缺乏调查数据的区域[23].

4.2.4 物种分布格局对模型预测能力的影响　模型预测能力评估必须考虑物种分布格局的因素，狭域分布的物种预测结果比广布种好[5,8,56].从统计学解释，出现-不出现位点判别能力的本质是环境变量在某一范围内对物种分布的限制程度，狭域分布物种受此限制更明显，因此模型表现更好[43,51].物种出现面积比率（relative occurrence area，ROA）也是影响模型评估的重要因素，ROA 越小，用于模型检验的不出现点距离出现点平均距离越远，受空间分类偏差影响，狭域分布物种更容易产生模型评估膨胀[5,51].使用 AUC、Kappa 检验评估模型表现时，在 ROA 接近 50%的情况下，遗漏误差和错判误差对模型评估的影响基本等效，ROA 偏离此数值越远，两类误差对模型评估影响效力差距越大，模型受判别失误的综合影响越小.但受观测遗漏的影响，ROA 大的物种更容易产生错判误差的高估，在使用假设缺失位点时，这种高估更明显，因此，狭域分布物种对判别失误耐受程度更高[5].除了物种的生态学特性，图层栅格的大小和研究区域的范围也会人为影响 ROA.在研究区域不变的情况下，图层栅格小，ROA 越小.出现位点不变的情况下，研究区域越大，ROA 越小.因此，在建模时应去除未进行调查区域以及物种无法到达的区域[7,10].

4.2.5 模型外推时预测表现的评估　缺乏检测数据是模型外推时经常面临的问题.严格地讲，只有封闭的模型可以被检验[43]，现有观测数据获得的评估结果并不适于作为外推时的模型选择标准[30].在进行潜在栖息地评估时，应选用现有分布区数据建模然后外推，而不是在全部研究区选取背景位点[41].模型外推前要评估环境变量是否超出建模区域范围，并选择复杂性低的模型，以降低遗漏误差[29-30].当环境变量超出建模数据范围时，应对预测结果持谨慎态度，特别是外推区域存在地理隔离或捕猎压力时，模型外推效果会大幅度下降[4].气候变化研究受

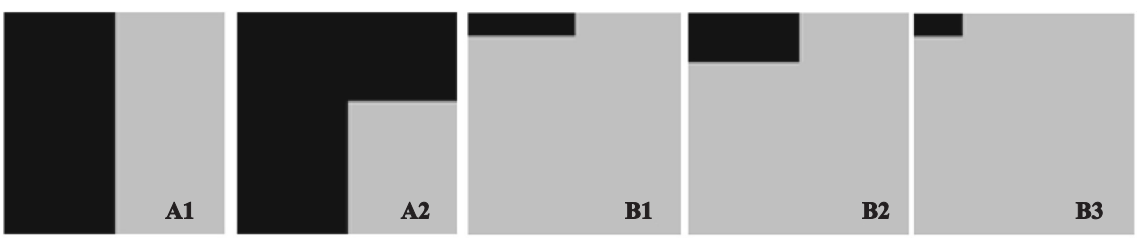

图 5　物种出现面积对模型表现的影响

Fig.5　Influence of relative occurrence area on model performance.

图中 A1、B1 为 2 个物种实际分布格局，A2、B2、B3 为预测分布格局，黑色表示物种分布区.A2 预测分布区超出实际分布 40%，B2 超出 102%，B3 遗漏 44%，而 3 个预测结果 Kappa 检验值均为 0.6 A1 and B1 represent the real distribution of two species with black color, and A2, B2, and B3 show the predicted distributions. A2 yields 40% commission error, B2 shows 102% commission error and B3 results 44% omission error. The Kappa score of A2, B2 and B3 are all 0.6 (Quoted from Jiménez-Valverde, *et al.*[5]).

变量波动影响更为严重,再加上常用的物候变量都有很强的相关性,模型外推时对变量波动的反映会很不稳定.因此,要去除与重要变量线性相关因子[29],在降低过度拟合风险的同时,减少环境变量超出建模范围的概率[42].

5 结 语

在 MaxEnt 的应用中,应该按照数据样本量、采样信息、空间分布情况,采用适宜的偏差校正方法降低模型过度拟合,根据模型应用范围选择适宜的复杂性,参考建模样本数据结构和研究目的选择适宜的出现点判定阈值,并分析模型评估指标的真实效力.虽然很多研究证实了 MaxEnt 模型优化的重要性,但受物种分布格局、模型条件、模型假设有效性以及优化措施适用性的约束,模型优化并不一定会提高预测能力[4,10,28].

应该注意的是,虽然 MaxEnt 模型在判别能力方面表现优异,但由于影响物种分布的因子很多,基于可得的分布数据建立的统计模型往往不能反映物种的生态本质,甚至基础生态位[5].一个在判别能力以及栖息地适宜度预测方面表现良好的模型可能在环境变量重要性的评估方面表现不佳[10].由于物种分布模型关注的重点是物种分布格局,预测结果与物种生态学特点,如种群密度、繁殖种群数量、繁殖成功率等因素之间的关系值得研究者进一步关注[43].

附表 最大熵模型优化相关开源软件
Appendix Open resources for MaxEnt model optimization

软件 Software	主要功能 Main function	下载地址 Resource acquisition
dismo	运行 MaxEnt 模型;计算 AUC、COR;确定 ESS、MSS、maxKappa、指定敏感度条件下的阈值	https://cran.r-project.org
ENMeval	对建模数据采用棋盘分组法等方法分组,并采用屏蔽地理结构分区法交叉验证;采用不同函数模式和调控系数组合建模,并计算 AIC、AUC、遗漏误差率	https://cran.r-project.org
ENMTools	计算 AIC、BIC、生态位相似度	http://purl.oclc.org/enmtools/
PresenceAbsence	计算 AUC	https://cran.r-project.org
pROC	判定出现阈值;计算模型敏感度、特异性、TSS	https://cran.r-project.org
psych	Kappa 检验	https://cran.r-project.org

参考文献

[1] Jaynes ET. Information theory and statistical mechanics. *Physics Reviews*, 1957, **106**: 620-630

[2] Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 2006, **190**: 231-259

[3] Guo J (郭 杰), Liu X-P (刘小平), Zhang Q (张琴), *et al*. Prediction for the potential distribution area of *Codonopsis pilosula* at global scale based on MaxEnt model. *Chinese Journal of Applied Ecology* (应用生态学报), 2017, **28**(3): 992-1000 (in Chinese)

[4] Phillips SJ, Dudík M, Elith J, *et al*. Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, 2009, **19**: 181-197

[5] Jiménez-Valverde A, Lobo JM, Hortal J. Not as good as they seem: The importance of concepts in species distribution modeling. *Diversity and Distributions*, 2008, **14**: 885-890

[6] Wintle BA, McCarthy MA, Parris KM, *et al*. Precision and bias of methods for estimating point survey detection probabilities. *Ecological Applications*, 2004, **14**: 703-712

[7] Elith J, Phillips SJ, Hastie T, *et al*. A statistical explanation of MaxEnt for Ecologists. *Diversity and Distributions*, 2011, **17**: 43-57

[8] Elith J, Graham CH, Anderson RP, *et al*. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 2006, **29**: 129-151

[9] Hernandez PA, Graham CH, Master LL, *et al*. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 2006, **29**: 773-785

[10] Fourcade Y, Engler JO, Rodder D, *et al*. Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. *PLoS One*, 2014, **9**(5): e97122, doi: 10.1371/journal.pone.0097122

[11] Luo M (罗 玫), Wang H (王 昊), Lyu Z (吕植). Evaluating the performance of species distribution models Biomod2 and MaxEnt using the giant panda distribtion data. *Chinese Journal of Applied Ecology* (应用生态学报), 2017, **28**(12): 4001-4006 (in Chinese)

[12] Warren DL, Seifert SN. Ecological niche modeling in MaxEnt: The importance of model complexity and the performance of model selection criteria. *Ecological Applications*, 2011, **21**: 335-342

[13] Syfert MM, Smith MJ, Coomes DA. The effects of sampling bias and model complexity on the predictive per-

formance of MaxEnt species distribution models. *PLoS One*, 2013, **8**(2): e55158, doi: 10.1371/journal.pone.0055158

[14] Dudík M, Schapire RE, Phillips SJ. Correcting sample selection bias in maximum entropy density estimation. International Conference on Neural Information Processing Systems, Vancouver, Canada, 2005: 323-330

[15] Reddy S, D'avalos LM. Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, 2003, **30**: 1719-1727

[16] Yackulic CB, Chandler R, Zipkin EF, *et al*. Presence-only modelling using MAXENT: When can we trust the inferences? *Methods in Ecology and Evolution*, 2013, **4**: 236-243

[17] Kadmon R, Farber O, Danin A. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, 2004, **14**: 401-413

[18] Reese GC, Wilson KR, Hoeting JA, *et al*. Factors affecting species distribution predictions: A simulation modeling experiment. *Ecological Applications*, 2005, **15**: 554-564

[19] Araujo MB, Guisan A. Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 2006, **33**: 1677-1688

[20] Zadrozny B. Learning and evaluating classifiers under sample selection bias. Proceedings of the Twenty-First International Conference on Machine Learning, Banff, Canada, 2004: 1-9

[21] Radosavljevic A, Robert PA. Making better MAXENT models of species distributions: Complexity, overfitting and evaluation. *Journal of Biogeography*, 2014, **41**: 629-643

[22] Anderson RP. Harnessing the world's biodiversity data: Promise and peril in ecological niche modeling of species distributions. *Annals of the New York Academy of Science*, 2012, **1260**: 66-80

[23] Shcheglovitova M, Anderson RP. Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecological Modelling*, 2013, **269**: 9-17

[24] Varela S, Anderson RP, García-Valdés R, *et al*. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, 2014, **37**: 1-8

[25] Osborne PE, Suárez-Seoane S. Should data be partitioned spatially before building large-scale distribution models? *Ecological Modelling*, 2002, **157**: 249-259

[26] Fourcade Y, Engler JO, Besnard AG, *et al*. Confronting expert-based and modelled distributions for species with uncertain conservation status: A case study from the Corncrake (*Crex crex*). *Biological Conservation*, 2013, **167**: 161-171

[27] Phillips SJ, Dudík M. Modeling of species distributions with MaxEnt: New extensions and a comprehensive evaluation. *Ecography*, 2008, **31**: 161-175

[28] Warton DI, Renner IW, Ramp D. Model-based control of observer bias for the analysis of presence-only data in ecology. *PLoS One*, 2013, **8**(11): e79168, doi: 10.1371/journal.pone.0079168

[29] Warren DL, Wright AN, Seifert SN, *et al*. Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern. *Diversity and Distributions*, 2014, **20**: 334-343

[30] Elith J, Kearney M, Phillips SJ. The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 2010, **1**: 330-342

[31] Warton DI, Shepherd LC. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Annals of Applied Statistics*, 2010, **4**: 1383-1402

[32] El-Gabbas A, Carsten FD. Improved species-occurrence predictions in data-poor regions: Using large-scale data and bias correction with down-weighted Poisson regression and MaxEnt. *Ecography*, 2018, **41**: 1161-1172

[33] Kramer-Schadt S, Niedballa J, Pilgrim JD, *et al*. The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, 2013, **19**: 1366-1379

[34] Austin MP. Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. *Ecological Modelling*, 2002, **157**: 101-118

[35] Rivers MC, Taylor L, Brummitt NA, *et al*. How many herbarium specimens are needed to detect threatened species? *Biological Conservation*, 2011, **144**: 2541-2547

[36] Brummitt NA, Bachman SP, Moat J. Applications of the IUCN Red List: Towards a global barometer for plant diversity. *Endangered Species Research*, 2008, **6**: 127-135

[37] Anderson RP, Gonzalez Jr I. Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with MaxEnt. *Ecological Modelling*, 2011, **222**: 2796-2811

[38] Moreno-Amat E, Mateo RG, Nieto-Lugilde D, *et al*. Impact of model complexity on cross-temporal transferability in MaxEnt species distribution models: An assessment using paleobotanical data. *Ecological Modelling*, 2015, **312**: 308-317

[39] Baldwin RA. Use of maximum entropy modeling in Wildlife Research. *Entropy*, 2009, **11**: 854-866

[40] Muscarella R, Galante PJ, Soley-Guardia M, *et al*. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for MAXENT ecological niche models. *Methods in Ecology and Evolution*, 2014, **5**: 1198-1205

[41] Phillips SJ. Transferability, sample selection bias and background data in presence-only modelling: A response to Peterson, *et al*. (2007). *Ecography*, 2008, **31**: 272-278

[42] Porfirio LL, Harris RM, Lefroy EC, *et al*. Improving the use of species distribution models in conservation planning and management under climate change. *PLoS One*,

2014, **9**(11): e113749, doi: 10.1371/journal.pone.0113749

[43] Lobo JM, Jimenez-Valverde A, Real R. AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 2008, **17**: 145-151

[44] Liu C, White M, Newell G. Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, 2013, **40**: 778-789

[45] Manel S, Dias JM, Ormerod SJ. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: A case study with a Himalayan river bird. *Ecology Modelling*, 1999, **120**: 337-347

[46] Liu C, Berry PM, Dawson TP, *et al*. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 2005, **28**: 385-393

[47] Liu C, Newell G, White M. On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and Evolution*, 2016, **6**: 337-348

[48] Braunisch V, Suchant R. Predicting species distributions based on incomplete survey data: The trade-off between precision and scale. *Ecography*, 2010, **33**: 826-840

[49] Smith AB. On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. *Diversity and Distributions*, 2013, **19**: 867-872

[50] Bahn V, McGill BJ. Testing the predictive performance of distribution models. *Oikos*, 2013, **122**: 321-331

[51] Hijmans RJ. Cross-validation of species distribution models: Removing spatial sorting bias and calibration with a null model. *Ecology*, 2012, **93**: 679-688

[52] MacKenzie DK, Nichols JD, Lachman GB, *et al*. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 2002, **83**: 248-255

[53] Gu W, Swihart RK. Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. *Biological Conservation*, 2004, **116**: 195-203

[54] Merckx B, Steyaert M, Vanreuse A, *et al*. Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. *Ecological Modelling*, 2011, **222**: 588-597

[55] Allouche O, Tsoar A, Kadmon R. Assessing the accuracy of species distribution models: Prevalence, Kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 2006, **43**: 1223-1232

[56] Raes N, ter Steege H. A null-model for significance testing of presence-only species distribution models. *Ecography*, 2007, **30**: 727-736

作者简介 孔维尧,男,1981 年生,博士研究生. 主要从事保护生物学研究. E-mail: kongweiyao@163.com

责任编辑 肖 红

孔维尧, 李欣海, 邹红菲. 最大熵模型在物种分布预测中的优化. 应用生态学报, 2019, **30**(6): 2116-2128
Kong W-Y, Li X-H, Zou H-F. Optimizing MaxEnt model in the prediction of species distribution. *Chinese Journal of Applied Ecology*, 2019, **30**(6): 2116-2128 (in Chinese)