arXiv:2601.06178v1 [cs.CY] 7 Jan 2026

利用遥感监测可持续发展目标的模型性能：三层元回归

Jonas Klingwort[1,*], Nina M. Leach[2], Joep Burger[1]

[1]荷兰统计局（CBS）研究与发展部 [2] 莱顿大学方法学与统计学系

*通讯邮箱：j.klingwort@cbs.nl

摘要

摘要： 机器学习（ML）是一种利用遥感数据监测和实现联合国可持续发展目标（SDGs）的工具。本文报告了一项荟萃分析，用于评估机器学习应用于遥感数据以监测可持续发展目标的性能。具体而言，我们旨在：1）估算平均性能；2）确定研究之间及研究内部的异质性程度；3）评估研究特征如何影响模型性能。依据PRISMA指南，我们在多个学术数据库中进行了检索，以确定可能相关的研究。三位审稿人对随机抽取的200项记录进行了筛选，最终纳入20项研究中的86项试验，涉及14个研究特征。总体准确率是报告最多的性能指标。我们采用双反正弦变换和三级随机效应模型对其进行了分析。最佳模型的平均总体准确率为0.90 $[0.86, 0.92]$。模型性能存在显著异质性，其中64%（第三级I^2）来自研究间差异。唯一显著的特征是多数类患病率，它解释了研究间异质性的61%（第三级 R^2）。其他13个特征均未提升模型表现。本文最重要的贡献是提出了以下两点见解。1）总体准确率是最常用的性能指标，但可以说也是洞见最少的指标。由于它对类别不平衡敏感，因此有必要对其进行标准化，而这远非普遍做法。2）该领域需要统一模型性能的报告规范。报告独立测试集的混淆矩阵，是比较不同研究中的机器学习分类器最重要的要素。这些发现凸显了机器学习应用亟需稳健且可比的评估指标，以确保获得可靠且可操作的见解，从而有效开展可持续发展目标监测与政策制定。

关键词：机器学习分类器、荟萃分析、嵌套数据、总体准确率、PRISMA指南

1 引言

2015年，联合国所有会员国通过了可持续发展目标（可持续发展目标），旨在应对气候变化、环境退化、贫困和不平等等全球挑战（Nations 2024年；联合国经济和社会事务部 2023；联合国全球地理信息管理专家委员会，2019年）。这项国际计划提出了17项全球目标，旨在创造更美好、更可持续的未来。可持续发展目标的实施时间已过半，但进展遭遇重大挫折，此时及时、高质量数据的重要作用比以往任何时候都更加显而易见。这些数据对于识别挑战、制定基于证据的解决方案、监测解决方案的实施情况以及进行必要的方向校正至关重要。国家统计局（国家统计局）监测可持续发展目标关键指标时，传统监测方法仍以家庭或田间层面的调查（地面采集数据）为主要数据来源（Burke 等，2021）。这些方法实施起来成本高、耗时长。因此，全球各地地面采集数据的采集频率差异显著；例如，全球24%的国家最近一次农业普查距今已超过15年（Burke 等，2021）。认识到这一挑战后，联合国经济和社会事务部（2023）和SIGD任务组（2017）均强调，应采用创新的方法学和数据来源，包括遥感和机器学习，以加强对可持续发展目标的监测与实现。

遥感数据通过卫星、飞机或无人机从远处采集，为监测范围广泛的地理区域提供了一种经济高效的方法（Khatami 等，2016；Maso 等，2023；联合国全球地理信息管理专家委员会，2019年；Zhao 等，2022）。数十年来，遥感影像一直局限于农业和社会经济领域的应用（Burke 等，2021；Lavallin and Downs，2021；Zhang 等，2022）。例如，遥感应用实验室（LARS）自20世纪60年代起便利用卫星数据和机器学习方法进行作物识别（Holloway 和 Mengersen，2018）。然而，近年来，遥感数据的空间、光谱和时间分辨率显著提高，同时，免费传感器数据的数量以及复杂数据分析所需的计算能力也大幅增长（Burke 等，2021；Thapa 等，2023；Zhang 等，2022）。应用潜力的扩大和遥感数据可用性的提升，使该领域发表的研究论文数量迅速增加（Burke 等，2021；Khatami 等，2016）。仅地球观测卫星就能测量42%的可持续发展目标指标（Zhang 等，2022）。欧盟地球观测计划哥白尼计划为陆地监测、应急管理、大气监测、海洋环境监测、气候变化和安全提供服务。地球观测数据有多个处理级别。原始数据包含光谱辐亮度（$mW\ mm^{-2}\ sr^{-1}\ nm^{-1}$），需要进行校正和定标。地球系统模型用于推导物理、化学和生物参数。还需进行更多建模，以整合不同空间和时间的数据。机器学习可用于学习经过处理的地球观测数据与通过关联登记数据或人工标注得到的标签之间的关系。本文探讨这些机器学习模型在监测可持续发展目标方面的表现。

鉴于以往研究中方法和情境的多样性，一个关键问题随之产生：哪些因素会影响使用遥感数据进行可持续发展目标监测的机器学习模型的性能？荟萃分析通过统计方式整合特定主题的大量证据，从而得出无偏的证据总结（Iliescu 等（2022年））。

本研究旨在探讨机器学习模型在应用于

可持续发展目标监测的遥感数据时表现如何。通过对该领域经同行评审的研究论文开展荟萃分析，本研究旨在：1）估计平均性能（总体效应量）；2）确定研究间及研究内的异质性程度；3）利用荟萃回归评估特定研究特征能否解释模型性能及异质性。这种方法能够对各项研究进行一致、透明且可复现的评估，这对于为参与可持续发展目标监测的决策者和利益相关方提供参考至关重要。通过考虑方法学异质性和情境因素，我们的方法有助于优先选择适合在现实场景和政策相关应用中部署的机器学习模型。

此前已有多项研究对遥感应用于可持续发展目标监测的相关研究进行了汇总。这些综述要么采用更接近不进行荟萃分析的综合方法（Campbell 等，2020）——例如 Thapa 等人（2023）以及 Ekmen 和 Kocaman（2024）——要么采用非加权元分析技术，例如 Khatami 等人（2016）和 Hall 等人（2023）。在非加权元分析中，无论样本量、质量或方差如何，所有研究都被一视同仁（Hall 和 Rosenthal，2018）。本研究采用随机效应方法，在汇总各项研究的结果时纳入原始研究的样本量。本研究旨在填补一项特定的研究空白：针对机器学习领域中利用遥感监测可持续发展目标的比例开展加权多层元分析。关于机器学习和遥感的一般文献综述，可参见 Estoque（2020）以及 Holloway 和 Mengersen（2018）等研究。

主要发现是，我们必须关注总体准确率，因为它是最常报告的性能指标。唯一显著的特征是最常见类别的占比，这一指标解释了大部分研究间异质性。这一发现表明，总体准确率的效用有限，因此有必要报告更加标准化的质量指标。本文在遥感与可持续发展目标领域的原创性和新贡献，在于结合荟萃分析与比例分析，并考虑数据的层级结构。

本文结构如下：第2节介绍遥感和机器学习的研究背景。第3节阐述所用方法。第4节呈现研究结果。第5节对研究进行讨论并总结。

2 研究背景

本节概述相关概念和方法，并介绍遥感与机器学习技术。

2.1 遥感

从最广义上说，遥感是指在不直接接触对象或现象的情况下获取相关信息（Campbell and Wynne 2011）。更具体地说，遥感是指利用安装在航空或卫星平台上的传感器，收集有关陆地或水面的数据；这些传感器记录地球表面反射或发射的电磁辐射（Campbell and Wynne 2011）。遥感的起源可追溯到19世纪摄影技术的发展，最早的航空或地球观测照片是用相机拍摄的

这些相机安装在气球、风筝、鸽子和飞机上（Burke et al. 2021; Campbell and Wynne 2011）。遥感首次大规模应用是在第一次世界大战期间，主要用于航空摄影。现代基于卫星的遥感时代始于1972年Landsat 1的发射，它是首颗专为地球观测设计的卫星（Campbell and Wynne 2011）。如今，遥感技术能够以全球覆盖的方式，频繁而系统地采集地球表面数据，彻底改变了我们监测和分析地球表面的能力（Burke et al. 2021; NASA 2024）。每年发射的卫星数量随时间不断增加，尤其是用于通信目的的卫星（如Starlink卫星），但用于地球观测的卫星数量也在增加（图1）。

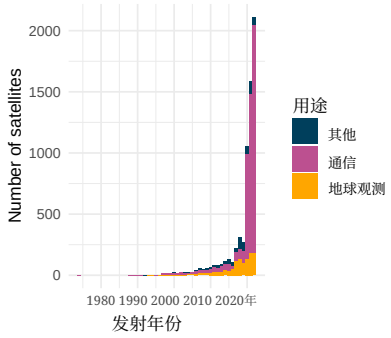

图1：按用途划分的每年发射卫星数量。数据来源：UCS 2023。

卫星等遥感设备上的传感器会测量地球表面对象反射的电磁辐射。这一过程有两种不同的方式：被动式和主动式。被动传感器依靠阳光等自然光源，记录地球表面反射的入射能量。主动传感器则自行产生能量，并将其发射出去，再测量其从地球表面反射回来的能量（美国国家航空航天局 2024年）。

地球表面的不同组成部分具有不同的光谱特征——也就是说，它们反射、吸收或透射的能量在数量和波长上各不相同（Campbell and Wynne 2011年）。遥感设备配备多个传感器，用于测量电磁波谱中特定范围的波长；这些范围称为光谱波段，例如红外线、可见光或紫外线（美国国家航空航天局 2024年）。通过获取特定波段的信息，可以利用地表的光谱特征识别地面上的对象。

2.2 机器学习

机器学习是一个研究领域，专注于开发和应用能够通过学习从数据中发现模式的算法。机器学习技术可应用于不同类型的数据。如果数据未标注，即仅包含特征而不包含感兴趣变量，则可应用无监督机器学习技术对实例进行分组（聚类）或对特征进行分组（降维）。如果数据已标注，即同时包含特征和感兴趣变量，则可应用监督机器学习技术来映射特征与标签之间的关系，并预测未标注实例的感兴趣变量，

如果感兴趣变量是分类变量，机器学习要解决的就是分类问题。如果感兴趣变量是数值型变量（离散型或连续型），机器学习要解决的就是回归问题。本文聚焦于机器学习分类器。

机器学习分类器（如神经网络（NN）、随机森林（RF）和支持向量机（SVM））长期以来一直应用于空间数据分析和地理建模（Lavallin and Downs 2021）。关于神经网络，我们参阅 Goodfellow 等（2016）、Ma 等（2019）、Reichstein 等（2019）和 Zhu 等（2017）；关于随机森林，我们参阅 Belgiu and Drăguţ（2016）、Breiman（2001）、Cutler 等（2007）、Rodriguez-Galiano 等（2012）和 Strobl 等（2007）；关于支持向量机，则参阅 Huang 等（2002）、Melgani and Bruzzone（2004）、Mountrakis 等（2011）和 Vapnik（1995）。我们还参阅 Kavvada 等（2020），他们对这些方法在各个可持续发展目标中的应用进行了全面概述。对于对这些模型的数学表示感兴趣的读者，我们推荐 Bishop（2006）、Hastie 等（2009）和 James 等（2021）编写的标准教材。关于可持续发展目标相关领域所用人工智能技术的近期综述，请参阅 Fan 等 2023；Greif 等 2024。

与单独使用光谱指数（例如植被指数）相比，机器学习技术提高了数据分析和解读过程的准确率与效率，使得有效分析大量数据成为可能。这对于处理遥感数据的高度复杂性和高维性尤为有用。近年来，随着大型数据集日益丰富以及计算能力不断提升，机器学习技术在遥感领域的应用迅速增长（联合国全球地理信息管理专家委员会，2019年；Zhang 等，2022）。

性能指标用于在独立测试集上检验模型预测的质量。对于分类任务，这些指标源自混淆矩阵：一种交叉制表，统计观测到的类别为 $r \in \{1, \cdots \ q\}$ 且预测为类别 $c \in \{1, \cdots \ q\}$ 的实例数量，其中 q 为类别数量（表 1）。在混淆矩阵中，正确分类的实例位于对角线上，而非对角线单元格则表明哪些类别容易混淆，即被错误分类。在遥感应用中，一个实例可以是一个像素、一组像素或一幅图像（Stehman and Foody，2019年）。

表 1：基于对 n个实例进行分类所得的单次总体准确率估计的混淆矩阵。

		预测				
	类别	1	$\cdots$	q	Sum	
观测到的	1	n_{11}	$\cdots$	n_{1q}	$n_{1.}$	R_1
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
	q	n_{q1}	$\cdots$	n_{qq}	$n_{q.}$	R_q
	Sum	$n_{.1}$	$\cdots$	$n_{.q}$	n	
		V_1	$\cdots$	V_q		p

根据该矩阵，可以推导出若干性能指标（Finegold 等人，2016年；Stehman and Foody，2019年；联合国全球地理信息管理专家委员会，2019年）。最简单的是总体准确率，即正确分类的实例所占比例（$p = k/n$，其中 $k = \sum_{r=1}^{q} n_{rr}$）。该指标并非类别特定指标。两个类别特定指标是召回率（又称敏感度或生产者精度）和精确率（又称可靠性或用户精度）。

类别 r 的召回率是类别 r 中观测到且被正确预测的实例所占比例（$R_r = n_{rr}/n_{r.}$）。类别 c 的精确率是预测为类别 c 且实际观测结果确为该类别的实例所占比例（$V_c = n_{cc}/n_{.c}$）。F1分数和平均精确率是两种类别特定指标，用于平衡某一类别的召回率与精确率之间的权衡。算术平均召回率或几何平均召回率，以及受试者工作特征曲线（ROC曲线）下面积（R_r 与 $R_{\neg r}$之间），都是用于平衡各类别召回率之间权衡的指标。需要认识到，可以从混淆矩阵中推导出许多性能指标；这些指标用途各异，而且大多数对类别不平衡（偏离离散均匀分布）较为敏感。

3 方法

3.1 文章筛选

本研究采用的方法按照Debray等人（2017）提出的框架分步骤进行描述。此外，我们还努力遵循系统综述和荟萃分析首选报告条目（PRISMA）指南（Page等人，2021年），尽管始终无法严格遵守。下文将详细介绍文章筛选标准。

我们于2024年1月15日和16日从包括ScienceDirect和Taylor & Francis Online在内的多个学术数据库中，收集了2018年1月至2023年12月期间发表的同行评审文章（图2）。为减少数据库覆盖范围可能造成的偏差，我们使用了多个学术数据库（Hansen 等，2022；Tawfik等人，2019年）。尽管谷歌学术可用于补充检索和查找灰色文献，但通常认为其不适合作为系统综述的主要来源（Gusenbauer和Haddaway，2020）。此外，谷歌学术的检索结果无法完全复现（Gusenbauer和Haddaway，2020），且检索结果中的参考文献无法批量下载。因此，我们决定不使用谷歌学术检索论文。

对于数据库查询，全文检索选用了三个搜索词：“遥感” AND “机器学习” AND “可持续发展目标”。之所以选择这些词，是因为我们认为，探讨利用遥感监测可持续发展目标的建模方法的论文很可能会包含这些词。起初，我们尝试加入常见缩写“ML”和“SDGs”。然而，最终决定不将其纳入搜索词，因为加入这些缩写会增加无关结果的比例（以 ScienceDirect 为例，加入缩写后，2018—2023 年的论文数量从 606 篇增至 1064 篇，增幅达 76%。例如，缩写“ML”会额外检索出涉及“最大似然”估计的论文。“SDG”也可能指梯度标准差（见于某些优化文献）或子采样双梯度（见于特定算法））。因此，最终仅使用本段开头提到的三个搜索词，未使用其他搜索词串。此次检索共获得 998篇论文。我们以 RIS格式 下载了这些数据库的搜索结果，并将其导入 Zotero 以作进一步处理。重复文章通过 Zotero 的“合并重复项”功能处理。

从998篇论文中抽取了一个包含200篇论文的简单随机样本，用于标题和摘要筛选。三位作者分别独立筛选了这些可能

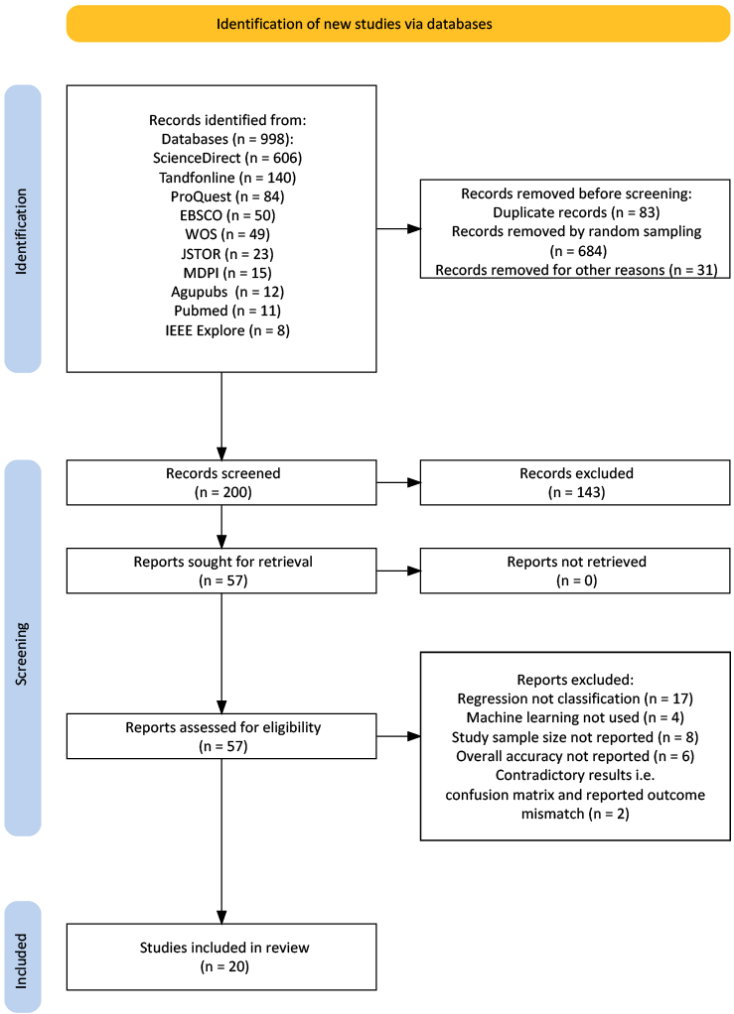

图2：PRISMA流程图。检索与研究筛选过程。

相关论文，使用R软件包 `metagear` (Lajeunesse 2016)。用户界面中突出显示的关键词有助于加快筛选流程，详见表2。论文根据以下标准进行筛选：(a) 使用遥感和机器学习技术的出版物；(b) 提及质量评估，例如总体准确率。在筛选的200篇摘要中，只有57篇被三位审稿人一致认为可能相关。

为获得可比较的性能指标，研究决定聚焦于分类相关论文。研究人员使用 `metagear` 筛查了57篇文章的标题和摘要，将其分为分类论文（40篇）和回归论文（17篇）。由于回归论文需要采用不同的质量指标，因此予以排除。另有两项研究因其报告的混淆矩阵与所报告的结果不符，被认为不可靠，故予以排除

表2： `metagear` 在标题和摘要筛选期间标出的关键词。

类别	关键词
一般	分析、实证、预测、结果、可持续发展
Data	地球观测、遥感、遥感、卫星
模型	人工神经网络、装袋法、贝叶斯、提升法、分类与回归树、分类、分类器， 决策树、深度学习、梯度、机器学习、神经网络， 随机森林、回归、回归树、随机森林、监督学习、支持- 向量机、支持向量机、测试集、训练集
质量指标	准确率，曲线下面积，决定系数，误差，F1分数，平均绝对误差， 平均绝对误差，均方误差，总体准确率，精确率，召回率， 均方根误差，受试者工作特征曲线，敏感度，特异度
省略	荟萃分析，综述，系统综述

。对研究层级特征报告不完整的研究则予以保留，并将该特征编码为“未知”（见第3.3节）。

在40篇分类论文中，总体准确率是最常报告的性能指标，因此决定纳入所有报告总体准确率的论文。只有一半入选的分类论文报告了荟萃回归所需的信息。大多数论文（$j = 1, \ldots\ldots, h$）报告了多个（m_j）来自不同试验的总体准确率估计值，因此，来自 $h = 20$ 篇论文的总体准确率估计值共有 $m = 87$个（表3）。 p_{ij} 是研究 j中试验 i 观察到的总体准确率。数据的层级结构在统计分析小节中进行处理。

3.2 特征

根据筛选出的论文和以往的系统综述，整理出一份潜在研究特征清单（表4）。这些特征提供了有关分类任务复杂程度（例如输出类别数量）以及多数类比例的信息，可用于指示可能影响分类模型性能的类别不平衡问题。所有论文均将已标记数据随机划分为训练集和测试集，因此多数类比例在两组间不存在系统性差异。我们还收集了遥感特有的信息，包括设备类型、光谱波段数量和像素分辨率，以评估数据采集对性能的影响。由于相关信息的未报告比例较高，后两项采用了类别化处理。引用次数通过 Local Citation Network 网络应用程序获取；该程序从 OpenAlex（科学论文的书目目录）收集文章元数据（Priem等，2022年）。

3.3 数据探索

为便于说明，图3按分类特征展示了各试验中观测到的总体准确率分布，图4则按数值特征和二元特征展示了该分布。这种数据探索仅限于观察层面，不得据此得出任何结论

表3：按每项研究的试验数量和发表年份排序的入选研究。

研究 j	参考文献	试验数量，m_j	p_{ij}
1	Nazir等人（2020年）	27	$p_{1,}{}^{1,\ \cdots\cdots,\ p_{27}}{}_{,1}$
2	Shen等人（2024年）	7	$p_{1,}{}^{2,\ \cdots\cdots,\ p_{7}}{}_{,2}$
3	Priyanka等人（2023）	7	$p_{1,}{}^{3,\ \cdots\cdots,\ p_{7}}{}_{,3}$
4	Jochem 等（2018）	7	$p_{1,}{}^{4,\ \cdots\cdots,\ p_{7}}{}_{,4}$
5	Solórzano 等（2023）	6	$p_{1,}{}^{5}, \ldots, p_{6,5}$
6	Aja 等（2022）	6	$p_{1,}{}^{6}, \ldots, p_{6,6}$
7	Agrillo 等（2021）	4	$p_{1,}{}^{7}, \ldots, p_{4,7}$
8	Shaharum 等（2020）	3	$p_{1,}{}^{8,\ \cdots,\ p_{3}}{}_{,8}$
9	Tao 和 You（2019）	3	$p_{1,}{}^{9,\ \cdots,\ p_{3}}{}_{,9}$
10	Al-Dousari 等（2023）	2	$p_{1,}{}^{10,\ p_{2}}{}_{,10}$
11	Fagua 等（2023）	2	$p_{1,}{}^{11,\ p_{2}}{}_{,11}$
12	Li 等（2022）	2	$p_{1,}{}^{12,\ p_{2}}{}_{,12}$
13	Xia 等（2022）	2	$p_{1,}{}^{13,\ p_{2}}{}_{,13}$
14	Mashaba-Munghemezulu 等（2021年）	2	$p_{1,}{}^{14,\ p_{2}}{}_{,14}$
15	Jia 等（2023）	1	$p_{1,15}$
16	Wang 等（2023）	1	$p_{1,16}$
17	Owers 等（2022）	1	$p_{1,17}$
18	Peng 等（2022）	1	$p_{1,18}$
19	Suryono 等（2022）	1	$p_{1,19}$
$h = 20$	Pareeth 等（2019）	1	$p_{1,20}$
		$m = \sum_j m_j = 86$	

。结论将基于第 3.4 节所述的统计分析得出。

表4对各项特征进行了说明。根据图3中的分类特征，可以得出以下具体观察结果：被动感知的总体准确率通常高于主动式或组合系统（设备）。基于树的模型和神经网络均能达到较高的准确率范围，而基于树的方法表现出更大的差异（模型）。就波段数量（波段数）而言，光谱丰富度似乎能提升模型性能。然而，总体准确率最高的是波段数量未报告的组别。不同分辨率组别未呈现明显规律（分辨率）。陆地卫星和Sentinel卫星的表现相近，且陆地卫星的差异较小（卫星）。不过，其他卫星的表现优于这两者，但也存在一些离群值。卫星类型未知的研究表现也不错，差异较小。可持续发展目标15（陆地生物）的表现差异较大（可持续发展目标），而有关可持续发展目标11（可持续城市）的研究总体准确率通常较高。可持续发展目标2（零饥饿）的总体准确率差异最小。就单位而言，面向对象分类的总体准确率通常高于基于像素的分类（单位）。

关于图 4 中的数值型和二元特征，可以得出以下具体观察结果：类别不平衡（以最大类别占比（maxprev）衡量）与总体准确率之间呈明显的正相关。这表明，在不平衡数据集上训练的模型可能会偏向主导类别，从而获得较高的准确率。引用次数（ncitations）与较高的准确率呈中等程度的相关性，这可能反映了方法学质量或发表偏倚。这一特征并非因果因素，而是作为控制变量，用来衡量论文的影响力或发表时间。相反，类别数量（nclasses）与准确率呈负相关，凸显了更复杂的分类任务难度更高。

表 4：研究层级特征。

特征 x_j	描述	取值
分类变量		
设备	遥感设备	被动式、主动式、组合式、未知
模型	机器学习模型类型	神经网络、基于树的、其他
波段数	光谱波段数量	[1,7], (7,14],未知
分辨率	像素分辨率（米）	(0,10], (10,30],未知
卫星	卫星类型	Sentinel、陆地卫星、其他、未知
sdg	联合国可持续发展目标	可持续城市（可持续发展目标11），陆地生物（可持续发展目标15），零饥饿（可持续发展目标2）
unit	分类单位	对象、像素、未知
数值型		
多数类患病率	多数类的流行率	0.142–0.995
引用次数	引用次数	0–68
类别数	类别数量	2–13
year	发表年份	2018–2023
二元		
辅助	辅助变量使用指示 除遥感数据外的辅助数据	0,1
混淆矩阵	混淆矩阵的报告指标矩阵	0,1
指数	遥感指数的使用 传感指数	0,1

年份变量显示，准确率可能随时间略有提升。使用辅助数据可显著提升性能，采用外部数据集的模型优于未采用此类数据集的模型。同样，纳入光谱指数也与较高的准确率相关。研究是否报告混淆矩阵（confmatrix）似乎不会影响所报告的准确率，这表明报告透明度未必与性能结果相关。

总而言之，根据这项目视检查，最有效的配置可能包括被动传感器、神经网络、更多光谱波段、Landsat影像、基于对象的分析、与可持续发展目标11相关的任务、在不平衡数据上训练的模型、辅助数据和指数的纳入，以及复杂度较低的分类任务。下一节将使用其中介绍的统计分析对这一配置进行检验。

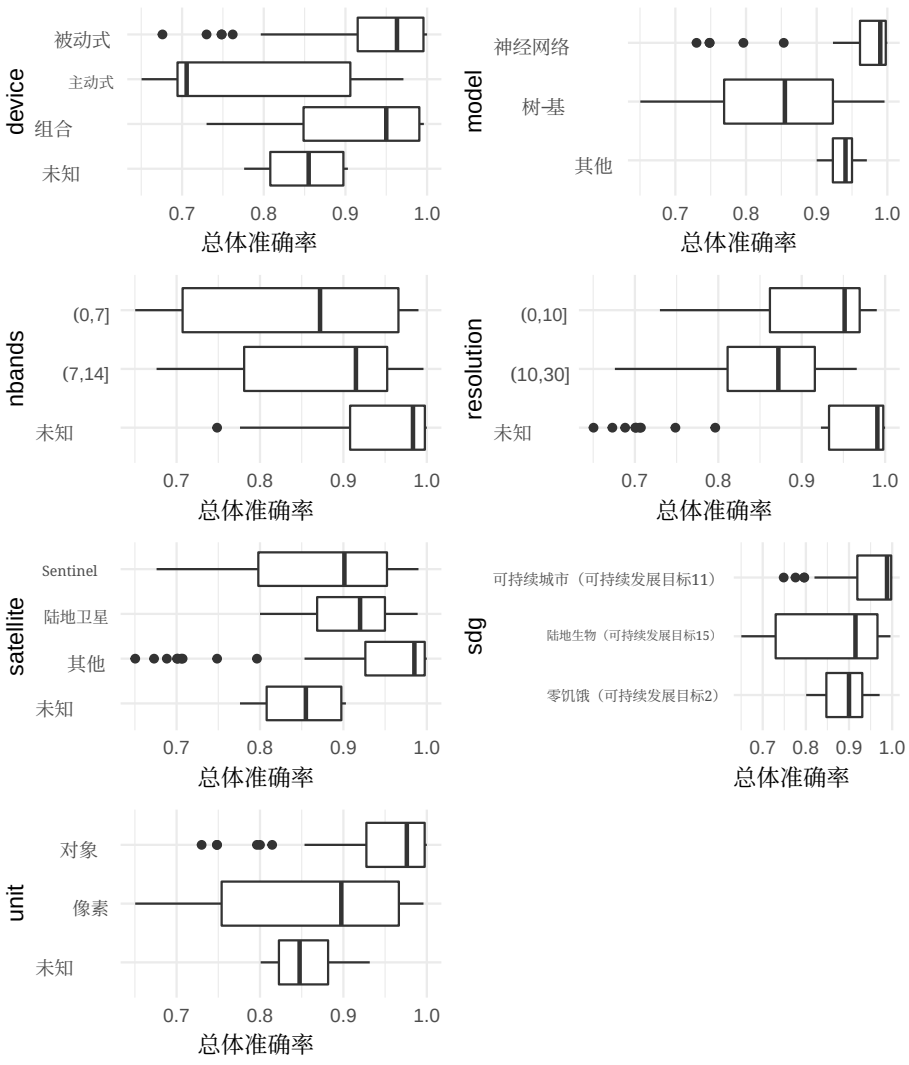

图3：按分类特征划分的总体准确率分布。

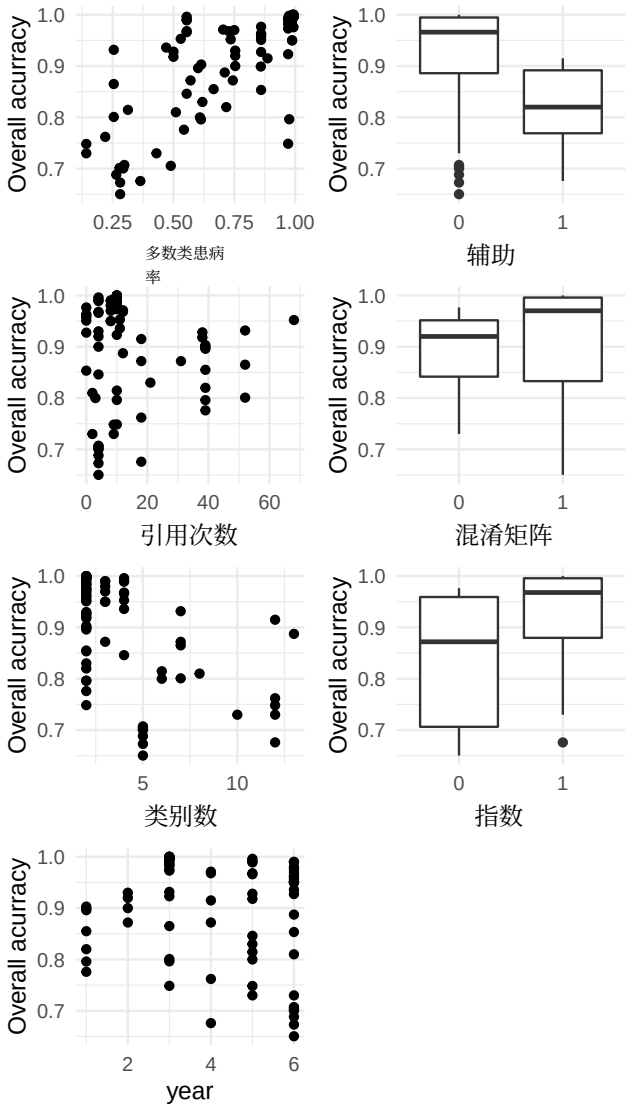

图 4：按数值特征（左列）和二元特征（右列）划分的总体准确率分布。

3.4 统计分析

变换

分析前，对观测到的总体准确率 p_{ij}应用 Freeman 和 Tukey（1950年）双反正弦变换（DA），得到观测到的个体效应量 $\tilde{}$及其抽样方差 v_{ij}:

$$p_{ij} = \frac{k_{ij}}{n_{ij}}, \tag{1}$$

$$\tilde{\theta}_{ij} = \mathrm{DA}(p_{ij}) = \frac{\sin^{-1}\sqrt{\frac{k_{ij}}{n_{ij}+1}} + \sin^{-1}\sqrt{\frac{k_{ij}+1}{n_{ij}+1}}}{2}, \tag{2}$$

$$v_{ij} = 1/(4n_{ij}+2). \tag{3}$$

其中， k_{ij} 是正确分类实例的数量， n_{ij} 是研究 j中试验 i的实例总数。经过一些争议后，这种变换被认为更适用于比例的荟萃分析（Doi 和 Xu 2021年）。根据 Shapiro-Wilk检验，与 DA 相比，不进行变换或采用其他变换（如对数、logit 和反正弦平方根变换）会导致更严重的偏离正态性。

三级元分析

拟合多层线性模型，以估计总体效应量以及研究之间和研究内部的异质性。该模型考虑到同一项研究中的个体效应量并非相互独立。

在个体层面（1），研究 j 中试验 i 的观测个体效应量。它被建模为（未观测到的）真实个体效应量 θ_{ij} 与误差成分 ϵ_{ij}之和。假设误差成分服从均值为 0、抽样方差为 v_{ij}的正态分布：

$$\begin{aligned} \tilde{\theta}_{ij} &= \theta_{ij} + \epsilon_{ij}, \\ \epsilon_{ij} &\sim N(0, v_{ij}). \end{aligned} \tag{4}$$

在研究层面（2），研究 j 中试验 i 的真实个体效应量 θ_{ij} 被建模为研究 j 中的研究效应量 κ_j 与研究内异质性 ζ_{ij}之和。假设这种异质性服从均值为 0、研究内方差为 σ_ζ^2 的正态分布：

$$\begin{aligned} \theta_{ij} &= \kappa_j + \zeta_{ij}, \\ \zeta_{ij} &\sim N(0, \sigma_\zeta^2). \end{aligned} \tag{5}$$

请注意，假设研究内方差 σ_ζ^2 在各项研究中保持恒定，这是一个常见假设。在总体层面（3），研究 j 中的研究效应量 κ_j 被建模为总体效应量 μ 与研究间异质性 ξ_j之和。假设这种异质性服从均值为 0、研究间方差为 σ_ξ^2的正态分布：

$$\kappa_j = \mu + \xi_j,$$
$$\xi_j \sim N(0, \sigma_\xi^2). \tag{6}$$

综上，观测到的个体效应量可分解为总体效应量、研究间异质性、研究内异质性和误差成分： $\tilde{\theta}_{ij} = \mu + \xi_j + \zeta_{ij} + \epsilon_{ij}$。其总方差为 $\text{Var}(\tilde{\theta}_{ij}) = \sigma_\xi^2 + \sigma_\zeta^2 + v_{ij}$。总体效应量 μ 是 $m = \sum_j m_j$ 观测到的个体效应量 θ_r 的加权和（ Cheung 2014 ; Konstantopoulos 2011）：

$$\mu = \frac{\sum_{r=1}^{m}(\sum_{c=1}^{m} w_{rc})\tilde{\theta}_r}{\sum_{rc} w_{rc}}, \tag{7}$$
$$\text{Var}(\mu) = \frac{1}{\sum_{rc} w_{rc}}. \tag{8}$$

权重 w_{rc} 并非简单地取方差倒数 $1/(\sigma_\xi^2 + \sigma_\zeta^2 + v_{ij})$，而是考虑了同一项研究中个体效应量之间的协方差。它是 r行 c列中的元素，属于 $m \times m$ 权重矩阵 $\boldsymbol{W}$，该矩阵是边际方差-协方差矩阵 $\boldsymbol{M}$的逆矩阵。该矩阵 $\boldsymbol{M}$在每项研究的对角线上包含 $\sigma_\xi^2 + \sigma_\zeta^2 + v_{ij}$，在非对角线上包含 σ_ξ^2:

$$\boldsymbol{W} = \boldsymbol{M}^{-1}, \tag{9}$$
$$\boldsymbol{M} = \oplus_{j=1}^{h} \left(\boldsymbol{J}_{m_j} \sigma_\xi^2 + \boldsymbol{I}_{m_j} \sigma_\zeta^2 + \boldsymbol{I}_{m_j} v_{ij} \right), \tag{10}$$

其中， $\oplus$ 是矩阵直和， $\boldsymbol{J}_{m_j}$ 是一个 m_j阶全1矩阵，而 $\boldsymbol{I}_{m_j}$ 是一个 m_j阶单位矩阵。因此，假定不同研究中的个体效应量彼此独立（$\text{Cov}(\theta_{ij}, \theta_{kl}) = 0$），而同一项研究中的个体效应量具有相同的协方差（$\text{Cov}(\tilde{\theta}_{ij}, \tilde{\theta}_{kj}) = \sigma_\xi^2$）（Cheung 2014年；Konstantopoulos 2011年）。参数采用限制性最大似然（REML）进行估计。还有其他选项，但如果目标变量是连续型，REML 是合适的。

为评估三个方差分量对总方差的相对贡献，将总方差简化为 $\text{Var}(\tilde{\theta}) = \sigma_\xi^2 + \sigma_\zeta^2 + \sigma_\epsilon^2$，其中 σ_ϵ^2 是 m 抽样方差 v_{ij} 的合并估计值（Higgins 和 Thompson，2002年）：

$$\sigma_\epsilon^2 = \frac{(m-1)\sum_{ij} 1/v_{ij}}{(\sum_{ij} 1/v_{ij})^2 - \sum_{ij} 1/v_{ij}^2}. \tag{11}$$

三层元回归

研究之间和研究内部的方差可能由特征（又称协变量、预测变量、自变量、调节变量）解释。为检验这一点，我们按以下方式扩展多层模型。设 $\boldsymbol{x}_j$ 为关于研究 j的 d 个特征组成的向量。也可以纳入子研究层面特征 $\boldsymbol{x}_{ij}$，但由于我们没有相关特征，此处将其忽略。模型变为:

$$\tilde{\theta}_{ij} = \mu + \boldsymbol{x}_j^\intercal \boldsymbol{\beta}_j + \xi_j + \zeta_{ij} + \epsilon_{ij}, \tag{12}$$

其中， β_j 为由 d 个回归系数组成的向量。异质性 $\xi_j + \zeta_{ij}$ 现在指真实个体效应量 θ_{ij} 之间未能由这些特征解释的变异。我们采用前向选择程序和赤池信息准则（AIC），选择相关特征，以在模型拟合与简约性之间取得适当平衡（Venables 和 Ripley, 2002年）。信息准则的任何下降都被视为改进。

在模型中加入特征会降低总体层面（ξ）和研究层面（ζ）的方差。降低幅度用 R^2 表示:

$$R^2_{\xi} = 1 - \frac{\sigma^2_{\xi}(X)}{\sigma^2_{\xi}(0)}, \tag{13}$$

$$R^2_{\zeta} = 1 - \frac{\sigma^2_{\zeta}(X)}{\sigma^2_{\zeta}(0)}, \tag{14}$$

其中， $\sigma^2(X)$ 是包含特征的模型对应的方差，而 $\sigma^2(0)$ 是不包含特征的模型（零模型）对应的方差。

反变换

建模基于双反正弦变换后的比例进行，但为了便于解释，结果会进行反变换。Freeman 和 Tukey（1950）提出的双反正弦变换的反变换最初由 Miller（1978）提出。为避免数值不稳定，Barendregt 等人（2013）将样本量的调和平均数（$\hat{n} = m/\sum_{ij} n_{ij}^{-1}$）替换为总体方差的倒数（$\hat{n} = 1/\mathrm{Var}(\mu) = \sum_{rc} w_{rc}$），作为样本量 n 的估计值:

$$\mathrm{DA}^{-1}(\gamma) = \frac{1 - \mathrm{sgn}\left(\cos(\gamma)\right)\sqrt{1 - \left[\sin(\gamma) + \frac{\sin(\gamma) - \frac{1}{\sin(\gamma)}}{\hat{n}}\right]^2}}{2}, \tag{15}$$

$$\overline{p} = \mathrm{DA}^{-1}(2\mu), \tag{16}$$

$$\mathrm{LCB}(\overline{p}) = \begin{cases} 0 & \text{if } \overline{p}\hat{n} < 2 \\ \mathrm{DA}^{-1}\left(2\left[\mu - t_{1-\alpha/2,h-1}\sqrt{\mathrm{Var}(\mu)}\right]\right) & \text{otherwise} \end{cases}, \tag{17}$$

$$\mathrm{UCB}(\overline{p}) = \begin{cases} 1 & \text{if } (1-\overline{p})\hat{n} < 2 \\ \mathrm{DA}^{-1}\left(2\left[\mu + t_{1-\alpha/2,h-1}\sqrt{\mathrm{Var}(\mu)}\right]\right) & \text{otherwise} \end{cases}, \tag{18}$$

其中， γ 是任意经双反正弦变换的比例，sgn 是符号函数， p 是反变换后的总体效应量，LCB 是置信下限，UCB 是置信上限，而 $t_{1-\alpha,\ \nu}$ 是显著性水平为 α 、自由度为 ν 时 Student t 分布的临界值。将总体方差的倒数用作有效样本量，比使用调和平均数更稳定，因为前者根据估计值的精确率对研究进行加权，而不是让较小样本量产生不成比例的影响。对于包含研究层级特征的模型， p 替换为 p_j， μ 替换为 $\mu + \boldsymbol{x}_j^{\intercal}\beta_j$，而 $\hat{n}$ 替换为 $\hat{n}_j$。

基准

假设只是简单地对每个类别 r 进行猜测，其概率等于该类别在训练集中的患病率 $n_{r.}/n$ （见第 2.2 节）。那么，基准总体准确率等于：

$$p_0 = \sum_{r=1}^{q} \left(\frac{n_{r.}}{n} \right)^2 . \tag{19}$$

例如，如果两个类别处于均衡状态，那么 $p_0 = 0.5$。但如果十个类别处于均衡状态（均匀分布），那么 $p_0 = 0.1$。如果两个类别不均衡，比例为9比1，那么 $p_0 = 0.82$。一般而言，类别越少、类别不平衡程度越高，基准就越高，也越容易获得较高的总体准确率。因此，在分析不同研究中的总体准确率时，将类别数量和类别不平衡程度作为特征纳入考量十分重要。另一种做法是用基准对总体准确率进行标准化，但这并非惯常做法。请注意，为了获得公平的测试结果，需要确保训练集和测试集中的类别分布是对总体类别分布的无偏估计，例如可以通过加权来实现。

实现

统计分析使用了以下 R函数： `escalc`、 `vcalc`、`rma.mv`、 `rma`、 `forest` 和 `regplot` ，它们来自软件包 `metafor` （Viechtbauer 2010）；以及来自软件包 `dmetar` （Harrer 等，2019）的 `var.comp` 。

4 结果

4.1 三级元分析

表5总结了结果。根据针对异质性的 Cochran Q 检验，观测效应量的变异性大于仅依据抽样方差所预期的水平。这表明真实个体效应量具有异质性。在不纳入特征的情况下，研究间方差 σ^2_ξ 占˜总方差 Var(θ_{ij}) 的64%，其余部分归因于研究内方差 σ^2_ζ，因此合并抽样方差 $\tilde{v}$ 的贡献可以忽略不计。表5还给出了双反正弦尺度上的总体效应量 μ及其方差，但将其反变换为 p后更易于解释：质量指标总体准确率的总体估计值为0.89，95% 置信区间为0.85–0.93。

–

图5展示了森林图。每一行代表一项 $h = 20$ 研究及其m_j 试验数量。接下来，图中展示了 p 质量指标总体准确率的 $\overline{h} = 20$ 研究层面估计值 κ_j 及其置信区间。方块的大小表示 κ_j的取值，并由 j的大小决定；也就是说，每项研究的试验越多，方块就越大。底部一行展示了质量指标总体准确率的总体估计值。虚线表示总体估计值，该值也落在所有研究的置信区间内。

表5：拟合模型汇总。 I^2 和 R^2 统计量以分数形式报告。

	荟萃分析	荟萃回归
Q	1.2×10^7 ($p < .0001$)	1.1×10^7 ($p < .0001$)
df	85	84
σ^2_ξ ($h = 20$)	0.0173 ($I^2_\xi = 0.64$)	0.0068（$I^2_\xi = 0.43$）
σ^2_ζ ($m = 86$)	0。0099 ($I^2_\zeta = 0.36$)	0.0091 ($I^2_\zeta = 0.57$)
σ^2_ϵ	10^{-8} ($I^2_\epsilon = 10^{-6}$)	10^{-8} ($I^2_\epsilon = 10^{-6}$)
R^2_ξ		0.61
R^2_ζ		0.08
μ (标准误)	1.2384 (0.0331)	0.9796 (0.0557)
β(多数类患病率) (SE)		0.4335 (0.0833)

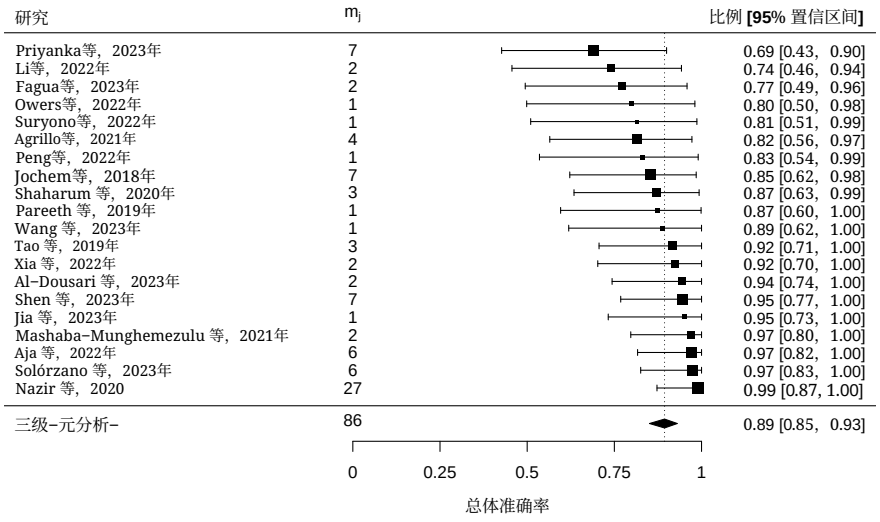

图5：荟萃分析的森林图。

4.2 三层元回归

图6展示了应用前向模型选择的结果。图的纵轴显示前向模型选择中的特征和特征组合，横轴显示不同质量指标的取值。每个面板代表一种不同的质量指标。本研究选用了三种广泛使用的模型评估指标：赤池信息准则（AIC）、贝叶斯信息准则（BIC）和均方根误差（RMSE）。AIC 平衡了以下两者之间的权衡：

拟合优度和模型复杂度。贝叶斯信息准则与赤池信息准则类似，但对模型复杂度施加了更严格的惩罚。均方根误差是预测误差的平均幅度，即预测值与实际值之间的平均距离。通过使用这三种不同的指标，我们从不同角度考察模型，旨在找出不仅准确，而且简约且具有泛化能力的模型。

对于所有质量指标，多数类患病率（maxprev）对报告的总体准确率都有显著影响。对于赤池信息准则和贝叶斯信息准则，增加任何额外特征都不会降低其取值（较低的赤池信息准则或贝叶斯信息准则取值表示模型更优）。根据均方根误差，可以加入辅助数据使用情况的指标，也可以加入遥感设备（例如被动式、主动式）的指标。不过，该指标侧重于模型拟合，未考虑模型复杂度。我们还观察到，根据赤池信息准则和贝叶斯信息准则，与零模型相比，若干仅使用单一特征的模型表现更差（这些特征的取值位于虚线右侧，因为虚线表示零模型的取值）。因此，这些特征只会增加噪声，无法捕捉数据中的任何信号。

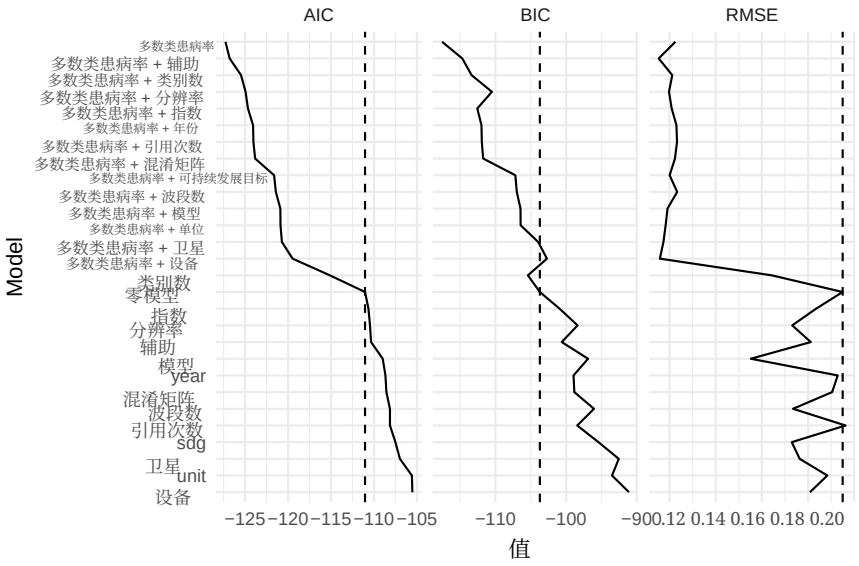

图6：前向特征选择。模型按赤池信息准则排序。虚线表示零模型（仅含截距）的取值。由于向仅包含多数类患病率（maxprev）的模型添加特征并未降低赤池信息准则，因此未拟合更复杂的模型。横轴显示赤池信息准则、贝叶斯信息准则或均方根误差的取值。较低的赤池信息准则和贝叶斯信息准则取值意味着模型在拟合度与复杂度之间的平衡更好。均方根误差反映预测性能，但未考虑复杂度。

此外，我们还进行了置换检验，以评估所采用特征选择方法的稳健性。最重要的发现是，置换特征重要性（Molnar 2025）证实，多数类患病率（maxprev）是最重要的特征（见附录A）。

将 maxprev 作为特征纳入模型，并不会改变 Cochran Q 检验对异质性的检验结果（表5）。现在，特征 maxprev 可以部分解释这种异质性。将 maxprev 纳入模型后，研究间方差所解释的总方差比例从 64% 降至 43%（I^2_ξ），相对减少了 61%（R^2_ξ）。由于 maxprev 是研究层面的特征，它对研究内方差的影响要小得多（$R^2_\zeta = 8\%$）。将 maxprev 纳入模型（表5中的 β(maxprev)）也缩小了总体估计值（针对平均 maxprev）和 $h = 20$ 研究层面估计值 κ_j的置信区间。这一点在图7中有所体现，可与图5进行比较。质量指标总体准确率的总体估计值现在为 0.90，95% 置信区间为 0.86–0.92。尽管如此，大多数研究的置信区间仍涵盖总体估计值。不过，现在有一项研究的估计值显著较低（Priyanka等，2023年），另有一项研究的估计值显著较高（Nazir 等，2020）。我们通过剔除包含 27 次试验的 Nazir 等人（2020）研究，进行了逐一剔除研究分析。模型选择结果显示，根据赤池信息准则，最佳模型包含 maxprev 和分辨率；而根据贝叶斯信息准则，最佳模型仍仅包含 maxprev。因此，即使不纳入 Nazir 等人（2020）的研究，本文的主要结论也不会改变。

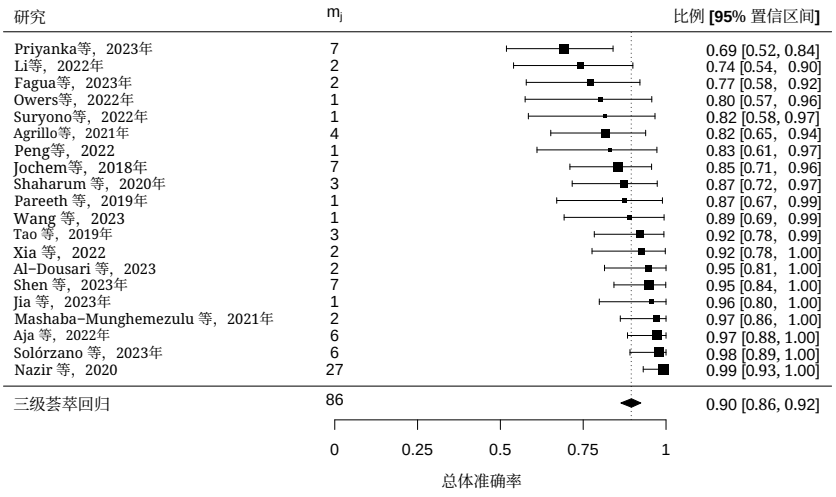

图7：荟萃回归森林图，其中多数类患病率（maxprev）是唯一显著特征。

总体准确率随着多数类患病率的升高而提高。DA尺度上的回归系数为正（表5）。图 8 在反变换尺度上展示了这一结果。纵轴表示反变换后的总体准确率，横轴表示多数类患病率。图中绘出了 $h = 20$ 研究及其 $m = 86$ 试验。黑色实线表示回归线，虚线表示其 95% 置信区间。因此，一个类别越占主导地位，总体准确率就越高。这一发现符合预期，且从理论上可以解

如第 3.4 节、公式 19 所述。

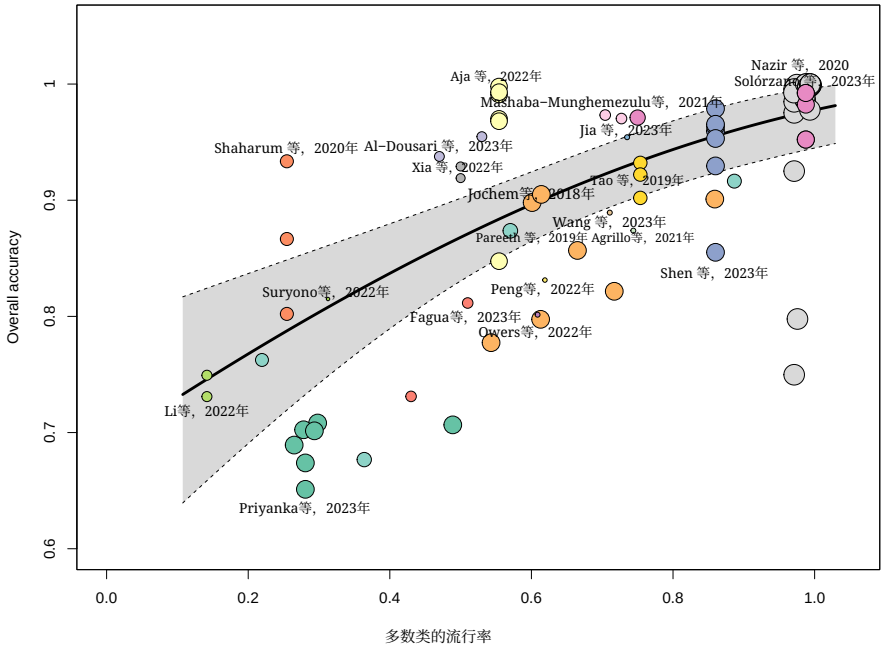

图 8: $m = 86$项试验（来自 $h = 20$ 项研究）的总体准确率与多数类患病率的关系。数据点按研究着色，点的大小取决于模型权重。回归线和 95% 置信区间来自 DA 反变换三层元回归模型。请注意，y 轴并非从零开始。

5 讨论

本荟萃分析评估了使用遥感数据进行可持续发展目标监测的机器学习模型性能。具体而言，本研究估算了平均性能，确定了研究间和研究内的异质性水平，并评估了哪些研究特征与模型性能相关。

结果显示，总体准确率始终较高。不含特征时，总体总体准确率估计为 $p = 0.89\ [0.85, 0.93]$；将多数类患病率（maxprev）作为唯一纳入的显著特征后，估计结果更为精确，为 $0.90\ [0.86, 0.92]$。结果还表明存在显著异质性，其中 64% 来自研究之间的差异，且 61% 可由 maxprev 特征解释。其余十三项研究特征均未产生显著的额外

影响。这与 Khatami 等人（2016）和 Hanadé Houmma 等人（2022）的研究结果形成对照；他们发现，使用辅助数据确实能提升模型性能。这也与 Khatami 等人（2016）的发现不同，后者发现机器学习模型类型会产生一定影响。值得注意的是，没有研究在评估组间性能差异时，明确对类别不平衡（多数类比例）进行校正。其他特征的影响被多数类患病率的强烈影响所掩盖。这或许可以解释我们的研究与 Khatami 等人（2016）和 Hanadé Houmma 等人（2022）的研究之间的差异。

局限性与未来研究

总体准确率是最常报告的指标，Jafarzadeh 等人（2022）和 Khatami 等人（2016）也发现了这一点。然而，总体准确率存在若干局限性（Foody 2020；Stehman 和 Foody，2019年）。最重要的是，它容易受到类别不平衡的影响，从而妨碍研究间比较。这解释了为什么多数类的比例是一个显著的预测因子。改为报告召回率算术平均值（平衡准确率）或几何平均召回率会有所改进。此外，也应考虑马修斯相关系数（MCC）或科恩卡帕系数等其他指标。若能报告多个经基准归一化的指标（Burger 和 Meertens 2020），效果会更好；也可以报告完整的混淆矩阵，从中推导出所有这些指标。遗憾的是，这尚未成为常见做法。可以使用网状荟萃分析对多个性能指标建模（Harrer 等，2021），这将是未来研究的一个有趣主题。聚焦于总体准确率，可能会让政策制定者对可持续发展目标的进展情况产生过于乐观的印象。未来研究还可以尝试纳入本研究未收集的子研究层面特征，例如特定的机器学习算法或超参数配置。

在随机抽取的 200 项研究中，只有 57 篇论文获得了三位审稿人的一致认可，而每位审稿人都认为其中有 77 至 81 项研究可以纳入。这凸显了筛选过程的主观性（可能是由于摘要含糊不清且涉及跨学科范围），也表明多位审稿人参与的重要性。使用开源工具 ASReview（van de Schoot 等，2021年）以及大语言模型进行特征提取（Mahuli 等，2023），可以提高效率，但无法解决该领域方法报告不一致或不清晰的问题。例如，在一些入选研究中，无法确定报告的性能是基于训练集、验证集、测试集，还是整个已标记数据集。此外，出于充分的理由，研究排除了谷歌学术（见第 3.1 节），但尚未评估这一决定的影响。

本分析共纳入了来自 20 项研究的 86 次试验。经过仔细筛除无关论文后，初始样本 200 项研究的低产出率（10%）由此产生，这也说明编制适当数据集的过程困难且耗费人力。

尽管一些模拟研究表明，三级元分析即使仅纳入 20 至 40 项研究，也能得出准确结果（Hedges 等，2010年；Polanin 和 Pigott，2014年），但本分析处于这一范围的下限。综述更多论文将提高统计功效，并为检验研究特征之间的交互作用、开展剂量-反应元分析（Viechtbauer，2024年），以及对相关抽样方差进行稳健方差估计（Harrer 等，2021）创造可能。

本研究仅考察了已发表的结果，因此存在发表偏倚——这是一种有充分文献记载的现象：取得正面结果的研究更有可能发表，

而负面或中性的发现则未能发表（Borenstein 等，2009；Bozada 等，2021；Hansen 等，2022；Harrer 等，2021）。这种偏倚可能导致总体准确率总体值被高估。我们决定仅纳入报告总体准确率的论文，因此排除了那些因总体准确率容易受类别不平衡影响而负责任地选择不予报告的研究。这也可能引入发表偏倚，不过这种偏倚更可能导致总体准确率总体值被低估。另一个无法最终核实的潜在发表偏倚来源是，研究可能只报告表现最佳模型的性能，而忽略其他受测模型的表现。本文并未量化潜在的发表偏倚。在比例的荟萃分析中使用漏斗图存在争议。我们在附录 B 中提供了一张漏斗图。传统的漏斗图方法并不适用于比例研究的荟萃分析（Hunter 等，2014）。评估潜在发表偏倚最常用的检验是 Egger 检验和 Begg 检验。然而，这些方法用于层级数据时同样存在问题。它们假设每项研究仅贡献一个独立的效应量估计值（Rodgers 和 P ustejovsky，2020；Sutton，2009），因此也会得出误导性的结论。此外，漏斗图不对称可能由发表偏倚以外的原因造成，例如研究间异质性（Fernández-Castilla 等，2020）。

结论

联合国可持续发展目标（可持续发展目标）的监测与实现对于应对当前社会挑战具有重要的现实意义。为此，机器学习是一种广泛采用和使用的工具。应用于遥感数据的机器学习模型总体准确率可视为良好。然而，总体准确率虽广受欢迎，却不是适合研究间比较的质量指标。总体准确率容易受到类别不平衡的影响，而不同研究中的类别不平衡程度各异。因此，若将总体准确率作为唯一的质量指标，可能会造成过于乐观的印象。政策制定者和国际组织必须在现实中的可持续发展目标监测工作中考虑这些发现及其政策影响。

例如，在监测森林砍伐率（可持续发展目标15）或城市扩张（可持续发展目标11）的项目中，若仅依赖总体准确率，可能会掩盖对少数类别的检测效果不佳，例如小片森林损失区域或非正规住区。这可能导致错误的政策决策、资源分配无效，以及错失及时采取干预措施的机会。因此，如果某个可持续发展目标监测项目获得资助，就必须在合同中规定，除总体准确率外，还须报告全面的模型性能指标，例如混淆矩阵和类别特定准确率。采用更规范化的指标，例如 F1 分数、精确率、召回率或科恩卡帕系数，可以更均衡地评估模型可靠性。

这些质量保证措施对于确保可持续发展目标监测工作的透明度、可比性和可信度至关重要。通过采用这些标准，政策制定者能够做出更明智的决策，更有效地制定干预措施，并最终推动可持续发展目标的实现。若没有如此严格的绩效报告要求，就可能会使用效果欠佳的模型，从而损害可持续发展目标监测举措的可信度和影响力。

免责声明

本报告所表达的观点属于作者观点，并不一定与荷兰统计局的政策一致。

参考文献

Agrillo, E., Filipponi, F., Pezzarossa, A., Casella, L., Smiraglia, D., Orasi, A., Attorre, F., & Taramelli, A. (2021). 利用地球观测和生物多样性大数据对森林生境类型进行分类和制图。遥感，*13*(7). https://doi.org/10.3390/rs13071231 Aja, D., Miyittah, M. K., & Angnuureng, D. B. (2022). 在加纳西部地区的湿地复合体中，结合光学影像和雷达影像量化红树林范围。《可持续性》，*14*(24). https://doi.org/10.3390/su142416687 Al-Dousari, A. E., Mishra, A., & Singh, S. (2023). 利用多层感知机神经网络（MLPNN）检测科威特都会区的土地利用/土地覆盖变化并预测城市扩张。《埃及遥感与空间科学杂志》，*26*(2), 381–392. https://doi.org/10.1016/j.ejrs.2023.05.003 Barendregt, J. J., Doi, S. A., Lee, Y. Y., Norman, R. E., & Vos, T. (2013). 患病率的荟萃分析。《流行病学与社区健康杂志》，*67*(11), 974–978. https://doi.org/10.1136/jech-2013-203104 Belgiu, M., & Drăguţ, L. (2016). 遥感中的随机森林：应用综述与未来方向。《摄影测量与遥感国际学会摄影测量与遥感杂志》，*114*，24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011 Bishop, C. M. (2006). 模式识别与机器学习。施普林格。 Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009).荟萃分析导论。威利。 Bozada, T., Jr., Borden, J., Workman, J., Del Cid, M., Malinowski, J., & Luechtefeld, T. (2021). Sysrev：一个用于数据整理和系统性证据综述的 FAIR 平台。《人工智能前沿》，*4*, 685298. https://doi.org/10.3389/frai.2021.685298 Breiman, L. (2001). 随机森林。机器学习，*45*(1), 5–32. https: // doi.org/https://doi.org/10.1023/A:1010933404324 Burger, J., & Meertens, Q. (2020). 算法与黑猩猩之争：论分类器性能指标的极小值。载于 L. Cao, W. Kosters, & J. Lijffijt（编者），*BNAIC/BeneLearn 2020*会议论文集 （第38–55页）。https://bnaic.liacs.leidenuniv.nl/bnaic2020proceedings.pdf Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. (2021). 利用卫星影像理解并促进可持续发展。《科学》，*371*(6535), eabe8628. https://doi.org/10.1126/science.abe8628 Campbell, J. B., & Wynne, R. H. (2011). 遥感导论（第5版

di-

（……）。Guilford 出版社。

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V.,Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). 系统综述中的无荟萃分析综合（SWiM）：报告指南。*BMJ*, *368*,l6890. https://doi.org/10.1136/bmj.l6890 Cheung, M. W. L. (2014). 使用三级荟萃分析对依赖效应量建模：一种结构方程建模方法。心理学方法, *19*(2), 211–229. https://doi.org/10.1037/a0032968 Cutler, D. R., Jr., T. C. E., Beard, K. H., Cutler, A., Hess,K. T., Gibson, J., & Lawler, J. J. (2007). 用于《生态学》分类的随机森林。《生态学》, *88*(11), 2783–2792. https://doi.org/https://doi.org/10.1890/ 07-0539.1 Debray, T. P. A., Damen, J. A. A. G., Snell, K. I. E., Ensor, J., Hooft, L., Reitsma, J. B., Riley, R. D., & Moons, K. G. M. (2017). 预测模型性能系统综述与荟萃分析指南。*BMJ*, *356*, i6460. https://doi.org/10.1136/bmj.i6460 Doi, S. A., & Xu, C. (2021). 用于比例荟萃分析的 Freeman-Tukey 双反正弦变换：近期批评存在严重误导。《循证医学杂志》, *14*(4), 259– 261. https://doi.org/10.1111/jebm.12445 Ekmen, O., & Kocaman, S. (2024). 面向联合国可持续发展目标的遥感：研究与合作的全球分析。《埃及遥感与空间科学杂志》, *27*(2), 329–341. https://doi.org/10. 1016/j.ejrs.2024.04.002 Estoque, R. C. (2020). 可持续性概念综述及利用遥感开展可持续发展目标监测的现状。遥感, *12*(11), 1770. https://doi.org/10.3390/rs12111770 Fagua, J. C., Rodríguez-Buriticá, S., & Jantz, P. (2023). 推进哥伦比亚高分辨率土地覆盖制图：采用适合当地的图例至关重要。遥感, *15*(10). https://doi.org/10. 3390/rs15102522 Fan, Z., Yan, Z., & Wen, S. (2023). 可持续性领域的深度学习与人工智能：可持续发展目标、可再生能源和环境健康综述。可持续性, *15*(18). https://doi.org/10.3390/su151813493 Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas,N., Onghena, P., & Van den Noortgate, W. (2020). 多个结局荟萃分析的可视化表示：对森林图、漏斗图和毛毛虫图的扩展。《方法学》, *16*(4), 299–315. https://doi.org/10. 5964/meth.4013 Finegold, Y., Ortmann, A., Lindquist, E., d'Annunzio, R., &Sandker, M. (2016).地图精度评估与面积估算：实用指南（工作论文）。联合国粮食及农业组织。罗马。http://www.fao.org/3/a-i5601e.pdf Foody, G. M. (2020). 解释卡帕系数为何不适用于专题地图精度的评估和比较……

通过图像分类获得。《环境遥感》，239，111630。https://doi.org/10.1016/j.rse.2019.111630 Freeman, M. F., & Tukey, J. W. (1950年). 与反正切和平方根相关的变换。《数学统计年刊》，21(4), 607–611。https://doi.org/10.1214/aoms/1177729756 Goodfellow, I., Bengio, Y., & Courville, A. (2016年). 深度学习 [http :/ / www.deeplearningbook.org]. 麻省理工学院出版社。Greif, L., Röckel, F., Kimmig, A., & Ovtcharova, J. (2024年). 一篇系统综述，回顾了当前用于可持续发展目标的人工智能技术。《国际环境研究杂志》，19(1)。https://doi.org/10.1007/s41742-024-00668-5 Gusenbauer, M., & Haddaway, N. R. (2020年). 哪些学术检索系统适用于系统综述或荟萃分析？对谷歌学术、PubMed及其他26种资源的检索质量进行评估。《研究综合方法》，11(2), 181–217。https://doi.org/10.1002/ jrsm.1378 Hall, J. A., & Rosenthal, R. (2018). 荟萃分析中如何选择随机效应模型：分析单位与所得结果的可推广性。《社会与人格心理学指南》，12(10), e12414。https://doi.org/10.1111/spc3.12414 Hall, O., Dompae, F., Wahab, I., & Dzanku, F. M. (2023). 机器学习与卫星影像在贫困预测中的应用综述：对发展研究与应用的启示。《国际发展杂志》，35(7), 1753–1768。https://doi.org/10.1002/jid. 3751 Hanadé Houmma, I., El Mansouri, L., Gadal, S., Garba, M., & Hadria, R. (2022). 农业干旱建模：大数据技术最新进展综述。《测绘学、自然灾害与风险》，13(1), 2737–2776。https://doi.org/10.1080/19475705.2022.2131471 Hansen, C., Steinmetz, H., & Block, J. (2022). 如何通过八个步骤开展荟萃分析：实用指南。《管理评论季刊》，72(1), 1–19。https://doi.org/10.1007/s11301-021-00247-4 Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2019).*dmetar*：指南《在*R*中进行荟萃分析》的配套*R*软件包。http : //dmetar.protectlab.org/ Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). 在*R*中进行荟萃分析：实践指南。Chapman & Hall/CRC。Hastie, T., Tibshirani, R., & Friedman, J. (2009). 统计学习的要素。施普林格。Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). 荟萃回归中依赖效应量估计的稳健方差估计。《研究综合方法》，1(1), 39–65。https://doi.org/10.1002/jrsm.5 Higgins, J. P. T., & Thompson, S. G. (2002). 量化荟萃分析中的异质性。《医学统计学》，21(11), 1539–1558。https://doi. org/10.1002/sim.1186

Holloway, J., & Mengersen, K. (2018). 统计机器学习方法与遥感在可持续发展目标中的应用：综述。遥感，10(9), 1365。https://doi.org/10.3390/rs10091365 Huang, C., Davis, L. S., & Townshend, J. R. G. (2002年). 对支持向量机用于土地覆盖分类的评估。国际遥感杂志，725–749。https://doi.org/https://doi.org/10.1080/01431160110040323 Hunter, J. P., Saratzis, A., Sutton, A. J., Boucher, R. H., Sayers, R. D., & Bown, M. J. (2014年). 在比例研究的荟萃分析中，漏斗图被发现并非评估发表偏倚的准确方法。《临床流行病学杂志》，67(8), 897–903。https://doi. org/10.1016/j.jclinepi.2014.03.003 Iliescu, D., Rusu, A., Greiff, S., Fokkema, M., & Scherer, R. (2022). 为什么测试与评估文献需要系统综述和荟萃分析。《欧洲心理评估杂志》，38(2), 73–77。https://doi.org/10.1027/1015-5759/a000705 Jafarzadeh, H., Mahdianpari, M., Gill, E. W., Brisco, B., & Mohammadi-manesh, F. (2022). 支持湿地监测的遥感与机器学习工具：三十年研究的荟萃分析。遥感，14 (23), 6104。https : / / doi . org / 10 . 3390 / rs14236104 James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021).统计学习导论。施普林格。Jia, M., Wang, Z., Mao, D., Ren, C., Song, K., Zhao, C., Wang, C., Xiao, X., & Wang, Y. (2023). 绘制全球红树林分布图，分辨率为10米。《科学通报》，68(12), 1306–1316。https://doi. org/10.1016/j.scib.2023.05.004 Jochem, W. C., Bird, T. J., & Tatem, A. J. (2018). 通过机器学习方法根据居住点识别居住社区类型。《计算机、环境与城市系统》，69, 104–113。https://doi.org/10.1016/j.compenvurbsys.2018.01.004 Kavvada, A., Metternicht, G., Kerblat, F., Mudau, N., Haldorson, M., Lalda-parsad, S., Friedl, L., Held, A., & Chuvieco, E. (2020年). 利用地球观测推动实现可持续发展目标。《环境遥感》，247, 111930。https: // doi. org/10.1016/j.rse.2020.111930 Khatami, R., Mountrakis, G., & Stehman, S. V. (2016年). 基于监督像元的土地覆盖图像分类过程的遥感研究荟萃分析：面向从业者的一般指南与未来研究。《环境遥感》，177, 89–100。https://doi.org/10.1016/j.rse.2016.02.028 Konstantopoulos, S. (2011年). 三级元分析中的固定效应与方差分量估计。《研究综合方法》，2(1), 61–76。https://doi.org/https://doi.org/10.1002/jrsm.35 Lajeunesse, M. J. (2016年). 使用R的metagear软件包，促进系统综述、数据提取和荟萃分析。《生态学方法》

与演化, *7*(3), 323–330. https://doi.org/10.1111/2041-210X. 12472 Lavallin, A., & Downs, J. A. (2021). 地理学中的机器学习：过去、现在与未来。*Geography Compass*, *15*(5), e12563. https://doi. org/10.1111/gec3.12563 Li, H., Zech, J., Hong, D., Ghamisi, P., Schultz, M., & Zipf, A. (2022). 利用OpenStreetMap和多模态遥感数据，通过联合深度学习检测污水处理厂。国际应用地球观测与地理信息杂志,*110*, 102804. https://doi.org/10.1016/j.jag.2022.102804 Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). 遥感应用中的深度学习：荟萃分析与综述。*ISPRS*摄影测量与遥感学报,*152*, 166– 177. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2019.04.015 Mahuli, S. A., Rai, A., Mahuli, A. V., & Kumar, A. (2023). 在系统综述和荟萃分析中应用ChatGPT。*BritishDental Journal*, *235*(2), 90–92. https://doi.org/10.1038/s41415-023- 6132-y Mashaba-Munghemezulu, Z., Chirima, G. J., & Munghemezulu, C. (2021). 利用多时相Sentinel-1数据绘制小农玉米农场分布图，以支持可持续发展目标。遥感, *13*(9). https://doi.org/10.3390/rs13091666 Maso, J., Zabala, A., & Serral, I. (2023). 面向可持续发展目标的地球观测。遥感, *15*(10), 2570. https://doi.org/ 10.3390/rs15102570 Melgani, F., & Bruzzone, L. (2004). 使用支持向量机对高光谱遥感图像进行分类。*IEEE*地球科学与遥感汇刊, *42*(8), 1778–1790. https://doi.org/ 10.1109/TGRS.2004.831865 Miller, J. J. (1978). Freeman–Tukey双反正弦变换的逆变换。*The American Statistician*, *32*(4), 138. https://doi.org/10. 1080/00031305.1978.10479283 Molnar, C. (2025). 可解释机器学习：让黑箱模型变得可解释的指南(第3版)。https://christophm.github.io/ interpretable-ml-book Mountrakis, G., I m, J., & Ogole, C. (2011). 支持向量机在遥感中的应用：综述。*ISPRS*摄影测量与遥感学报, *66*(3), 247–259. https://doi.org/10.1016/j.isprsjprs. 2010.11.001 美国国家航空航天局。(2024). 遥感 [访问日期：2024-11-19]。Nations, U. (2024).《*2024*年可持续发展目标报告》(报告)。联合国。纽约。https://unstats.un.org/sdgs/report/2024/ The-Sustainable-Development-Goals-Report-2024.pdf Nazir, U., Mian, U. K., Sohail, M. U., Taj, M., & Uppal, M. (2020). Kiln-net：一种用于检测南亚砖窑的门控神经网络。*IEEE*应用地球观测与

遥感,*13*, 3251–3262. https://doi.org/10.1109/JSTARS. 2020.3001980 Owers, C. J., Lucas, R. M., Clewley, D., Tissott, B., Chua, S. M. T., Hunt, G., Mueller, N., Planque, C., Punalekar, S. M., Bunting, P., Tan, P., & Metternicht, G. (2022). 利用开放数据立方体对澳大利亚进行业务化大陆尺度土地覆盖制图。国际数字地球杂志, *15*(1), 1715–1737. https: / / doi . org/ 10 . 1080 / 17538947.2022.2130461 Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). PRISMA 2020声明：系统综述报告指南的更新。*BMJ*,*372*, n71. https://doi.org/10.1136/bmj.n71 Pareeth, S., Karimi, P., Shafiei, M., & De Fraiture, C. (2019). 利用随机森林分层方法从陆地卫星8号时间序列绘制农业土地利用模式。遥感, *11*(5). https://doi. org/10.3390/rs11050601 Peng, Y., Sengupta, D., Duan, Y., Chen, C., & Tian, B. (2022). 基于Sentinel-2时间序列图像的生物物理参数，精确绘制中国沿海水产养殖池塘分布图。《海洋污染通报》, *181*, 113901. https://doi.org/10.1016/j.marpolbul.2022. 113901 Polanin, J. R., & Pigott, T. D. (2014). 元分析统计显著性检验的应用。《研究综合方法》, *6*(1), 63–73. https://doi.org/ 10.1002/jrsm.1124 Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex：一个完全开放的学术作品、作者、出版场所、机构和概念索引。载于 N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (编者),第*26*届科学、技术与创新指标国际会议。https://doi.org/10.48550/arXiv.2205.01833 Priyanka, Rajat, Avtar, R., Malik, R., Musthafa, M., Rathore, V. S., Kumar, P., & Singh, G. (2023). 基于全极化时间平均 SAR 散射功率的森林种植物种分类。遥感应用：社会与环境, *29*, 100924. https://doi.org/10. 1016/j.rsase.2023.100924 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). 面向数据驱动地球系统科学的深度学习与过程理解。《自然》, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1 Rodgers, M. A., & Pustejovsky, J. E. (2020). 在存在依赖效应量的情况下，评估用于检测选择性报告的元分析方法。《心理学方法》。https://doi.org/10.1037/met0000300

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., &Rigol-Sanchez, J. P. (2012). 随机森林分类器在土地覆盖分类中的有效性评估。*ISPRS*摄影测量与遥感学报，*67*，93–104。https://doi.org/10.1016/j.isprsjprs.2011.11.002 Shaharum, N. S. N., Shafri, H. Z. M., Ghani, W. A. W. A.K., Samsatli, S., Al-Habshi, M. M. A., & Yusuf, B. (2020). 利用 Google Earth Engine 和机器学习算法绘制马来西亚半岛油棕分布图。《遥感应用：社会与环境》，*17*，100287。https://doi.org/10.1016/j.rsase.2020.100287 Shen, S., Chen, J., Cheng, D., Liu, H., &Zhang, T. (2024). 通过深度交换数字正射影像图和数字表面模型特征进行崩岗分割。《国际水土保持研究》，*12*(3)，589–599。https://doi.org/10.1016/j.iswcr.2023.11.004 SIGD Task Team. (2017). 用于官方统计的地球观测 （报告）。联合国。纽约。https://unstats.un.org/bigdata/task- teams/earth-observation/UNGWG_Satellite_Task_Team_Report_WhiteCover.pdf Solórzano, J. V., Mas, J. F., Gallardo-Cruz, J. A., Gao, Y., &Fernández- Montes de Oca, A. (2023). 利用结合合成孔径雷达和多光谱图像的时空深度学习方法检测森林砍伐。*ISPRS*摄影测量与遥感学报，*199*，87–101。https://doi.org/10.1016/j.isprsjprs.2023.03.017 Stehman, S. V., & Foody, G. M. (2019). 土地覆盖产品严格准确率评估中的关键问题。《环境遥感》，*231*，111199。https://doi.org/10.1016/j.rse.2019.05.018 Strobl, C., Boulesteix, A.-L.,Zeileis, A., & Hothorn, T. (2007). 随机森林变量重要性度量中的偏差：示例、来源及解决方案。《*BMC*生物信息学》，*8*(1)，1–21。https://doi.org/10.1186/ 1471-2105-8-25 Suryono, H., Kuswanto, H., & Iriawan, N. (2022). 利用 Google Earth Engine，根据数据不平衡情况采用随机森林算法进行水稻物候分类。《计算机科学程序汇刊》，*197*，668–676。https: //doi.org/10.1016/j.procs.2021.12.201 Sutton, A. J. (2009). 出版偏倚。载于 H. Cooper, L. V. Hedges, & J. C. Valentine（编），研究综合与荟萃分析手册 （第1版，第435–445页）。Sage出版社。Tao, Y., & You, F. (2019). 利用卫星衍生数据，通过机器学习模型预测覆盖作物采用情况。《*IFAC*在线论文集》，*52*(30)，137–142。https://doi.org/10.1016/j.ifacol.2019.12.511 Tawfik, G. M., Dila, K. A. S., Mohamed, M. Y. F., Tam, D.N. H.,Kien, N. D., Ahmed, A. M., & Huy, N. T. (2019). 使用模拟数据开展系统综述和荟萃分析的分步指南。《热带医学与健康》，*47*(1)，46。https://doi.org/10.1186/s41182-019-0165-6

Thapa, A., Horanont, T., Neupane, B., & Aryal, J. (2023). 用于遥感图像场景分类的深度学习：综述与荟萃分析。*Remote Sensing*, *15*(19), 4804。https://doi.org/10.3390/rs15194804 UCS. (2023). *UCS*卫星数据库 （报告）。忧思科学家联盟。马萨诸塞州剑桥。https://www.ucsusa.org/resources/satellite-database UNDESA. (2023). 《*2023*年可持续发展目标报告：特别版》 （报告）。联合国。纽约。https : / / doi . org/ 10 . 18356/9789210024914 UN-GGIM. (2019). 可持续发展目标指标的地域维度：地理空间数据分析及其与统计数据的整合 （报告）。葡萄牙国家统计局。里斯本。https://un-ggim-europe.org/wp-content/uploads/2019/05/UN_GGIM_08_05_2019-The-territorial-dimension-in-SDG-indicators-Final.pdf van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Wei- jdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tum- mers, L., & Oberski, D. L. (2021). 一种用于高效、透明地开展系统综述的开源机器学习框架。*NatureMachine Intelligence*, *3*, 125–133。https://doi.org/10.1038/s42256-020- 00287-7 Vapnik, V. (1995). 统计学习理论的本质。施普林格。Venables, W. N., & Ripley, B. D. (2002). 现代应用统计学（使用*S*）（第四版）。施普林格。Viechtbauer, W. (2010). 使用metafor软件包在R中开展荟萃分析。*Journal of Statistical Software*, *36*(3), 1–48。https : / / doi . org/10.18637/jss.v036.i03 Viechtbauer, W. (2024). 软件包*"metafor"*：*R*的荟萃分析软件包。https://cran.r-project.org/web/packages/metafor/metafor.pdf Wang, M., Mao, D., Wang, Y., Xiao, X., Xiang, H., Feng, K., Luo, L., Jia, M., Song, K., & Wang, Z. (2023). 基于Sentinel-1/2影像，利用两阶段面向对象随机森林和层次决策树算法绘制东亚湿地分布图。《环境遥感》， *297*, 113793。https://doi.org/10.1016/j.rse.2023.113793 Xia, Z., Li, Y., Chen, R., Sengupta, D., Guo, X., Xiong, B., & Niu, Y. (2022). 利用遥感绘制中国西北部光伏电站的快速发展图景。*Energy Reports*, *8*, 4117–4127。https://doi.org/10.1016/j.egyr.2022.03.039 Zhang, Y., Liu, J., & Shen, W. (2022). 遥感应用中集成学习算法综述。*Applied Sciences*, *12*(17), 8654。https://doi.org/10.3390/app12178654 Zhao, Q., Yu, L., Du, Z., Peng, D., Hao, P., Zhang, Y., & Gong, P. (2022). 地球观测卫星数据应用概述：影响与未来趋势。*Remote Sensing*, *14*(8), 1863。https://doi.org/10.3390/rs14081863 Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., & Xu, F. (2017). 遥感中的深度学习：全面综述及研究清单

来源。*IEEE*地球科学与遥感杂志，*5*(4)，8–36。
https://doi.org/10.1109/MGRS.2017.2762307

附录A 置换特征重要性

置换特征重要性根据 Molnar（2025，第23章）计算。特征重要性是指置换（即随机打乱）某个特征后，均方根误差的比例增幅。打破特征与目标变量之间的关系对预测误差的影响越大，该特征就越重要。

为获得样本外预测误差，首先将数据随机划分为 $K = 5$ 折。使用全部 $d = 14$ 个特征（表4）的模型（公式12）在其中 $K - 1 = 4$ 折（训练集）上训练，并用于预测剩余一折（测试集）的目标变量。将预测的个体效应量与观测值进行比较，$\tilde{\theta}_{ij}$ 即可得到原始均方根误差。随后，在测试集的副本中对一个特征进行置换，并使用同一训练好的模型预测目标变量，从而得到置换后的均方根误差。置换特征重要性（PFI）定义为置换后的均方根误差与原始均方根误差之比。对每一折，都会重新训练模型，并对每个特征进行 $B = 200$ 次置换（在 $(m/K)! = (86/5)! \approx 10^{14}$ 种可能的置换中）。

图9显示了各特征在 $KB = 1000$ 次重复中的 PFI 分布。特征按平均 PFI 排序。根据此方法，最重要的特征是多数类的最大流行率（maxprev）。这是唯一一个 PFI 的第 2.5 百分位数超过 1 的特征。这证实了使用惩罚似然进行特征选择的结果（图6）。对 maxprev 进行置换，RMSE 平均增加 1.47 倍。其他特征的重要性则与图6的结果不一致。请注意，图6显示的是向零模型添加单个特征对惩罚似然的影响，而图9显示的是对完整模型中的某个特征进行置换对预测误差的影响。PFI 方法的一个缺点是，随机打乱特征可能会产生不切实际的特征组合，进而导致不切实际的预测。

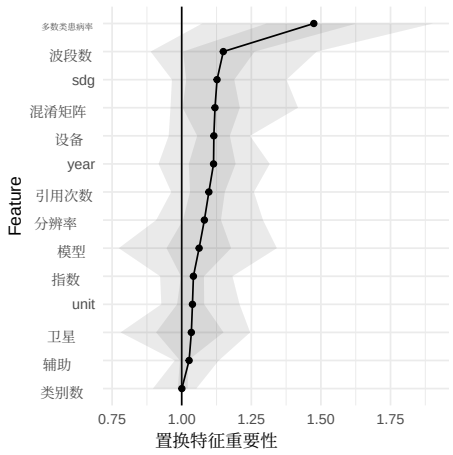

图 9：置换特征重要性。5 折 $\times$ 100 次重复样本的均值（点）、第 25 和第 75 百分位数（深灰色），以及第 2.5 和第 97.5 百分位数（浅灰色）。垂直线表示 PFI $= 1$，即无影响。

附录B 漏斗图

如讨论部分所述，在比例的荟萃分析中使用漏斗图是一个有争议的话题。不过，我们绘制了一张漏斗图（见图10）。可以推测存在轻微的正向发表偏倚。根据 Fernández-Castilla 等（2020），无法从该图得出明确结论。我们建议未来研究重点关注在比例的荟萃分析和层级数据背景下开发发表偏倚方法学。

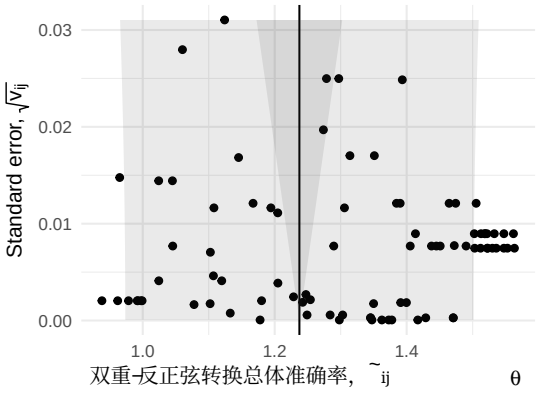

图10：漏斗图。每个点表示研究 j 内试验 i 的估计值，垂直线表示总体效应量（μ），阴影区域表示 95% 置信区间 $\mu \pm t_{0.975,\ h-1}\sqrt{Var\ (\tilde{\theta}_{ij})}$，其中仅包含抽样方差（$Var\ \tilde{\theta}_{ij}\ = v_{ij}$ $Var\ \tilde{\theta}_{ij}\ = \sigma_{\xi}^2 + \sigma_{\zeta}^2 + (\ )$，深灰$v_{ij}$色），以及包含所有方差分量（( )，浅灰色）。