arXiv:2303.02088v1 [stat.ME] 3 Mar 2023

用于整合标准化探测-未探测数据与机会性仅出现数据的新型空间模型：应用于估计与电力线致鸟类死亡相关的风险因素

JorgeSicacha-Parada *[1]，Diego Pavon-Jordan †[2]，Ingelin Steinsland ‡[1]，Roel May §[2]，以及 B˚ardStokke ¶[2]

[1]数学科学系。挪威科技大学（NTNU）。特隆赫姆。挪威

[2]陆地生态学系。挪威自然研究所（NINA）。Torgarden 邮政信箱 5685。N-7485。特隆赫姆。挪威

摘要

能源消费持续增长，使得扩展能源输配网络成为必要。电力线的铺设会对鸟类种群构成风险。因此，要减少输电线路的影响，深入了解电力线导致鸟类死亡的情况及其背后的因素至关重要。针对这一问题，我们可以利用专业调查和公民科学数据。前一种数据类型由专家通过昂贵的标准化抽样方案，在空间的较小区域内采集；后一种数据则由公民科学家在广阔区域内机会性地收集。本文建立了完整的贝叶斯空间模型，以：1）融合专业调查与公民科学数据；2）明确考虑影响专业

*jorge.sicacha@ntnu.no
†diego.pavon-jordan@nina.no
‡ingelin.steinsland@ntnu.no§roel.may@nina.no

¶bard.stokke@nina.no

调查数据的偏好抽样，以及影响公民科学数据质量的因素。所提出的模型属于潜在高斯模型这一类，因为两类数据均被视为稀疏化空间点模式，并以对数高斯 Cox 过程进行建模。这些模型的设定假定，存在一个共同的潜在空间过程，作为两类数据观测的基础。

所提出的模型既应用于模拟数据，也应用于挪威特伦德拉格郡电力线导致鸟类死亡的真实数据。模拟研究清楚地表明，融合两类数据，并恰当考虑影响数据收集过程的因素后，参数估计的准确性会提高。对输电线路致死情况的研究表明，电力线网络的密度与输电线路对鸟类种群构成的风险之间存在明显关联。模型的选择会影响从这一案例研究中得出的结论，因为不同模型对输电线路致死风险与暴露鸟类数量之间关联的估计各不相同。

1 引言

预计到2050年，能源消费将增长约50%（Conti 等人，2016），因此全球能源输送网络也必须扩展，尤其是为了实现联合国可持续发展目标（SDG）7：“普及可负担、可靠和现代化的能源服务”（联合国，2018）。除这意味着高额支出外（Bernardino 等人，2018），此类基础设施发展和土地节约也将带来巨大的环境代价（Biasotto 和 Kindel，2018）。不断累积的证据表明，输电线路对许多鸟类物种构成重要威胁（Martin，2011）：架空导线会割裂鸟类利用的空域，增加碰撞死亡风险（Bernardino 等人，2018）；许多物种会利用杆塔作为栖息结构，从而导致触电死亡率上升（Hern´ández-Lambra˜no 等人，2018）。Davis（2002）估计，电力线造成的鸟类死亡轻易就可达到每年十亿只。此后，电力线网络每年增长5%（Jenkins 等人，2010），尽管当地采取了有效的风险缓解措施（Pav´on-Jord´an 等人，2020；Barrientos 等人，2012），死亡数量很可能仍有所增加。

从保护的角度出发，并与可持续发展目标9“建设具有韧性和可持续性的基础设施”、可持续发展目标11“使城市和人类住区实现可持续发展”、可持续发展目标12“确保可持续的消费和生产模式”以及可持续发展目标15“保护、恢复陆地生态系统，促进其可持续利用，并遏制生物多样性丧失”（联合国，2018）保持一致，深入了解输电线路在何种情况下（即如何、何地以及何时）会带来较高的死亡风险至关重要。这有助于采取缓解措施，并规划新的输电线路，从而降低其生态影响。

迄今为止，关于输电线路死亡影响的研究均基于标准化的探测—未探测数据（参见 Miller 等人（2019）），通常在特定地点和相对较小的空间尺度上开展（例如 Bevanger（1995）；Bevanger 和 Brøseth（2001）），和/或聚焦于受到关注的目标物种（例如 JANSS 和 FERRER（2001）；L´opez-L´opez 等人（2011））。这限制了我们对这一公认的人兽冲突所涉及因素得出更广泛结论的能力。扩大此类数据集中所涵盖的物种、栖息地及地理范围的一种方法

是利用公民科学（CS）门户网站所包含的大量信息（例如 https://ebird.org、www.artsobservasjoner.no）。其中一些公民科学平台允许公民科学参与者在观测记录中补充信息，包括发现的是否为死亡动物，甚至可能的死亡原因（例如触电、与电力线碰撞、与围栏碰撞、道路致死）。然而，这两种信息来源（标准化的探测—未探测数据和机会性仅出现数据）无法直接比较，因为它们具有不同的固有偏差，尤其是在调查力度方面（Botella 等人，2021）。

鉴于整合多种数据类型可能带来的潜在益处，近年来许多研究尝试克服整合多个数据源所面临的挑战（Koshkina 等人，2017；Pacifici 等人，2017；Miller 等人，2019；Gelfand and Shirota，2019；Zipkin 等人，2021；Wang等人，2021）。整合多种数据类型能够提升推断和预测性能的证据正在不断积累（Simmonds 等人，2020；Gelfand and Shirota，2019）。这些益处包括：与分别对各数据类型进行建模的结果相比，可降低预测变量的不确定性，并提高参数估计的准确性。然而，要使融合模型如预期般发挥作用，对数据收集过程中潜在偏差的考量至关重要（Simmonds 等人，2020）。

本文的研究动机源于一个案例研究，其目标是确定哪些因素与鸟类电力线致死的高风险相关，进而突出挪威特伦德拉格输电线路网络中风险较高的区域。为解决这一问题，可获得两类数据：由挪威自然研究所（NINA）开展的专业调查所收集的数据，以及来自两个来源的机会性记录：i）Artsobservasjoner，即挪威生物多样性信息中心（NIBC）的一个数据库；ii）挪威鸟类环志中心。该领域的大量研究表明，可见度、土地利用、输电线路密度以及可能与输电线路发生碰撞的鸟类数量等因素，均与输电线路致死风险相关（Bevanger 等，2014；Drewitt and Langston，2008；Martin and Shaw，2010）。

在生态学研究的专业调查中收集的检出—未检出数据，通常会使

用占域模型（MacKenzie和Kendall，2002）及地统计空间模型（Banerjee等人，2015）进行分析，因为数据位置在抽样设计阶段已固定并确定。在我们的案例研究中，NINA收集的数据包含输电线路致死事件发生位置的精确坐标。由于选定一条电力线进行抽样后，会对其周边区域开展普查，因此我们将这些数据视为稀疏点模式（Illian等人，2008），其稀疏化概率取决于点所属的区域。如果从原始点模式中移除点是由取决于生态过程的非随机缺失（MNAR）机制（Little和Rubin，2019）所驱动，那么这种稀疏化就是由偏好抽样设计（Diggle等人，2010）造成的。否则，假定点是随机移除的。

公民科学数据也被视为一种稀疏点模式，但由于存在非随机缺失机制（Little 和 Rubin，2019），该机制取决于采样力度、可探测性、报告力度和/或错误分类等因素。**Peprah、Sicacha-Parada，2022**提出了一个用于生成公民科学数据的通用灵活框架。该框架将点模式的稀疏操作（Illian 等人，2008）与公民科学数据中的偏差联系起来，为公民科学数据建模提供了新视角。该框架基于共享过程模型这一理念，用于对通过非随机缺失机制生成的数据进行建模（Little 和 Rubin，2019）。因此，该模型假设一个共同的潜在效应驱动多个观测数据。在本研究中，我们假设这一共同的潜在效应同时影响观测数据和缺失过程。对于本案例研究，公民科学数据来自两个来源，二者都存在一些共同偏差，例如不均衡的采样力度，这可能受到可达性和/或土地利用等因素的影响（Monsarrat 等人，2019；Sicacha-Parada 等人，2021）；可探测性差异，这可以由土地利用、栖息地类型、死亡鸟类的大小和/或观测时间来解释（Dom´ınguez del Valle 等人，2020）；以及不均衡的报告力度（August，2020）。

因此，本文旨在介绍一种建模框架，根据 Simmonds 等人（2020）的建议，在考虑每种数据类型收集过程中存在的偏差的同时，整合专业调查和公民科学数据。作为所提出建模框架的自然结果，我们还希望突出显示输电线路致死风险较高的区域。该框架扩展了

数据整合模型的现有技术水平，因为它同时对公民科学数据及其偏差、专业调查数据及其抽样过程进行建模，而后者可能具有偏向性（Diggle 等人，2010）。该建模框架被规定为一组依赖共享空间随机效应的贝叶斯模型，属于潜在高斯模型（LGMs；Rue 等人，2009）这一类别。由于这些模型属于潜在高斯模型（LGMs）家族，因此适合使用积分嵌套拉普拉斯近似（INLA）（Rue 等人，2009）和随机偏微分方程（SPDE）（Lindgren 等人，2011）进行拟合；这两种方法能够高效近似模型参数的后验分布，以及模型设定中涉及的高斯随机场（GRF）（Simpson 等人，2016）。这是一个灵活的框架，可用于整合两种现有数据类型，并考虑专业调查和公民科学数据中的多种偏差来源。因此，我们提出的模型不仅可以直接应用于评估输电线路对动物死亡的影响（我们的案例研究），还可用于评估道路和风电场等其他人类基础设施的影响。在当前可再生能源开发相关新项目迅速增加的时期，这类评估至关重要，因为这些项目造成的死亡正在叠加于传统的道路致死（Barrientos 等人，2021）和输电线路致死（Bernardino 等人，2018）之上，同时还面临强烈的公众反对（Serrano 等人，2020）。

我们展示了所提出的模型在模拟数据和案例研究中的表现。通过模拟研究，我们展示了整合两类数据的意义、对参数估计的潜在益处以及面临的挑战。在案例研究中，尽管我们不知道真实值，但仍希望找出输电线路致死风险较高的区域、决定这一风险的因素，并比较所提出的各个模型。

本文安排如下。第2节介绍案例研究中可用的两类数据，以及与建模框架设定相关的数据收集过程详情。第 3 节介绍该框架，并阐述用于整合两类数据和校正其偏差的模型的技术细节。第 4 节介绍用于评估模型特性，以及检验校正偏差和整合两类数据之必要性的模拟研究。第 5 节包含电力线导致的鸟类死亡分析及其结果。最后，在第 6 节中，我们

讨论所提出的框架，并提出其未来的扩展方向。

2 数据与案例研究：挪威特伦德拉格输电线路导致的鸟类死亡

本文使用两个不同的数据集，研究因输电线路致死的鸟类。首先，我们使用挪威自然研究所（NINA）开展的专业调查所得的标准化数据，调查旨在找出输电线路特定区段下方的所有鸟类尸体（Bevanger 等，2014）。其次，我们使用两类机会性、仅记录出现情况的公民科学数据：1）自2016年以来，在挪威公民科学门户网站 www.artsobservasjoner.no 上发现并报告为“输电线旁的死亡鸟类”的记录；2）报告给挪威鸟类环志中心的死亡鸟类记录。图1展示了两类数据的出现记录。由于在距输电线路约100米范围内都可能探测到鸟类，且出于计算便利，案例研究的空间域采用输电线路网络的缓冲区版本（两侧各100米）。本研究使用挪威中部特伦德拉格的数据。该郡环境条件的多样性，以及存在鸟类丰度较高的区域（Sicacha-Parada 等，2022），使其适合作为本案例研究区域。

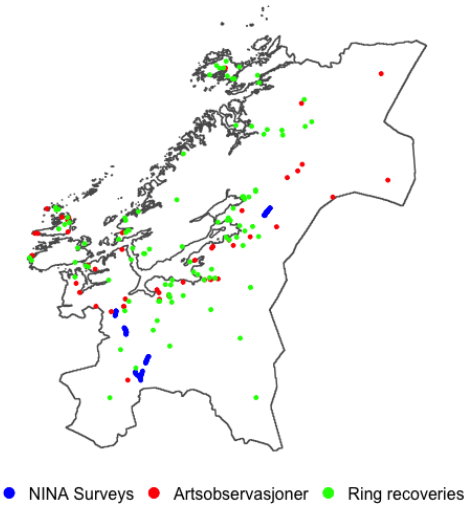

图 1：用于研究输电线路导致鸟类死亡风险及其空间分布的数据来源。蓝色：*NINA* 收集的观测记录的空间分布。红色：公民科学家在 *Artsobservasjoner* 报告的机会性记录。绿色：通过挪威鸟类环志中心报告的机会性记录。

2.1 鸟类死亡的专业调查

电力线导致鸟类死亡的专业调查数据采集自 OPTIPOL 项目（Bevanger 等，2014）。2011—2014年，研究人员使用一只训练犬（瓦赫特尔犬），在一段长 7.1 公里的高压输电线路（300 kV 输电线路）下方搜寻尸体。每年，这只狗都会按照同一调查规程巡查同一段线路，在皆伐区内沿输电线路下方来回穿行（详情及研究设计参见 Bevanger 等人（2014））。为避免重复计数，找到的尸体均被移走。本研究使用了 147 条点位级尸体观测记录，因为每条记录都有精确的地理位置。

2.2 鸟类死亡的机会性记录

我们从两个来源获取了机会性（仅记录出现）记录。第一个来源是挪威生物多样性信息中心（NBIC）的数据库 Artsobservasjoner（www.artsobservasjoner.no），任何人——无论志愿者还是专业人员——都可以报告任何物种的出现记录，并附上地点、日期/时间，以及观察者认为与该观测记录相关的重要补充信息。例如，观察者可以报告观测到的个体是否已死亡，以及在死因明确时报告死因（如鸟类触电死亡、因碰撞导致翅膀折断）。第二个机会性记录来源是挪威鸟类环志中心，由斯塔万格自然历史博物馆运营（https://www.museumstavanger.no/en/forskning/den-norske-ringmerkingssentralen-1）。其数据库登记了 800 多万条记录，包括环志数据（即给鸟类佩戴带有唯一标识符的金属环）和环志回收数据。回收数据（即发现佩戴金属环的鸟死亡并向国家环志办公室报告）使我们能够收集地点和死因等信息（前提是观察者报告了死因，但并非总是如此）。由于此类数据的每条报告都包含精确的地理位置信息，因此这些数据按点位级数据处理。

总体而言，在我们针对特伦德拉格郡的案例研究中，我们使用了98条环志回收记录中的观测记录（发现带有金属环的死亡鸟类并将其报告给环志中心）以及46条观测记录

或的公民科学门户网站 Arts Observasjoner 中被电力线致死的鸟类（$n = 144$）.

2.3 解释变量

拟合下文提出的模型需要两组解释变量。第一组候选变量旨在解释输电线路致死背后的生态过程，第二组则用于解释专业调查和机会性记录的抽样过程。

在用于解释生态过程的候选变量组中，我们纳入海拔（数字高程模型；DEM）、平均气温和降水量（挪威气象研究所）、根据挪威常见鸟类监测计划估算的鸟类丰度（https://tov-e.nina.no；另见 Sicacha-Parada 等人（2022））、电力线密度（挪威水资源与能源局；NVE）、云量（https://www.earthenv.org/cloud）和土地覆盖（AR50, https://www.nibio.no/tema/jord/arealressurser/ar50）。所有协变量信息均栅格化至 1 x 1 千米的空间尺度。第二组变量用于解释专业调查和公民科学数据的抽样过程。用于解释抽样过程的候选协变量包括到最近（三级）道路的距离（道路网络数据来自 OpenStreetMap https://www.openstreetmap.org），以及到最近水体的距离（由于鸟类丰度较高，公民科学家常去水体附近观鸟）。为解释产生观测到的专业调查数据的抽样过程，我们使用海拔梯度。

2.4 探索性分析

尽管 NINA 的专业调查采用了标准化的采样力度，但这些数据仍存在一些局限。首先，由于开展此类调查既耗时又昂贵，其空间覆盖范围较小。其次，调查地点（输电线路）的选择并非完全随机（Bevanger 等，2014），因为需要依据专家知识来确定应调查哪些输电线路。现在，我们来考察公民科学项目中的采样地点，并判断这些数据是否存在偏好抽样的迹象。为此，我们构建了两个数据集：一个是在特伦德拉格电力线网络100米缓冲区内，以100米×100米网格布设的点；另一个是在 NINA 已调查的输电线路上，以100米×100米网格布设的点

。这些比较聚焦于两个协变量：电力线密度和云量。结果见图2。请注意，这两个协变量均根据其在特伦德拉格的取值进行了标准化。

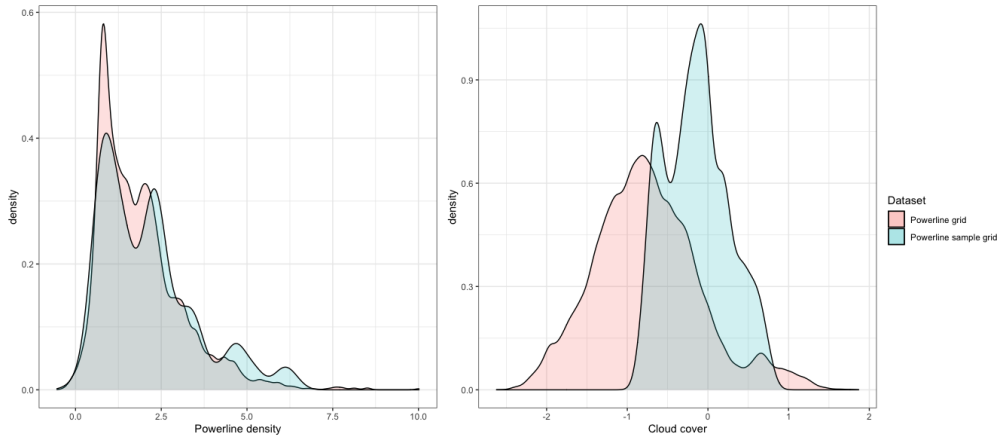

图2：前文定义的两个数据集的电力线密度和云量密度图。红色表示针对覆盖整个电力线网络的网格绘制的密度图，蓝色表示针对为抽样输电线路定义的点网格绘制的密度图。

尽管两种电力线密度分布看起来相似，但我们注意到，在专家选定的采样地点中，较密集的输电线路更为常见。另一方面，图2右图显示，专家选定的输电线路所在位置的云量高于电力线网络沿线的通常水平。众所周知，电力线密度和可见度（以云量作为代理指标）都是与电力线导致的死亡率相关的因素（Drewitt and Langston，2008）。因此，NINA 专家实施的抽样设计显然存在偏好抽样的迹象。

Artsobservasjoner 数据和环志回收记录都可视为公民科学数据。二者都存在一些偏差，例如采样力度不均、可探测性差异以及报告力度不均。后者可能是一个重要的偏差来源，因为对于罕见的出现记录（例如死亡鸟类），人们往往不如报告活体个体那样方便地进行报告。现在，我们将探究现有的公民科学数据，并判断是否有迹象表明，抽样设计受到可达性或土地利用等因素的影响。在这种情况下，我们再次定义两个数据集。第一个是覆盖特伦德拉格、由点构成的 100 米×100 米网格；第二个是发生输电线路致鸟死亡的位置。结果见图 3。

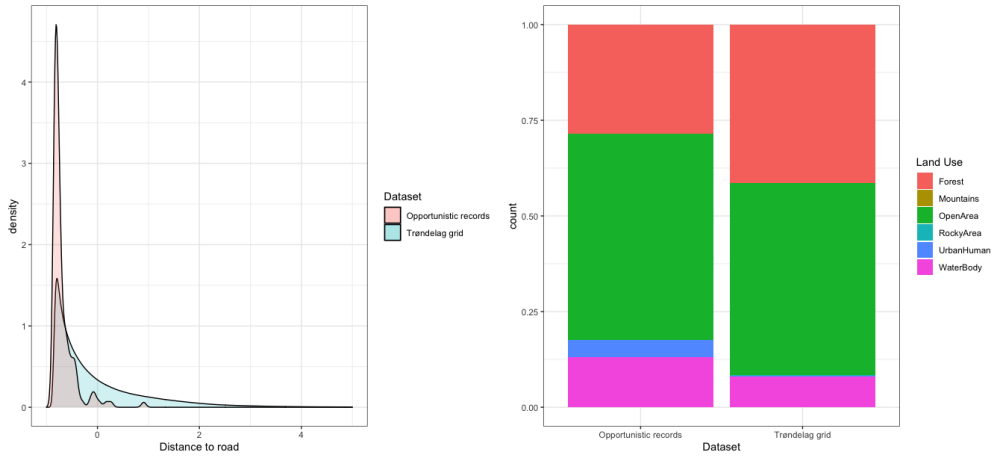

图 3：两个已定义数据集到（三级）道路的距离密度图，以及土地利用分区。左图：蓝色表示覆盖特伦德拉格全郡网格的密度图；红色表示机会性记录所在位置的密度图。右图：左侧柱形表示机会性记录所在位置的土地利用分区；右侧柱形表示特伦德拉格全郡的土地利用分区

从图 3 左图可以看出，公民科学家报告死亡鸟类出现记录的位置，明显比研究区域内所有其他位置更靠近三级道路。关于土地利用，从图 3 右图可以看出，开阔地中死亡鸟类报告所占的比例，高于特伦德拉格开阔地所占的比例。另一方面，在森林区域内报告的比例，远低于特伦德拉格实际被森林覆盖的面积比例。已知可达性与输电线路致鸟死亡的风险无关，因此，公民科学数据的抽样过程中似乎存在由可达性造成的偏差，正如 Fithian 等人（2015）；Monsarrat 等人（2019）；Sicacha-Parada 等人（2021）先前所提到的。与可达性不同，土地利用可能与电力线致死风险差异有关。然而，它也是一个与可探测性高低相关的因素。因此，在接下来的章节中，我们将考察土地利用是否应被视为与输电线路致死风险差异及可探测性差异相关的因素。

3 模型

与特伦德拉格电力线网络相关的死亡鸟类出现记录的生成过程，是本研究的目标模型。我们将沿输电线路网络的死亡鸟类出现记录建模为

$P \subset \mathbb{R}^2$ 具有强度 $\lambda_{true}(\mathbf{s})$ 的点模式。该强度进一步建模为已知环境变量 $X(\mathbf{s})$ 的线性组合与空间高斯随机场 $\omega_1(\mathbf{s})$ 之和，例如电力线密度和鸟类丰度：

$$\log(\lambda_{true}(\mathbf{s})) = X(\mathbf{s})\beta + \omega_1(\mathbf{s}) \tag{1}$$

因电力线死亡的鸟类的真实强度及其组成部分是我们关注的量。在我们的案例研究中，有两种可用的数据来源，即第2节所述的公民科学（CS）数据和专业数据。每种数据类型都有两种不同的观测过程，我们将在下文中为其提出模型。

3.1 专业调查的数据过程建模

我们希望在专业调查模型中纳入两个方面：1）采样地点可能是偏好选择的；2）采样力度很高，因此我们假设该过程得到了完全抽样。因此，我们将专家观测到的数据解释为 Diggle 等人（2010）所描述的偏好抽样过程的一个实现。每条 m缓冲电力线（见第2节）都有 $\phi_i(\mathbf{s})$；$i = \{1, \ldots, m\}$ 被抽样的概率。如果所访问线路的选择完全随机，我们假设 $\phi_1(\mathbf{s}) = \ldots = \phi_m(\mathbf{s})$。然而，正如我们的案例研究所示，专家通常对在哪里能够找到他们想要调查的对象具有先验知识，因此我们假设 $\mathbf{\Phi} = [\phi_1(\mathbf{s}) \ldots, \phi_m(\mathbf{s})]^T$中的概率与潜在过程 $\omega_1(\mathbf{s})$（见公式1）以及最终与 $X(\mathbf{s})$中的某些协变量具有随机依赖关系。除 $\omega_1(\mathbf{s})$外， $\mathbf{\Phi}$ 中的概率还由已知变量 $Z(\mathbf{s})$解释，这些变量可能包括 $X(\mathbf{s})$中的某些协变量，如下所示：

$$logit(\phi_i(P_i)) = Z(P_i)\gamma + \zeta\omega_1(P_i) \tag{2}$$

系数 ζ 与此设定相关，因为它决定了被抽样的输电线路的抽样选择与定义鸟类撞击或触电热点区域的真实生态过程之间的依赖程度（Gelfand和Shirota，2019）。

选定待抽样的输电线路后，正如本小节开头所述，我们假设在每个选定区域内，对死亡鸟类进行全面普查，并且完全检测。因此，我们观测到一个强度为 $\lambda_{PA}(\mathbf{s}) = \lambda_{true}(\mathbf{s})\phi_k(\mathbf{s})$ 的点模式 $\mathbf{P}_{PA}$，其中 k 表示位置 $\mathbf{s}$ 所属的输电线路。图 S.1.以图示方式说明了专业调查中采用的观测过程。

值得注意的是，由于我们使用的是网络的缓冲区版本 P，式 (2) 中的模型属于区域数据模型（Banerjee等人，2015）。在避免可辨识性问题的同时考虑偏好抽样，意味着要联合建模 $\lambda_{PA}(\mathbf{s})$ 和选择概率 $\phi_i(\mathbf{s})$。关于这一建模框架部分的更多细节见第 3.1.1 节。

3.1.1 偏好抽样建模

如前所述，在我们的案例研究中，对偏好抽样进行建模意味着需要整合点过程与区域数据。两类数据都依赖于高斯随机场 $\omega_1(\mathbf{s})$。然而，高斯随机场的连续版本会生成点过程，而其离散化版本则产生区域数据。此前，Roksv˚ag等人（2020）和Wang等人（2021）分别通过水文学和流行病学应用，研究了两种或更多类型空间数据的整合。一般而言，区域 $\mathbf{A}$ 上的高斯随机场可表示并近似为：

$$\omega_1(\mathbf{A}) = \frac{1}{|\mathbf{A}|}\int_{u\in\mathbf{A}} \omega_1(u)du \approx \frac{1}{H}\sum_{\mathbf{s}\in\mathbf{A}} \omega_1(\mathbf{s}) \tag{3}$$

与$\mathbf{s} = \{s_1, \ldots, s_H\}$ 中一组采样点$\mathbf{A}$。然而，由于在公式（2）中，$\omega_1(P_i)$ 是$\phi_i(P_i)$ 的非线性函数，因此，对于我们的案例研究，我们将 $\omega_1(\mathbf{A})$表示并近似为：

$$\omega_1(\mathbf{A}) = logit\left(\frac{1}{|\mathbf{A}|}\int_{u\in\mathbf{A}} \frac{\exp(\omega_1(u))}{1+\exp(\omega_1(u))}du\right) \approx logit\left(\frac{1}{H}\sum_{\mathbf{s}\in\mathbf{A}} \frac{\exp(\omega_1(\mathbf{s}))}{1+\exp(\omega_1(\mathbf{s}))}\right) \tag{4}$$

如Wang等人（2021）所述，以避免生态偏差（Greenland，1992）。

3.2 公民科学数据处理模型

我们假设，公民科学数据是对死亡鸟类真实点模式进行稀疏化的结果，该过程包含三个组成部分：采样力度、可探测性和报告力度。由于采样力度不均（即公民科学家对不同地点进行采样的概率存在差异）、事件出现地点不同导致可探测性存在差异，以及观察者之间报告概率不同，这些因素都可能造成偏差。基于此，我们假设公民科学数据遵循点过程模型，具体而言，是对数高斯考克斯过程（LGCP），$\mathbf{P_{CS}}$，其强度为 λ_{CS}(s)：

$$\lambda_{CS}(\mathbf{s}) = \lambda_{true}(\mathbf{s}) \cdot \tau(\mathbf{s}) \cdot \psi(\mathbf{s}) \cdot \delta(\mathbf{s}) \tag{5}$$

其中，λ_{true}(s)表示电力线导致的死亡鸟类真实出现事件的强度，τ(s)表示位置s 属于公民科学家抽样地点的概率，ψ(s)表示在位置s 已被访问的前提下探测到出现事件的概率，而 δ(s)表示在公民科学家已访问该地点且探测到出现事件的前提下，在s 报告出现事件的概率。

我们重点考虑三种偏差来源：采样力度不均、可探测性以及报告力度。不过，该建模框架具有足够的灵活性，也可以纳入可能需要考虑的任何其他偏差来源。

3.2.1 采样力度

公民科学数据的抽样过程并未标准化（Isaac 等人，2014年），且往往偏向可达性较高的地点，或观察者预期出现记录更多的地点，即优先抽样的地点（Fithian 等人，2015；Monsarrat 等人，2019）。因此，我们提出将项 τ(s) 建模为：

$$logit(\tau(\mathbf{s})) = Z(\mathbf{s})\alpha + \omega_2(\mathbf{s}) \tag{6}$$

其中， $Z(\mathbf{s})$ 是旨在解释公民科学家最有可能前往哪些地点的协变量，而 $\omega_2(\mathbf{s})$ 是空间随机效应。

3.2.2 可探测性的差异

在许多现实情境中，不能假定感兴趣事件（在我们的案例中是死亡鸟类）的探测概率恒定。对于我们的案例研究，影响能否探测到死亡鸟类的一个相关因素是土地利用。正如先前研究所指出的（Dom´ınguez del Valle 等，2020），开阔地等栖息地类型使此类事件更容易被探测到。因此，可探测性的差异被视为死亡鸟类实际点模式的第二个稀疏化因素。我们将某地点s 在已被访问的条件下，探测到死亡鸟类出现记录的概率定义为：

$$logit(\psi(\mathbf{s})) = W(\mathbf{s})\nu \tag{7}$$

其中， $W(\mathbf{s})$ 表示用于解释影响可探测性因素的协变量，在我们的案例中是土地利用或栖息地类型。

3.2.3 报告力度

要估计 $\delta(\mathbf{s})$ 所代表的报告力度，几乎总是需要无法获得的信息。不过，我们在此提出两种方法，用于处理报告力度不均所导致的噪声，这种噪声会影响观测到的死亡鸟类报告点模式。

简单报告误差模型

该方案假设，导致报告误差差异的随机效应不存在任何结构。因此， $\delta(\mathbf{s})$被假定取决于一个超参数 θ。 $\delta(\mathbf{s})$与θ 之间的关系如下：

$$\delta(\mathbf{s}) = \frac{\exp(\theta)}{1 + \exp(\theta)} \tag{8}$$

其中， θ服从正态先验 $\pi(\theta) = N(0, 1)$。

观察者特异性报告努力模型

这是一种更复杂的方法，因为我们假设与每条出现记录相关的报告概率取决于其位置，以及在其周边开展公民科学活动的观察者（即，根据观察者在该位置附近的其他已报告观测，判断其活跃程度）。在这种情况下，我们将 δ（s）表示为：在某个位置 s，已知一位公民科学参与者曾到访该地且检测到了某条出现记录时，报告该出现记录的概率：

$$\delta(\mathbf{s}) = \sum_{j \in \{obs\}} \psi_j(\mathbf{s}) \kappa_j(\mathbf{s}) \tag{9}$$

其中，ψ_j(s) 是抽样到的某条出现记录的观察者为观察者j 的概率；κ_j(s) 是观察者 j 已探测到的这条记录最终被报告的概率。式（9）可解释为一种加权平均，即每位观察者看到一只死亡鸟类后报告其出现记录的概率的加权平均。权重 ψ_j(s) 取决于观察者的特征，例如其与s的距离，或观察者的活跃程度（即在公民科学门户网站上的观测记录数量）。在此建模框架中，ψ_j(s) 被视为确定性输入，而模型旨在估计 κ_j(s)。关于如何估计 ψ_j 的更详细说明，请参见补充信息。

3.3 先验设定

式（1）和式（6）中的空间高斯随机场 ω_1(s) 和 ω_2(s) 假定服从 Mat´ern 协方差函数，其形式如下：

$$\frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}(\kappa\|s_i - s_j\|)^{\nu} K_{\nu}(\kappa\|s_i - s_j\|) \tag{10}$$

其中，$\|s_i - s_j\|$ 表示两个位置 s_i、$s_j \in D$之间的欧几里得距离。σ^2 表示边际方差，K_{ν} 表示阶数为 $\nu > 0$的第二类修正贝塞尔函数。ν是决定该过程平滑程度的参数，而 $\kappa > 0$ 是一个尺度参数。参数 ν 固定为 1。空间范围 ρ 表示为$\rho = \sqrt{8}/\kappa$。参数 ρ 和 σ 的先验分布采用惩罚复杂度（PC）先验指定（Fuglstad 等人，2019）。参数向量 $\boldsymbol{\beta}$、$\boldsymbol{\gamma}$、$\boldsymbol{\alpha}$ 和 $\boldsymbol{\nu}$ 服从均值为 0、精度为 0.01 的正态先验分布。

3.4 拟合模型：inlabru R 包

由于我们的建模框架属于潜在高斯模型（LGM；Rue 等人（2009））框架，因此可以使用 INLA-SPDE 方法（Rue 等人，2009；Lindgren 等人，2011）方便地拟合我们的模型。INLA 能够快速、可靠地进行推断，其目标是对模型参数和超参数的边际后验分布进行数值近似。SPDE 方法基于 SPDE 的解，该解可通过定义在空间域离散化网格上的基函数表示进行近似。Simpson 等人（2016）提出采用基于空间剖分的 SPDE 方法，以高效地近似 LGCP 的似然。

该建模框架在线性预测子中纳入非线性项，以校正公民科学数据中的偏差。处理这些非线性的一种替代方法是对其进行迭代线性化。此方法使用 inlabru R 包（Bachl 等，2019）实现：先在所谓的线性化点处对非线性函数进行线性化，以此作近似；然后在该点处通过二阶泰勒级数近似来近似后验分布。通过不动点迭代法（Burden 等人，2015）寻找线性化点。更多细节和示例请参见 *https://inlabru-org.github.io/inlabru/articles/method.html*。

4 模拟研究

为了探究在分析公民科学数据时考虑抽样过程的重要性，以及公民科学家之间报告力度的不均衡性，我们基于特伦德拉格案例研究开展了一项模拟研究，使用该地区的输电线路以及第2.3节中定义的一些解释变量。我们设置了四种模拟情景，分别代表专家访问输电线路时的不同选择方案（随机抽样或偏好抽样），以及公民科学家发现死亡鸟类后报告的不同意愿（低或高）。对于每种情景，我们生成100个数据集，并将其拟合到一组模型中。这些模型涵盖了分别使用各数据来源的模型，以及整合专业调查和机会性记录、同时考虑两类数据收集过程中部分偏差的模型。

4.1 数据集模拟模型

4.1.1 真实出现记录的模拟模型

为了开始模拟研究，我们生成了100个强度为 $\lambda_{true}\mathbf{s}$（$\mathbf{s}$）的空间点模式，其代表真实风险强度，具体如下：

$$\log(\lambda_{true}(\mathbf{s})) = -2 + 0.75\ CLOUDCOVER(\mathbf{s}) + \omega_1(\mathbf{s}) \tag{11}$$

其中，表示高斯随机场 ω_1(s) 的 100 次实现中的每一次，其范围为 $\rho_1 = 1$ ，方差为$\sigma_1^2 = 0.3$。

4.1.2 公民科学模拟模型

按照公式 (11) 的设定生成的点，经过了稀疏化，稀疏化概率取决于公民科学家抽样过程（见第 3 节）。在本例中，公民科学家的抽样过程由强度为 λ_{CS}(s) 的点过程表示，其定义如下：

$$\log(\lambda_{CS}(\mathbf{s})) = -4 - 2\ DISTANCE(\mathbf{s}) + \omega_2(\mathbf{s}) \tag{12}$$

其中，距离 表示到最近三级道路的距离， ω_2(s) 表示范围为$\rho_2 = 100$ 、方差为 $\sigma_2^2 = 1.3$ 的高斯随机场。由于协变量 距离 的系数为负，位于公民科学家更容易到达地点的出现事件更有可能被保留。下一阶段的稀疏化取决于死亡鸟类出现记录的探测概率，该概率受土地利用等因素影响；土地利用会影响观察者到达有出现记录的地点后探测到死亡鸟类的可能性。保留出现记录的概率 ψ(s) 表示为：

$$\text{logit}(\psi(\mathbf{s})) = 1 - 2\text{山地}(\mathbf{s}) + 1.2\text{开阔地}(\mathbf{s}) + 1.4\text{岩石地}(\mathbf{s}) + 1.8\text{城市}(\mathbf{s}) - 3\text{水域}(\mathbf{s}) \tag{13}$$

该模型中的每个协变量都是第2.3节所定义的各类土地利用的指示变量。稀疏化的最后阶段取决于以下概率：在死亡鸟类已被发现且观察者已到达尸体所在地点的情况下，报告该死亡鸟类出现记录的概率

。假设公民科学参与者群体由10名观察者组成，并按照第 3.2.3 节中的定义，将 ϕ_j(s)建模为：

$$\text{logit}(\phi_j(\mathbf{s})) = 10 - 0.3DISTTOOBS_j(\mathbf{s}) + \zeta ACTLEVEL_j \tag{14}$$

其中， $DISTTOOBS_j$(s)表示观察者$j = \{1, \ldots, 10\}$的公民科学活动质心到位置s 的距离，$ACTLEVEL_j$表示一个分类有序变量，取值为每位观察者的活动水平（1–5）。向量$\zeta = [0, 0.5, 1, 1.5, 2]$决定了各活动水平与 ϕ_j(s)之间关系的强度。根据公民科学参与者群体的报告意愿，κ_j(s)有两种定义。当使用 $Beta$(2,5)分布（即均值为 ≈ 0.29）生成κ_j(s)的值时，假设报告意愿较低；当使用 $Beta$(5,1.5)分布（即均值为 ≈ 0.77）时，假设报告意愿较高。补充信息中的图 S.1 总结了各数据集的生成过程。

4.1.3 专业调查模拟模型

这些数据经过模拟，旨在仿照 NINA 开展的专业调查。对于每个模拟数据集，我们假设电力线周围的缓冲区 P_i 以概率ϕ_i(P_i)被抽样。假设采用随机抽样时，所有缓冲区的这一概率均相同（ϕ_i(P_i) $= p$）；而假设采用偏好抽样时，该概率为：

$$\text{logit}(\phi_i(P_i)) = -2.5 - 1.5ELEVGRADIENT(\mathbf{s}) + 3.5CLOUDCOVER(\mathbf{s}) \tag{15}$$

在这种情况下，我们假设 ω_1（s）不属于偏好抽样模型，以降低模拟的计算复杂度。一旦选定一条路线段，便假定缓冲区内的所有出现事件都能被观测、检测并报告，因为这些调查采用训练犬进行抽样，采样力度较高。图 S1 展示了这种抽样方式的示意图

4.2 模拟情景

对于每种模拟情景，根据公式（11）对第 节中生成的 100 个点模式进行稀疏化，从而生成 100 个观测数据集。这些点模式的稀疏化方式定义了图4所示的四种模拟情景。

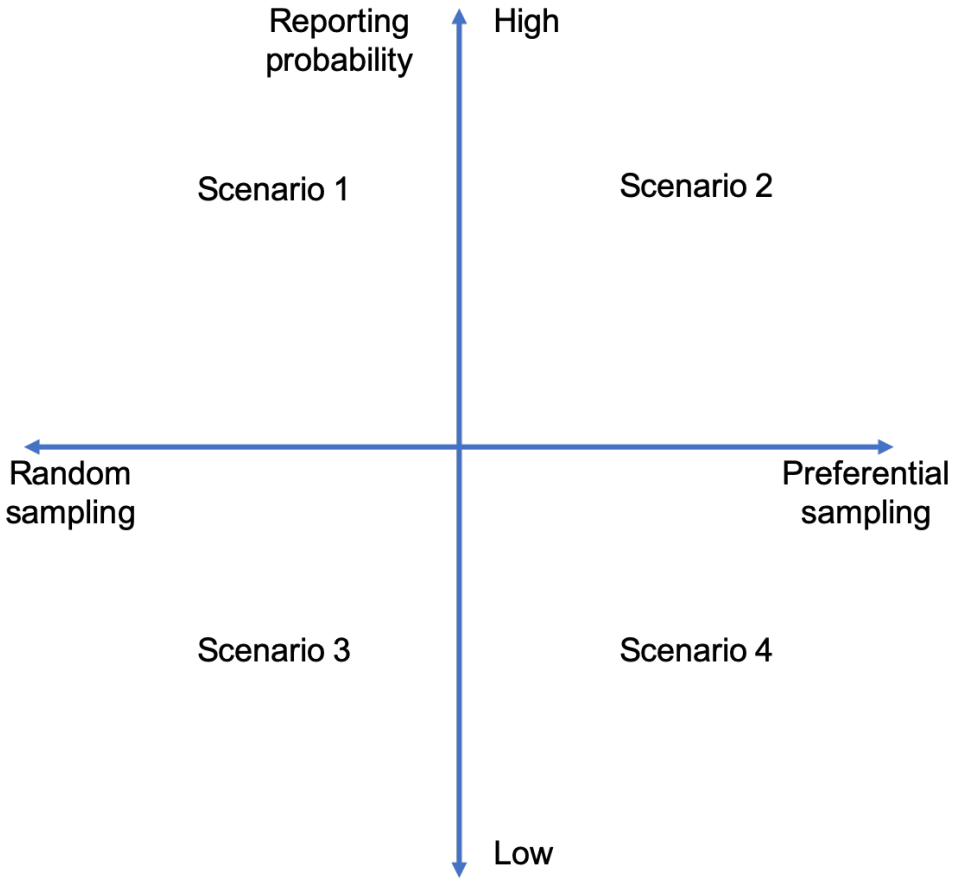

图4：第 4 节模拟研究的情景。

四种模拟情景由两个因素决定：i）探测到的发生事件的报告概率是高还是低（见第 4.1.2 节）；ii）专业调查中输电线路的选择是完全随机还是具有偏向性（见第 4.1.3 节）。在情景1和情景2中，我们假设观察者探测到的出现事件的报告概率较高；但在情景1中，输电线路的抽样被假设为完全随机，而在情景2中，输电线路的抽样则像公式（15）所示具有偏向性。情景3和情景4之间也存在类似的关系。不过，这些情景的报告概率被假设为较低。

4.3 模型比较

我们希望比较对先前所述不同模拟情景生成的数据集拟合不同模型的结果（即仅使用一种数据类型的模型，以及融合两类数据的模型）。我们在表1中定义了八个模型，并将其拟合到生成的 400 个数据集中的每一个（100 个真实发生点模式 $\times$ 4 模拟情景）。由于模型 7 和 8 的计算开销较大，且专门针对报告概率较低的情景（情景3和情景4）设计，因此未在情景1和情景2中拟合。

表1：模拟研究中拟合的模型描述。在数据来源列中：PS代表专业调查，CS代表公民科学。

模型	数据来源	描述
1	PS	仅来自专业调查的数据（公式（1））
2	PS	仅使用考虑偏好抽样的专业调查数据（方程（1）和（2））
3	CS	仅使用来自公民科学报告且包含位置信息的数据（方程（1））
4	CS	与模型3相同，但考虑公民科学的抽样过程（方程（5） $\delta(\mathbf{s})=1$）
5	CS	与模型4相同，但还考虑了检测过程（方程（5）与 $\delta(\mathbf{s})=1$）
6	PS +公民科学	整合模型2和模型5，同时考虑抽样和检测过程公民科学和偏好抽样。（方程（1）、（2）和（5）与 $\delta(\mathbf{s})=1$）
7	专业调查 + 公民科学	整合模型2和模型5，并考虑抽样、探测以及公民科学的报告过程和偏好抽样。（方程（1）、（2）和（5），其中 $\delta(\mathbf{s})$ 与方程（8）中的相同）
8	专业调查 + 公民科学	整合模型2和模型5，并考虑抽样、探测以及公民科学的报告过程与偏好抽样。（方程（1）、（2）和（5），其中 $\delta(\mathbf{s})$ 与方程（9）中相同）

4.4 结果

本模拟研究的结果针对模型的每个参数，使用偏差和均方根误差进行汇总。协变量对生态状态的影响对于将其用于决策支持至关重要。因此，在图5中，我们展示了参数β_1的性能指标；在本模拟研究中，该参数表示协变量云量 对电力线致死风险的影响。

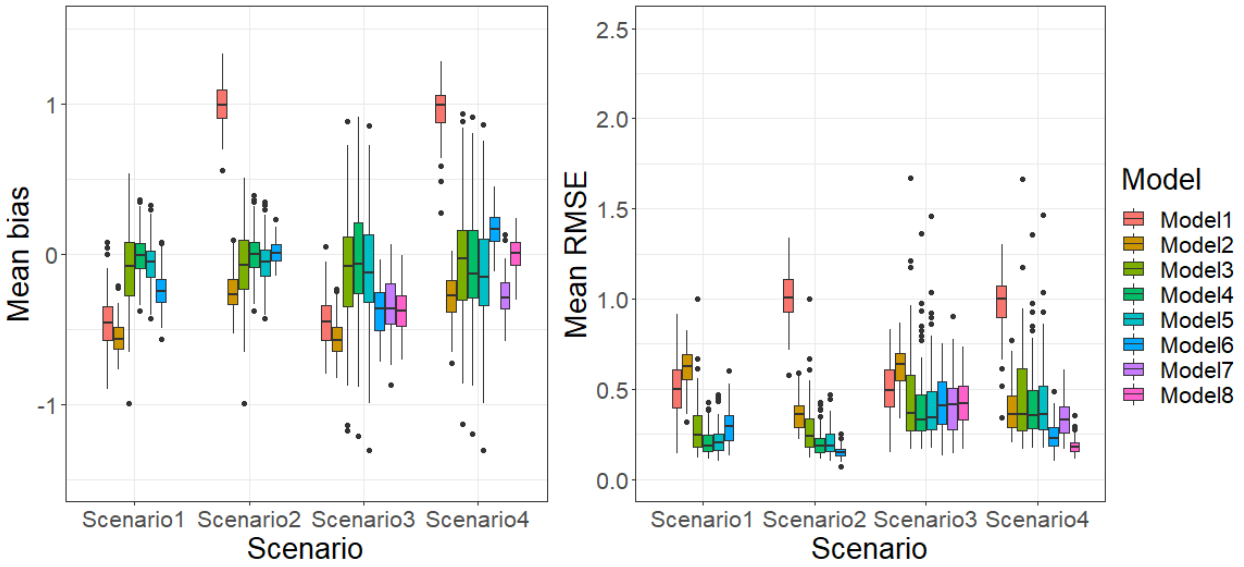

图5：所提出的八个模型各自的平均偏差和均方根误差 β_1 .

我们首先观察到，模型1（仅基于专业调查且未考虑偏好抽样）在存在偏好抽样的情景（情景2 和 4）中表现较差，但在随机抽样的情景（情景 1 和 3）中优于模型2（仅基于专业调查且考虑了偏好抽样）。然而，在所有模拟情景中，这两个模型都不如所有基于公民科学数据的模型（模型3–8）。仅基于公民科学数据的模型（模型 3、4 和 5）在随机选择抽取输电线路的情景（情景 1 和 3）中表现更好。另一方面，在假设对输电线路进行偏好选择且报告意愿较高的模拟情景（情景2）中，模型6（两类数据，同时考虑偏好抽样以及公民科学数据收集偏差，但不包括报告力度）明显优于其他竞争模型；此外，该模型的参数 β_1后验均值始终表现出较小的变异性。在存在偏好抽样且报告意愿较低的情景（情景4）中，模型8（两类数据，同时考虑偏好抽样以及公民科学数据收集中的所有偏差）在偏差和均方根误差方面均优于所有竞争模型。在该情景中，模型6优于模型7（两类数据，并采用简单方法处理报告力度），这可能是因为模型7在解释报告力度差异方面缺乏结构。最后需要注意的是，对于所有模型而言，在报告意愿较高的情景（情景1和情景2）中，β_1 的后验均值变异性都明显小于报告意愿较低的情景

（情景3和情景4）因为公民科学数据中包含的信息要丰富得多。

固定效应和空间超参数的汇总结果均列于补充信息的表 S1 中。对于 β_0 ，我们发现，在所有模拟情景中，不考虑数据收集过程中的任何偏差的模型（模型1和模型3）都会低估其值；而对于情景1和情景2，模型5（仅基于公民科学数据，并考虑可达性和可探测性）的表现最佳。此外，在报告意愿较高的情景中（情景3和情景4），模型6、7和8（融合两类数据的模型）的表现优于所有其他竞争模型。请注意，对于最复杂的模型（模型2、6、7和8），高斯随机场 $\omega_1(s)$ 的空间超参数估计并不准确。为更详细地探究这种偏差的影响，我们固定了模拟研究中一个数据集的空间超参数，并拟合了模型6、7和8。补充信息中的图 S.11 和 S.12 展示了情景3和情景4下 β_1 的边际后验分布。对于 β_1 的边际后验分布，我们得到了类似的准确性结果。不过，我们注意到，与未固定空间超参数时相比，这些后验分布更加精确。鉴于模型6、7和8中的空间超参数后验分布对其先验设定十分敏感，使用信息量更丰富的先验信息，或有助于获得固定效应和空间超参数更准确的后验分布。

我们提出的框架不仅能够推断驱动生态过程的参数的后验分布，还能帮助我们了解导致公民科学数据和专业调查数据产生偏差的过程。在补充信息的表 S.1–S.4 中，我们列出了将真实点模式稀疏化以生成公民科学和专业调查数据所涉及参数的比较指标。我们发现，模型 4 至 8 对抽样过程模型参数的估计结果相近，α_0 和 α_1的偏差均较小。另一方面，与可探测性相关的参数存在偏差，且均方根误差较大。这可能与这些参数的可辨识性较差有关；采用信息性先验信息，或进一步了解土地利用与可探测性之间的关系，或许能够改善这一问题。最后，在情景 2 和 4 中，偏好抽样模型的参数均得到了准确

估计，这与预期一致。模型 8 的一个特点是，它能够对每位观察者报告输电线路致死事件的意愿进行后验推断。

最后，我们从预测性能方面比较所提出的模型。比较主要关注两个方面：预测的准确性以及预测的不确定程度。这些预测仅基于线性预测子的固定效应。为比较各模型生成预测的准确性，我们计算了每个模型在每种模拟情景下，对特伦德拉格电力线网络沿线一个密集网格上的预测概率的均方根误差。各模型在各情景下的均方根误差图见图 S.2、S.4、S.6 和 S.8。结果表明，在报告意愿较高的情景（情景1和情景2）中，模型5（仅基于公民科学数据，并考虑可达性和可探测性）的表现优于所有其他模型。在报告意愿较低的情景（情景3和情景4）中，在研究区域的大片区域内，融合两类数据的模型（模型6–8）表现优于模型5。

关于不同模拟情景下生成的预测的不确定性，我们计算了上述密集网格上各模型 95% 预测区间的平均宽度。预测区间长度的地图见图 S.3、S.5、S.7 和 S.9。模型 2（仅使用专业调查，同时考虑偏好抽样）产生的不确定性最大；而最复杂的模型（模型 6 至 8）的不确定性高于仅基于公民科学数据的模型（模型 3–5）。

5 特伦德拉格输电线路鸟类死亡案例研究

我们的真实数据应用研究旨在更深入地了解景观（环境协变量）在形成鸟类电力线致死高风险区域方面所起的作用。根据这些因素对输电线路致死风险的影响，可以绘制输电线路致死风险地图，为电力公司和水资源与能源局（NVE）提供参考。为实现这一目标，我们使用了两种信息来源（见第2节）：1）由 NINA 开展的专业调查——

以及 2）由公民科学家收集的机会性记录。本节将对第 4 节介绍的模型 1 至 7 进行拟合，然后展示鸟类电力线致死风险的模型预测结果，最后比较使用所提出模型得出的预测结果。

为解释两种观测数据类型背后的过程而选取候选变量，是基于专家知识以及第2节中开展的探索性分析。由于预计鸟类暴露程度越高、电力线网络越密集，输电线路致死事件就越多，因此我们将电力线密度 （挪威水资源与能源局；NVE）和鸟类丰度 （Sicacha-Parada 等人，2022）作为首选候选变量。我们还考虑了土地利用 （AR50, https://www.nibio.no/tema/jord/arealressurser/ar50），它能说明电力线周边的土地类型；以及云量 （https://www.earthenv.org/cloud），它可作为鸟类飞行时可见度的替代指标。这两者也被纳入候选变量考虑范围，但由于土地利用与选定的两个协变量相关，且加入云量并未改善现有模型，因此最终将其舍弃。正如 Fithian 等人（2015）和 Monsarrat 等人（2019）此前所述，可达性是决定公民科学家更可能在何处收集生物多样性信息的主要因素之一。因此，我们考虑使用协变量到（三级）道路的距离 和到水体的距离 （海洋、湖泊和河流），来解释公民科学家在何处收集观测记录。我们未将土地利用用于解释公民科学家前往何处收集观测记录，因为这一协变量也与距离变量相关。另一个需要考虑的因素是死亡鸟类的可探测性。这会受到鸟类体型（ Borner 等人，2017；Ponce 等人，2010）、观测记录发生的时间（Bevanger，1995）、尸体所在地点的土地利用（Philibert 等人，1993；Schutgens 等人，2014；Dom´ınguez del Valle 等人，2020）等因素的影响。在本研究中，我们将不同的土地利用类型作为解释可探测性差异的替代指标。最后，公民科学家是否愿意报告死亡鸟类仍是一个尚无定论的问题。因此，我们将其视为影响公民科学数据库中观测结果的另一个相关因素，并通过模型7对此加以考虑。

偏好抽样也被视为 NINA 专业调查数据收集过程中的一种潜在缺陷，因为负责这些项目的专家拥有先验

知识，在决定在哪里收集观测记录时可能会用到这些知识。第2节的结果表明，电力线密度和云量可能是导致偏好抽样的因素。根据 NINA 提供的信息，海拔梯度也是选择要巡查哪些线路时的相关因素。鉴于海拔梯度与云量相关（皮尔逊相关系数， $\rho = 0.76$），因此选择海拔梯度和电力线密度作为解释偏好抽样的变量。

在第 3 节中，我们指出，随着待估计参数数量大幅增加，在尝试校正公民科学数据中的偏差时可能会出现可辨识性问题。因此，除了两种观测数据类型外，我们还纳入了一个包含公民科学报告位置的点模式，以便识别抽样过程的参数。对于关联不同土地利用类型与死亡鸟类探测的参数，由于目前没有研究解释挪威土地利用与可探测性之间的关系，我们依据专家知识提出了信息性先验分布（Dom´ınguez del Valle et al., 2020）。我们拟合了所提出的七个模型，并整合了数据收集过程中各类偏差来源对应的可用数据类型。图6以图形方式展示了生态过程所涉及各固定效应的后验汇总。

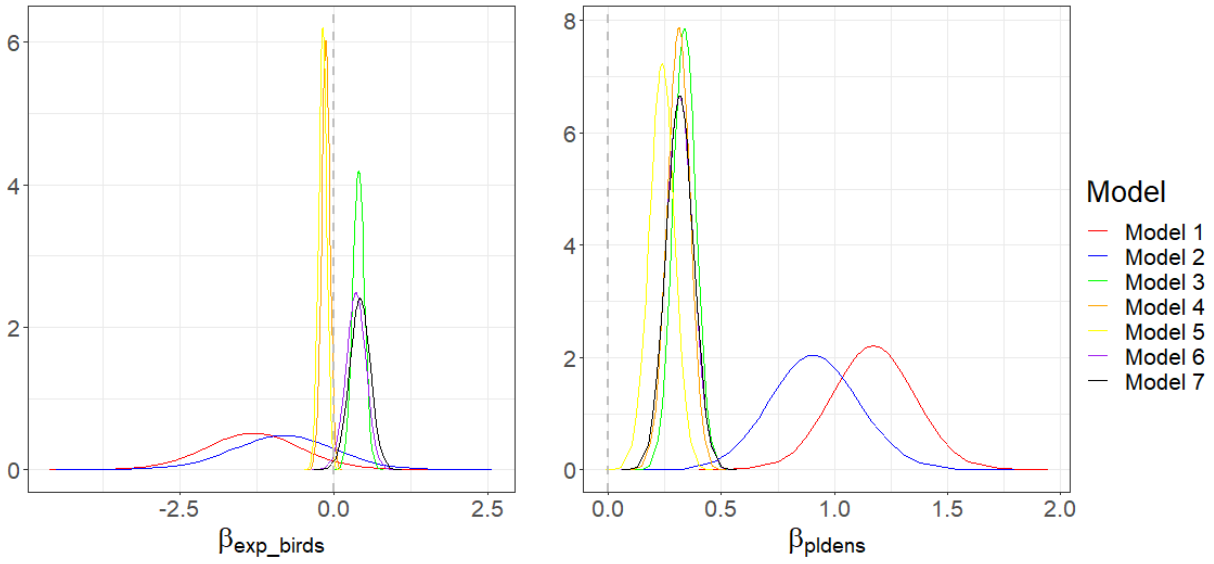

图6：所拟合七个模型中固定效应的后验分布

对于协变量 暴露鸟类 的影响，我们观察到所提出的模型之间存在相当大的不一致。模型 1、2、4 和 5 的边际后验分布中心位于 0 以下，

而模型 3、6 和 7 的该分布则主要位于 0 以上。对于协变量 电力线密度 的影响，所有所提出的模型给出的结果更加一致，其边际后验分布均集中在相同数值附近，但模型 1 和 2 除外。值得注意的是，与其他模型相比，模型 1 和 2 的后验分布表现出高得多的不确定性。这些参数以及模型中其他参数的后验汇总见补充信息。需要注意的是，与模拟研究中一样，模型 2、6 和 7 的空间范围参数和边际方差均显著高于模型 3、4 和 5。

使用提出的七个模型分别预测了特伦德拉格电力线网络沿线的电力线致死风险。风险的后验中位数如图 7 所示

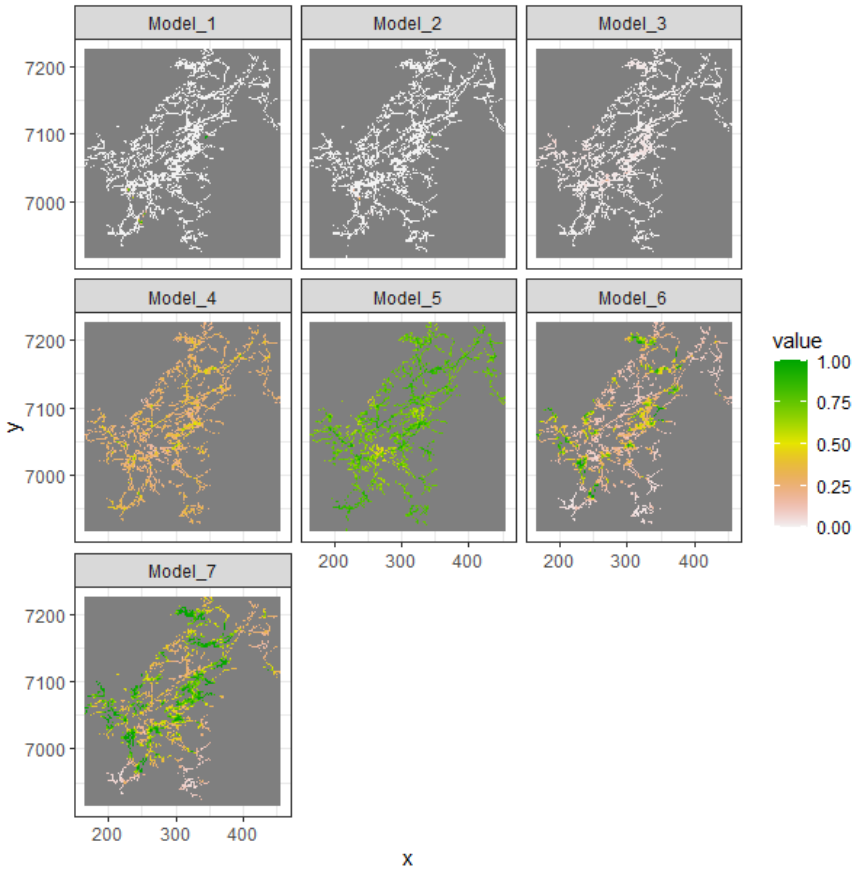

图 7：电力线致死风险的后验中位数

如表 S.5 所示，模型 1、2 和 3 中截距的后验均值使

预测风险的后验中位数取值偏低，但已进行抽样的地点除外。模型 4-7 得出的预测更为现实。模型4仅在研究区域内突出显示了极少数地点，而模型5在考虑可探测性差异后，预测整个区域的风险都较高。整合两类数据后（模型 6 和 7），被标记为电力线致死风险较高的区域减少了。尽管两者相似，模型7预测的概率高于模型6，因为前者考虑了报告力度的差异。预测的不确定性通过预测风险的标准差计算，并见图 8。

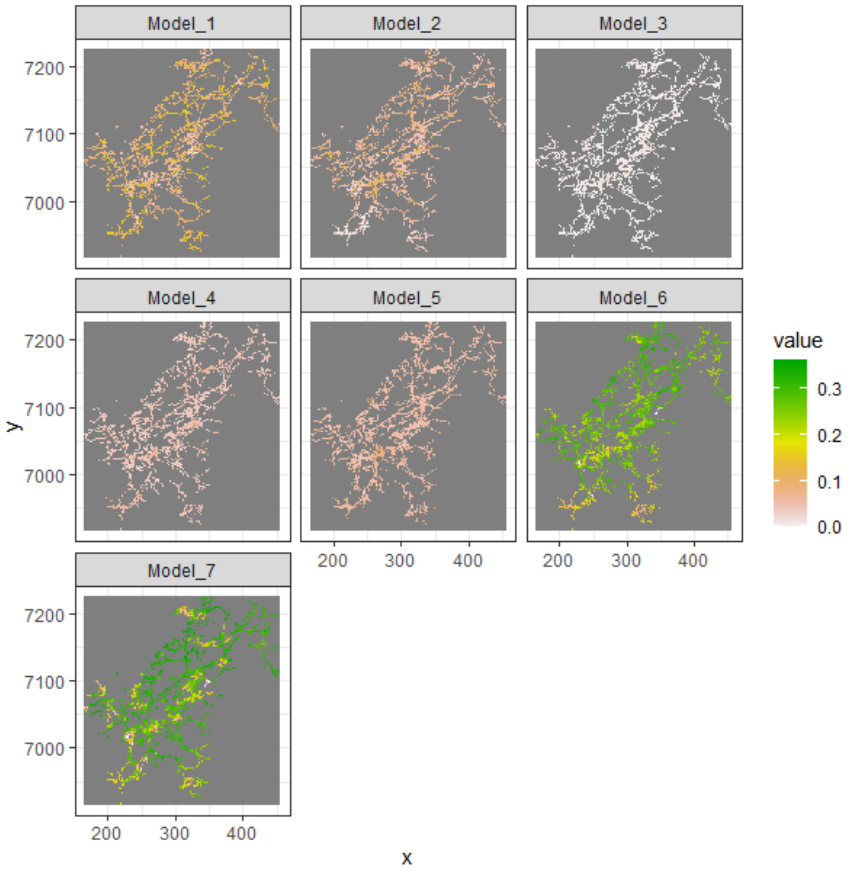

图 8：电力线致死预测风险的标准差 hs

正如预期，仅基于专业调查的预测所覆盖的研究区域较小，因此其不确定性高于仅使用公民科学数据中的机会性记录所得的预测。如图 S.10. 所示，对于模型 6 和 7，整个研究区域内术语 ω_1 (s) 的不确定性都相当大，而对于模型 1 和

2，固定效应的不确定性更大。转换为电力线致死风险尺度后，模型 6 和 7 的不确定性最大。

这些结果与第 4 节中针对模拟数据集作出的预测一致：仅使用公民科学数据的模型所生成的预测不确定性较低。然而，仅使用专业调查数据的模型 6 和模型 7，其预测的不确定性高于所有其他模型。

6 讨论

本文提出并评估了一种方法框架，用于整合多种数据类型，以探讨生态学和生物多样性方面的问题。我们重点关注两类待整合的数据：专业调查数据，以及由公民科学家收集的机会性记录。本研究的一个基本假设是，两类数据中的观测记录都源自相同的潜在生态过程，但其观测过程和报告过程不同，因此产生了空间覆盖范围和观测力度各异的观测记录。我们首先通过对决定各数据类型生成的因素进行建模，着手应对整合这些数据类型的挑战。具体而言，专家收集的数据受到开展调查的专业人员先验知识的影响，因此我们假设采用偏好抽样设计，使某些输电线路被访问的概率高于其他线路。另一方面，公民科学数据受到可达性、可探测性以及公民科学家之间报告力度不均等因素的影响。我们将两类数据都视为稀疏点过程的实现，但稀疏化实际生态过程的机制各不相同。我们假设公民科学数据受到发生在点位级的过程影响，而专业调查数据则受到发生在区域级的选择概率影响。

为考虑影响两类数据收集过程的因素并避免可识别性问题（Fithian 等人，2015），我们使用了额外的信息来源，以了解稀疏化所涉及的过程。由于研究区域内没有可用于说明土地利用/栖息地与死亡鸟类可探测性之间关系的既有研究，

因此，我们提出了基于专家知识的先验。我们提出了能够模拟各类数据生成过程、且属于潜在高斯模型范畴的贝叶斯空间模型，因此采用 INLA-SPDE 方法进行拟合，因为该方法在拟合此类贝叶斯空间模型时计算效率较高。然而，如第 3 节所示，用于解释收集过程中的偏差如何影响观测到的点模式生成的线性预测子项，并不符合 INLA 方法对数高斯 Cox 过程所要求的对数线性条件。因此，作为对传统 INLA-SPDE 方法的补充，我们采用了 R 包 $inlabru$ 中的方法（Bachl 等，2019），该方法基于对线性预测子中的非线性项进行迭代线性化。当高斯随机场属于线性预测子的非线性部分时，这一额外步骤可能会增加计算负担，因为 SPDE 方法带来的稀疏性，以及将高斯随机场近似为高斯马尔可夫随机场所带来的稀疏性，可能会有所降低。对于最复杂的模型（第 4 节中定义的模型 7 和 8），我们遇到了一些数值问题，这些问题可能与这种近似有关，因为模型未能收敛。

为研究第 3 节所提出模型的特征和性质，我们开展了一项模拟研究和一项案例研究。我们针对 100 个不同的数据集拟合了使用一种或两种数据类型、并以不同程度考虑数据收集过程中的偏差的模型。数据集涵盖了以下情景：专业调查优先或随机选择输电线路，以及公民科学家报告输电线路致死事件的意愿高或低。模拟研究结果表明，在存在偏好抽样的情景中，整合两类数据十分重要：在情景 2 和 4（存在偏好抽样的情景）中，模型 6–8 对固定效应的估计更准确。这些情景也最接近案例研究中的实际情况。在情景 1 和 3（不存在偏好抽样的情景）中，考虑偏好抽样的模型表现不如在其他情景中出色，这可能是因为这些模型引入的大量参数缺乏可辨识性。

对所提出模型的预测性能评估表明，整合两类数据的模型（模型 6-8）优于仅基于一种数据类型的模型（模型 1-5）。关于预测的不确定性，仅基于专业调查的模型（模型 1 和 2）所作预测的不确定性

大于仅基于公民科学数据的模型（模型 3-5），这可能是由于公民科学数据覆盖的区域更广。此外，整合两类数据的模型（模型 6-8）所作预测的不确定性也大于仅基于公民科学数据的模型（模型 3-5）。在我们的案例研究中也观察到了类似的模式：使用模型 1、2、6 和 7 得出的预测，其不确定性高于模型 3-5 的预测。

在模拟研究中，我们发现，与仅使用公民科学数据获得的估计相比，最复杂模型（模型 2 和 6-8）的空间超参数后验估计偏差大得多。具体而言，在模拟研究和案例研究中，这些模型的 $\omega_1(s)$ 空间范围都大得多。可以说，对于仅基于专业调查的模型（模型 1 和 2），先验分布的选择影响更大，因为专业调查的空间覆盖范围较小。

通过模拟研究，我们注意到，所提出的框架有助于同时考虑生态过程背后的固定效应及其成因中的空间自相关。此外，该框架也已证明有助于考虑并量化影响公民科学数据和专业调查数据收集过程的因素。因此，我们的模型估计结果可用于制定信息更充分的公民科学数据抽样方案，将采样力度集中在不确定性较高或采样力度较低的区域。模拟研究还表明，了解公民科学参与者的偏好十分重要。若能更多地了解这些偏好，就可以针对公民科学参与者开展更有针对性的活动，以解决特定研究问题；同时，也能更好地将报告力度差异纳入推断。

这篇论文以挪威特伦德拉格地区输电线路致死案例研究为缘起。通过这一案例研究，我们展示了提出结合多种数据类型的方法的重要性，并说明了诸如暴露鸟类数量等因素的影响在不同模型中会有所差异。此外，生成的预测地图标示出了输电线路致死风险较高的区域。保护项目可利用这些地图，将针对性缓解措施用于风险较高的输电线路。

本文提出的方法也通过代码开放提供，使生态学和生物多样性领域的从业者能够利用这些方法，结合多种信息来源并考虑其他偏差来源，以解答更多问题；这些方法可灵活设定专业调查和公民科学数据的模型。

参考文献

August, T., Fox, R., Roy, D. B. 和 Pocock, M. J.（2020）。描述野外公民科学家行为的数据衍生指标，为项目设计和建模偏差提供了见解。*Scientific reports*，10(1)：1–12。

Bachl, F. E., Lindgren, F., Borchers, D. L. 和 Illian, J. B.（2019）。inlabru：一个基于生态调查数据进行贝叶斯空间建模的 R软件包。《生态学与进化方法》，10(6)：760–766。

Banerjee, S., Carlin, B. P. 和 Gelfand, A. E.（2015）。《空间数据的层次建模与分析》第二版。Chapman & Hall/CRC 统计学与应用概率专著。CRC Press，第2版。

Barrientos, R., Ascens˜ao, F., D'Amico, M., Grilo, C. 和 Pereira, H. M.（2021）。消失的道路：交通网络是否危及野生动物种群持续存续？*Perspectives in Ecology andConservation*，19(4)：411–416。

Barrientos, R.、Ponce, C.、Palac´ın, C.、Mart´ın, C. A.、Mart´ın, B. 和 Alonso, J. C.（2012）。导线标记可使输电线路鸟类死亡率小幅但显著下降：一项BACI设计研究。*PLOS ONE*，7(3)：e32569。

Bernardino, J.、Bevanger, K.、Barrientos, R.、Dwyer, J.、Marques, A.、Martins, R.、Shaw, J.、Silva, J. 和 Moreira, F.（2018）。鸟类与输电线路碰撞：研究现状与重点研究领域。《生物保护》，222：1–13。

Bevanger, K.（1995）。挪威高压输电线路碰撞导致的松鸡科死亡率估算及其种群影响。*Journal of Applied Ecology*，32(4)：745–753。

Bevanger, K., Bartzke, G., Brøseth, H., Dahl, E., Gjershaug, J., Hanssen, F., Jacobsen, K.-O., Kleven, O., Kvaløy, P., May, R., Me ̊as, R., Nyg ̊ard, T., Refsnæs, S., Stokke, S. 和 Thomassen, J.（2014）。输电线路的优化设计与路径规划：生态学、技术和经济视角（optipol）。最终报告：2009—2014年的研究发现。

Bevanger, K. 和 Brøseth, H.（2001）。输电线路上的鸟类碰撞——一项针对岩雷鸟属的实验。《生物保护》，99(3):341–346。

Biasotto, L. D. 和 Kindel, A.（2018）。输电线路及其对生物多样性的影响：系统综述。《环境影响评估评论》，71:110–119。

Borner, L.、Duriez, O.、Besnard, A.、Robert, A.、Carrere, V. 和 Jiguet, F.（2017）。鸟类与输电线路碰撞：估算地面搜索调查中尸体留存时间和探测率。《生态圈》，8(11):e01966。

Botella, C.、Joly, A.、Bonnet, P.、Munoz, F. 和 Monestiez, P.（2021）。利用机会性仅出现数据联合估算多个物种的空间抽样工作量和栖息地适宜性。《生态学与进化方法》，12(5):933–945。

Burden, R. L., Faires, J. D., and Burden, A. M. (2015). 数值分析。圣智学习。

Conti, J., Holtberg, P., Diefenderfer, J., LaRose, A., Turnure, J. T., and Westfall, L. (2016). 《2016年国际能源展望》，包含至2040年的预测。技术报告，美国能源部能源信息署（能源信息署），华盛顿特区（美国）……

Davis, G. (2002). 关于加利福尼亚州输电线路鸟类碰撞问题的PIER研究路线图。

Diggle, P. J., Menezes, R., and Su, T.-l. (2010). 偏好抽样下的地统计推断。《英国皇家统计学会会刊：*C*辑（应用统计）》，59(2):191–232。

Dom´ınguez delValle, J., Cervantes Peralta, F., and JaqueroArjona, M. I.(2020). 使用犬和人工搜索人员在风电场进行尸体探测的影响因素。*Journal of Applied Ecology*，57(10):1926–1935。

Drewitt, A. L. 和 Langston, R. H. (2008)。风力发电机及其他障碍物对鸟类的碰撞影响。《纽约科学院年刊》，1134(1):233–266。

Fithian, W.、Elith, J.、Hastie, T. 和 Keith, D. A. (2015)。物种分布模型中的偏差校正：整合多个物种的调查数据和采集数据。《生态学与进化方法》，6(4):424–438。

Fuglstad, G.-A., Simpson, D., Lindgren, F. 和 Rue, H.（2019）。构建惩罚高斯随机场复杂度的先验。美国统计学会会刊，114(525):445–452。

Gelfand, A. E. 和 Shirota, S.（2019）。针对存在/缺失数据的偏好抽样，以及将存在/缺失数据与仅存在数据融合。生态学专论，89(3):e01372。

Greenland, S.（1992）。生态学研究与个体层面研究中的差异性偏差。医学统计学，11(9):1209–1223。

Hern´andez-Lambra˜no, R. E., S´anchez-Agudo, J. ´A. 和 Carbonell, R.（2018）。从何处着手？开发一种空间工具，优先改造对猛禽构成危险的输电线杆。*Journal of Applied Ecology*，55(6):2685–2697。

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). 空间点模式的统计分析与建模，第70卷。约翰·威立父子公司。

Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P., and Roy, D. B. (2014年). 公民科学的统计方法：从嘈杂的生态数据中提取变化信号。《生态学与进化方法》，5(10):1052–1060。

JANSS, G. F. and FERRER, M. (2001). 西班牙杆塔设计和邻近栖息地与鸟类触电死亡率的关系。《国际鸟类保护》，11(1):3–12。

Jenkins, A. R., Smallie, J. J., and Diamond, M. (2010). 鸟类与输电线路碰撞：成因和缓解措施的全球综述，并以南非为视角。《国际鸟类保护》，20(3):263–278。

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., 和 Stone, L.（2017）。综合物种分布模型：将存在-背景数据和地点占据数据与不完全检测相结合。《生态学与进化方法》，8(4):420–430。

Lindgren, F., Rue, H., 和 Lindstr¨om, J.（2011）。高斯场与高斯马尔可夫随机场之间的明确联系：随机偏微分方程方法。英国皇家统计学会期刊：*B*辑（统计方法学），73(4):423–498。

Little, R. J. 和 Rubin, D. B.（2019）。含缺失数据的统计分析，第 793 卷。约翰·威立父子公司。

L´opez-L´opez, P.、Ferrer, M.、Madero, A.、Casado, E. 和 McGrady, M.（2011）。解决人为造成的大规模保护问题：西班牙帝雕与输电线路。*PLOS ONE*，6(3):e17196。

MacKenzie, D. I. 和 Kendall, W. L.（2002）。在相对丰度估计中应如何纳入探测概率？生态学，83(9):2387–2393。

Martin, G. 和 Shaw, J.（2010）。鸟类与输电线路碰撞：未能看清前路？《生物保护》，143(11):2695–2702。

Martin, G. R.（2011）。理解鸟类与人造物体的碰撞：一种感觉生态学方法。《朱鹭》，153(2):239–254。

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., 和 Reich, B. J. (2019). 用于估算物种分布的数据整合方法：近期进展与未来展望。《生态学与进化方法》,10(1):22–37.

Monsarrat, S., Boshoff, A. F., 和 Kerley, G. I. H. (2019). 将可达性地图作为预测历史生物多样性出现记录中抽样偏差的工具。《生态地理学》, 42(1):125–136.

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., 和 Collazo, J. A. (2017). 物种分布建模中多种数据来源的整合：数据融合框架 *。《生态学》，98(3):840–850.

Pav´on-Jord´an, D., Stokke, B. G., ˚Astr¨om, J., Bevanger, K., Hamre, Ø., Torsæter, E., and May, R. (2020年). 鸟类是否会对高压输电线路架空导线上的螺旋标记器作出反应？来自专用鸟类雷达的见解。《全球生态学与保护》, 24:e01363.

Philibert, H., Wobeser, G., and Clark, R. (1993年). 统计死亡鸟类：对相关方法的考察。《野生动物疾病杂志》, 29(2):284–289.

Ponce, C., Alonso, J. C., Argando˜na, G., Garc´ıa Fern´andez, A., 和 Carrasco, M.（2010）。食腐动物移除尸体和搜索准确性会影响输电线路鸟类死亡率估计。《动物保护》，13(6):603–612。

Roksv°ag, T., Steinsland, I., 和 Engeland, K.（2020）。在地统计学框架内利用长期空间模式和短期记录估算年径流量。《水文学与地球系统科学》，24(8):4109–4133。

Rue, H., Martino, S., 和 Chopin, N.（2009）。使用积分嵌套拉普拉斯近似对潜在高斯模型进行近似贝叶斯推断。英国皇家统计学会期刊：*B*辑（统计方法学），71(2):319–392。

Schutgens, M., Shaw, J. M., 和 Ryan, P. G.（2014年）。估算南非卡鲁地区输电线路大型鸟类碰撞死亡调查中的食腐动物偏差和搜索偏差。《鸵鸟》，85(1):39– 45。

Serrano, D., Margalida, A., P´érez-Garc´ıa, J. M., Juste, J., Traba, J., Valera, F., Carrete, M., Aihartza, J., Real, J., Ma˜nosa, S., 等（2020）。西班牙的可再生能源威胁生物多样性。《科学》，370(6522):1282–1283。

Sicacha-Parada, J., Pavon-Jordan, D., Steinsland, I., May, R., Stokke, B., 和 Øien, I. J.（2022）。用于采用不同抽样方案开展监测调查的空间建模框架：斯堪的纳维亚中部鸟类丰度案例研究。《农业、生物与环境统计学杂志》，第1–30页。

Sicacha-Parada, J., Steinsland, I., Cretois, B., 和 Borgelt, J.（2021）。在公民科学数据中校正因可达性导致的空间采样力度差异：挪威驼鹿案例研究。《空间统计学》，42:100446。

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B., 和 O'Hara, R. B.（2020）。数据越多就总是越好吗？关于综合分布模型的优势与局限的模拟研究。《生态地理学》，43(10):1413–1422。

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., 和 Rue, H.（2016）。摆脱网格：对数高斯 Cox 过程的计算高效推断。*Biometrika*，103(1):49–70。

联合国（2018）。《2018年可持续发展目标报告》。技术报告，联合国。

Wang, C.、Furrer, R.、Group, S. S. 等（2021）。利用基于过程的空间融合模型整合异质空间数据集：一个统一框架。计算统计与数据分析，161：107240。

Zipkin, E. F.、Zylstra, E. R.、Wright, A. D.、Saunders, S. P.、Finley, A. O.、Dietze, M. C.、Itter, M. S. 和 Tingley, M. W.（2021）。应对数据整合挑战，以跨尺度关联生态过程。生态学与环境前沿，19(1)：30–38。