arXiv:2305.01989v1 [stat.AP] 3 May 2023

在多物种分布模型中对分类过程的异质性进行建模，可以提高预测性能。

Kwaku Peprah Adjei[1,2*†] Robert B. O'Hara[1,2‡] Wouter Koch[2,4§] 和 Anders Finstad [2,3¶]

[1]数学科学系，挪威科技大学，特隆赫姆，挪威[2]生物多样性动态中心，挪威科技大学，挪威[3]自然历史系，挪威科技大学，挪威[4]挪威生物多样性信息中心，特隆赫姆，挪威

摘要

1. 基于大规模生物多样性数据构建物种分布模型和分布图，对于保护管理至关重要。目前的一个问题是，生物多样性数据容易出现分类学误分类。在多物种分布模型中处理这些误分类的方法，均假设分类概率在整个研究过程中保持不变。实际上，分类概率可能会随多个协变量而变化。未能考虑这种异质性可能会导致参数估计出现偏差。

2. 本文提出了一种通用的多物种分布模型，可考虑分类过程中的异质性。所提出的模型假设分类混淆矩阵服从多项广义线性模型。我们比较

*通讯作者。
†电子邮箱：kwaku.p.adjei@ntnu.no
‡电子邮箱：bob.ohara@ntnu.no §电子邮箱：wouter.koch@artsdatabanken.no
¶电子邮箱：anders.finstad@ntnu.no

异质分类模型与同质分类模型的表现，评估它们对模型参数的估计效果，以及在留出样本上的预测性能。我们将该模型应用于来自全球生物多样性信息机构的挪威、丹麦和芬兰鸥类数据。

3. 我们的模拟研究表明，在分类过程中考虑异质性使精确率提高了 30%，但准确率和召回率降低了 6%。将该模型框架应用于鸥类数据集时，由于被错误分类的样本量较小，异质模型与同质模型之间的预测性能并未得到改善。然而，当使用机器学习预测分数作为权重，为物种分布模型提供有关分类过程的信息时，精确率提高了 70%。

4. 当数据中包含相对较多的误分类样本时，我们建议使用多重多项式回归来建模分类过程中的变异。当数据中包含相对较少的误分类样本时，则应使用机器预测分数。

关键词： 贝叶斯模型、公民科学、假阳性、机器学习、误分类、多物种分布模型

1 引言

物种分布模型是生态学和保护管理中的重要工具，可用于预测自然因素和人为因素如何影响生物多样性（Elith 和 Leathwick，2009年；Vermeiren 等，2020年）。随着科学家能够获取越来越多来自多物种调查的生物多样性数据，多物种分布模型（下文简称多物种分布模型）已广泛用于分析这些数据。这些多物种分布模型在群落层面上对数据进行建模，识别驱动物种共现的重要变量，并预测群落中物种的分布（Hui 等，2015年；Ovaskainen 和 Soininen，2011年；Pollock 等，2014年）。

然而，从这些调查中获得的生物多样性数据可能存在观测误差。这些观测误差包括假阳性（将不存在的物种记录为存在，以及将一个物种误认成另一个物种）和假阴性（未能记录存在的物种）。这些假阳性通常是由物种被错误分类造成的。如果观测结果在被识别为假阳性后能够得到正确分类，而不是被丢弃，将会带来很大优势。假阳性可能源于不完善的分类器（Spiers 等人, 2022年；Wright 等人, 2020 年）、观察者误差以及许多其他因素。在本研究中，我们将所有假阳性和误认统称为错误分类。本研究使用“真实状态”一词来描述我们希望建模的正确或实际观测对象身份。占域模型通常通过在观测模型中对不完全检测进行建模，以处理假阴性。未能处理或纠正这些误差，会导致对占域概率、协变量效应和相对活动度等状态变量的推断产生偏差（Clare 等人, 2021；Ferguson 等人, 2015；Wright 等人, 2020），进而损害决策制定（Hoekman, 2021）。

处理生物多样性数据中的错误分类的方法可分为数据审查方法和基于模型的方法（Clare et al., 2021）。数据审查方法要求采用完整、恰当的数据收集与处理方法。由于错误分类难以控制，这一过程可能非常繁琐。因此，在处理来自

大型生物多样性数据供应商（如全球生物多样性信息机构，以下简称 GBIF；GBIF.Org, 2022）的大规模数据集时，基于模型的方法更为常用。基于模型的方法会在估计感兴趣的真实状态变量的同时，估计分类概率。目前，在多物种占域模型中尝试处理错误分类的基于模型的方法包括：通过探测异质性对错误分类进行建模（Clement et al., 2022; Ferguson et al., 2015; Louvrier et al., 2018）；将多位观察者的记录与距离抽样和 N-混合模型等其他方法相结合（Hoekman, 2021）；利用观测确认或核验（Ferguson et al., 2015; Guillera-Arroita et al., 2017）、地点确认（Clare et al., 2021）及其他校准方法所提供的额外信息进行监督。这些方法需要从验证过程中获取额外数据，以便在半监督情境下估计错误分类概率（Spiers et al., 2022），并使模型参数可识别（Guillera-Arroita et al., 2017）。上述研究使用的数据包括：在地点级收集的经核验数据（已知某物种在某地点的占域状态，但未知其在个体样本级的状态；Chambert 等人, 2018b年）；利用带有地点协变量的多项式模型对个体样本级数据进行汇总后的数据（Wright 等人, 2020年）；以及个体样本级验证数据，后者有助于将非物种身份（形态种）建模为物种身份（Spiers 等人, 2022年）。还值得指出的是，一些研究探讨了丰度模型（Conn et al., 2013）和捕获—再捕获模型（Augustine et al., 2020）中的错误分类问题。

此外，这些先前的研究假设研究中的错误分类概率是均一的（常数）。实际上，分类概率可能会随环境协变量（例如野外条件；Conn et al., 2013）或观察者经验（尤其是了解每位观察者在公民科学项目中对报告进行分类的准确程度会很有帮助；Arazy and Malkinson, 2021; Johnston et al., 2022）、使用样线数据时与样线的距离（Conn et al., 2013）、图像质量等因素而变化。对分类过程中的异质性进行建模的一种尝试，是将这一影响作为真实状态模型（例如丰度或占域）的一部分进行建模。然而，这种方法可能无法解决分类中的异质性问题。

分类过程，因为真实状态观测过程参数的估计值仅作为分类过程的有信息先验分布（Spiers 等人，2022年）。

对这种异质性进行建模的一种更恰当的方法，是在分类过程中加入协变量效应。一些关于动态假阳性单物种占域模型的研究将年份作为协变量，对假阳性的时间变化进行建模（Kéry and Royle，2020年；Miller 等，2013年；Sutherland 等，2013年），表明可以对错误分类随时间变化的趋势进行建模。本研究尝试使用多项广义线性模型，对真实状态身份进行分类的概率建模，从而刻画多物种分布模型中分类概率的变化。据我们所知，此前尚无相关研究。若未能考虑错误分类概率的异质性，将导致真实状态观测过程的参数（如物种丰度、丰富度和占域概率）估计产生偏差，并降低模型的预测性能。

此外，近期为正确分类生物多样性调查观测数据所做的努力，依赖于机器学习（以下简称 ML）算法（Borowiec 等，2022年；Keshavan 等，2019；Koch 等，2022年；Lotfian 等，2021年；Saoud 等，2020年；Suzuki-Ohno 等，2022年；Willi 等，2019）。这些机器学习算法利用观测数据的声音和/或图像来预测单条观测的真实身份，并且可以通过训练来模仿专家对观测数据的核验（Keshavan 等，2019；Langenkämper 等，2019；Ponti 和 Seredko，2022年）。这些机器学习算法使用预测得分（一种表示将观测数据预测为其他类别的权重的数值），来预测单条已报告观测数据可能对应的真实身份列表。这些预测得分和可能的真实身份列表提供了有关每条观测分类过程的信息，可用于对分类过程中的异质性建模。本研究首次利用预测得分对已报告观测数据的分布进行加权，并据此预测实际观测身份的分布，从而对分类过程中的异质性进行建模。

在此，我们提出一种联合模型，同时对感兴趣的真实状态变量（如占域和丰度）以及分类过程中的异质性进行建模。我们的模型设定扩展了 Wright 等人（2020）和 Spiers 等人

（2022）的工作，具体包括：a）允许分类概率随协变量变化；b）允许不完全分类中有更多形态状态类别（例如，将“未识别的啄木鸟”纳入分类观测）；c）使用机器学习预测分数作为权重，以考虑分类过程中的异质性；d）对分类过程协变量进行变量选择，以检查 mSDM 是否存在潜在的过拟合。我们通过模拟研究，将本模型的分类表现与 Wright 等人（2020）以及 Spiers 等人（2022）提出的假设分类概率均一的模型进行了比较。我们使用从 GBIF（GBIF.Org，2022）下载的 iNaturalist（Matheson，2014）挪威、芬兰和丹麦鸥类公民科学数据，对模型进行参数化。

2 方法

2.1 模型框架

所提出的框架假设存在一个模型，用于描述真实状态的身份“存在”于何处（本文中，我们使用“状态”来指代分类单元身份，以及任何其他鉴定类别或形态状态，例如未识别组）。真实状态可以处于任何分类学层级。然而，真实状态往往未知，并被视为潜在的。个体经过采样后，会以一定概率被归类为另一种状态，即所谓的“报告”或“分类”状态。报告状态也可以处于任何分类学层级，或属于任何未识别组（例如纳入形态种的研究；Spiers 等人，2022年）。我们进一步假设这些个体可以核验（并且我们掌握有关验证过程的信息），且已验证状态近似于真实状态身份。

2.1.1 真实状态观测过程模型

我们为真实状态观测过程指定了一个相对丰度模型（具体而言，是一个多项逻辑回归模型）。我们的目标是展示如何对分类过程中的异质性进行建模，而不是对真实状态的丰度进行推断，因此我们选择了一个更易于拟合和理解的模型，以解释真实状态的分布

状态。

设 λ_{is} 为位置 i =1，2，……，R处 s = 1、2、……、S 真实个体的平均数量，用以描述真实状态在空间中的丰度。这意味着，强度可以建模为非齐次点过程（PPM），并假设数据依赖于环境协变量；也可以建模为对数高斯 Cox 点过程（LGCP），并假设观测数据存在空间依赖（Renner 等人，2015年）。这里，平均强度采用非齐次点过程进行建模，定义如下：

$$\ln(\lambda_{is}) = \beta_{0s} + \sum_{j=1}^{n_e} x_{ij}\beta_{js}, \tag{1}$$

其中，β_{0s} 是状态 s的截距，β_{js} 是协变量 j 对真实状态 s强度的影响，x_{ij} 是影响位置 i处状态观测的 j^{th} 协变量，n_e 是观测过程模型中的协变量数量。请注意，我们假设相对丰度模型中不存在物种间相互作用；这些相互作用也可以作为随机效应加入真实状态强度的定义中（公式（1））。

设 p_{is} 为真实状态 s 位于位置 i 的概率。我们根据平均强度估计此概率，如下所示：

$$p_{is} = \frac{\lambda_{is}}{\sum_s \lambda_{is}}, \tag{2}$$

其中，λ_{is} 由式 (1) 给出。

位置 $i = 1、2, 2, \ldots, R$ 处的真实状态 $s = 1, 2, \ldots, S$，是从概率为 p_{is} 的分类分布中抽取的一次实现。

这里所使用的生态过程模型设定与 Spiers 等人（2022）使用的占域动态和遭遇率模型相似，假设占据概率等于 1；同时也与 Wright 等人（2020）使用的模型相似，假设泊松计数的强度为 λ_{is}（有关我们的模型框架与 Spiers 等人（2022）以及 Wright 等人（2020）模型之间的联系，请参阅表 3）。

2.1.2 分类过程模型

对于每次观测，我们假设分类并不完全准确，并且真实状态身份可以被分类为 $k = 1$、2、. . . 、 K 个状态。假设真实状态分布代表所研究物种的真实点格局，则每条报告的信息都可以视为从所考虑的 K 个状态中的任意一个状态中，按照给定概率抽取的一次结果。如上所述，这些状态可以处于任何分类学层级，也可以包含任何未识别组。例如，可以有四种真实状态：普通鸥、银鸥、奥杜安鸥和灰鸥。这些物种可以被报告为三种状态：大型白头鸥、大型黑头鸥和其他鸥。分类概率（混淆矩阵）的示例如表1所示。

	报告状态		
真实状态	大型白头鸥	大型黑头鸥	其他鸥
普通鸥	0.8	0.1	0.1
银鸥	0.9	0	0.1
奥杜安鸥	0	0.9	0.1
灰鸥	0	1	0

表1 适用于我们模型的混淆矩阵示例。

令 Ω_{sk} 表示单个真实状态 $s \in \{1, \ldots, S\}$ 被分类为状态$k \in \{1, \ldots, K\}$的概率。所有可能的 k 状态对应的概率之和为 1。在分类概率均一的研究中，分类的混淆矩阵可表示为：

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_1 \\ \mathbf{\Omega}_2 \\ \vdots \\ \mathbf{\Omega}_S \end{bmatrix} = \begin{bmatrix} \Omega_{11} & \Omega_{11} & \cdots & \Omega_{1k} & \cdots & \Omega_{1K} \\ \Omega_{21} & \Omega_{22} & \cdots & \Omega_{2k} & \cdots & \Omega_{2K} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \Omega_{S1} & \Omega_{S2} & \cdots & \Omega_{Sk} & \cdots & \Omega_{SK} \end{bmatrix} \tag{3}$$

其中，行对应真实状态 s，列对应报告状态 k。

我们通过对$\mathbf{\Omega}$ 的每一行拟合一个多项广义线性模型（Fahrmeiret al., 2013 年），来建模分类概率的异质性，该矩阵在方程

（3）中定义。我们将这种方法称为多重多项广义线性模型（以下简称 MMGLM）。我们将 MMGLM 的线性预测变量定义为:

$$\begin{aligned}\zeta_{isk} &= \omega_{0sk} + \sum_{j=1}^{n_c} z_{ij}\omega_{jsk} \\ &= \begin{bmatrix} \omega_{011} & \omega_{012} & \cdots & \omega_{01K} \\ \omega_{021} & \omega_{022} & \cdots & \omega_{02K} \\ \vdots & \vdots & \cdots & \vdots \\ \omega_{0S1} & \omega_{0S2} & \cdots & \omega_{0SK} \end{bmatrix} + \ldots + \begin{bmatrix} \omega_{n11} & \omega_{n12} & \cdots & \omega_{n1K} \\ \omega_{n21} & \omega_{n22} & \cdots & \omega_{n2K} \\ \vdots & \vdots & \cdots & \vdots \\ \omega_{nS1} & \omega_{nS2} & \cdots & \omega_{nSK} \end{bmatrix} * z_{in_c}\end{aligned}$$

并用矩阵表示:

$$\boldsymbol{\zeta}_i = \boldsymbol{\omega_0} + \sum_{j=1}^{n_c} \boldsymbol{\omega_j} z_{ij}; \tag{4}$$

其中， $\boldsymbol{\omega_0}$ 是一个由截距组成的 $S \times K$ 矩阵， $\boldsymbol{\omega_j}$ 是一个由协变量 j效应组成的 $S \times K$ 矩阵，而 z_{ij} 是驱动位置 i处分类过程的 j^{th} 协变量。将方程（4）作为线性预测变量的定义，我们估计的参数 ω_{0sk}和 ω_{jsk} 相对于一个报告状态是可识别的。因此，对于每个真实状态身份 s，相对于状态 K 的每个报告状态 $k = 1, 2, \ldots, K-1$ 的分类概率建模如下:

$$\ln\left(\frac{\Omega_{sk}}{\Omega_{sK}}\right) = (\omega_{0sk} - \omega_{0sK}) + (\omega_{1sk} - \omega_{1sK}) * \mathbf{z}_1 + \ldots + (\omega_{nsk} - \omega_{nsK}) * \mathbf{z}_n; \tag{5}$$

其中 $\mathbf{z}_1, \ldots, \mathbf{z}_n$ 是协变量向量。

这一通用框架有 $S \times (K-1) \times (n_c+1)$ 个待估计参数，其中 S 是真实状态的数量， K 是报告状态的数量，而 n_c 是分类过程中协变量的数量。随着真实状态、报告状态和协变量数量的增加，这些参数的估计在计算上可能会非常昂贵，并且需要大量被错误分类的状态样本才能完成估计。因此，我们探讨了公式（5）中广义模型的简化形式。

方程（5）的一个简化情形假设分类协变量$\mathbf{z}$ 仅影-

影响正确分类真实状态的概率。在这种情况下，矩阵 $\boldsymbol{\omega_j}$ 的非主对角元素 $\omega_{jsk} = 0$ 对协变量 j 的 $s \neq k$，且这些参数不进行估计（这是我们研究中的“固定协变量”情景，见表 2）。这一简化使分类过程的待估参数数量减少 $n_c \times （K - S - 1）$，其中 S 为真实状态的数量，K 为报告状态的数量，n_c 为分类协变量的数量。进一步的简化还可以假设，正确分类真实状态的平均概率不随物种变化，即对于所有 $s = k$，ω_{0sk} 都相同（这是我们研究中的“固定交互协变量”情景，见表 2）。分类过程的异质性由对所有 $s = k$ 都适用的协变量效应 ω_{1sk} 表示。后一种做法进一步使待估参数数量减少 $S - 1$。当各状态非常相似时，最后这种简化尤其有用，因为此时可以预期它们的平均分类概率相同。

给定位置 i 处存在的真实状态（s），该位置的报告状态以概率$\boldsymbol{\Omega}$从 K 个状态中抽取。

$$Y_i | V_i = s \sim \text{Categorical}(\Omega_{is\cdot}), \tag{6}$$

其中，$\Omega_{is\cdot}$ 指的是：给定位置 i 处的真实状态身份为状态 s 时，被分类为 K 个状态中任一状态的概率。

2.1.3 变量选择

拟合一个包含如此多参数的复杂模型，可能会导致模型过拟合。过拟合模型会捕捉训练数据中的模式和噪声，但在验证数据或测试数据上的表现较差（Montesinos López et al., 2022年）。在某些情况下，真实状态观测和分类过程协变量可能高度相关。这些相关的协变量会增大估计参数的标准误（降低精确率）（Caradima et al., 2019年；Roberts et al., 2017年；Yu et al., 2015年）。为避免模型过拟合，需要进行变量选择。本研究采用了贝叶斯变量选择，具体而言，是对分类过程协变量使用尖峰与平板先验分布。

变量（有关贝叶斯变量选择的综述，参见 O'Hara 和 Sillanpää，2009年）。对于每个分类过程协变量，我们将公式（4）中的线性预测变量重新定义为：

$$\zeta_{isk} = \omega_{0sk} + \sum_{j=1}^{n_c} \psi_j z_{ij} \omega_{jsk}; \quad \text{with} \quad \psi_j \sim Bernoulli(q_j) \tag{7}$$

其中，ψ_j 表示变量 j 以预期概率 q_j被选中的指示变量。公式（7）中指定的变量选择指示变量会联合选择影响模型中所有真实状态的变量，但也可以针对特定状态进行设置（Ovaskainen and Abrego，2020年）。概率 q_j 越接近1，表示该变量对模型的贡献越大，应当被选中；越接近0，则表示该变量的贡献越小，可以舍弃。

分类 type	**模型**	**真实状态观测过程**	**分类过程**
异质的	变量 / 系数变量	$\ln(\lambda_{is}) = \beta_{0s} + x_{1i}\beta_{1s} + x_{2i}\beta_{2s}$	$\ln(\frac{\zeta_{isk}}{\zeta_{isK}}) = \omega_{0sk} + \psi_i z_{1i}\omega_{1sk}$
	固定的-协变量	$\ln(\lambda_{is}) = \beta_{0s} + x_{1i}\beta_{1s} + x_{2i}\beta_{2s}$	$\ln(\frac{\zeta_{isk}}{\zeta_{isK}}) = \omega_{0sk} + \psi_i z_{1i}\omega_{1ss}$, 其中 $\omega_{1sk} = 0$ 用于 $s \neq k$
	固定交互协变量	$\ln(\lambda_{is}) = \beta_{0s} + x_{1i}\beta_{1s} + x_{2i}\beta_{2s}$	$\ln(\frac{\zeta_{isk}}{\zeta_{isK}}) = \omega_{0sk} + \psi_i z_{1i}\omega_{1ss}$, 其中 $\omega_{1sk} = 0$ 对于 $s \neq k$ 且 ω_{0sk} 对于 $s = k$ 相同
同质	截距	$\ln(\lambda_{is}) = \beta_{0s} + x_{1i}\beta_{1s} + x_{2i}\beta_{2s}$	$\ln(\frac{\zeta_{isk}}{\zeta_{isK}}) = \omega_{0sk}$
	常数	$\ln(\lambda_{is}) = \beta_{0s} + x_{1i}\beta_{1s} + x_{2i}\beta_{2s}$	$\Omega_{sk} \sim Dirichlet(\alpha_{sk})$, 其中 $\alpha_{sk} \sim Exp(1)$
	main	$\ln(\lambda_{is}) = \beta_{0s} + x_{1i}\beta_{1s} + x_{2i}\beta_{2s} + \psi_i z_{1i}\beta_{3s}$	$\Omega_{sk} \sim Dirichlet(\alpha_{sk})$, 其中 $\alpha_{sk} \sim Exp(1)$

表 2 MMGLM 的变体：由方程（7）定义的分类过程和由方程（1）定义的观测过程，被用作我们的研究情景。异质模型有三种：协变量、固定协变量和固定交互协变量；同质模型有三种：截距、常数和主效应。

2.2 使用机器学习预测分数对异质性进行建模

一些研究会考虑所报告观测或状态的真实身份，例如，因为它们使用机器学习对观测进行分类。这些预测得分（例如 F_1 得分、均方误差和对数损失）并非分类概率，而是用于表示机器学习算法在测试样本中对数据分类效果的数值。我们可以利用这些信息对分类过程中的异质性建模，并按如下方式预测真实状态身份：

$$\begin{aligned} V_i|Y_i &\sim \text{Categorical}(p_{iks}); \\ \text{with} \quad p_{iks} &= \frac{\lambda_{ik} w_{iks}}{\sum_k \lambda_{ik} w_{iks}} \end{aligned} \tag{8}$$

其中， λ_{ik} 是地点 i 处报告状态 k 的平均丰度， w_{iks} 是采样单元 i处报告状态 k^{th} 对应真实状态 s 的预测得分。

需要注意的是，这种方法是一种用于处理分类过程中异质性的非参数方法，而 MMGLM 则是一种参数方法。此外，利用机器学习预测分数对分类过程中的异质性进行建模，是对报告状态的预期丰度中的协变量效应进行建模，并使用预测分数作为权重对其进行校正，从而获得真实状态的相对丰度。然而，MMGLM 对真实状态的预期丰度中的协变量效应进行建模，并使用参数模型估计分类过程中的异质性。真实状态身份的预测则是通过将真实状态的预期强度与估计得到的分类协变量相结合进行加权来实现的。

2.3 模型框架的泛化性

所提出的框架推广了现有用于处理占域模型中误分类问题的多物种分布模型。Wright 等人（2020）提出了一个用于处理分类过程同质性的框架；我们的模型通过在观测过程中利用多项分布与泊松分布之间的关系（Steel 等人，1953 年），以及采用物种特异性常数模型，与该框架相联系，适用于那些

分类过程（表 3）。Wright 等人（2020年）进一步提出论据，认为其提出的模型是以下模型的广义形式：用于两个物种二元检测的模型（Chambert 等人，2018a年）、具有计数检测的单物种模型（Chambert 等人，2015年）以及具有二元检测的单物种模型（Chambert 等人，2018b年）。由于我们提出的框架可以视为 Wright 等人（2020年）模型的异质性版本，因此，我们的框架也是 Chambert 等人（2018a年、2015年、2018b年）所提出模型的一种可推广形式。Spiers 等人（2022）提出了一种个体层面的半监督方法，该方法利用占据动态和遭遇率来估计物种错误分类。如果我们假设分类过程是同质的（即 $\boldsymbol{\omega_1} = \mathbf{0}$），并假设占据概率为 1，那么我们的模型便与该方法相关（表 3）。

2.4 模拟研究

为了展示我们提出的模型如何运作及其在预测中的用途，我们针对 $S = 2$ 种真实状态和 $K = 3$ 种报告状态，在 $R = 1000$个地点开展了模拟研究。我们为真实状态观测过程模拟了两个协变量，为分类过程模拟了一个协变量，所有协变量均来自均值为 0、方差为 1 的正态分布。真实强度根据公式 (1) 进行模拟。两个真实状态模型的截距分别设为 $\beta_{01} = -1$和 $\beta_{02} = 0$，每个真实状态的协变量效应分别设为 $\beta_{11} = 4$、$\beta_{12} = -2$、$\beta_{21} = 0$ 和 $\beta_{22} = 0$（也就是说，状态 2 被用作参照类别）。分类过程的截距和协变量效应设定如下：

$$\omega_0 = \begin{bmatrix} 2 & 0.5 & 0 \\ 1 & 1 & 0 \end{bmatrix}; \omega_1 = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 1 & 0 \end{bmatrix}.$$

选择这些数值是为了获得数量充足的误分类状态样本。我们使用表 2 中的变量模型，模拟了 200 个分类过程具有异质性的 数据集（称为**“完整模型”**）；在公式（5）中令 $\boldsymbol{\omega_1} = \mathbf{0}$（一个全零矩阵），模拟了 200 个分类过程具有同质性的数据集（称为**“简化模型”**）；此外，还模拟了 200 个数据集，其中分类过程存在协变量效应（使用

作者	模型框架	链接到我们的模型
Wright et al.（2020年）和模型的所提出的框架推广了此类 as Chambert et al.（2018a年, 2015年、2018b年）	真实状态观测过程：*ab-* 溶质计数 z_{is} ~伯努利分布(ψ_{is}) $[V_{ijs}\|z_{is}=1] \sim$ 泊松分布(λ_{ijs})	真实状态观测过程：*rel-* 相对丰度 选择 $\psi_{is}=1$，然后 $z_{is}=1$ 适用于所有地点 i 和真实状态 s。 $[V_{ijs}\|z_{is}=1] \sim$ 使用 Poisson(λ_{ijs}) 之间的关系 之间 多项式-mial 和泊松分布（Steel 等人，1953 年）。
	分类过程： $[Y_{ijs}\|V_{ijs}=v_{ijs}, z_{is}=1] \sim$ 多项分布(v_{ijs}, Ω_{sk})， 其中 $\Omega_{sk} \sim$ 狄利克雷(α)	分类过程： $[Y_{ijs}\|V_{ijs}=1, z_{is}=1] \sim$ 分类分布(Ω_{sk})， 其中 Ω_{sk} 可以选择为任意一种同质模型，参见表 2 中所述 表 2 中所述
Spiers 等人（2022）	真实状态观测过程：*oc-* 占域动态与遭遇机制 比率 z_{ist} ~伯努利分布(ψ_{ist}) $[V_{ijst}] \sim Categorical\left(\frac{\lambda_{ijst} z_{ist}}{\sum_s \lambda_{ijst} z_{ist}}\right)$;	真实状态观测过程：相对丰度 选择 $t=1$，$\psi_{ist}=1$，然后选择 $z_{ist}=$ 所有地点 i 和真实状态 s均为 1。 $[V_{ijs}] \sim Categorical\left(\frac{\lambda_{ijs}}{\sum_s \lambda_{ijs}}\right)$;
	分类过程： $[Y_{ijs}\|V_{ijs}] \sim$分类分布(Ω_{sk})， 其中 $\Omega_{sk} \sim$ 狄利克雷(α)	分类过程： $[Y_{ijs}\|V_{ijs}=1] \sim$分类分布($\Omega_{sk}$)， 其中 Ω_{sk} 可以选择为任意形式同质模型的— 如表 2 所述。

表 3 Wright 等人（2020）和 Spiers 等人（2022）开展的分类过程同质研究中所提模型的扩展。该表列出了 Wright 等人（绝对丰度模型）、Spiers 等人（占据动态与遭遇率模型）以及我们（相对丰度模型）的真实状态观测过程模型；还列出了 Wright 等人和 Spiers 等人所用的分类过程模型（分类概率由狄利克雷分布模拟得到的同质分类过程），以及我们采用的表 2 中所述异质模型中的分类过程模型。由于我们的框架扩展了 Wright 等人（2020）的研究，因此可以说，他们的模型也是 Chambert 等人（2018a年、2015年、2018b年）模型的推广形式。

表 2 中与真实状态观测过程协变量相关的变量模型（称为**"相关模型"**）。第一个模拟数据集探讨了异质性对分类过程的影响，而后一个则探讨了分类过程与真实状态观测过程的协变量存在相关性时的影响。第二个数据集评估了分类过程模型过拟合的影响（即在分类过程本应同质时，为其加入异质性）。此外，我们

评估了误分类样本数量对多物种分布模型预测性能的影响。我们将 ω_0 的主对角线分量增加 6，以减少模拟生成的误分类样本数量。

我们从每个模拟数据集中随机抽取了两百个真实状态身份，作为验证样本。验证样本的数量没有变化，因为 Spiers 等人（2022）发现，验证样本数量对模型预测能力的影响较小。我们按照表 2 中描述的各种情景拟合模型，并在验证样本上评估模型的预测性能。

2.5 案例研究：鸥类数据集

我们使用所提出的模型分析了从 GBIF 下载的鸥类数据集（GBIF.Org，2022年）。该数据库收录了超过二十亿条出现记录，观察者人数远超一百万（网站访问日期：2023年2月17日）。我们关注 iNaturalist 的记录，因为这些记录经过社区验证（Matheson，2014年）。观察者收集这些出现记录，并上传带有图像和/或声音的观测，以便进行核验。报告的观测经过 iNaturalist 社区验证；当三分之二的社区成员认可分类单元鉴定时，这些观测便被认定为研究级（Ueda，2020年），随后发布到 GBIF。我们假设社区认可的分类单元名称即为真实状态**V**。我们查看了 iNaturalist 平台，以追踪观测的鉴定过程，并将首次报告的鉴定结果作为报告状态**Y**。

我们获取了2015年年至2022年年期间丹麦、芬兰和挪威部分鸥类物种的观测数据。具体而言，我们选择了大黑背鸥（*Larus marinus*）、银鸥（*Larus argentatus*）、普通鸥（*Larus canus*）和小黑背鸥（*Larus fuscus*），因为 iNaturalist 网站报告指出，这些物种经常被误分类为其他物种。除上述物种外，报告中的任何其他物种均被标记为“其他”。我们在模型中使用年降水量（从 raster 软件包获取；Hijmans et al., 2015）作为真实状态观测过程的协变量，因为部分文献指出，年降水量会影响海鸟的分布，例如.....

鸥（Algimantas 和 Rasa，2010年；Jongbloed，2016年）。

获得的数据为仅存在记录。探索性分析显示，所选物种在同一位置没有多次观测。因此，我们假设这些位置是离散的，并将数据视为标记点过程：在某一位置报告的物种赋值为 1，本研究中的其他物种则赋值为 0。如果我们有物种名录，就可以将其建模为同一位置上的重复标记点过程，并将地点作为随机效应处理。

众所周知，公民科学数据会受到多种偏差来源的影响。一些常见偏差包括空间偏差（Johnston 等，2022年；Tang 等，2021年）和误分类（Johnston 等，2022年；Tulloch 等，2013年）。本研究仅对误分类进行了考量，因为我们关注的是解释分类过程，而非推断鸥的丰度。据报道，随着公民科学家在报告物种方面积累经验，他们能够更准确地对物种进行分类（Vohland 等，2021年）。因此，我们将每位观察者提交的报告数量作为分类过程中的协变量，以此对分类过程的变化进行建模。如果某位观察者的观测次数超过十次，则将超出的观测次数按 10 进行校准。我们将观察者的观测次数用作经验的衡量指标（Johnston 等，2018年；Kelling 等，2015年），不过还有其他指标可用于衡量公民科学家的调查努力或经验（Santos-Fernandez 和 Mengersen，2021年；Vohland 等，2021年）。

我们还使用机器学习算法，为下载的数据中可能的真实状态身份获取预测得分（具体为 F_1 scores）。该机器学习算法是一种卷积神经网络（Koch 等人于 2022年提出的一种改进形式），使用挪威境内所有物种的公民科学观测数据进行训练。由于该机器学习算法使用 GBIF 中挪威境内所有鸟类数据进行训练，我们使用挪威境内所选鸥类物种的全部数据，以及芬兰和丹麦 2022年之前报告的全部数据，训练了表 2 中总结的全部六个研究情景模型；并将芬兰和丹麦在 2022年报告的全部数据用作验证样本。训练样本和验证样本中分类结果（真阳性和假阳性）的汇总见补充信息

2（S2-2、S2-3）。

2.6 模型拟合与评估

我们使用R软件（R Core Team，2022年），通过NIMBLE软件包（de Valpine等，2017）中的马尔可夫链蒙特卡洛方法，在贝叶斯框架下开展了所有分析。所有真实状态观测参数的先验分布均设为均值为0、标准差为10的正态分布；分类过程参数的先验分布则设为均值为0、标准差为1的正态分布。对于常数情景和主模型，我们将混淆矩阵（Ω）的先验分布设为参数为alpha（α）的狄利克雷分布，而alpha的先验分布为均值为1的指数分布。

我们运行了3条链，每条链迭代10000次；前5000次迭代作为预热期。我们通过目视检查轨迹图来判断模型是否收敛，并剔除Gelman-Rubin统计量（Brooks和Gelman，1998）大于.1的1个模型。我们保留了每条链剩余样本的五分之一。

我们采用准确率（验证样本所有预测中，被预测为真实状态的比例）、精确率（验证样本中不匹配的真实状态里，根据预测被正确分类的比例）和召回率（从预测中检索到的验证样本正确真实状态的比例）作为性能指标。我们采用贝叶斯方法，得到了参数的后验分布。验证指标值越高，表明模型越优。我们还通过估计偏差（真实值与估计值之差）和参数精确率，检验模型对真实状态观测过程参数（$\boldsymbol{\beta_0}$、$\boldsymbol{\beta_1}$ 和 $\boldsymbol{\beta_2}$）以及分类过程参数（$\boldsymbol{\omega_0}$ 和 $\boldsymbol{\omega_1}$）的估计效果。

3 结果

3.1 模拟研究

3.1.1 预测性能

我们通过模型预测的准确率、召回率和精确率展示了模型性能的提升（图 1 A 和 B）。当数据由完整模型和相关模型模拟时，结果有力表明，将分类过程中的变异性纳入模型能够提升多物种分布模型的预测性能。也就是说，对于完整模型和相关模型，“变量”模型表现最佳，其准确率、召回率和精确率均为最高（图 1 A (i - ii)、B (i - ii)）。然而，简化的异质模型（固定交互协变量和固定协变量）表现并不优于同质模型（图 1 A (i - ii)、B (i - ii)）。这表明，简化异质分类模型并不能提升预测性能，而能够捕捉全部变异性的异质模型（本例中为变量模型；表 2）才是最佳预测模型。当分类过程协变量作为观测过程的一部分进行建模时（主模型），模型的预测性能也与同质模型相近（图 1 A 和 B）。

使用异质分类过程对同质分类过程过拟合，并未影响 mSDM 的预测性能（图 1 A iii) 和 B iii)）。我们原以为过拟合的异质模型预测性能会较差（Montesinos López 等，2022年），但异质模型与同质模型的表现相似（在全部六个研究模型中，召回率、准确率和精确率均相同）。贝叶斯变量选择概率表明，同质分类模型更优（异质模型中纳入分类协变量的概率为 0.359 ± 0.012；补充信息 2；S2-1）。尽管简化的异质模型未能提升预测性能，但其在变量选择过程中与变量模型表现相似。

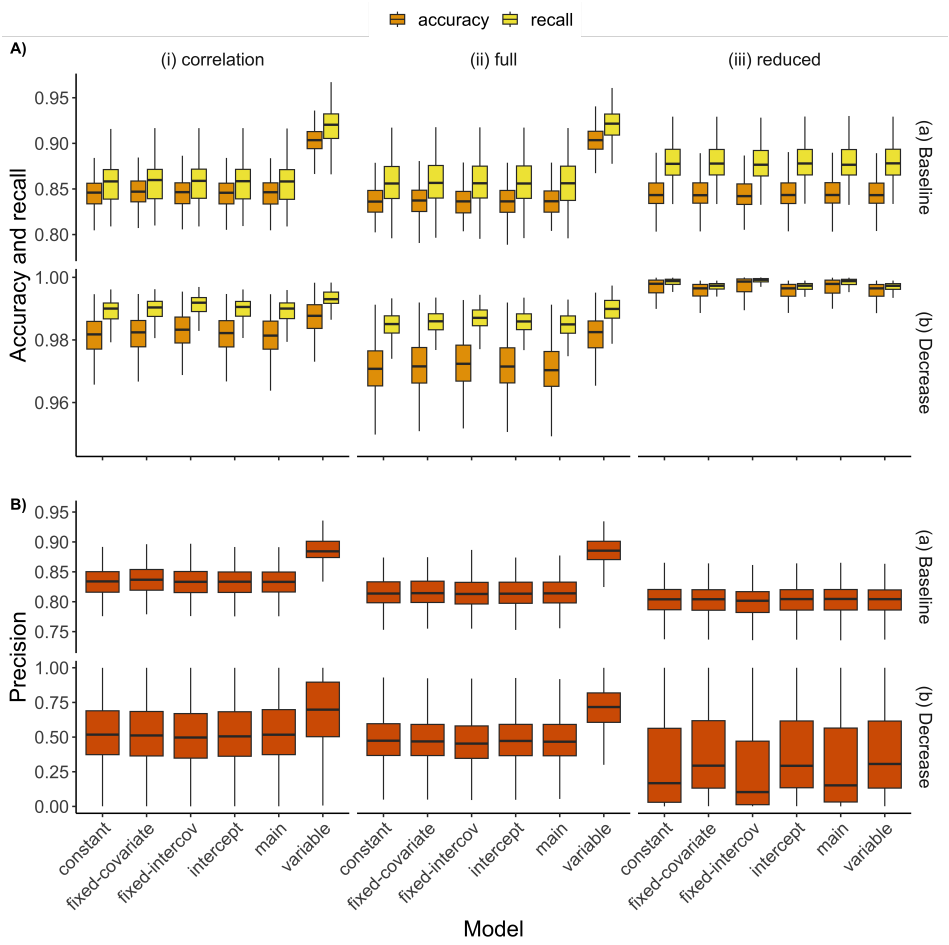


图 1：表 2 中定义的六种研究模型在每个模拟数据集中模拟的 1000 个样本中留出的 200 个（200）样本上的验证指标（准确率、精确率、召回率）箱线图。准确率是被正确分类的留出样本所占的比例，召回率是从留出样本中检索出的正确分类样本所占的比例，精确率是被正确分类的误分类样本所占的比例。每个箱线图显示中位数和四分位距（第 25–75 百分位数）。每列显示用于模拟数据集的模型类型：“full”指使用表 2 中的变量/协变量模型，“reduced”指使用表 2 中的截距模型，“correlation”指使用表 2 中的变量模型，但真实状态观测与分类过程协变量之间存在相关性。各行对应于对模拟数据集中误分类样本数量所做的变更：“Baseline”指使用第 2.4 节中定义的数值，“Decrease”指按照第 2.4 节所述，将误分类样本数量按 ω 的对角元素减少 6

3.1.2 错误分类样本数量的影响

随着我们增加模拟数据中误分类样本的数量，精确率平均提高了30%，而准确率和召回率降低了6%（图1 A（i - ii）、B（i - ii））。准确率和召回率的下降可能归因于：随着误分类样本数量增加，模拟数据中的正确分类数量减少。此外，当误分类样本数量较高时，分类过程参数的估计效果更好，从而使预测结果具有较高的精确率（补充信息 2；S2-1）。这表明，当存在大量误分类样本时，我们提出的模型将具有显著益处。

3.1.3 观测和分类过程参数中的偏差

尽管在多物种分布模型中未考虑错误分类可能导致真实状态观测过程参数产生偏差（Spiers 等人，2022年；Wright 等人，2020年），但在多物种分布模型中用于处理错误分类的任何方法，对真实状态观测过程参数的准确率和精确率影响都很小。在六种模型中，观测过程参数的偏差始终较低；在完全模型和简化模型下，所有情景的覆盖率都更高（补充信息 2；S2-1）。所有研究情景都以某种方式考虑了错误分类，从而校正了观测参数估计值的偏差（Spiers 等人，2022年；Wright 等人，2020年）。与其他模型相比，变量模型对分类过程参数的估计更准确、更精确（补充信息；S2-1）。但只有在误分类样本数量充足时，这一点才有可能实现。这表明，如果研究目的是利用多物种分布模型预测真实状态身份，那么对完全异质性进行建模可以提高预测性能；如果目的是推断真实状态分布，那么异质模型可能并不比同质模型有任何优势。

方法	准确率	精确率	召回率	变量选择-tion 概率-ity
变量 / 协-变量	0.97	0.1	0.99	0.71
常数	0.97	0.1	0.99	-
截距	0.97	0.1	0.99	-
Main	0.97	0.1	0.99	0.29
固定交互协变量	0.97	0.1	0.99	0.70
固定协变量	0.97	0.1	0.99	0.71
机器学习-ing	0.89	0.8	0.90	-

表4 研究模型在留出鸥类数据集上的验证指标。准确率是正确分类的验证数据所占的比例，精确率是错误匹配的身份中被正确匹配的比例，召回率则是成功找回的正确匹配身份所占的比例。验证样本共384个，其中有10个样本存在错误匹配。

3.2 案例研究：鸥类数据集

六种研究情景模型的预测性能相当，准确率和召回率较高，但精确率较低（表4）。精确率较低并非因为分类协变量（观察者经验）在解释分类过程异质性方面不显著，因为变量选择概率接近 1（表4）；而是由于误分类样本量较小（补充信息 1；S2-2、S2-3）。不过，使用机器学习算法的预测得分来考虑分类过程中的异质性后，精确率从 10% 提高到了 80%（表4）。机器学习算法的预测得分因观测而异，为观测过程模型提供了直接信息。然而，六种分类模型能否捕捉分类过程中的异质性，取决于误分类样本量。这表明，对于误分类样本极少的数据，改善多物种分布模型预测性能的一种方法是使用机器学习权重来考虑多物种分布模型中的异质性。

尽管研究情景模型的精确率较低，但观察发现，

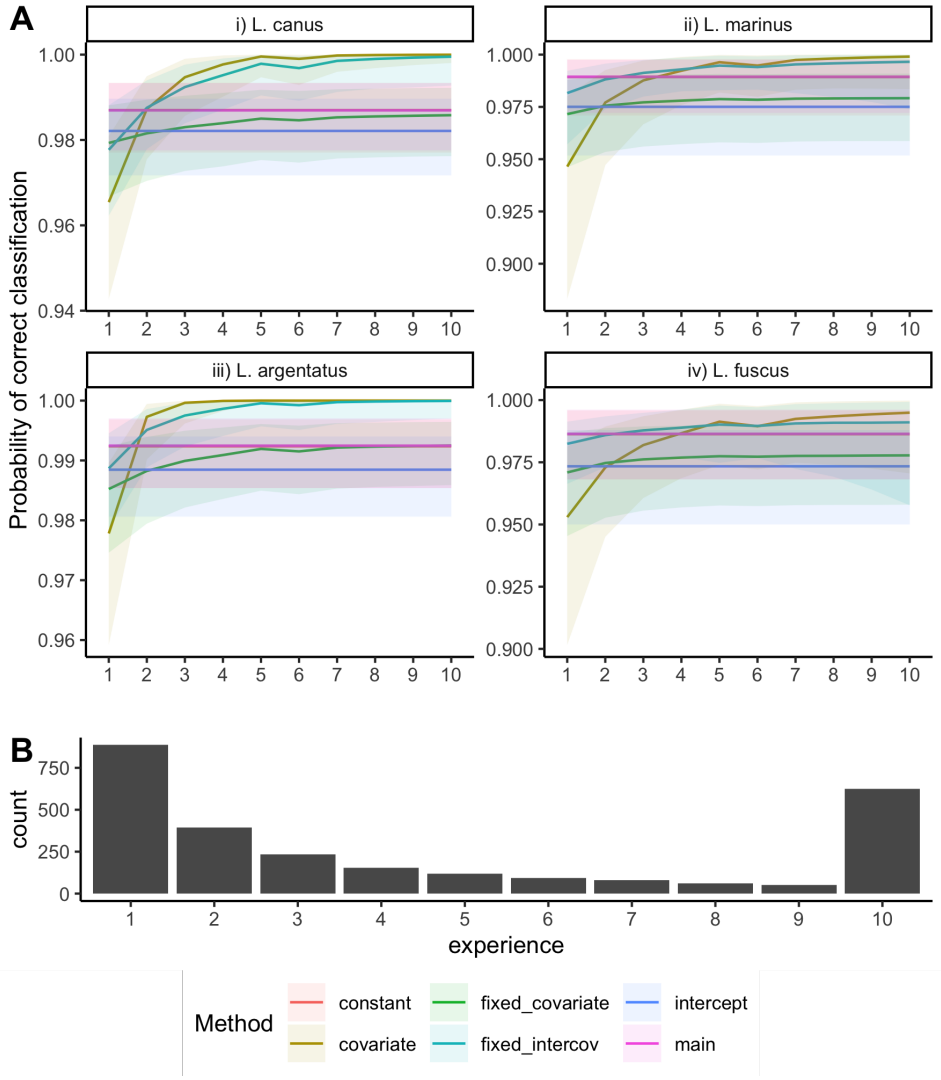


图 2：鸥类数据集模型拟合结果概要，显示 A）普通鸥（*Laruscanus*）、银鸥（*Larus argentatus*）、大黑背鸥（*Larusmarinus*）和小黑背鸥（*Larus fuscus*）的正确分类概率，以及 B）建模所用观察者的经验分布。正确分类概率估计值周围的带状区域表示估计值的 95% 可信区间。

在丹麦、芬兰和挪威，正确分类海鸥物种的概率会随着观察者经验的增加而提高（图 2）。这一模式表明，观察者在最初几次报告中很可能出错，而随着报告数量增加，他们的分类能力也会逐渐提高（Vohland 等人，2021）。

4 讨论

本文的主要目标是提出一个通用框架，用于在多物种分布模型（mSDM）中处理来自不完全分类（例如调查中的分类结果）和不确定分类（来自自动分类器）的错误分类。本研究基于 Spiers 等人（2022）和 Wright 等人（2020）的先前工作，在允许分类类别多于已核实物种（例如未知物种、形态种等）的同时，考量分类概率的异质性。此外，我们评估了对同质分类过程进行过拟合对多物种分布模型预测表现的影响，并提出了检查分类过程模型是否过拟合的方法。

我们的研究通过对分类过程中的异质性进行建模，弥补了文献中关于如何在多物种分布模型中考虑分类错误的知识空白。生物多样性数据中假阳性不可避免（Kéry and Royle，2020年；Miller 等人，2013年）。为对本研究中的这些假阳性进行建模，我们以分层形式将真实状态观测过程和分类过程分别表示为两个模型。我们将每种真实状态身份对应的分类概率建模为多项广义线性模型。这一设定将分类过程的建模推广到可将协变量的影响设为固定效应、随机效应或二者兼有。例如，针对志愿者采集的数据，可以假设观察者具有随机效应，从而估计每位观察者的分类概率。这种对分类过程的表述也缓解了将狄利克雷分布用作分类概率先验分布时所面临的建模问题（Spiers 等人，2022年）。

此外，在所提出的框架中，为真实状态观测过程单独指定一个状态空间模型，使我们能够采用各种多物种

观测模型（例如联合物种分布模型（Ovaskainen and Abrego, 2020年；Tobler et al., 2019），用于丰度的 Royle-Nichols 模型（Royle and Nichols, 2003）等）来对真实状态的分布进行建模。我们的模拟研究表明，所提出的模型框架能够估计真实状态的相对丰度参数（估计参数的偏差接近于零；补充信息 2；S2-1）。此前采用观测确认设计对 mSDM 中的错误分类进行建模的研究也报告了这一发现（Kéry and Royle, 2020年；Spiers 等人, 2022年；Wright 等人, 2020年）。我们已表明，真实状态观测过程模型是考虑错误分类的占域与丰度模型的一种简化形式（表 3），因此我们认为，所提出的框架可扩展至任何用于收集和核实真实状态数据的设计（例如点过程、距离抽样、地点确认（多方法）设计等），以及任何用于拟合数据的模型（例如多状态占域模型（Kéry and Royle, 2020年）、联合物种分布模型（Ovaskainen and Abrego, 2020年；Tobler et al., 2019））。

如果在对真实状态观测过程进行建模时，采用比本研究所用相对丰度模型更复杂的模型，就会为真实状态观测过程增加另一个层级结构（例如，对探测概率或真实占域状态进行建模）。这种复杂性可能导致分类过程参数与观测过程参数之间出现混杂问题；采用频率学派估计方法时，还可能使似然呈多峰分布（Kéry and Royle，2020年）。本研究未探讨这些问题，未来可以对此开展进一步研究。在贝叶斯框架下，为避免此类混杂问题，一种可能的解决方案是使用不同的协变量对各个过程分别建模，为 mSDM 参数选择合适的先验分布，并利用重复调查访问数据对真实状态观测过程建模（Kéry and Royle，2020年）。此外，也可以通过使用包含大量探测信息和假阳性探测信息的数据来处理可识别性或混杂问题，例如声学调查所得数据（Clement et al.，2022年）；还可以将不易发生误分类的占域数据（如相机陷阱所得数据）与存在误分类的占域数据整合起来（Kéry and Royle，2020年）。

考虑分类过程中的异质性会提升预测

性能。对于分类过程中的变化，同质分类模型有时可能无法加以解释（Conn 等，2013），从而因过拟合导致模型预测性能不佳（Montesinos López 等，2022）。模拟研究显示，在多物种分布模型中考虑分类过程的异质性后，精确率提高了 30%，准确率和召回率则分别下降了 6%（图 1 A、B）。然而，当异质分类模型对同质分类过程发生过拟合时，预测性能并未改变（图 1 A、B）；这是由于分类协变量效应较小，这一点可从参数估计的偏差和较低的贝叶斯变量选择概率中看出（图 1 A、B）。由于真实状态身份的后验预测概率高度依赖错误分类概率的权重（补充信息 1），未能考虑这一点就意味着后验概率会被错误估计。错误预测的概率会导致对真实状态覆盖范围的预测被低估，也可能导致对其丰度的预测被低估（Molinari-Jobin 等，2012）。

在真实状态观测模型中加入一个解释分类过程异质性的协变量，使模型变得更加复杂，但这并未提供足够的信息来提升多物种分布模型的预测性能。以往研究表明，观测过程模型的估计值有助于估计分类概率（Spiers 等人，2022年），但无法根据观测过程模型的变异性推断分类过程的变异性（图 1 A(i -iii)，补充信息 1）。因此，生态学家应在其分类过程模型中对分类的变异性进行建模，以提升多物种分布模型的预测性能。

我们的模型使用志愿者收集的鸥类数据进行参数化。这些志愿者收集的数据在生成过程中存在多种偏差来源，例如空间偏差和物种误识别等。我们承认数据中可能存在所有这些偏差来源，但我们仅对物种误识别进行了建模，并在分类过程模型中使用此前收集的数据量作为观察者经验的代理指标。由于误分类样本较少，齐性模型与异质模型的预测性能大致相同——

（训练数据中有 1382 个样本，其中 19 个被误分类（补充信息 S2-2）；验证数据中有 378 个样本，其中 10 个被误分类（补充信息 S2-3））。然而，估计得到的协变量效应让我们得以了解经验如何影响对新观测进行分类的概率。具体而言，随着观察者经验的增加，正确识别物种的概率也会提高，正如一些文献所指出的那样（Johnston 等，2018年；Kelling 等，2015年；Santos-Fernandez 和 Mengersen，2021；Vohland 等，2021）。因此，当误分类样本相对较少时，模型正确分类不匹配数据的能力（精确率）与理解驱动分类过程的协变量效应之间存在权衡。

在多物种分布模型中纳入机器学习预测分数，以考量分类过程中的异质性，使我们的预测精确率提高了70%（表4）。这些机器学习预测分数针对具体观测，能够提供大量有关分类过程的信息，从而提高模型的精确率。机器学习提供的信息不取决于错误分类样本量，而取决于图像质量（Koch et al., 2022年），因此，当错误分类样本量较小时（例如我们的鸥类数据），利用这些信息考量分类过程中的异质性具有优势。

本研究提出的模型框架具有灵活性，可以推广至任何物种分布模型和整合分布模型。所提出的框架与 Wright 等人（2020）和 Spiers 等人（2022）提供的框架，以及其研究所推广的任何框架相适配。我们提出的分类过程模型 MMGLM 改善了多物种分布模型的预测性能，但其高度依赖于错误分类样本的数量。此外，模型框架中定义的混淆矩阵允许对不同的分类群进行分类，而不是仅使用 Wright 等人（2020）提出的逐物种混淆矩阵，并且还将形态种纳入分类类别（Spiers 等人，2022）。这将使公民科学数据分析人员能够在数据收集过程的任何层级考虑数据的错误分类。我们建议在分析过程中进行变量选择或模型选择，以检查是否存在过拟合。此外，生态学家应探索使用机器学习预测分数

（在可以获得预测得分的情况下）作为多物种分布模型中的权重，旨在预测真实状态分布，尤其是在数据中错误分类样本数量较少时。

5 数据可用性

本文使用的海鸥数据可从 GBIF（https://doi.org/10. 15468/dl.h24bp5）下载。本文所用的所有代码均可通过 https: //doi.org/10.5061/dryad.0rxwdbs51 在 Dryad 数据库获取。

6 利益冲突

作者声明不存在利益冲突。

7 作者贡献

所有作者均参与了构思和论文撰写；KPA 和 RBO 参与了方法学开发；WK 运行了机器学习算法，KPA 则完成了本文所用的全部分析并主导了论文撰写。

参考文献

Algimantas, P. 和 Rasa, U.（2010年）立陶宛波罗的海沿岸鸥（鸥科）的丰度及季节性迁徙。*Daugavpils*大学生物学学报，**10**，147–164。

Arazy, O. 和 Malkinson, D. (2021) 公民科学生物多样性监测中基于观察者的偏差框架：使非结构化的生物多样性监测规程实现半结构化。《生态学与进化前沿》，435。

Augustine, B. C., Royle, J. A., Linden, D. W. 和 Fuller, A. K.（2020年）空间邻近性调节遗传标记研究中的基因型不确定性。美国国家科学院院刊，**117**，17903–17912。

Borowiec, M.L., Dikow, R. B., Frandsen, P.B., McKeeken, A., Valentini, G.和 White, A. E.（2022年）深度学习作为生态学与进化研究的工具。《生态学与进化方法》，**13**，1640–1660。

Brooks, S. P. 和 Gelman, A.（1998）用于监测迭代模拟收敛性的一般方法。《计算与图形统计学杂志》，**7**，434–455。

Caradima, B., Schuwirth, N. 和 Reichert, P.（2019）从单物种到联合物种分布模型：模型复杂度与预测性能的比较。《生物地理学杂志》，**46**，2260–2274。

Chambert, T., Grant, E. H. C., Miller, D. A., Nichols, J. D., Mulder, K. P. 和 Brand, A. B.（2018a年）考虑物种误识别和未探测情况的双物种占域建模。《生态学与进化方法》，**9**，1468–1477。

Chambert, T., Miller, D. A. 和 Nichols, J. D.（2015年）不同研究设计下物种出现记录中假阳性探测的建模。《生态学》，**96**，332–339。

Chambert, T., Waddle, J. H., Miller, D. A., Walls, S. C. 和 Nichols, J. D.（2018b年）自动声学物种检测数据分析的新框架：占域估计与录音后处理优化。《生态学与进化方法》,**9**, 560–570.

Clare, J. D., Townsend, P. A. 和 Zuckerberg, B.（2021）基于广义模型的物种检测/未检测数据中假阳性错误的解决方案。《生态学》，102，e03241。

Clement, M. J.、Royle, J. A. 和 Mixan, R. J.（2022年）在存在误分类和检测异质性的情况下，根据自动录音单元数据估计占域。《生态学与进化方法》。

Conn, P. B.、McClintock, B. T.、Cameron, M. F.、Johnson, D. S.、Moreland, E. E. 和 Boveng, P. L.（2013年）在样带调查中考虑物种鉴定错误。《生态学》，**94**，2607–2618。

Elith, J. 和 Leathwick, J. R.（2009年）物种分布模型：跨空间和时间的生态学解释与预测。《生态学、进化与系统学年度评论》，**40**，677–697。

Fahrmeir, L.、Kneib, T.、Lang, S. 和 Marx, B.（2013年）类别回归模型，325–347。柏林、海德堡：Springer Berlin Heidelberg。URL：https://doi.org/10. 1007/978-3-642-34333-9 6._

Ferguson, P. F., Conroy, M. J. 和 Hepinstall-Cymerman, J.（2015年）用于存在假阳性和假阴性错误、且地点与调查间存在异质性的数据的占域模型。《生态学与进化方法》，**6**，1395–1406。

GBIF.Org（2022）GBIF：全球生物多样性信息机构（年份）什么是 GBIF？网址：https://www.gbif.org/what-is-gbif。

Guillera-Arroita, G., Lahoz-Monfort, J. J., van Rooyen, A. R., Weeks, A. R. 和 Tingley, R.（2017）在多个层级处理有关物种出现的假阳性和假阴性错误。《生态学与进化方法》，**8**，1081–1091。

Hijmans, R. J., Van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J. A., Lamigueiro, O. P., Bevan, A., Racine, E. B., Shortridge, A. 等（2015年）软件包“raster”。*Rpackage*，**734**。

Hoekman, S. T.（2021）用于估计不确定物种鉴定的多观察者方法。*Ecosphere*，**12**，e03648。

Hui, F. K., Warton, D. I. 和 Foster, S. D.（2015年）使用惩罚回归混合模型进行多物种分布建模。《应用统计年鉴》，**9**，866–882。

Johnston, A., Fink, D., Hochachka, W. M. 和 Kelling, S.（2018年）基于公民科学数据的物种分布估计：观察者专业能力的提升作用。《生态学与进化方法》，**9**，88–97。

Johnston, A.、Matechou, E. 和 Dennis, E. B.（2022年）利用公民科学数据开展生物多样性监测面临的重大挑战及未来方向。《生态学与进化方法》。

Jongbloed, R. H. (2016年) 海鸟的飞行高度：一项文献研究。

Kelling, S., Johnston, A., Hochachka, W.M., Iliff, M., Fink, D., Gerbracht, J., Lagoze, C., La Sorte, F. A.、Moore, T., Wiggins, A. 等（2015年）能否使用物种累积曲线估计公民科学家的观测技能？《公共科学图书馆·综合》，**10**，e0139600。

Kéry, M. 和 Royle, J. A.（2020年）《生态学中的应用层次建模》：使用 *R* 和 *BUGS* 分析分布、丰度和物种丰富度：第 *2* 卷：动态与高级模型。《学术出版社》。

Keshavan, A.、Yeatman, J. D. 和 Rokem, A.（2019）结合公民科学与深度学习，扩大神经影像学领域的专业能力。《神经信息学前沿》，**13**，29。

Koch, W.、Hogeweg, L.、Nilsen, E. B. 和 Finstad, A. G.（2022年）最大化公民科学家对自动物种识别的贡献。《科学报告》，**12**，1–10。

Langenkämper, D.、Simon-Lledo, E.、Hosking, B.、Jones, D.O. 和 Nattkemper, T. W.（2019）公民科学衍生的数据质量对基于深度学习的海洋图像分类的影响。《公共科学图书馆·综合》，**14**，e0218086。

Lotfian, M., Ingensand, J. 和 Brovelli, M. A.（2021）公民科学与机器学习的结合：参与、数据收集和数据质量方面的益处、风险与未来挑战。《可持续发展》，13，8087。

Louvrier, J., Chambert, T., Marboutin, E. 和 Gimenez, O.（2018年）使用隐马尔可夫模型，在占域研究中考虑误识别和异质性。《生态建模》，387，61–69。

Matheson, C. A.（2014年）iNaturalist。《参考评论》。

Miller, D. A.、Nichols, J. D.、Gude, J. A.、Rich, L. N.、Podruzny, K. M.、Hines, J. E. 和 Mitchell, M. S.（2013年）在出现假阳性时确定出现动态：利用公众调查数据估计狼的分布范围动态。《公共科学图书馆·综合》，**8**，e65808。

Molinari-Jobin, A.、Kéry, M.、Marboutin, E.、Molinari, P.、Koren, I.、Fuxjäger, C.、Breitenmoser-Würsten, C.、Wölfl, S.、Fasel, M.、Kos, I. 等（2012年）存在物种误识别时的监测：欧亚猞猁在阿尔卑斯山的案例。《动物保护》，**15**，266–273。

Montesinos López, O. A.、Montesinos López, A. 和 Crossa, J.（2022年）过拟合、模型调优和预测性能评估。载于《基因组预测的多变量统计机器学习方法》，109–139。施普林格。

O'Hara, R. B. 和 Sillanpää, M. J.（2009年）贝叶斯变量选择方法综述：是什么、如何进行以及选择哪种方法。《贝叶斯分析》，**4**，85–117。

Ovaskainen, O. 和 Abrego, N.（2020年）评估模型拟合并在多个模型中进行选择，217–252。《生态学、生物多样性与保护》。剑桥大学出版社。

Ovaskainen, O. 和 Soininen, J.（2011年）充分利用稀疏数据：物种群落的分层建模。《生态学》，92，289–295。

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R.B., Parris, K. M., Vesk, P. A. and McCarthy, M. A.（2014年）通过使用联合物种分布模型（jsdm）同时对物种进行建模来理解共现。《生态学与进化方法》,**5**, 397–406。

Ponti, M. 和 Seredko, A.（2022年）公民科学中的人类与机器学习整合及任务分配。《人文与社会科学通讯》，**9**，1–15。

R 核心团队（2022年）*R*：用于统计计算的语言和环境。R 统计计算基金会，奥地利维也纳。网址：https://www.R-project.org/。

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G. 和 Warton, D. I.（2015年）用于仅存在数据分析的点过程模型。《生态学与进化方法》，**6**，366–379。

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauen- stein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W. 等（2017）具有时间、空间、层级或系统发育结构的数据的交叉验证策略。生态地理学，**40**，913–929。

Royle, J. A. 和 Nichols, J. D.（2003）利用重复出现–缺失数据或定点计数估算丰度。《生态学》，84，777–790。

Santos-Fernandez, E. 和 Mengersen, K.（2021）使用项目反应模型理解公民科学观测数据的可靠性。《生态学与进化方法》，**12**，1533–1548。

Saoud, Z.、Fontaine, C.、Loïs, G.、Julliard, R. 和 Rakotoniaina, I.（2020）使用机器学习检测公民科学平台用于生物多样性监测时的错误识别。《生态信息学》，**60**，101176。

Spiers, A. I.、Royle, J. A.、Torrens, C. L. 和 Joseph, M. B.（2022）利用占据动态和遭遇率估计物种错误分类：一种半监督的个体层面方法。《生态学与进化方法》。

Steel, G. 等（1953）泊松分布与多项分布之间的关系。

Sutherland, C.、Elston, D. 和 Lambin, X.（2013）考虑瞬时个体引发的假阳性检测错误。《野生动物研究》，**40**，490–498。

Suzuki-Ohno, Y., Westfechtel, T., Yokoyama, J., Ohno, K., Nakashizuka, T., Kawata, M. 和Okatani, T.（2022年）深度学习提高了公民科学项目中公民拍摄的生物照片的可用性。《科学报告》，12，1–10。

Tang, B., Clark, J. S. 和 Gelfand, A.E. (2021) 通过 eBird 数据库对存在空间偏差的公民科学调查力度进行建模。环境与生态统计，**28**，609–630。

Tobler, M. W., Kéry, M., Hui, F. K., Guillera-Arroita, G., Knaus, P. 和 Sattler, T.（2019）具有物种相关性和不完全检测的联合物种分布模型。《生态学》，100，e02754。

Tulloch, A. I.、Possingham, H. P.、Joseph, L. N.、Szabo, J. 和 Martin, T. G.（2013年）实现公民科学监测项目的全部潜力。《生物保护》，**165**，128–138。

Ueda, K.（2020）iNaturalist 研究级观测记录。iNaturalist.org。出现记录数据集 *https://doi.org/10.15468/ab3s5x [于2020年5月26日通过 GBIF.org 访问, https://www.gbif.org/occurrence/2294559370]*。

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T. 和 Bodik, R. (2017) 使用模型编程：利用 NIMBLE 为一般模型结构编写统计算法。《计算与图形统计学杂志》，**26**，403–413。

Vermeiren, P., Reichert, P. 和 Schuwirth, N. (2020) 将关于生态偏好的不确定先验知识纳入多物种分布模型：模型复杂度对预测性能的影响。《生态建模》，**420**，108956。

Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R. 和 Wagenknecht, K. (2021)公民科学的科学。施普林格·自然。

Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M. 和 Fortson, L. (2019) 使用深度学习和公民科学识别相机陷阱图像中的动物物种。《生态学与进化方法》，**10**，80–91。

Wright, W. J., Irvine, K. M., Almberg, E. S. 和 Litt, A. R.（2020年）估计占域和相对活动度时，多物种声学数据中的误分类建模。《生态学与进化方法》,**11**, 71–81.

Yu, H., Jiang, S. 和 Land, K. C.（2015年）分层线性模型中的多重共线性。社会科学研究,**53**, 118–136.