arXiv:2305.01907v1 [stat.AP] 3 May 2023

疟疾地统计建模新计算方法的比较

Spencer Wong[1], Jennifer A. Flegg[1],Nick Golding[2],和 Sevvandi Kandanaarachchi[3]

[1]数学与统计学院，墨尔本大学，Parkville VIC 3010[2]Telethon Kids Institute，Perth Children's Hospital，15 Hospital Ave，Nedlands WA 6009[2]Curtin University，Kent St，Bentley WA 6102[3]CSIRO 的 Data61，Research Way，Clayton VIC 3168

摘要

背景 健康数据的地统计分析日益广泛地用于对疟疾患病率、疾病负担及其他指标的空间变异进行建模。传统的地统计建模推断方法以计算密集著称，因此推动了更新、更近似的方法的发展。随着所建模区域的范围和空间位置数量不断增加，更快速方法的优势尤为显著。

方法 我们对四种提出的"快速"地统计建模方法及其配套实现软件进行了应用比较：集成嵌套拉普拉斯近似（INLA）、结合高斯过程与混合效应模型的树提升法（GPBoost）、固定秩克里金法（FRK）和空间随机森林（SpRF）。我们通过在两种不同的空间尺度——国家和大陆——上估算疟疾患病率，展示这四种方法的应用。我们从准确性、计算时间和实施难易程度等方面比较了这四种方法在这些数据上的表现。

结果 随着数据规模的增加，其中两种方法——SpRF 和 GPBoost——无法良好扩展，因此可能不适用于更大规模的分析问题。然而，剩余的两种方法——INLA 和 FRK——在计算方面具有良好的扩展性，但所得的模型拟合结果对用户的建模假设和参数选择非常敏感。INLA 通常用于疾病患病率制图的二项观测分布无法处理疟疾患病率数据中存在的小尺度过度离散，这可能导致较差的预测。必须选择适当的替代方案（如贝塔-二项分布），才能获得可靠的模型拟合结果。FRK 中的小尺度随机效应项克服了这一缺陷，但 FRK 模型的估计结果高度依赖于所提供基函数的数量是否充足以及配置是否适当。遗憾的是，随着基函数分辨率的提高，FRK 的计算时间会迅速增加。

结论 INLA 和 FRK 均支持对疟疾患病率数据进行可扩展的地统计建模。然而，在使用这两种方法时，必须谨慎评估模型对数据的拟合程度以及预测的合理性，以便选择适当的模型假设和近似参数。

1 引言

空间邻近性往往在支配地理过程的传播方面发挥重要作用，包括传染病患病率或发病率的波动。因此，地统计建模是绘制传染病及其影响的流行病学地图时广泛使用的方法。这在疟疾制图领域尤为明显：研究人员借助明确考虑空间因素的统计模型，在广阔的空间范围内绘制重要的流行病学指标，以弥补数据的空间稀疏性。例如，研究者已发表利用地统计模型生成的预测地图，展示疟疾患病率（Weiss 等，2019）、死亡率（Gething 等，2016）和疟疾干预措施的使用情况（Bertozzi-Villa 等，2021）

以及抗疟药物耐药性（Flegg 等，2013、2022）。这些指标受多种空间过程影响，包括环境因素（例如降雨量和温度）、医疗保健服务可及性的差异以及人类迁移。在尚未全面了解所有这些过程的情况下，空间统计建模旨在描述感兴趣指标的空间变异，而这种变异是由潜在的空间过程所导致的。

在他们的著作中，Diggle & Ribeiro Jr. (2007) 提出了一种用于地理空间数据建模的基本范式，将以往的空间建模方法与基于模型的统计分析结合起来。感兴趣的量或响应变量 y_i 在整个连续的研究区域内定义，而样本位置$\mathbf{x}_i$ 处的每次测量都是随机变量 Y_i 的一次实现，其分布取决于位置$\mathbf{x}_i$ ，也取决于与其他数据位置相关联的随机变量。也就是说，空间中的随机变量会根据彼此的邻近程度相互依赖。因此， n 个位置处的观测响应被建模为一个联合的n维随机变量向量，其中这种依赖关系可通过将空间随机效应纳入广义线性地统计模型来建模。空间相关的随机变量被建模为高斯过程（GP），而空间协变量（如生物气候和环境图层）通常会作为额外的回归变量纳入模型，以捕捉总体趋势。

高斯过程广泛应用于时空建模，包括疟疾患病率制图研究（Hay & Snow 2006年，Gething 等，2011年，Bhatt 等，2017）。随着机器学习研究的迅猛发展，GP 的受欢迎程度在理论和应用领域都显著提高（Hensman 等，2013）。Rasmussen 和 Nickisch（2010）发布了一个名为 *GPML* 的工具箱，用于机器学习回归和分类任务。大规模回归中的 GP（Park & Apley 2018年）以及稀疏近似中的 GP（Quiñonero-Candela 和 Rasmussen 2005）都是 GP 在机器学习中应用的例子。这些新进展使 GP 成为一种可行的工具，能够对远超疟疾制图领域规模的超大型数据集进行建模（Datta 等，2016年）。

这些建模方法在推断程序（例如贝叶斯或频率学派）和计算技术（例如模拟与优化）方面各不相同。这些较新的方法有一个共同点：它们有望避免使用“完整”（无近似）高斯过程（GP）进行计算，因为完整 GP 模型的计算规模会随训练数据中不同位置的数量呈立方级增长。也就是说，数据集中不同空间位置的数量增加 10 倍，计算时间就会增加 1000 倍。因此，对于大型数据集（例如用于国家和大陆尺度疟疾制图的数据集），完整 GP 的计算可能变得不可行。对于此类大型数据集，使用马尔可夫链蒙特卡罗（MCMC）等渐近精确的贝叶斯方法拟合模型也可能代价过高，因此研究人员也探索了这类模拟方法的确定性近似。在使用全球尺度数据集时（Moraga 等人 2021，Pfeffer 等人 2018），通常会采用更新的空间随机效应近似方法和推断方法近似技术，以克服这一限制。

除了完整的贝叶斯/频率学派推断和完整的高斯过程之外，还有许多可用的近似技术，它们同样能够拟合疾病制图中使用的显式空间模型。本文篇幅有限，无法评述所有这类技术。本文针对一个疟疾患病率制图问题，对四种此类近似方法进行比较：采用高斯马尔可夫随机场近似高斯过程的集成嵌套拉普拉斯近似推断（INLA）、通过提升算法拟合的高斯过程（GPBoost）、空间随机森林（SpRF）以及固定秩克里金法（FRK）。选择这四种方法，是因为它们在空间相关结构的底层模型以及推断方法上各不相同。

本比较面向两类读者。我们希望它既能引起有意进入空间（或时空）制图领域、希望了解当前可用方法的研究人员的兴趣，也能为已经身处该领域、希望应用更快速方法的研究人员提供参考。我们首先以国家尺度进行分析，选择肯尼亚作为研究对象。随后，我们将分析扩展至非洲大陆尺度。我们在第2节简要介绍这四种方法，并说明各自采用的模型。由于这些方法在模型设定和推断程序方面存在根本的数学差异，因此很难直接比较结果。为缓解这一问题，我们采用交叉验证方案，将点预测和区间预测与观测数据进行比较。此外，我们还探讨各模型生成的预测性空间地图。我们在第3.1节和第3.2节讨论国家尺度和大陆尺度的结果，并在第3.3节比较各方法在不同尺度下所需的计算时间。作为结语，我们在第4节简要讨论分析中揭示的模型差异。本文的编程脚本

可见于 https://github.com/sevvandi/supplementary_material/tree/master/stcompare .

2 方法

Diggle & Ribeiro Jr.（2007）从以下不含任何协变量的基本地统计模型入手。他们考虑由（$\mathbf{x}_i$, y_i）给出的数据，其中 $i \in \{1, \ldots\ldots, n\}$；这里，$\mathbf{x}_i$ 表示空间位置（即坐标），而 y_i 表示该位置处所关注量的测量值（例如，$\mathbf{x}_i$处的疟疾发病率）。他们针对正态分布响应数据描述了一个采用平稳高斯过程的模型（即在整个分析区域内，过程会回归到相同的平均值），其形式为

1. $\{S(\mathbf{x})$：$\mathbf{x} \in \mathbb{R}^2\}$ 其中，$S(\mathbf{x})$ 是一个高斯过程，其均值为 μ（研究区域上的平均值），在每个位置 σ^2 = 处的过程方差或振幅为 var{$S(\mathbf{x})$}，相关函数为 $\rho(u)$ =cor{$S(\mathbf{x})$, $S(\mathbf{x}')$}，其中 $u = \|\mathbf{x} - \mathbf{x}'\|$，而 $\|\cdot\|$ 表示欧几里得距离（该距离根据响应变量之间的距离远近控制其相似性）；

2. 以及 y_i 是在 $\{S(\mathbf{x})$ 条件下相互独立的高斯随机变量 Y_i 的实现值：$\mathbf{x} \in \mathbb{R}^2\}$（即在考虑空间相关性之后，每个 y_i 都相互独立且服从正态分布）。

该模型可由以下方程描述

$$\begin{aligned} Y_i &\sim N(z_i, \tau^2), \\ z_i &= S(\mathbf{x}_i) \quad \text{for} \quad i = 1, \ldots, n\,. \end{aligned} \tag{1}$$

常见的相关函数（纳入方差项σ^2时称为协方差函数）包括 Matérn 函数、指数函数和平方指数函数。

INLA、GPBoost 和 FRK 可视为 Diggle & Ribeiro Jr.（2007）在方程（1）中描述的基本模型的不同改编版本；而 SpRF 是随机森林算法的空间版本。随机森林算法源自机器学习，采用了截然不同的方法。

我们使用以下四种方法的实现：

1. INLA：集成嵌套拉普拉斯近似，通过 R 包 INLA 实现（Rue 等，2009年；Lindgren 等人，2011年）。2. GPBoost：结合高斯过程和混合效应模型的树提升法，通过 R 包 gpboost 实现（Sigrist 2022）。3. SpRF：空间随机森林，通过 R 包 ranger 实现（Hengl 等人，2018年）。4. FRK：固定秩克里金法，通过 R 包 FRK 实现（Zammit-Mangion & Cressie，2021 年）。

我们选取的国家尺度数据集是来自肯尼亚的2009年恶性疟原虫 患病率数据，取自Malaria Atlas Project疟疾患病率数据集的开放获取部分。之所以选择肯尼亚，是因为该国的调查总数最多，且2009年开展的调查数量最多。扩展到大陆尺度后，我们使用了2009年非洲各地可获得的调查数据，在不同分析中保持年份一致。我们使用R软件包malariaAtlas（Pfeffer 等人，2018年）下载疟疾流行率调查数据。由于我们的目标是评估统计模型在疟疾制图方面的性能，而不是生成可靠的地图，因此我们没有对这些数据点进行任何额外的验证、校正或筛选。对于每条记录，我们仅提取空间坐标、接受筛查的个体数量，以及其中恶性疟原虫检测呈阳性的个体数量。

2.1 INLA

INLA（积分嵌套拉普拉斯近似）是一种近似贝叶斯推断方法，相较于马尔可夫链蒙特卡罗等渐近精确方法，它能够提高计算速度。INLA不通过模拟来估计高维联合后验分布，而是获得模型参数单变量后验边缘分布的近似值。INLA仅适用于能够表示为潜在高斯马尔可夫随机场的模型类别。不过，许多常用模型都可以采用这种形式表示，包括广义线性地统计模型。这种推断方法与将空间高斯过程近似为覆盖研究区域的离散“网格”上的高斯马尔可夫随机场（GMRF）十分契合，并可通过分段线性插值获得落在该“网格”节点之间的任意位置的值。当高斯过程具有Matérn类型的协方差函数时，可以使用GMRF的随机偏微分方程（SPDE）表示；与完整的高斯过程方法相比，这使得在大型空间数据集上对空间过程进行评估的速度大幅提升。

多年来，INLA（Rue 等，2017）不断得到更新，以拓展其适用范围并促进多样化问题求解任务。更多详情请参阅其网站 https://www.r-inla.org/home。

使用 INLA 进行推断时，需要结合一系列假设和拉普拉斯近似，以计算模型参数和潜在效应的边缘后验分布。INLA 假设响应向量 $\mathbf{y}$取决于潜变量 $\boldsymbol{\eta}$ 和超参数 $\boldsymbol{\theta}_1$，其密度为 $\pi(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\theta}_1)$。例如，潜变量可能包括线性预测子、截距、回归系数以及任意随机效应的值。重要的是，$\boldsymbol{\eta}$ 被假设为均值为$\mathbf{0}$ 的高斯马尔可夫随机场，其精度矩阵为$\mathbf{Q}(\boldsymbol{\theta}_2)$（连续空间模型中$\mathbf{Q}$ 的构造见（Bakka 等，2018）），其中 $\boldsymbol{\theta}_2$ 是超参数向量。超参数通常合并为单个向量 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$，其先验分布为 $\pi(\boldsymbol{\theta})$。随后，INLA 对边缘后验分布 $\pi(\eta_i|\mathbf{y})$ 和 $\pi(\theta_k|\mathbf{y})$ 进行如下近似。

第一步是写出超参数的联合后验分布，如下所示

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\pi(\boldsymbol{\eta}, \boldsymbol{\theta}|\mathbf{y})}{\pi(\boldsymbol{\eta}|\boldsymbol{\theta}, \mathbf{y})}, \\ &\propto \frac{\pi(\boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y})}{\pi(\boldsymbol{\eta}|\boldsymbol{\theta}, \mathbf{y})}.\end{aligned}$$

拉普拉斯近似将应用于分母，用高斯分布替代分母，从而得到近似式

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \left.\frac{\pi(\boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\boldsymbol{\eta}|\boldsymbol{\theta}, \mathbf{y})}\right|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*(\boldsymbol{\theta})}, \tag{2}$$

其中，$\boldsymbol{\eta}^*(\boldsymbol{\theta})$ 是 $\pi(\boldsymbol{\eta}|\boldsymbol{\theta}, \mathbf{y})$ 的众数，而 $\tilde{\pi}_G(\boldsymbol{\eta}|\boldsymbol{\theta}, \mathbf{y})$ 是其高斯近似。随后可得到超参数的近似后验边缘分布，如下所示

$$\tilde{\pi}(\theta_k|\mathbf{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_{-k}.$$

潜在效应的精确边缘分布

$$\pi(\eta_i|\mathbf{y}) = \int \pi(\eta_i|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta},$$

可通过数值积分近似为

$$\tilde{\pi}(\eta_i|\mathbf{y}) = \sum_{k=1}^{K} \tilde{\pi}\left(\eta_i|\boldsymbol{\theta}^{(k)}, \mathbf{y}\right) \tilde{\pi}\left(\boldsymbol{\theta}^{(k)}|\mathbf{y}\right) \Delta_k, \tag{3}$$

其中，$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ 与式（2）中的定义相同，$\tilde{\pi}(\eta_i|\boldsymbol{\theta})$ 是 $\pi(\eta_i|\boldsymbol{\theta})$ 的近似。INLA 提供了三种主要方法来计算 $\tilde{\pi}(\eta_i|\boldsymbol{\theta})$，分别称为高斯分布、拉普拉斯和简化拉普拉斯策略，此外还有自适应和自动策略。每种策略都会对不同的条件

分布应用拉普拉斯近似或级数展开，在效率和准确性方面各有取舍。关于这些方法的完整细节，可参见例如（Rue 等，2009年；Gómez-Rubio，2020年；Wang 等，2018年）。

在 INLA 中，预测与模型拟合同时进行，计算每个预测位置处响应变量的后验预测分布（Gómez-Rubio，2020年）。INLA 软件提供预测分布的汇总统计量，包括均值、中位数、标准差和分位数。

2.1.1 INLA 模型

我们构建了一个与 Kang 等人（2018年）和 Moraga（2019）类似的模型，使用 INLA 预测疟疾患病率。令 H_i 表示位置 $\mathbf{x}_i$ 在 $i = 1, \ldots, n$ 的阳性结果数量（例如，在我们的案例中为疟疾感染数量），令 p_i 表示位置 $\mathbf{x}_i$ 的建模患病率，且 $\mathbf{p}$ 为所有位置的建模患病率向量。然后，我们使用二项分布对 H_i 建模，如下所示。

$$H_i \sim \text{Binomial}\,(N_i, p_i)\,.$$

二项分布的标准链接函数是 logit 函数，它将 $[0, 1]$ 中的概率展开为 $(-\infty, \infty)$ 中的实数。因此，我们得到，

$$\text{logit}\,(p_i) = \beta_0 + S(\mathbf{x}_i)\,, \tag{4}$$

其中，β_0表示 截距，S 是服从零均值高斯过程的空间随机效应，其
具有 Matérn函数协方差

$$\text{cov}\left(S(\mathbf{x}_i), S(\mathbf{x}_j)\right) = \frac{\sigma^2}{2^{\lambda-1}\Gamma(\lambda)}\left(\kappa\|\mathbf{x}_i - \mathbf{x}_j\|\right)^{\lambda} K_{\lambda}\left(\kappa\|\mathbf{x}_i - \mathbf{x}_j\|\right)\,. \tag{5}$$

其中，λ 是平滑度参数，σ^2 表示方差，K_{λ} 是第二类修正贝塞尔函数。参数 κ 控制相关性随距离衰减的速度。

我们根据（Moraga，2019年）中提供的示例，为我们的模型确定了实现方式和参数设置。建立模型的第一步是构建一个用于求解随机偏微分方程的三角网格。该软件根据用户提供的限制条件构建此网格；通常，数据附近是由较小三角形组成的区域，外围则扩展为较粗的三角形，以避免边界效应（Lindgren 和 Rue，2015年）。使用肯尼亚数据时，我们将内部区域的最大三角形边长设为 0.5，将扩展区域的最大边长设为 4。截断参数设定了一个距离，在构建网格顶点时，距离小于该值的点会被归为一组。我们将其设为 0.01；此外，min.angle 和 offset 参数仍采用默认值，分别为 21 度和 −0.1，这两个参数分别决定三角形允许的最小角度和扩展区域的大小。使用非洲数据时，我们采用了单位球面上的网格，并将上述参数值转换为弧度。

平滑的 Matérn 协方差函数 (5) 中的光滑度参数 λ 必须通过 alpha 参数来选择。 r

$$\lambda = \alpha - \frac{d}{2}\,,$$

其中 d 是第 th e 空间的维度（即空间模型为 2）。我们将 alpha 设为其默认值。 2.

用户设置还会控制推断过程中的近似方法。我们使用了默认的自动策略来近似 $\tilde{\pi}(\eta_l|\boldsymbol{\theta}, \mathbf{y})$。随后，int.strategy 参数确定用于式（3）中数值积分的点 $\boldsymbol{\theta}_{(k)}$ 的选取方式，而我们使用了更快的经验贝叶斯策略，该策略选择单个点，即 $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ 的众数，因此不会对超参数不确定性下的预测取平均，而这通常会在马尔可夫链蒙特卡罗推断过程中发生。对于点预测，我们选择使用预测分布的中位数，不过也可以使用均值等其他量。

2.2 GPBoost

GPBoost 将树提升与高斯过程和混合效应模型相结合。其旨在充分发挥树提升算法的优势，例如处理复杂的非线性关系、不连续性以及更高的...

阶次交互作用，并兼具高斯过程的灵活性（Sigrist 2022）。它支持使用混合效应模型，尤其是包含分组随机效应的模型。

GPBoost模型的一般方程为

$$\begin{aligned} Y_i &\sim N(z_i, \tau^2) \quad \text{for} \quad i = 1, ..., n, \\ \mathbf{z} &= F(X) + Z\mathbf{S}, \\ \mathbf{S} &\sim \mathcal{N}(0, \Sigma), \end{aligned} \tag{6}$$

其中， Y_i 是位置 $\mathbf{x_i}$处的响应变量。矩阵 $X \in \mathbb{R}^{n \times p}$ 是固定效应预测变量矩阵，其中第 i行包含位置 $\mathbf{x}_i$ 处的协变量。协变量 F的固定效应函数是非线性的，并通过提升法学习。$\mathbf{S} \in \mathbb{R}^m$ 包含协方差矩阵为 $\Sigma \in \mathbb{R}^{m \times m}$的随机效应，而 $Z \in \mathbb{R}^{n \times m}$是随机效应预测变量矩阵，通常用于定义分组随机效应。

在高斯过程模型中，随机效应 $\mathbf{S} = (S(\mathbf{x}_1), S(\mathbf{x}_2), \ldots, S(\mathbf{x}_m))$ 是高斯过程 $S(\mathbf{x})$ 的有限维版本，其协方差函数为

$$\operatorname{cov}(S(\mathbf{x}), S(\mathbf{x}')) = c(\mathbf{x}, \mathbf{x}'), \ \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d .$$

这里， c 是一个协方差函数，通常参数化为

$$c(\mathbf{x}, \mathbf{x}') = \sigma_1^2 r(\|\mathbf{x} - \mathbf{x}'\|/\rho) ,$$

其中， r 是各向同性自相关函数， $\sigma_1^2 = \operatorname{var}(S(\mathbf{x}))$， ρ 是范围参数，用于确定 r 随距离衰减的速度。GPBoost目前支持指数、高斯分布、Matérn、幂指数、Wendland和截尾指数协方差函数。在高斯过程模型中，Z 通常编码为对角矩阵，因此$\mathbf{S}$ 中的每个元素都包含该位置的空间随机效应。

在默认设置下，GPBoost 不会对高斯过程应用近似方法。为提高效率，软件提供了 Vecchia 近似，该近似假设基于距离的响应之间具有条件独立性，从而在计算过程中生成稀疏矩阵（Sigrist 2022）。

GPBoost 的推断通过联合优化非线性固定效应函数 F 以及方差和协方差参数 $\boldsymbol{\theta}$ （即 τ^2、 σ_1^2 和 ρ）来实现。在高斯过程情形下，优化的目标是最小化 风险泛函

$$R(F, \boldsymbol{\theta}) = L(\mathbf{y}, F(X), \boldsymbol{\theta}) ,$$

其中， $\mathbf{y} = (y_1, ..., y_n)$ 是位置 $\mathbf{x}_1, ..., \mathbf{x}_n$ 处的观测响应。这里， $L(\mathbf{y}, F(X), \boldsymbol{\theta})$ 是在给定观测协变量矩阵 X 和模型参数 $\boldsymbol{\theta}$ 的情况下获得观测响应 $\mathbf{y}$ 的负对数边际似然，

$$L(\mathbf{y}, F(X), \boldsymbol{\theta}) = \frac{1}{2}(\mathbf{y} - F(X))^T \Psi^{-1}(\mathbf{y} - F(X)) + \frac{1}{2}\log\det(\Psi) + \frac{n}{2}\log(2\pi) ,$$

其中， $\Psi = Z\Sigma Z^T + \tau^2 I$。风险泛函通过迭代更新 F 和 $\boldsymbol{\theta}$ 来最小化。在第 k 步，固定 F_{k-1} ，并使用梯度法或拟牛顿法计算 $\boldsymbol{\theta}_k = \operatorname{argmin}_{\theta}(L(y, F_{k-1}(X), \boldsymbol{\theta}))$。取该 $\boldsymbol{\theta}_k$ 值后，通过提升算法执行一步更新 F 。

优化后，GPBoost 以类似于高斯过程回归的方式进行预测。首先构建观测响应和预测响应的联合分布，并以观测响应为条件。然后，将所得条件分布的均值作为响应变量的预测值。

由于我们的模型中未使用协变量，因此树提升仅用于求得截距。虽然这确实忽略了 GPBoost 学习协变量非线性函数的功能，但我们仍将 GPBoost 纳入分析，以便希望在可能受益于树提升的更复杂情境中应用该方法的用户使用。

2.2.1 GPBoost模型

GPBoost 支持响应变量服从高斯分布、伯努利-概率单位、伯努利-对数几率、泊松分布和伽马分布，但与 INLA 不同，目前不支持二项响应。因此，我们通过对方程（6）进行定制，将疟疾患病率建模如下:

$$H_i/N_i \sim N(z_i, \tau^2),$$
$$z_i = \beta_0 + S(\mathbf{x}_i),$$

其中，β_0 为截距，H_i 和 N_i 分别表示位置 $\mathbf{x}_i$处的阳性结果数量和接受检测的人数。需要注意的是，为了简化处理，我们选择使用阳性检测的直接比例，而非经验Logit，并将预测值限制在 $[0, 1]$ 范围内，以绘制患病率地图。我们使用指数协方差函数 $r(\|\mathbf{x} - \mathbf{x}'\|/\rho) = \exp(-\|\mathbf{x} - \mathbf{x}'\|/\rho)$，这是该软件中的默认选择。需要指出的是，该模型并未充分利用 GPBoost 学习协变量非线性函数的全部能力，但其构建方式与我们不在任何模型中使用协变量的选择保持一致。

我们的模型参数设置遵循软件包作者的示例（Sigrist 2020）。对于模型中的空间随机效应，我们使用完整的高斯过程，不进行近似，并将 vecchia_approx 参数设为 FALSE。软件中的其他参数用于控制学习固定效应函数 F 时使用的树和提升算法。我们将提升轮数设为 247，学习率设为 0.01，使用的参数为 nrounds 和 learning_rate。模型的其他设置包括 num_leaves $= 1024$、max_depth $= 6$ 和 min_data_in_leaf $= 5$，这些参数分别用于控制树的大小。

2.3 SpRF

空间随机森林（SpRF）（Hengl 等人 2018年）通过将观测点之间的距离用作解释变量，将经典随机森林扩展到空间域。也就是说，使用 SpRF 拟合模型时，对于每个点 $\mathbf{x}_i$（其中给定 y_i），所用的协变量表示该点与其他每个观测点之间的距离。因此，模型这部分的设计矩阵就是所有观测位置两两之间的距离矩阵。为获得不确定性估计，SpRF 的作者使用分位数回归森林来估计条件分布 $Y_i|X_i$（Meinshausen 2006年）的指定分位数，其中 X_i 是第 i个响应变量的协变量；相比之下，经典随机森林不提供不确定性估计。

SpRF 模型的一般方程为

$$Y_i = f\left(X_{G_i}, X_{R_i}, X_{P_i}\right),$$

其中，Y_i 表示位置 $\mathbf{x}_i$处的响应变量，X_{G_i} 表示查询点 $\mathbf{x}_i$ 到各观测位置的距离向量（包括该点到自身的距离 0，该距离位于向量的第 i个位置）；X_{R_i} 和 X_{P_i}表示两类协变量——地表反射率和基于过程的协变量。函数 f由随机森林学习。与我们讨论的其他方法不同，SpRF 不使用协方差函数。

SpRF 基于用于随机森林的 ranger 软件包，该软件包实现了分位数回归森林，其训练过程由 Meinshausen（2006年）提出。点预测由分位数回归森林估计的中位数给出。

2.3.1 SpRF 模型

与（Hengl 等人，2018年）一样，我们对响应变量增加了正态假设，以构建 simpleSpRF 模型

$$H_i/N_i \sim N(z_i, \tau^2),$$
$$z_i = f(X_{G_i}) \quad \text{for} \quad i = 1, ..., n,$$

其中，H_i 和 N_i 如上文所定义，X_{G_i} 包含从各观测点到$\mathbf{x}_i$的距离。

SpRF 的用户参数决定了随机森林的结构以及每棵树的生长规则，包括通过 num.trees 和

mtry 参数设置树的数量，以及每个节点用于分裂的变量数量。我们将每个参数都保留为默认值，因此得到一片包含 500 棵树的森林,¯其中每个节点都使用 $\sqrt{n_v}$ 个变量进行分裂（n_v 是输入随机森林的变量总数）。其他用于进一步调整树和森林结构的参数也保留为默认值；我们的 SpRF 代码基于该方法作者编写的一份教程（Hengl 等，2021 年）。

2.4 FRK

固定秩克里金法（FRK）（Zammit-Mangion 与 Cressie，2021 年）是一种专为大型数据集构建的时空建模框架。它采用空间随机效应（SRE）模型，通过空间基函数的线性组合，将具有空间相关性的零均值随机过程进行分解。利用数量相对较少的基函数进行降维，确保了 FRK 的计算效率。空间域D 被划分为 M 个子集， $A_1, ..., A_M$，称为基本面积单元（BAU），其质心为 $\mathbf{x}_1, ..., \mathbf{x}_M$。SRE 模型基于这些 BAU 构建，BAU 决定模型的粒度；并假设该过程在各 BAU 上分段连续。

具有高斯响应的 FRK 一般方程可写为

$$
\begin{aligned}
Y_i &\sim N(z_i, \tau^2) \quad \text{for} \quad i = 1, ..., n, \\
\mathbf{z} &= C_Z \boldsymbol{\zeta}, \\
\zeta_j &= \mathbf{t}(\mathbf{x}_j)^T \boldsymbol{\beta} + v(\mathbf{x}_j) + \xi(\mathbf{x}_j) \quad \text{for} \quad j = 1, ..., M.
\end{aligned}
$$

这里， $Y_i, i = 1, ..., n$ 是观测位置处的响应变量， $\boldsymbol{\zeta} = (\zeta_1, ..., \zeta_M)^T$ 是潜在空间过程在各 BAU 处的取值，其质心为 $\mathbf{x}_1, ..., \mathbf{x}_M$； C_Z 是一个 $n \times M$ 矩阵，用于关联观测位置与 BAU 位置。向量 $\mathbf{t}(\mathbf{x}_j)$ 是 BAU j处协变量的集合， $\boldsymbol{\beta}$是回归系数向量；而 $v(\mathbf{x}_j)$ 是小尺度、空间相关的随机效应。最后， $\xi(\mathbf{x}_j)$ 是细尺度随机效应，假设其在各 BAU 之间不相关（Zammit-Mangion 与 Cressie，2021 年）。

FRK 通过将观测分布替换为指数族中的一个成员，并使用链接函数将潜在过程转换为均值过程，从而将非高斯数据引入模型。此类模型的一般结构为

$$
\begin{aligned}
Y_i | \mu_i, \psi &\sim \text{EF}(\mu_i, \psi) \quad \text{for } i = 1, ..., n, \\
\boldsymbol{\mu} &= C_Z \boldsymbol{\mu}', \\
g(\boldsymbol{\mu}') &= \boldsymbol{\zeta}, \\
\zeta_j &= \mathbf{t}(\mathbf{x}_j)^T \boldsymbol{\beta} + v(\mathbf{x}_j) + \xi(\mathbf{x_j}) \quad \text{for } j = 1, ..., M,
\end{aligned}
\tag{7}
$$

其中， ψ 是指数族 EF 中上下文相关成员的离散参数， $\boldsymbol{\mu}$ 称为均值过程， $g(\cdot)$为链接函数。我们用μ表示观测位置处的均值过程，而 μ' 表示 BAU 处的均值过程。

空间相关随机效应 $v(\mathbf{x})$ 被分解为

$$
v(\mathbf{x}) = \sum_{l=1}^{r} \phi_l(\mathbf{x}) \eta_l,
$$

其中， ϕ_1、.....、 ϕ_r 是在空间域上定义的一组固定基函数， $\boldsymbol{\eta} = (\eta_1$、.....、 $\eta_r)^T$ 是一个协方差矩阵为$\mathbf{K}$的 r 维高斯随机变量。为了估计模型参数，包括系数 α、细尺度随机效应的方差参数 ξ 以及协方差矩阵$\mathbf{K}$的协方差参数，FRK 采用最大似然估计。对于非高斯数据，使用拉普拉斯近似来近似边际似然，然后通过拟牛顿法对其进行最大化。

默认情况下，FRK 会为每个 BAU 预测均值过程 $\mu(\cdot)$。预测值和不确定性通过蒙特卡洛抽样方法生成，而每个 BAU 中 μ 的预测值取样本的平均值。

2.4.1 FRK 模型

与 INLA 一样，我们使用二项分布对阳性检测数量 H_i 进行建模

$$H_i \sim \text{Binomial}(N_i, p_i),$$
$$\mathbf{p} = C_Z \mathbf{p}',$$

其中，p_i 表示第 i 个观测位置的患病率。向量 $\mathbf{p}'$ 表示 BAU 上的患病率，并通过 C_Z 矩阵转换为观测位置上的患病率；该矩阵的构造详见（Sainsbury-Dale 等人，2021 年）。随后在方程（7）中使用 logit 函数作为链接函数 g，即

$$\text{logit}(p'_j) = \zeta_j .$$

由于我们没有使用任何协变量，BAU 上的潜在过程 ζ_j 可写为

$$\zeta_j = \beta_0 + v(\mathbf{x}_j) + \xi(\mathbf{x}_j),$$

其中，β_0 表示截距。

我们的模型使用两个不同尺度的高斯基函数对空间随机效应 v（$\mathbf{x}$）进行分解，这些基函数按规律布设在空间域内，其布设方式分别由 type、nres 和 regular 参数控制。这些基函数的空间尺度由 regular 参数和 scale_aperture 共同决定；我们将它们保留为默认值，分别为 1 和 1.25。随机系数 $\boldsymbol{\eta}$ 的假设相关结构由 K_type 参数控制。使用非高斯模型时，该参数默认取值为 precision，基于 Leroux 模型使用精度矩阵$\mathbf{Q}$ 对系数间的依赖关系进行建模（Sainsbury-Dale 等人，2021 年）。进行预测时，用户可以指定要抽取的蒙特卡洛样本数，我们保留了默认值 400。我们的 FRK 模型所用的代码和参数选择基于软件包作者提供的示例（Zammit-Mangion & Cressie，2021 年；Sainsbury-Dale 等人，2021 年）。

2.5 国家尺度分析的方法

在国家尺度上，使用其预测地图对四个模型进行了定性比较，同时使用交叉验证比较其预测性能。为了生成患病率预测地图，我们使用来自malariaAtlas R 包的肯尼亚2009年所有可用的 恶性疟原虫患病率调查对每个模型进行拟合。这包括分布于全国各地的382项调查，其位置如图 1(a)所示。拟合后的模型在覆盖肯尼亚的网格上生成了患病率点估计和不确定性，每个网格单元在经度和纬度方向上均覆盖名义上的0.1度（在赤道处约为11x11 km）。

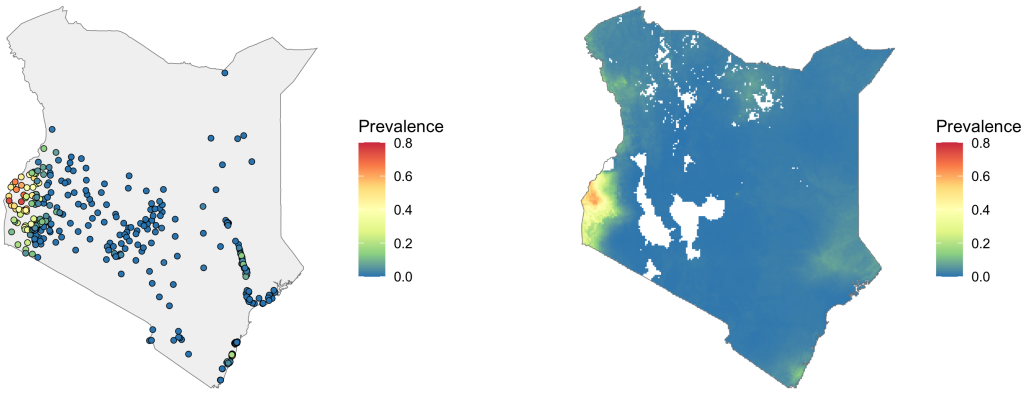
模型拟合和预测是在一台 2014 年款 MacBook Pro 上运行的，该设备配备双核 2.8GHz Intel Core i5 处理器，运行 macOS 10.13.6。我们使用单线程运行每个模型，以获得基准性能对比，同时将重点放在模型预测上。我们指出，每个模型均提供并行化选项，这些选项可能有助于提升性能。记录的运行时间是执行模型 R 脚本的总耗时，包括拟合和预测两个阶段。

为了评估这些模型，我们使用 10 折和 50 折进行空间分块交叉验证（CV）（Roberts 等人，2017）。在空间情境下，随机将点分配到交叉验证折中并不有效，因为相邻的点可以充当代理。我们根据患病率调查的空间坐标，使用 k-均值聚类（Likas 等人，2003）来选取各折，从而形成一系列由空间相邻点组成的“区块”。图 2 显示了两组交叉验证折中各点的位置，每种颜色代表一折。10 折和 50 折交叉验证衡量的是这些方法不同方面的能力。50 折交叉验证量化短距离插值能力，而 10 折交叉验证量化跨较长距离进行插值的能力。

使用10折和50折交叉验证，我们研究以下内容:

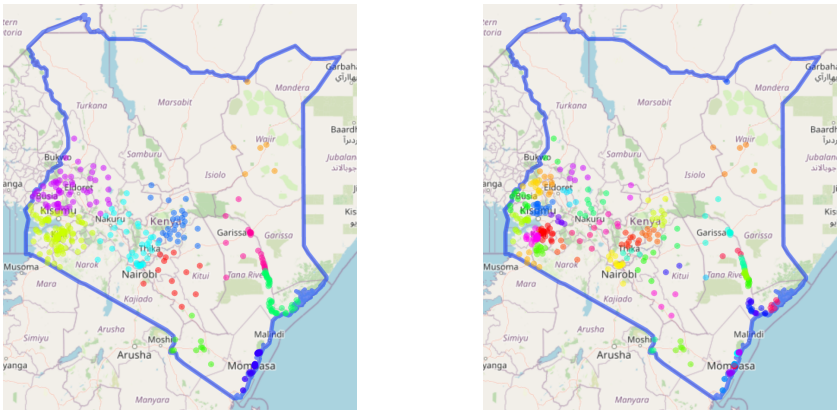
1. 点预测分析，包括使用多种指标对预测结果与样本外患病率值进行比较，

2. 对各模型的不确定性界限进行分析，以及



(a) (b)

图 1：2009年肯尼亚恶性疟原虫 患病率数据。(a) 显示患病率调查结果，(b) 显示疟疾地图项目预测的患病率。



(a) 10折交叉验证位置 (b) 50折交叉验证位置

图 2：2009年肯尼亚恶性疟原虫患病率调查位置。颜色表示 10折交叉验证（a）和50折交叉验证（b）中不同的交叉验证折。

3. 分析预测与抽样位置密度之间的关系（附录A）。

由于各模型输出的不确定性度量各不相同，不确定性分析较为复杂。INLA 包含有关拟合模型后验边际密度摘要的信息，并可计算预测的标准差和不同分位数。GPBoost 提供每个预测的方差。在线性高斯情形下，FRK 会预测每个预测值的标准差。对于非高斯情形，它通过蒙特卡罗方法提供预测（Sainsbury-Dale 等人，2021）。SpRF 使用分位数回归，且可在 ranger 软件包中指定分位数。为了将 SpRF 与其他方法进行比较，我们假设响应变量服从正态分布，与（Hengl 等人，2018年）中的做法相同，并估计 SpRF 预测的标准差，如下所示。

$$\mathrm{SD} \approx \mathrm{IQR}/1.34898\,.$$

Hengl 等人（2018年）指出，这一假设并不总是成立，因此我们只能粗略地将 SpRF 模型的不确定性与其他三个模型进行比较。

对于每个模型，我们测量有多少观测到的患病率值落在预测的不确定性区间内。令 $\hat{y}_i$ 表示观测 y_i 的预测响应均值。我们定义

$$\text{Within 1SD}(\hat{y}_i) = \text{TRUE if } |y_i - \hat{y}_i| \leq \text{SD}(\hat{y}_i)$$
$$\text{Within 2SD}(\hat{y}_i) = \text{TRUE if Within 1SD}(\hat{y}_i) = \text{FALSE and } |y_i - \hat{y}_i| \leq 2\text{SD}(\hat{y}_i)$$

其中 SD 表示标准差。由于患病率值介于 0 和 1 之间，若上下界超出此范围，我们便对其进行截断。请注意，$\hat{y}_i$ 对应于 GPBoost 和 FRK 模型的患病率预测，但不对应于 INLA 和 SpRF 模型，因为这两个模型使用中位数进行预测。

2.6 大陆尺度的方法

在大陆尺度上，我们重点关注预测地图，分别使用非洲的三组患病率数据拟合各个模型。第一组数据包含 868 项 2009 年的恶性疟原虫患病率调查，可通过 malariaAtlas R 包获取。图 3(a) 展示了这些数据，调查点集中在肯尼亚和索马里。我们还使用两类模拟数据拟合各个模型，以便将预测结果与已知真实值进行比较，并比较模型在插值和外推任务上的表现；最后，还评估空间稀疏性和噪声等数据属性如何影响模型预测。

模拟数据使用 Malaria Atlas Project (MAP) 制作的 2009 年恶性疟原虫患病率栅格生成，如图 3(d) 所示（Weiss 等，2019）。我们在 2009 年调查点对应的位置从栅格中抽取患病率值，并结合各位置的检测人数，生成阳性人数的二项样本。在 868 个观测位置中，有 28 个点位于患病率栅格的空缺处，因此将其排除；第二个数据集最终包含 840 个点，如图 3(b) 所示。这组空间聚集的模拟数据可用于评估各模型在数据很少或没有数据的区域进行外推的能力。尽管该数据集与观测数据使用相同的位置，其患病率所含噪声明显较少。

第二组模拟数据是在 MAP栅格上随机选取 1000 个点生成的，以便比较模型在空间覆盖良好的数据上训练时的插值性能。在每个位置生成阳性检测数量的二项样本，其中受检人数设为 85，约为 2009年调查中的平均人数。该模拟数据集中的患病率如图 3(c) 所示。

我们采用了与国家尺度分析中相同的参数设置，不过鉴于数据具有更大的空间范围，在可能的情况下，我们使用了能够计算点之间适当球面距离的设置。对于使用大圆距离的 SpRF，以及允许在单位球面上构建网格的 INLA，这是可行的。在撰写本文时，GPBoost 尚不具备此功能，不过与软件包作者的交流表明，他们希望未来能够添加此功能。虽然 FRK 确实支持在某些模型中使用大圆距离，但对于具有二项响应的模型，目前对该功能的支持并不完善，并且在我们的测试中无法运行。因此，对于 GPBoost 和 FRK，我们均使用了坐标之间的欧几里得距离。

使用 INLA 和球面几何时，需要在球面的一个子集上构建网格。尽管文献中采用了多种构建此网格的方法（Lindgren 和 Rue，2015年；Bakka 等，2018；Humphreys 等，2017），但这些方法得出的结果相似，因此我们采用了 Lindgren 和 Rue（2015年）所述的方法。

模型拟合和预测在墨尔本大学高性能计算集群 Spartan 的 Physical 分区上，使用一台 3.00GHz Intel(R) Xeon(R) Gold 6154 CPU 完成；每次模型运行分配了 32 GB 随机存取存储器。与国家尺度数据一样，每个模型均使用单个线程运行。预测采用网格生成，网格单元边长为 0.15 度（在赤道处约为 16.7 km）。

3 结果

本节介绍对肯尼亚的分析，包括预测地图和交叉验证结果；还包括大陆尺度分析，其中包含基于上述三个不同输入数据集训练的模型。

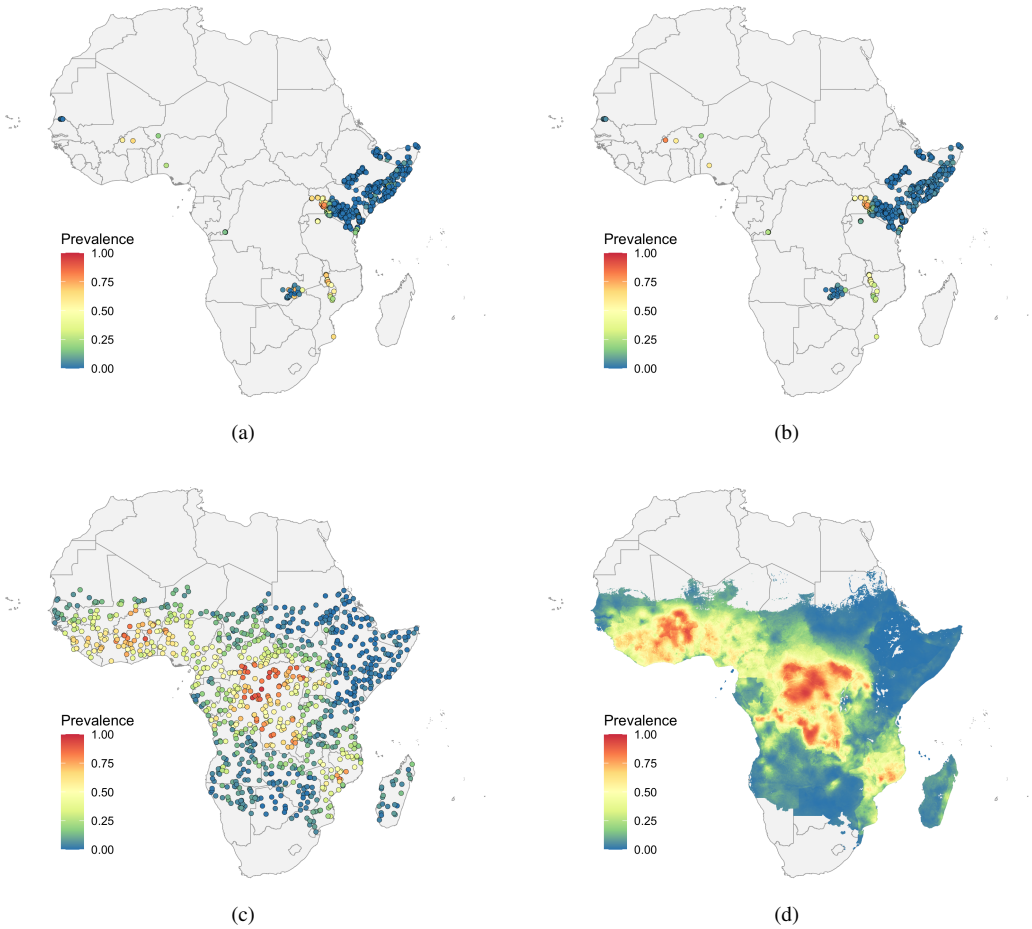


图 3：用于在大陆尺度拟合四个模型的恶性疟原虫 患病率数据。3(a) 展示了 2009年在 868 个位置观测到的数据。3(b) 展示了根据观测位置的二项抽样生成的患病率。3(c) 展示了根据 1000 个均匀随机位置的二项抽样生成的患病率。3(d) 是疟疾地图项目基于 2009年预测的患病率栅格，用于生成 3(b) 和 3(c) 中的样本。

3.1 案例研究：肯尼亚

在国家尺度上，我们采用两种方法验证模型：1. 预测地图；2. 10折和50折交叉验证预测。生成地图时，我们使用了全部数据；而进行交叉验证时，我们在每一折中都留出了一部分数据。

3.1.1 预测地图

图 4 展示了四种模型在使用 2009年 肯尼亚患病率数据进行训练后得出的预测和不确定性。从最宏观的尺度来看，各模型的预测结果相似：肯尼亚西部患病率较高，东部有若干患病率较高的簇，而该国大部分其他地区的患病率较低。对于每种模型，远离数据的区域患病率预测值都很快降至零，这表明其空间范围可能比预期更小。这一点在 INLA 模型中尤为明显，或许表明数据存在过度离散。

一个显著特征是，SpRF 在图 3(c) 中的预测结果显示肯尼亚西北部有一条弧形的高患病率带，这一点将在第 3.2 节进一步讨论。GPBoost 的预测结果也在一定程度上显示该地区患病率较高，FRK 的预测结果则略有体现。这片预测患病率较高的区域位于一片广阔的无患病率数据区域内，因此体现了四种模型在外推方面的不同方法。

3.1.2 交叉验证结果

表1：四种模型的交叉验证结果，各类别中表现最佳的结果以粗体显示。

模型	10折-RMSE	50折RMSE	训练相关性-tion	10折相关性-tion	50折相关性-tion	% 点 在1个标准差范围内（10折）
INLA	0.181	0.124	0.909	0.235	0.683	75
GPBoost	0.127	**0.11**	0.873	0.646	**0.751**	84.211
SpRF	0.132	0.121	**0.912**	0.641	0.702	37.105
FRK	**0.125**	0.123	0.902	**0.661**	0.702	83.421

表1给出了交叉验证结果。就交叉验证均方根误差和相关性而言，FRK在10折交叉验证中表现最佳，而GPBoost在50折交叉验证中表现最佳。SpRF的预测与用于训练模型的数据相关性最高，但与样本外数据的相关性较低，这表明该模型可能对训练数据过拟合。就10折均方根误差和相关性而言，INLA表现较差。

从表1可以看出，在10折交叉验证中，SpRF仅有37.105%的点位于1个标准差范围内，远低于其他模型；我们将在附录A中进一步讨论这一点。就位于1个标准差范围内的点所占比例而言，GPBoost表现最佳。不过，这些结果需要结合实际情况来看，因为较大的标准差会提高这一比例。

图5展示了四种方法在10折交叉验证中的区间预测和点预测。绿色表示位于1个标准差范围内的点，蓝色表示位于2个标准差范围内的点，其余点以红色表示。我们发现，INLA的许多预测值都接近于零，并将在附录B中探讨这一问题。Cressie（2015）详细介绍了其他交叉验证指标，这些指标会考虑预测误差除以其标准差后的结果。不过，我们不考虑这些指标，因为某些方法得出的标准差非常小，因此会产生非常大的数值。附录A进一步介绍了按簇和位置密度划分的交叉验证结果。

3.2 大陆尺度结果

地统计制图通常在大陆或全球尺度上进行，并且经常使用包含大量观测值的数据集。由于推断和预测的计算时间可能会随着数据量和区域范围的增大而表现出较差的扩展性，因此，评估近期方法的性能十分重要，这既包括预测能力，也包括计算时间。为考察四种方法在更大尺度上的表现，并了解潜在的模型假设违反情况、观测点的聚类和稀疏性如何影响预测，我们将研究区域扩展至2009年的整个非洲大陆，并使用图3所示的三个患病率数据集分别拟合每个模型。图6展示了模型在每个输入数据集上训练后得到的患病率预测结果。相应的不确定性见附录C。

扩展到大陆尺度后，模型之间的差异变得明显，而这些差异在较小尺度上并不明显。尽管图4中的国家尺度患病率地图总体上相似，但存在轻微的带状差异

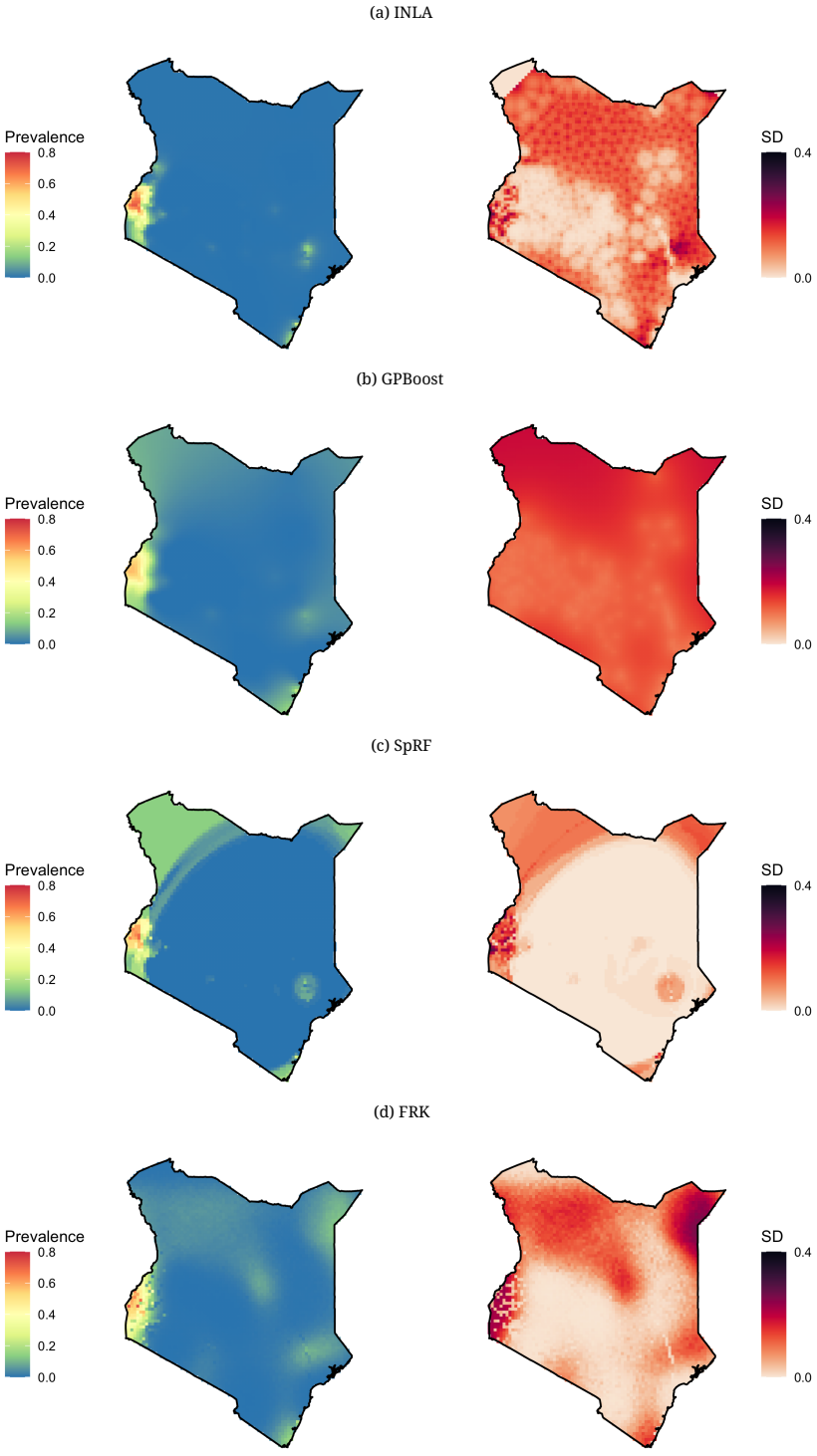
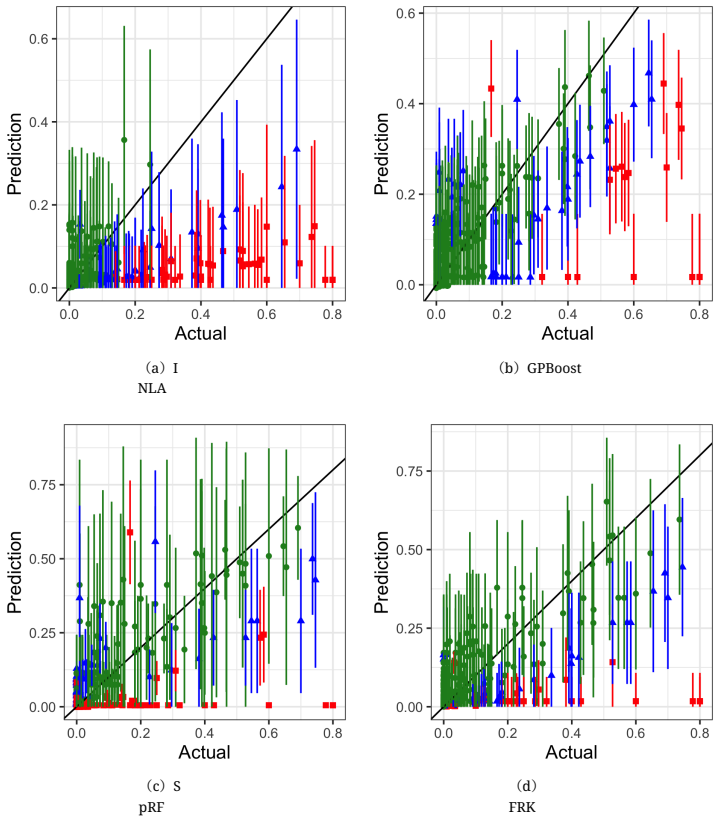


图 4：使用 2009年肯尼亚的恶性疟原虫患病率数据训练四种模型后得到的预测患病率和不确定性。请注意，这些地图仅用于说明在小样本数据上拟合时模型预测的差异，不太可能准确反映该年度全国范围内的疟疾患病率。



(a) INLA (b) GPBoost (c) SpRF (d) FRK

图5：肯尼亚国家层面数据10折交叉验证的区间预测。点表示每个模型的预测均值，区间表示均值上下一个标准差的范围。

由于SpRF的预测中出现了这种影响，图6中的患病率地图存在显著差异，其中几幅地图出现了伪影。

总体而言，与在大范围无数据区域进行外推相比，这四种模型更擅长局部插值。使用随机分布数据生成的图6(iii)中的预测，比使用较稀疏的非均匀数据生成的图6(ii)中的预测，更忠实地还原了图3(d)中MAP栅格的患病率结构。这种表现符合预期，因为众所周知，疟疾患病率具有高度异质性，而且我们的模型未使用协变量数据。

SpRF的预测呈现出明显的条带效应，在国家和大陆尺度地图中均可见：高患病率的连续弧形带同时出现在点估计和不确定性估计中。这或许是因为SpRF根据与已知数值点之间的距离来建模研究对象——在本例中即疟疾患病率。因此，我们会观察到距观测值簇不同半径处出现高或低患病率带，而随机森林的分段常数特性会使各条带之间形成明显的阶跃。图6(ci)和图6(cii)中的条带效应尤为明显，因为这些图中的点被聚集在较小区域内；而当数据点具有良好的空间覆盖时，这种效应则不太明显，例如图6(ciii)中没有横跨整片大陆的条带。研究发现，进一步增加模拟点的数量能够进一步减弱这些条带的显著程度（结果未示出）。

尽管SpRF生成的地图具有这一不理想的特征，但其交叉验证点估计相当准确。从附录A中的表3可以看出，在10折交叉验证中，SpRF的绝对误差小于0.05和0.1的点所占比例最高，而与50折交叉验证相比，10折交叉验证对算法而言是一项更具挑战性的任务。因此，我们可以认为，尽管SpRF可能会产生……，但它在各个点上的预测结果仍然是可靠的。

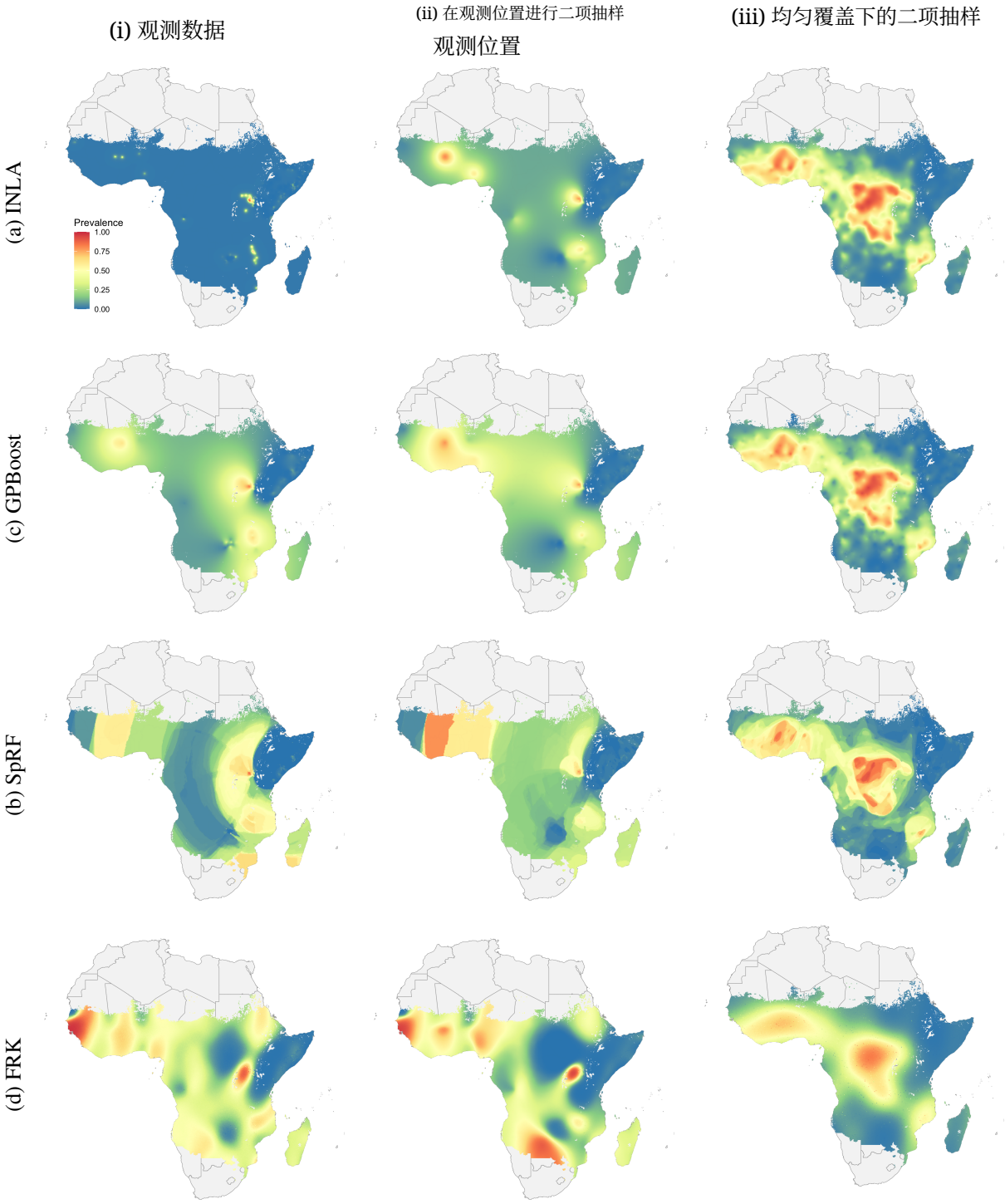


图6：使用三个不同数据集拟合模型时的恶性疟原虫 患病率预测。第(i)列中的模型使用图 3(a)所示的2009年非洲调查数据进行拟合。第(ii)列中的模型使用从疟疾地图集患病率栅格中抽取的二项样本进行拟合，抽样位置与图 3(b)所示的调查位置相同。第(iii)列中的模型使用从栅格中抽取的二项样本进行拟合，抽样位置为整个大陆范围内均匀选取的1000个位置，如图 3(c)所示。输出结果已使用图 3(d)中的疟疾地图项目栅格进行掩膜处理。16

在没有样本点的区域，预测地图可能会产生误导。

图6(ai)由 INLA 利用观测数据生成，显示出患病率在远离观测值的位置时骤然下降，导致覆盖大陆大部分区域的预测值均呈现接近于零的平坦分布。这似乎是数据中稀疏性和噪声共同作用的结果，而不仅仅是数据的聚集特性所致。图6(aii)使用的位置几乎完全相同，但显示出更高的患病率值，并且这些值从观测值向外扩散的范围更广。在附录B中，我们概述了相关证据，表明这种效应源于观测数据中未被考虑的过度离散。具体而言，数据中噪声的增加似乎会降低空间随机效应的估计范围，导致模型在远离观测位置时恢复为恒定预测值。这一行为与 INLA 在第3.1.2节的10折交叉验证分析中的较差表现一致，在该分析中，模型对每个留出折的疟疾患病率均预测为接近于零。

我们观察到，FRK 的预测结果高度依赖于基函数的排列方式，而这些基函数通常由软件根据数据位置以及第2.4.1节中引入的用户参数进行布置。例如，图6(di)和(dii)中的 FRK 预测结果在数据稀少或完全没有数据的区域呈现出伪振荡现象，但这些振荡实际上与基函数的周期性布置相对应。我们测试的其他排列方式则导致整个大陆范围内的预测结果均呈现平坦状态（未展示结果）。在图6(diii)中，由于输入数据具有良好的空间覆盖，因此并未出现这类伪影。然而，该地图看起来只是输入数据经过平滑处理后的版本，无法解析 MAP 表面中更精细的结构。基函数的排列方式及其数量对预测地图的影响将在第3.4.2节中进一步详细讨论。

对于肯尼亚和非洲，GPBoost 生成的患病率地图没有出现其他模型输出中的伪影。然而，附录C中的图 20(ci)-(cii)所示的不确定性地图显示，总体不确定性水平较高，无论各区域的调查点数量多少都是如此。附录A中的图 16也进一步证实了这一点：区间宽度几乎保持不变，极少随调查点的密度而波动。尽管 GPBoost 目前仅计算坐标之间的欧几里得距离，但非洲和肯尼亚的患病率地图看起来都合理。不过，对于全球模型而言，使用经度和纬度坐标之间的欧几里得距离并非最优选择。

3.3 计算结果

表2展示了在各个数据集上运行每个模型所需的时间。虽然FRK始终是运行速度最快的模型，但INLA在非洲各数据集上的运行时间差异很大，从使用均匀模拟数据时的不到10分钟，到使用观测数据时的69.11分钟不等。附录B对INLA的这一时间差异进行了进一步分析。

表2：在每个数据集上训练模型并生成预测地图所需的时间（分钟）。肯尼亚：观测值列对应图4中的地图。非洲：观测值、非洲：模拟观测和非洲：模拟均匀数据集 列分别对应图6的第 (i)、(ii) 和 (iii) 列。请注意，用于运行肯尼亚和非洲数据集模型的计算机不同。

		数据集			
		肯尼亚：观测	非洲：观测	非洲：模拟观测	非洲：模拟均匀
模型	INLA	0.34	69.11	11.49	7.05
	GPBoost	0.99	11.27813	6.56	13.85
	SpRF	0.44	24.54	24.64	27.60
	FRK	0.35	3.28	3.41	3.09

每个模型还额外使用了随机选取的、包含1000至10000个点的模拟患病率数据集进行训练。图7展示了随着数据集规模的变化，拟合各个模型并生成预测所需的时间。两者

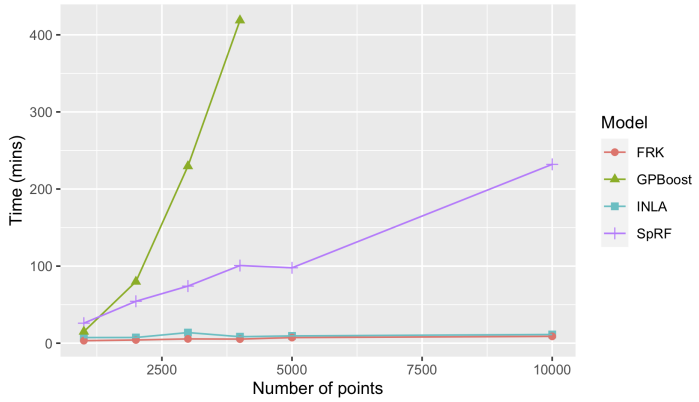


图7：各模型在均匀分布的模拟数据集上所用的时间。由于计算时间可能很长，GPBoost 未在 5000 或 10000 个点的数据上运行。

INLA 和 FRK 在较大的数据集上仍然非常快速，所需时间几乎没有变化。相比之下，SpRF 的运行时间似乎随数据集大小线性增长，而 GPBoost 在较大的数据集上速度迅速下降，这反映了使用未近似高斯过程所需的计算资源。正如第 2.2 节所述，GPBoost 提供了 Vecchia近似，可提高计算效率。附录 D 中，我们将考察应用此近似对计算和模型预测的影响。

3.4 FRK 和 INLA 对参数选择的敏感性

在这四种方法中，INLA 和 FRK 的计算效率表现出良好的潜力；如图7所示，与 SpRF 和 GPBoost 相比，它们具有更优的扩展性。此外，虽然 SpRF 输出中的伪影似乎源于其将距离作为输入的方式，但 INLA 和 FRK 的患病率地图中的伪影究竟是由特定模型参数造成的，还是根本上由各自方法所采用的近似引起的，则不太明确。因此，我们将更仔细地考察这两种方法，并测试其预测对模型参数的敏感性。

3.4.1 INLA

INLA 预测地图中最明显的伪影，是模型拟合观测数据时出现的平坦、接近于零的预测值，如图6(ai)所示。附录B列举的证据表明，这一特征是由过度离散造成的；这意味着二项响应变量不适合用于模拟观测到的疟疾数据的变异性，尽管在有关将 INLA 应用于疾病制图问题的教程中，二项响应变量很常见。

可以对模型进行多种调整来处理这种过度离散，例如采用贝塔-二项分布或高斯响应（直接作用于阳性比例，或作用于其经验Logit变换），或者在线性预测子中加入一个独立误差项。所有这些选项都会在模型中增加一个参数，用于捕捉观测层面的误差方差。图 8展示了一个采用贝塔-二项分布响应的INLA模型对观测数据进行拟合后得到的预测。图6(ai)中的平坦预测明显不再出现，这表明贝塔-二项分布能够有效解决过度离散问题。此外，我们还测试了高斯响应，发现它同样能够处理观测数据中的变异性，结果见附录B.1。这些结果表明，将INLA与二项响应用于疾病制图问题时需要谨慎，并且强调了检查过度离散的重要性。

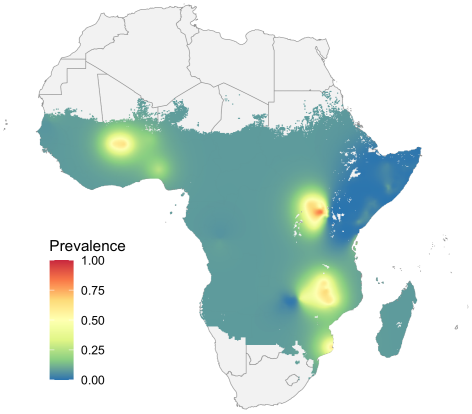


图 8：基于 INLA 模型的患病率预测。该模型采用贝塔-二项分布响应，并拟合图 3(a) 中的观测数据。

3.4.2 FRK

FRK是四种方法中速度最快的，但在使用两个空间稀疏数据集进行拟合时，其大陆尺度预测会呈现一种虚假的“斑点状”模式；而在随机选取位置的模拟数据上进行拟合时，生成的地图细节则少得多（图6）。这些特征似乎源于FRK使用少量基函数来近似高斯过程。本节将探讨增加此类函数的数量能否消除FRK输出中的伪影。

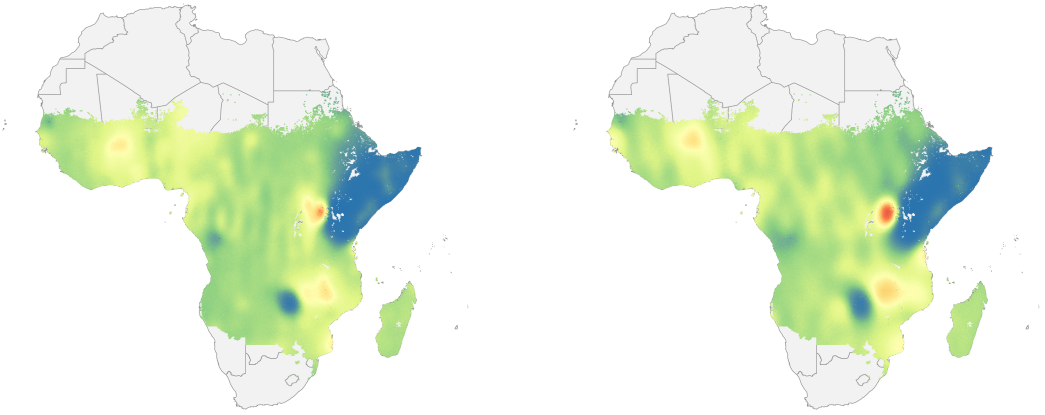
FRK 近似中使用的基函数数量主要由 nres 和 regular 参数控制。增大 nres 参数会增加一个具有更精细空间尺度的“分辨率”层或基函数层；增大 regular 的值则会缩小每个基函数的尺度，并为其排列增加行和列。有关这些参数影响的详细信息，请参阅软件文档（Zammit-Mangion 和 Sainsbury-Dale 2023）。第 3.1–3.3 节中使用的模型将这些参数设置为 nres= 2 and regular= 1。

图 9(a) 展示了 FRK 在拟合观测数据时的预测结果：nres 增至 3，regular 保持为 1，所得模型包含 1338 个基函数。尽管该模型中的大尺度斑驳现象不那么明显，但在中非的一些地区，可以清楚地看到更精细尺度的振荡。这一有限的改进付出了显著的计算代价：模型运行超过 55 分钟，需要 106GB 随机存取存储器；相比之下，当 nres 设为 2、regular 设为 1 时，运行仅需 3.28 分钟和 4.77GB 随机存取存储器。图 9(b) 展示了 nres 保持为 2、regular 增至 2 时的预测结果。这些设置产生了 600 个基函数，运行需要 22.74GB 内存和 9 分钟，所需资源显著少于增大 nres 参数时。然而，在数据很少或没有数据的区域，精细尺度的振荡明显更加突出。

虽然 FRK 计算效率的关键在于将空间随机效应分解为少量基函数，但这些结果表明，在实践中很难平衡这种效率与模型输出中出现伪影的风险，尤其是在数据稀疏的大规模制图问题中。

4 讨论

我们对四种地统计建模方法进行了实际比较，发现其中两种（SpRF 和 GPBoost）在可扩展性或准确性方面不足，无法应用于大规模疟疾患病率建模问题。SpRF 的空间预测呈现出明显的“条带”伪影，乍看之下，SpRF 模型似乎过拟合了（与训练数据高度吻合，但对留出数据的预测较差）。然而，



(a) (b)

图 9：使用 FRK 模型得到的恶性疟原虫 患病率预测：(a) nres= 3 和常规 = 1，以及 (b) nres = 2 和常规 = 2。

仔细检查后（附录A），我们发现，SpRF 在低密度区域较窄的不确定性区间可能造成了这种印象。与其他方法不同，SpRF 不采用协方差函数，而是将坐标之间距离矩阵的各列作为协变量纳入随机森林。我们指出，实际上可以在将距离矩阵纳入模型之前对其应用协方差函数。这种做法的含义与我们所考虑的基于高斯过程的模型不同，但可以让 SpRF 在建模未观测空间效应时考虑基于距离的衰减。不过，SpRF 中的随机森林推断机制无法估计此类函数的参数（例如随距离增加的衰减率），而这些参数似乎也难以事先合理设定。鉴于这些拟合问题，以及 SpRF 的计算时间大致随数据规模线性增长，这种方法不太可能适用于疟疾数据的实际地统计建模。

GPBoost 对留出数据的预测效果相当不错：在全国尺度比较中，它在 50 折空间分块交叉验证中表现最佳（表明其具备在短距离上外推的良好能力），在 10 折交叉验证中则仅次于 FRK，排名第二（表明其具备在较长距离上外推的能力）。然而，使用默认 GPBoost 设定时，随着数据量增加，计算时间急剧增长。这是因为 GPBoost 默认对完整（未经近似的）高斯过程进行推断，而推断程序的每一步都需要进行一次 $\mathcal{O}(n^3)$ 协方差矩阵求逆。无论是对高斯过程超参数进行最大似然推断，还是对截距（以及使用时的协变量效应）进行提升推断，都无法减轻这一计算负担。然而，正如附录 D 所示，即使采用 GPBoost 提供的 Vecchia 高斯过程近似，也未能解决这些问题。Vecchia 近似虽然缩短了计算时间，但计算时间仍随数据量线性增长，而且模型预测中还出现了严重伪影。提高近似的复杂度（纳入更多邻近点）可以消除这些伪影，但代价是所需计算时间和内存用量大幅增加。我们无法找到一组近似参数与提升法参数，使计算时间降至与 INLA 和 FRK 相当的水平。我们还注意到，GPBoost 是一种相对较新的技术，未来版本或许会包含更快速的近似方法。

与 SpRF 和 GPBoost 相比，INLA 和 FRK 在处理不断增大的数据集时具有显著更好的可扩展性，拟合 10,000 个数据点仅需几分钟。INLA 不仅具有良好的计算可扩展性，而且是广泛应用的疟疾数据地统计建模方法及软件；然而，使用常见的、通常建议用于患病率数据的二项分布来实现 INLA（例如 Moraga，2019年和 Moraga 等人，2021 年的建议），在 10折和 50折交叉验证测试中都导致了不合理的预测和较差的外推能力。我们已经证明，这是因为所建模的疟疾患病率数据相对于二项抽样假设和仅包含空间效应的模型而言存在过度离散。也就是说，

这一假设并不成立：给定某个位置估计的患病率，同一样本中每个人的感染状态彼此独立。从流行病学视角来看，这并不令人意外，因为某个地点的感染并非独立发生——每次感染都是由另一次感染引起的。这会产生局部噪声，其表现可能出现在像素或像素组层面（该位置可能存在平滑空间模型未考虑的某种风险因素），也可能出现在观测层面（采样当天，该人群的患病率可能高于或低于通常水平）。在这种情况下，INLA 试图通过一个非常“起伏”的空间随机效应来捕捉这些小尺度变化，即该效应会随距离增加而迅速衰减。对于过度离散数据，它倾向于采用这种参数配置，因为使用二项似然时观测方差是固定的，且其大小不足以解释数据。观测层方差与高斯过程长度尺度之间可识别性较差的问题此前已有描述（Rasmussen & Williams（2006），见图 5.4），在经典地统计学（及基于模型的地统计学）中，可通过对每个观测或每个观测位置加入独立的“块金”效应来解决（Diggle & Ribeiro Jr.（2007））。尽管 FRK 同样使用二项观测分布，却没有遇到这一问题，因为其“小尺度”效应参数中包含一种空间块金效应。

对于使用 INLA 进行疟疾患病率建模，我们建议采用一种比标准二项观测模型更可靠的“默认”模型，即纳入额外观测层级随机噪声的模型。这可以通过使用贝塔-二项分布或高斯分布（针对观测患病率或经验对数几率尺度）来实现。这两种方案都包含一个额外的观测层级方差参数，可用于解释相对于二项分布的过度离散。其中，贝塔-二项分布最有可能普遍适用于疟疾患病率数据，因为在受测个体中仅有极少数感染的常见情形下，它能够准确地考虑观测误差。不过，我们也注意到，在 INLA 中使用高斯响应进行拟合的计算效率要高得多，因此，如果计算时间是主要限制因素，这种方法可能更可取。另一种方法是在模型设定中纳入独立的观测层级随机效应。

FRK 能够很好地扩展至大型数据集（通常所需时间略少于 INLA），并且在 10折和 50 折外推比较中均表现良好；然而，对于大陆尺度建模，我们无法通过适度修改模型，构建出既能实现计算可扩展性、又能避免基函数产生虚假振荡效应的模型。虽然不那么明显，但在国家尺度分析中，肯尼亚西北部和最东北部的部分地区也能看到类似的模式；这些地区没有可用于指导此类预测的数据。鉴于这些问题，我们认为，将 FRK 用于绘制稀疏疟疾测量数据的地图时必须格外谨慎，以避免由计算上便捷的近似而非数据所驱动的虚假预测。

比较四种方法论上不同的技术存在一定局限。其中一个局限是，这些方法固有的差异使得比较有些困难。例如，各方法的似然函数不同，其底层模型结构和／或协方差函数也有所不同。因此，每种方法都有自己的评价指标，我们无法将 INLA 的拟合优度指标与 SpRF 的指标进行比较，反之亦然。为缓解这一问题，我们着重关注输出结果——预测地图和交叉验证结果。

另一个值得关注的方面是参数设置。每种方法都有许多不同的参数设置。本研究采用了常用的（默认）参数设置；尽管我们也探索了若干不同的参数设置，但并未对这些算法的参数空间进行全面探索。虽然默认参数设置适用于肯尼亚，但我们预计，在非洲尺度上运行模型时，算法可能会受益于定制参数。地理区域范围之广凸显了参数选择的局限性。探索最优参数选择是另一个研究方向。此外，也可能存在其他参数设置，使不同模型的推断近似结果更具可比性。

从实践者的角度来看，采用新的空间建模方法颇具挑战，主要是因为学习方法并编写代码以生成有意义的输出需要很长时间。这是一项重要的入门障碍。如果本文讨论的方法能够提供调优函数，探索参数空间并选出一组可帮助实践者构建优质模型的参数，就能提高这些方法的易用性。

深入研究这些模型的优势与不足，将是另一个值得探索的方向。

一种方案是构建一个元模型，根据不同位置的特征预测最优模型（Wang 等，2009年）。这种元模型可以结合各种模型的优势，从而做出更准确的预测。本文的研究结果可供创建、解读或处理空间数据的人士参考，并为新的计算地质统计模型提供基准比较。

5 致谢

我们谨此感谢 Håvard Rue、Andrew Zammit-Mangion、Matthew Sainsbury-Dale、Fabio Sigrist 和 Noel Cressie 在通信交流以及模型的搭建与故障排查方面提供的帮助。本研究得到了墨尔本大学研究计算服务（Research Computing Services）和千兆级校园倡议（Petascale Campus Initiative）的支持。J.A. Flegg 的研究得到了澳大利亚研究理事会（DP200100747、FT210100034）和国家卫生与医学研究委员会（APP2019093）的资助。

参考文献

Bakka, H.、Rue, H.、Fuglstad, G.-A.、Riebler, A.、Bolin, D.、Illian, J.、Krainski, E.、Simpson, D. 和 Lindgren, F.（2018年），《使用 R-INLA 进行空间建模：综述》，《*Wiley* 跨学科评论：计算统计学》**10**(6), e1443。

Bertozzi-Villa, A., Bever, C. A., Koenker, H., Weiss, D.J., Vargas-Ruiz, C., Nandi, A. K., Gibson, H. S., Harris, J., Battle, K. E., Rumisha, S. F. 等（2021），《2000—2020年非洲杀虫剂处理蚊帐的可及性、使用情况及人均蚊帐数量的地图和指标》，《自然通讯》**12**(1), 1–12。

Bhatt, S., Cameron, E., Flaxman, S. R., Weiss, D. J., Smith, D. L. & Gething, P. W. (2017), '利用高斯过程堆叠泛化提高疾病风险制图的预测准确性',《英国皇家学会界面期刊》**14**(134)。

Cressie, N. (2015), 空间数据统计学，约翰·威利父子公司。

Datta, A., Banerjee, S., Finley, A. O. & Gelfand, A.E.(2016), '用于大型地质统计数据集的层次近邻高斯过程模型',《美国统计学会会刊》**111**(514), 800–812。

Diggle, P. & Ribeiro Jr., P. J. (2007), 基于模型的地质统计学，施普林格。

Flegg, J. A., Humphreys, G. S., Montanez, B., Strickland, T., Jacome-Meza, Z. J., Barnes, K. I., Raman, J., Guerin, P. J., Hopkins Sibley, C. & Dahlström Otienoburu, S. (2022), '1990—2020年磺胺多辛-乙胺嘧啶耐药性恶性疟原虫突变在非洲的时空传播',《*PLoS* 计算生物学》**18**(8), e1010317。

Flegg, J. A., Patil, A. P., Venkatesan, M., Roper, C., Naidoo, I., Hay, S. I., Sibley, C. H. & Guerin, P. J. (2013), '非洲恶性疟原虫dhps基因突变的时空数学建模',《疟疾杂志》**12**(1),1–9.

Gething, P. W., Casey, D. C., Weiss, D. J., Bisanzio, D., Bhatt, S., Cameron, E., Battle, K. E., Dalrymple, U., Rozier, J., Rao, P. C. et al. (2016年), '绘制1990年至2015年非洲恶性疟原虫死亡率地图',《新英格兰医学杂志》**375**(25),2435–2445.

Gething, P. W., Patil, A. P., Smith, D. L., Guerra, C. A., Elyazar, I. R. F., Johnston, G. L., Tatem, A. J. & Hay, S. I. (2011年), '全新的疟疾世界地图：2010年恶性疟原虫流行状况',《疟疾杂志》**10**.

Gómez-Rubio, V.（2020），使用*INLA*进行贝叶斯推断，Chapman & Hall/CRC出版社，佛罗里达州博卡拉顿。Hay, S. I. & Snow, R. W.（2006），《疟疾地图项目：开发全球疟疾风险地图》，《*PLoS*医学》**3**(12)，2204–2208。

Hengl, T.、Nussbaum, M. & Wright, M. N.（2021），《GeoMLA》，GitHub代码库, https://github.com/ thengl/GeoMLA，访问于09/21。Hengl, T.、Nussbaum, M.、Wright, M. N.、Heuvelink, G. B. & Gräler, B.（2018），《随机森林作为空间和时空变量预测建模的通用框架》，*PeerJ***6**，e5518。Hensman, J.、Fusi, N. & Lawrence, N. D.（2013），《面向大数据的高斯过程》，*arXiv*预印本 *arXiv:1309.6835* 。Humphreys, J. M.、Elsner, J. B.、Jagger, T. H. & Pau, S.（2017），《贝叶斯地统计方法：在预测气候变暖情景下对微叶海金沙全球分布进行建模》，《生态建模》**363**，192–206。Kang, S. Y.、Battle, K. E.、Gibson, H. S.、Ratsimbasoa, A.、Randrianarivelojosia, M.、Ramboarina, S.、Zimmer- man, P. A.、Weiss, D. J.、Cameron, E.、Gething, P. W. 等（2018），《对马达加斯加疟疾指标调查结果进行时空制图，以评估2011年至2016年间恶性疟原虫地方性流行趋势》，《*BMC*医学》**16**(1)，1–15。Likas, A.、Vlassis, N. & J. Verbeek, J.（2003），《全局k均值聚类算法》，《模式识别》**36**(2)，451–461。Lindgren, F. & Rue, H.（2015），《使用R-INLA进行贝叶斯空间建模》，《统计软件杂志》**63**，1–25。Lindgren, F.、Rue, H. & Lindström, 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》，《英国皇家统计学会杂志：*B*辑（统计方法论）》**73**(4)，423–498。Meinshausen, N.（2006），《分位数回归森林》，《机器学习研究杂志》**7**，983–999。Moraga, P.（2019），地理空间健康数据：使用*R-INLA*和*shiny*进行建模与可视化，Chapman & Hall/CRC生物统计学丛书。Moraga, P.、Dean, C.、Inoue, J.、Morawiecki, P.、Noureen, S. R. & Wang, F.（2021），《使用INLA和SPDE方法对地统计数据进行贝叶斯空间建模：预测莫桑比克疟疾风险的案例研究》，《空间与时空流行病学》**39**。Park, C. & Apley, D.（2018），《用于大规模高斯过程回归的拼接克里金法》，《机器学习研究杂志》**19**，1–43。Pfeffer, D. A.、Lucas, T. C.、May, D.、Harris, J.、Rozier, J.、Twohig, K. A.、Dalrymple, U.、Guerra, C. A.、Moyes, C. L.、Thorn, M.、Nguyen, M.、Bhatt, S.、Cameron, E.、Weiss, D. J.、Howes, R. E.、Battle, K. E.、Gibson, H. S. & Gething, P. W.（2018），《MalariaAtlas：连接托管于疟疾地图项目的全球疟疾测量数据的R接口》，《疟疾杂志》**17**(1)。Quiñonero-Candela, J. & Rasmussen, C. E.（2005），《稀疏近似高斯过程回归的统一视角》，《机器学习研究杂志》**6**，1939–1959。Rasmussen, C. E. & Nickisch, H.（2010），《机器学习高斯过程（GPML）工具箱》，《机器学习研究杂志》**11**，3011–3015。Rasmussen, C. E. & Williams, C. K. I.（2006），《机器学习高斯过程》，第14卷，麻省理工学院出版社。Roberts, D. R.、Bahn, V.、Ciuti, S.、Boyce, M. S.、Elith, J.、Guillera-Arroita, G.、Hauenstein, S.、Lahoz-Monfort, J. J.、Schröder, B.、Thuiller, W.、Warton, D. I.、Wintle, B. A.、Hartig, F. & Dormann, C. F.（2017），《适用于具有时间、空间、层级或系统发育结构数据的交叉验证策略》，《生态地理学》**40**(8)，913–929。

Rue, H., Martino, S. & Chopin, N. (2009), ‘使用集成嵌套拉普拉斯近似对潜在高斯模型进行近似贝叶斯推断’, 《英国皇家统计学会杂志：*B*辑（统计方法论）》**71**(2), 319–392. Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P. & Lindgren, F. K. (2017), ‘基于INLA的贝叶斯计算：综述’, 《统计学及其应用年度评论》 **4**, 395–421. Sainsbury-Dale, M., Zammit-Mangion, A. & Cressie, N. (2021), ‘使用FRK对大型异质非高斯空间及时空数据进行建模’, *arXiv*预印本 *arXiv:2110.02507*. Sigrist, F. (2020), ‘GPBoost’, GitHub代码库，https://github.com/fabsig/GPBoost，访问于10/21。Sigrist, F. (2022), ‘高斯过程提升’, 《机器学习研究杂志》**23**(232), 1–46. Wang, X., Ryan, Y. Y. & J. Faraway, J. (2018), 基于*INLA*的贝叶斯回归建模, Chapman & Hall/CRC出版社, 博卡拉顿，佛罗里达州。Wang, X., Smith-Miles, K. & Hyndman, R. (2009), ‘用于预测方法选择的规则归纳：单变量时间序列特征的元学习’, 《神经计算》**72**(10-12), 2581–2594. Weiss, D. J., Lucas, T. C., Nguyen, M., Nandi, A. K., Bisanzio, D., Battle, K. E., Cameron, E., Twohig, K. A., Pfeffer, D. A., Rozier, J. A. 等 (2019), ‘绘制恶性疟原虫全球患病率、发病率和死亡率地图，2000–17：一项空间及时序建模研究’, 《柳叶刀》**394**(10195), 322–331. Zammit-Mangion, A. & Cressie, N. (2021), ‘FRK：用于大型数据集空间及时空预测的R软件包’, 《统计软件杂志》**98**(4), 1–48. Zammit-Mangion, A. & Sainsbury-Dale, M. (2023), ‘软件包“FRK”’, 软件包文档，https://cran.r-project.org/web/packages/FRK/FRK.pdf，访问于04/23。

A 详细交叉验证结果

A.1 点预测

表3：四个模型的交叉验证结果，其中各类别中结果最佳者以粗体显示。

Fold	模型	RMSE	相关性	绝对误差小于指定值的数据点百分比		
				0.05	0.1	0.2
10折	INLA	0.181	0.235	69.211	76.316	86.053
	GPBoost	0.127	0.646	52.632	73.947	**93.158**
	SpRF	0.132	0.641	**69.474**	**80.263**	91.053
	FRK	**0.125**	**0.661**	57.105	76.579	92.105
50折	INLA	0.124	0.683	**69.474**	**80.789**	87.895
	GPBoost	**0.11**	**0.751**	65.789	**80.789**	90.0
	SpRF	0.121	0.702	67.632	78.947	**90.526**
	FRK	0.123	0.702	66.053	79.211	90.263

表3给出了10折和50折交叉验证的结果。对于每组折，表中给出了均方根误差（RMSE）、预测值与实际值之间的相关系数，以及绝对误差（|预测值 –实际值|）小于指定阈值的观测值百分比（阈值为0.05、

（考虑0.1和0.2）。如前所述，对于10折和50折交叉验证，FRK和GPBoost的均方根误差及相关性数值分别最佳。就绝对误差小于0.05和0.1的观测值百分比而言，SpRF的表现最佳。与其他模型相比，INLA在10折交叉验证中的表现较差，均方根误差较高，相关系数显著较低。不过，其绝对误差小于三个阈值的观测值百分比很高。这是因为大量观测值的患病率值较低。图 10进一步说明了这一点，该图展示了使用10折交叉验证时各模型的实际值和预测值。

图 10 展示了图 2(a) 中确定的各交叉验证折所包含的点。由于各折是通过 k-均值聚类确定的，因此每折中的观测值彼此接近。我们看到，大多数折中的数据点具有相似的患病率值。然而，第 3、8 和 9 折中的数据点取值范围较广。对于四种模型而言，分配到第 8 折的点都难以预测。这些点位于海岸沿线、蒙巴萨市附近，并且与其他簇相对孤立，这可能是造成这一现象的原因之一。

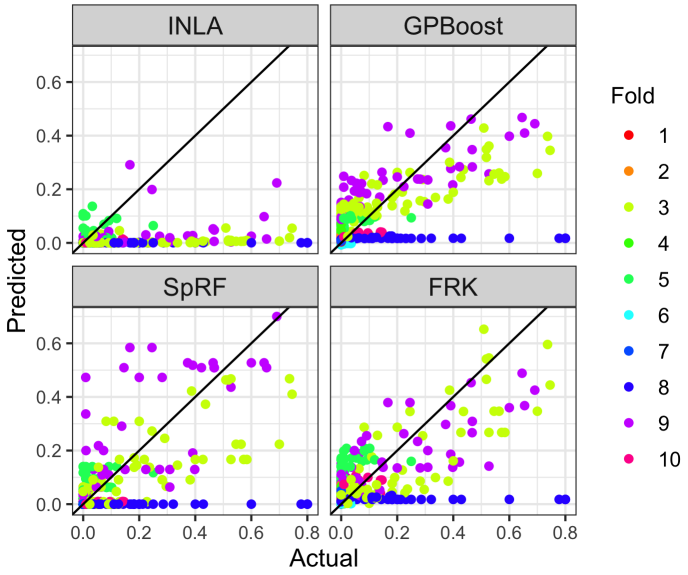


图 10：使用 10 折交叉验证时，四种模型的模型预测与实际患病率对比。

图 11 展示了四种模型的 50 折交叉验证结果，图 12 则展示了它们的区间预测。从表3可以看出，就均方根误差和相关性而言，GPBoost 的表现更好。在绝对误差小于 0.1 的观测值占比最高这一指标上，它与 INLA 表现相同。绝对误差小于 0.05 的观测值占比最高的是 INLA，而绝对误差小于 0.2 的观测值占比最高的是 SpRF。从图 11 可以看出，某些折的表现不佳。这些折与 10 折交叉验证情景中表现不佳的折所在位置相对应。另一个有趣的观察是，尽管 GPBoost 在两组折上的表现都不错，但它对高患病率观测值的预测较差，而 FRK 和 SpRF 似乎没有这一局限。

A.2 按位置密度进行点预测

我们进一步利用抽样位置的密度分析这些结果，即某些模型是否难以预测低密度区域中的观测值？图 13 分别在两张地图上展示了疟疾患病率和抽样位置的核密度估计。图 14 展示了患病率与密度的散点图，点的颜色表示所属折。对于 10 折交叉验证，我们发现第 8 折所在区域靠近蒙巴萨市，患病率值的范围较广，但密度相对较低。这解释了误差较高的原因

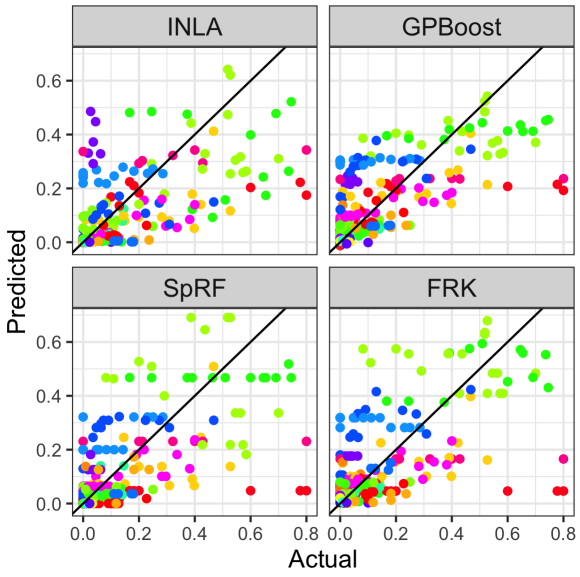
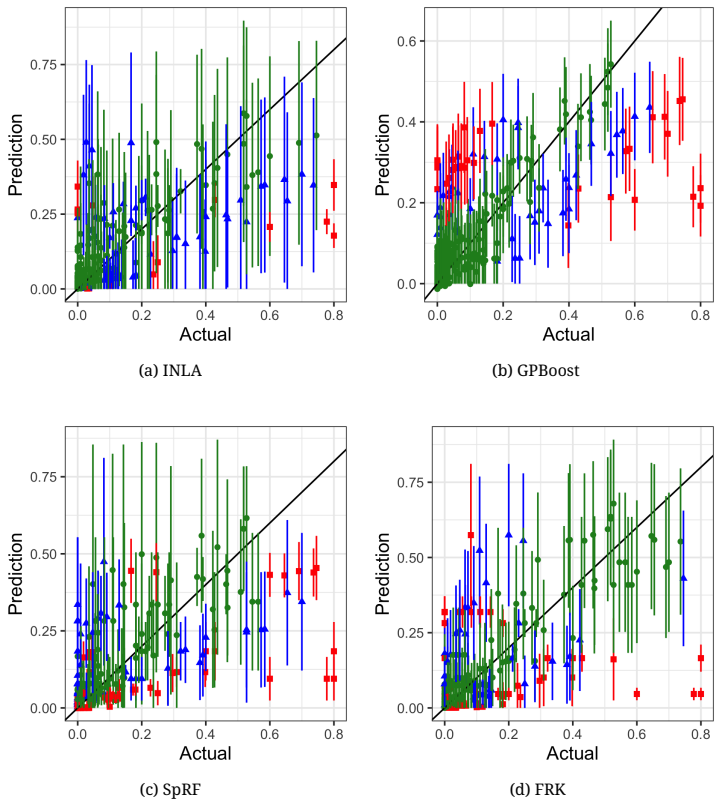


图 11：使用 50折交叉验证的四种模型预测与实际患病率对比，不同颜色表示不同折。



(a) INLA

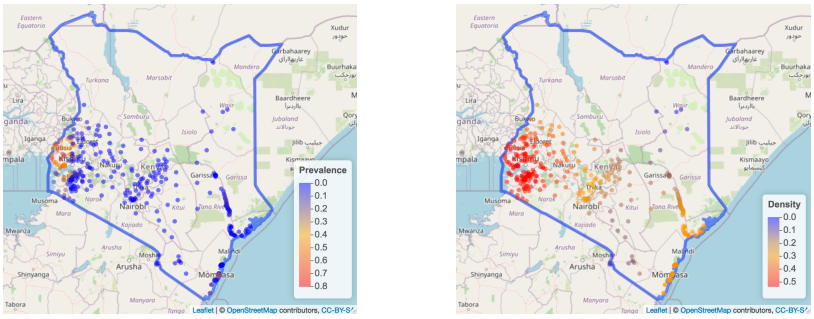
(b) GPBoost

(c) SpRF

(d) FRK

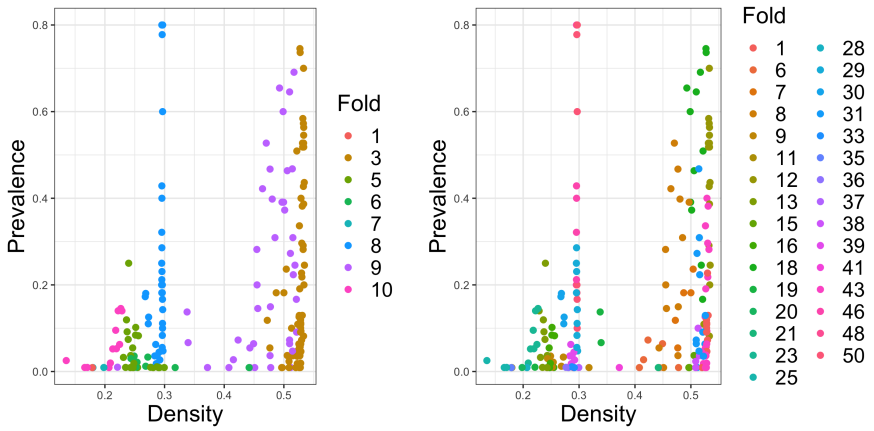
图 12：50折交叉验证的区间预测。

对于第 8 折（图 10）。当采样点彼此相距较远（密度低）且患病率值变化较大时，模型很难准确预测。



(a) 疟疾患病率　　(b) 抽样位置的密度

图13：恶性疟原虫 2009年肯尼亚的患病率以及抽样位置的核密度估计。



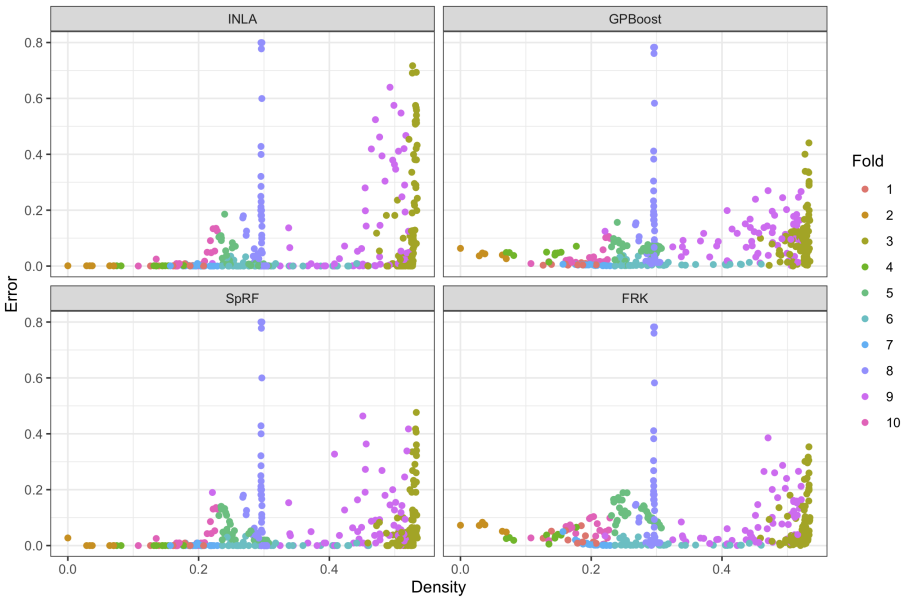
(a) 10 折　　(b) 50 折

图14：移除零患病率观测值后，不同簇（折）的恶性疟原虫 患病率和核密度估计。

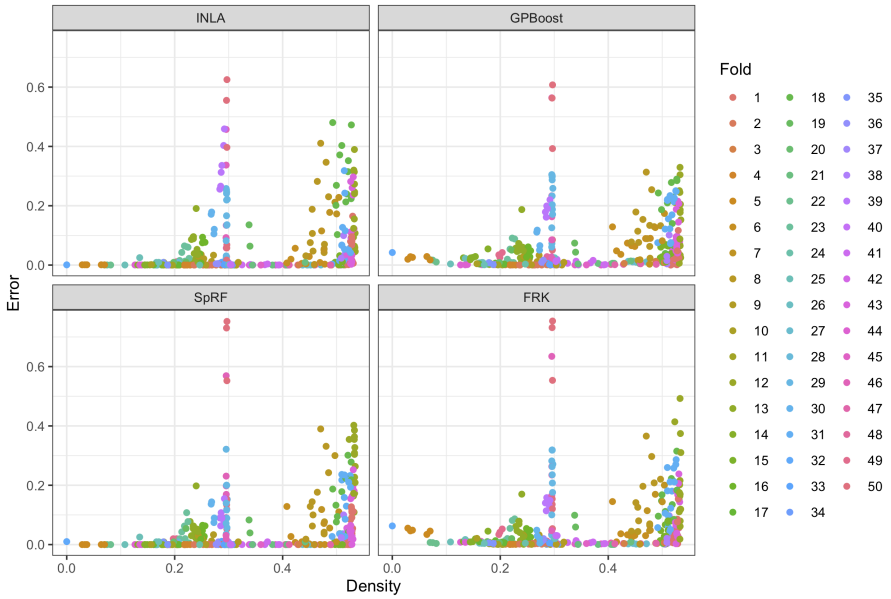
图15展示了在10折和50折交叉验证中，四种模型的绝对误差随密度变化的情况。对于10折交叉验证，我们看到第8折的四种模型误差率都很高。若对50折交叉验证使用相同的一组点，可以看到FRK和SpRF的误差率相近（最大值为 ≈ 0.75），而INLA和GPBoost的误差率相对较低（最大值为 ≈ 0.6）。因此，在此例中，增加折数对GPBoost和INLA的益处大于对FRK或SpRF的益处。

接下来考察密度很低的点（< 0.2），从图14可以看到，这些点的患病率值相对较低。图15(a)显示，与FRK和GPBoost相比，INLA和SpRF在这些低密度点上的误差较低。在图15(b)所示的50折交叉验证情形中，当点的密度低于0.2时，也可以观察到类似结果。

如图14所示，高密度区域（$> 0。4$）中的点，其患病率变化范围更大，为0到0.75。图15(a)中的10折交叉验证结果表明，对于这些高密度点，FRK表现最佳，误差为 $< 0。4$，其次是GPBoost和SpRF。INLA在10折交叉验证中的这些点上表现不佳。对于50折交叉验证（图15(b)），GPBoost在高密度点上表现最佳，其次是SpRF，而FRK和INLA表现相近。



(a) 使用10折交叉验证的四种模型的密度和绝对误差



（b）使用50折交叉验证的四种模型的密度和绝对误差

图15：10折和50折交叉验证中四种模型的位置及其绝对误差的核密度估计。

表4列出了10折和50折交叉验证中不同密度组的指标。我们将低密度定义为密度 ≤ 0。2，中密度定义为0。2 < ≤密度 ≤ 0。4，高密度定义为密度 > 0。4。可以看出，四种模型在低密度点上的绝对误差都很小。在10折和50折交叉验证数据集中，绝对误差小于0.05、0.1和0.2的观测值比例都相当高。在这两组交叉验证数据中，INLA的均方根误差最低，SpRF紧随其后。对于10折和50折交叉验证，实际值与预测值之间的相关系数均为负值。这表明大多数患病率值都接近

在低密度位置降至零。这也可以解释为什么这些区域采集的样本点较少，因为通常会从高患病率区域采集更多样本点。

表4：按抽样位置密度分组的交叉验证结果，最佳结果以粗体显示。

Fold	密度	模型	RMSE	相关系数	绝对误差小于指定值的数据点百分比		
					0.05	0.1	0.2
10折	Low	INLA	**0.005**	-0.234	100	100	100
		GPBoost	0.028	-0.189	96	100	100
		SpRF	0.006	-0.059	100	100	100
		FRK	0.048	-0.021	62	98	100
	中等	INLA	0.141	-0.003	75.723	82.659	93.642
		GPBoost	**0.138**	-0.021	60.694	85.549	94.22
		SpRF	0.143	-0.025	72.832	80.347	93.642
		FRK	0.145	-0.035	58.382	76.301	94.22
	High	INLA	0.24	0.305	52.229	61.783	73.248
		GPBoost	0.134	0.788	29.936	52.866	89.809
		SpRF	0.14	0.737	56.051	73.885	85.35
		FRK	**0.119**	0.828	54.14	70.064	87.261
50折	Low	INLA	**0.005**	-0.128	100	100	100
		GPBoost	0.017	-0.202	98	100	100
		SpRF	0.006	-0.077	100	100	100
		FRK	0.021	-0.17	96	100	100
	中等	INLA	0.123	0.501	73.41	83.815	90.173
		GPBoost	**0.117**	0.511	73.41	84.393	91.908
		SpRF	0.121	0.435	71.676	85.549	96.532
		FRK	0.133	0.31	75.723	83.237	91.329
	High	INLA	0.143	0.776	55.414	71.338	81.529
		GPBoost	**0.119**	0.809	47.134	70.701	84.713
		SpRF	0.14	0.742	52.866	64.968	80.892
		FRK	0.129	0.784	45.86	68.153	85.987

在 10 折交叉验证中，与高密度点集相比，GPBoost、SpRF 和 FRK 在中密度点集上的均方根误差更高。在 50 折交叉验证中，FRK 也表现出类似的情况。这是由于第 8 折中存在较大的绝对误差，正如前文所讨论的那样。对于中密度点，按照均方根误差衡量，在 10 折交叉验证和 50 折交叉验证中，GPBoost 均为首选。对于高密度点，在 10 折交叉验证中，FRK 是首选，而在 50 折交叉验证中，GPBoost 是首选。就绝对误差小于 0.05 和 0.1 的点所占百分比而言，在 10 折交叉验证中，SpRF 优于其他模型，而在 50 折交叉验证中，INLA 表现最佳。在 10 折和 50 折交叉验证中，与另外两种方法相比，INLA 和 SpRF 在低密度点上的表现更好，而 GPBoost 和 FRK 在高密度点上的表现更好。

A.3 区间预测

如第3节所述，每种方法都有不同的不确定性量化机制，但我们估算了各个模型的预测标准差，以便进行比较。

表5给出了全部四种模型的区间预测结果。图5和图12展示了10折和50折交叉验证的区间预测。对于10折和50折交叉验证，SpRF的平均不确定性区间最窄，且均值的一个或两个标准差范围内的点数最少。

令人意外的是，在10折和50折交叉验证中，SpRF的区间宽度分别有148个和164个预测值为零。这些点的预测值（中位数）均为零患病率，且大多数点的观测患病率也为零。然而，几乎所有这些位置的响应变量均值虽小但不为零，因此这些点的患病率不在均值的任何标准差范围内，导致表5中SpRF的百分比较低。

表5：四种模型的区间预测结果。均值宽度指预测标准差的平均值，标准差宽度指这些预测标准差的标准差。

Fold	模型	Mean 宽度	标准差 宽度	范围内的点（%）	
				1SD	2SD
10折	INLA	0.102	0.048	75	87.105
	GPBoost	0.136	0.014	84.211	95.526
	SpRF	0.071	0.106	37.105	55
	FRK	0.1	0.072	83.421	92.368
50折	INLA	0.096	0.091	81.053	95.789
	GPBoost	0.112	0.014	81.053	91.053
	SpRF	0.056	0.086	28.158	42.895
	FRK	0.062	0.068	74.474	85

对于 10 折和 50 折交叉验证，GPBoost 的区间宽度均值最大，但标准差最小，这表明它对大多数观测值都能稳定地预测出较宽的区间。对于 10 折交叉验证，落在一个或两个标准差范围内的点所占比例因此最高。对于 50 折交叉验证，INLA 中落在各类区间内的点所占比例更高，这可能是因为 INLA 的区间宽度变化更大。

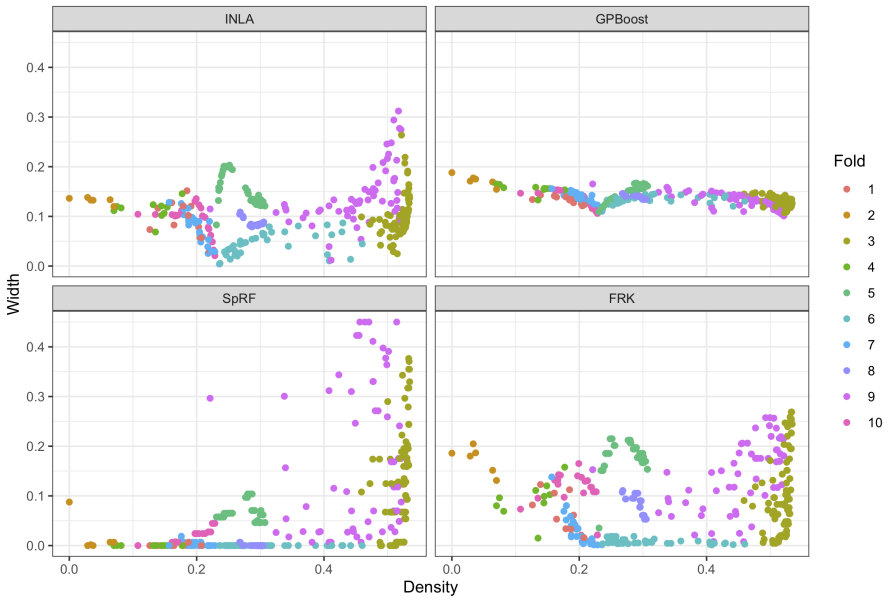
图 16 展示了 10 折和 50 折交叉验证中四种模型的位置及其相应区间宽度的核密度估计。对于 10 折和 50 折交叉验证，我们看到 GPBoost 对所有观测值给出的区间宽度相近。低密度点的区间宽度略有增加，但区间宽度随密度的变化并不明显。相比之下，FRK、INLA 和 SpRF 在不同折中的区间宽度各不相同。我们看到，高密度位置的区间宽度变化很大，这很可能是患病率变化所致。对于 10 折和 50 折交叉验证，FRK 对低密度点给出的区间宽度相对较大。相反，INLA 和 SpRF 对低密度点给出的区间宽度较小。与点预测类似，对于 FRK、INLA 和 SpRF，我们看到中等密度点（密度 ≈ 0.3）的区间宽度变化很大。

B 输入噪声对 INLA 的影响

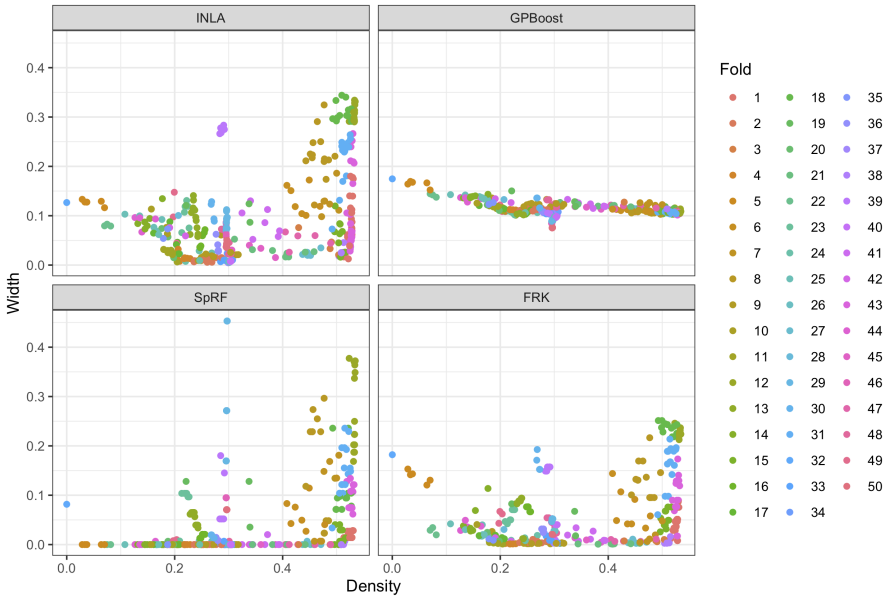
图6(ai)显示，使用观测数据进行训练时，INLA预测非洲大部分地区的患病率接近于零且几乎不变，而将模型拟合到任一组模拟数据时，都不会出现这种情况。这种现象可能是由观测数据中的噪声造成的，在图 3(a) 中尤以乌干达的情况最为明显。相比之下，图 3(b) 所示的相同位置的模拟数据看起来平滑得多。

我们通过向模拟数据中添加高斯噪声来检验这一工作假设。根据观测点确定位置，从图 3(d) 中的 MAP栅格抽取患病率值，并使用 logit 函数进行转换。随后，向转换后的值添加具有指定标准差的高斯白噪声，再通过逆 logit 函数将其还原为 0 到 1 之间的值。然后，利用这些患病率抽取阳性检测数量的二项抽样，并使用 INLA 对这些数据进行拟合。

图17展示了拟合于三个不同噪声水平数据的 INLA 预测。图18展示了拟合模型的截距、范围和方差的后验均值及四分位距，分别为。



(a) 使用10折交叉验证的四种模型的密度和区间宽度

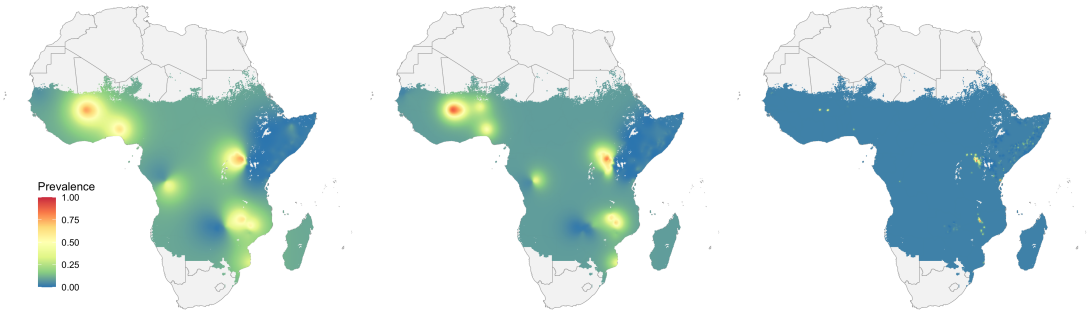


(b) 使用50折交叉验证的四种模型的密度和区间宽度

图 16：10折和50折交叉验证中四种模型的位置及区间宽度的核密度估计。

随着添加噪声的标准差增大，拟合每个模型并生成预测所需的时间也会增加。

随着添加噪声的标准差增大，截距和空间范围均会降低，且不同位置之间的预测相关性会减弱。当标准差超过 0.6 时，所需时间会大幅增加，模型可能难以收敛。噪声越大，模型越会根据截距的取值，在远离模拟数据的区域预测出平坦的患病率；图 17(c) 中的输出与.....相似



(a) 未添加噪声　(b) 标准高斯噪声偏差 0.4　(c) 标准高斯噪声的情况下偏差 1.2

图17：将 INLA 拟合到观测位置处的模拟数据时，在添加不同程度高斯噪声的情况下预测的患病率。

基于观测数据训练的模型在图6(ai)中的预测结果。这些结果表明存在过度离散，且所用的 INLA 模型可能设定有误。事实上，方程（4）中的模型不包含独立误差项。解决方法包括在模型中添加观测随机效应，或使用贝塔-二项分布响应。

B.1 使用高斯响应的 INLA

由于将 INLA 模型拟合到观测数据时存在过度离散，我们测试了另一个采用高斯响应的 INLA 模型。该模型的预测结果如图 19 所示，且不存在平坦预测值，这表明该响应能够解决过度离散问题。

C 预测不确定性

图20显示了图6中非洲各患病率地图对应的预测不确定性。

D 采用Vecchia近似的 GPBoost

由于采用完整的高斯过程，GPBoost模型在较大数据集上的计算时间最不理想，这并不令人意外。为提高效率，该软件提供了 Vecchia近似，用于近似响应变量的分布，其形式为

$$
\begin{aligned}
p(\mathbf{y}|F(X),\boldsymbol{\theta}) &= \prod_{i=1}^{n} p(y_i|y_{i-1},...,y_1,F(X),\boldsymbol{\theta}) \\
&\approx \prod_{i=1}^{N} p(y_i|y_{N(i)},F(X),\boldsymbol{\theta}),
\end{aligned}
$$

根据 Sigrist (2022)。此处，$y_{N(i)}$ 是$\{y_1,...,y_{i-1}\}$的子集，其中包含距离 y_i最近的 m_v 个最近邻；“最近邻”是根据响应变量对应位置之间的距离确定的。参数 m_v 决定拟合期间使用的邻居数量，而另一个参数 $m_{v,p}$则控制预测时使用的邻居数量。此外，该近似还需要选择观测响应 $\{y_1,...,y_n\}$的排序方式，默认情况下采用输入数据的原始顺序。

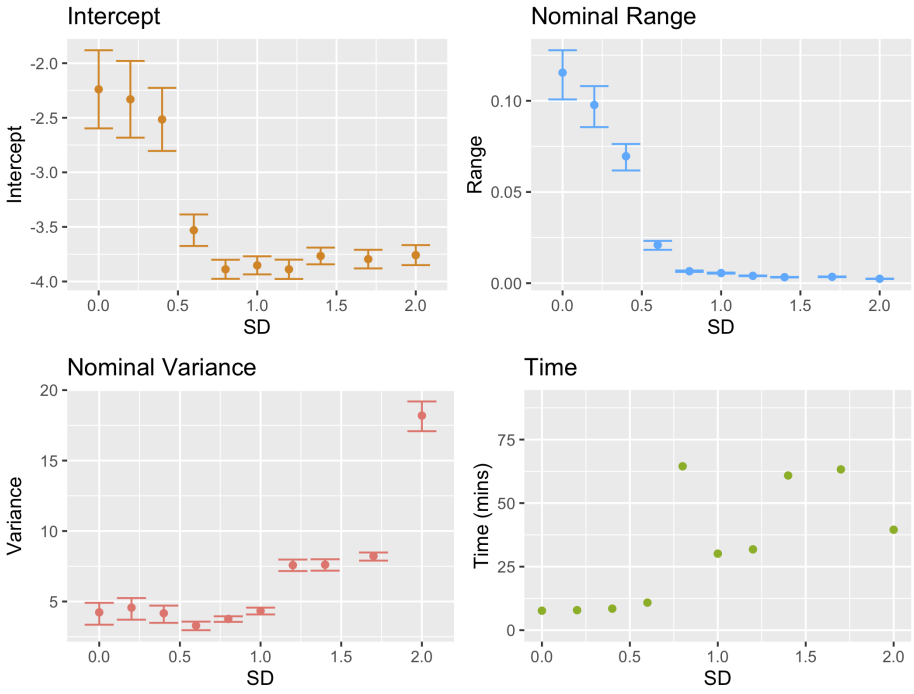


图18：使用观测位置处添加了不同标准差高斯噪声的模拟数据拟合 INLA 模型后，截距、范围和方差的后验均值。右下图显示了对每个数据集拟合 INLA 模型所用的时间。误差条表示后验四分位距。

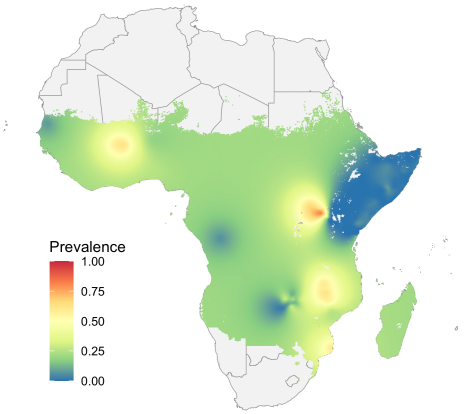


图19：使用观测数据拟合的高斯响应 INLA 模型的预测值。数值已截断至 $[0, 1]$ 范围内。

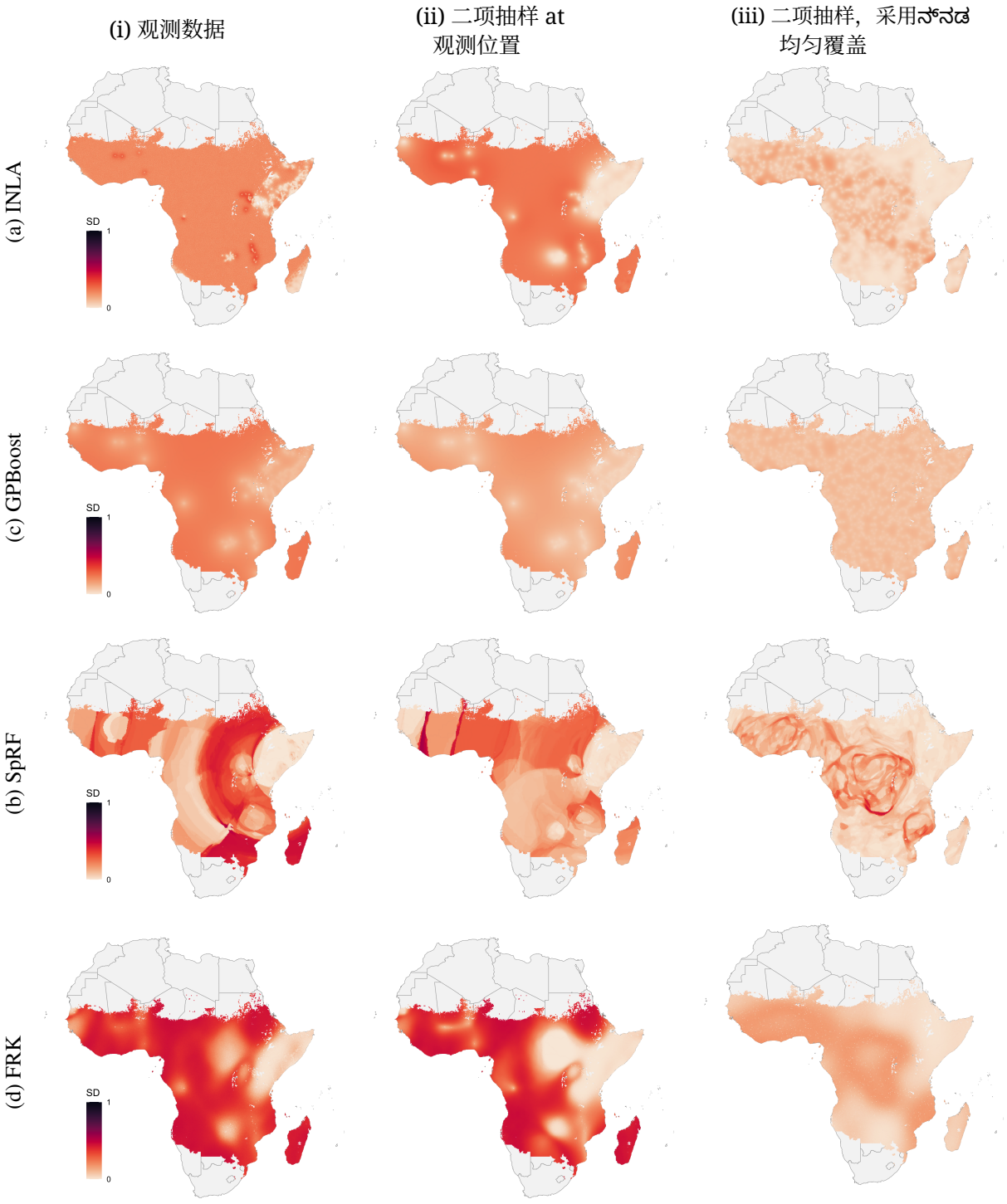


图20：图6所示各地图的预测标准差。

图21展示了在使用 Vecchia近似并采用 m_v 和$m_{v、p}$的多个取值时，GPBoost 的预测结果。由于目前软件在使用 Vecchia近似时对不确定性支持不足，因此未生成不确定性预测。

应用 Vecchia近似会在 GPBoost 的预测结果中引入若干伪影。图 21(a) 和 21(d)显示出明显的不连续现象和噪声预测；每当使用较低的m_v 和$m_{v、p}$ 时，这两种现象都十分突出。使用肯尼亚数据进行的实验表明，通过增大 m_v 和$m_{v、p}$ 可以避免不连续现象和噪声（结果未显示）；然而，对于大陆尺度的数据集，这样做会带来显著的计算成本。将$m_{v、p}$ 从 30 增大到 150，同时保持 m_v固定为 30，对计算时间的影响相对较小：计算时间从 7.4 分钟增加到 9 分钟，但所需内存却从 1289MB 激增至 20546MB。与此同时，将 m_v 和$m_{v、p}$ 都增大到 150，则会大幅增加计算时间，运行所需时间超过 3.4 小时，远长于未应用 Vecchia近似时的运行时间。此外，这种模型配置还需要 20679MB 的随机存取存储器。这些示例表明，增大 m_v 主要会增加所需的计算时间，而不会影响内存用量；增大$m_{v、p}$ 则会增加所需内存，对计算时间的影响较小。尽管计算需求增加了，但对这些参数进行任何一种调整都无法完全消除噪声和不连续现象。

使用 Vecchia近似的一个优点是能够改善扩展性，尽管单个数据集所需的计算资源很大程度上取决于 m_v 和 $m_{v,\ p}$的选择。图22展示了图7中的计算结果，并额外绘制了使用 Vecchia近似的 GPBoost 结果。参数 m_v 和 $m_{v,\ p}$ 分别固定为 30 和 150，模型的计算时间随观测值数量增加而线性增长。然而，其扩展性仍然比 INLA 和 FRK 更差。这些结果凸显了将 GPBoost 应用于大规模数据的一大障碍。使用完整的高斯过程可能导致计算时间很长，而采用 Vecchia近似则会引入额外的伪影，需要牺牲计算效率才能消除。

GPBoost 的计算时间会因 nrounds 参数取值较高而增加；根据现有教程，该参数被设为 247。如第2.2.1节所述，该参数控制拟合过程中的优化步骤数。将模型拟合到本文中使用的疟疾数据集时，对数似然通常在 5 到 10 步后便不再增加，这表明对于我们的数据而言，247 个训练步骤过多。将 nrounds 降至约 10，可以显著改善图22中使用 Vecchia近似的 GPBoost模型的扩展性斜率。然而，进一步实验发现，减少迭代轮数对 $m_{v,\ p}$取值较大时较高的随机存取存储器需求几乎没有影响；而较大的取值可能是尽量减少预测中的不连续性和噪声所必需的。减少 nrounds 也能提高未使用 Vecchia近似时 GPBoost模型的效率，但不会改变整体扩展性表现。

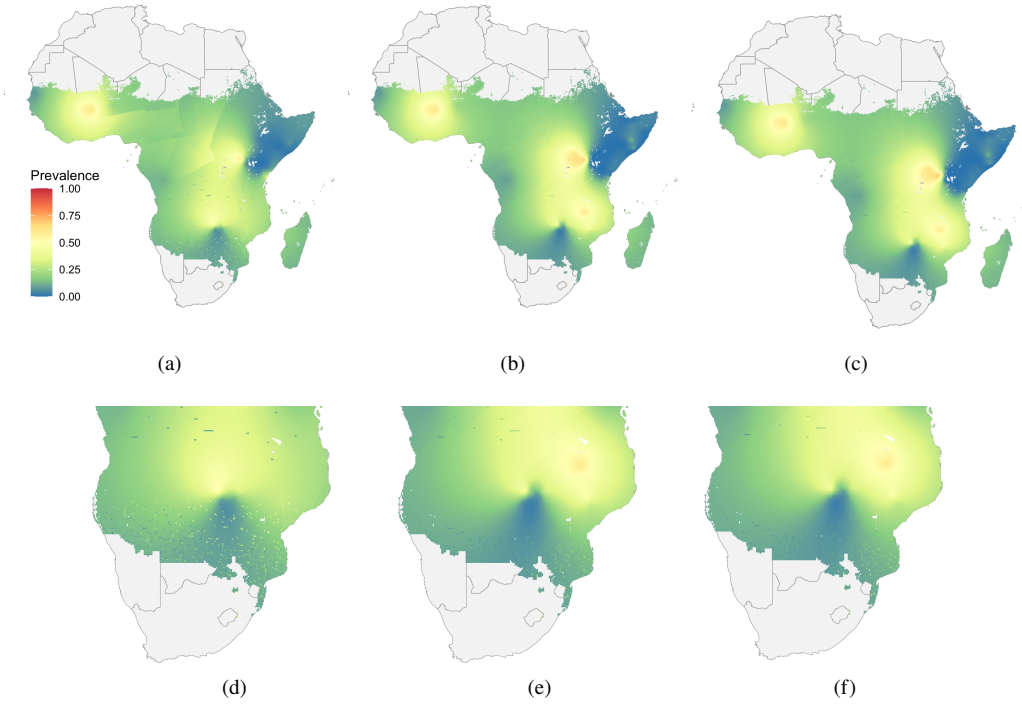


图21：使用不同最近邻参数值 m_v 和$m_{v,\ p}$ 时，GPBoost 采用 Vecchia近似得到的恶性疟原虫 患病率预测。(a) 使用 $m_v = m_{v,\ p} = 30$，(b) 使用 $m_v = 30$和 $m_{v,\ p} = 150$，而 (c) 使用$m_v = m_{v,\ p} = 150$。(d)–(f) 展示了上述图中南部区域的情况，其中预测中的噪声更加明显。

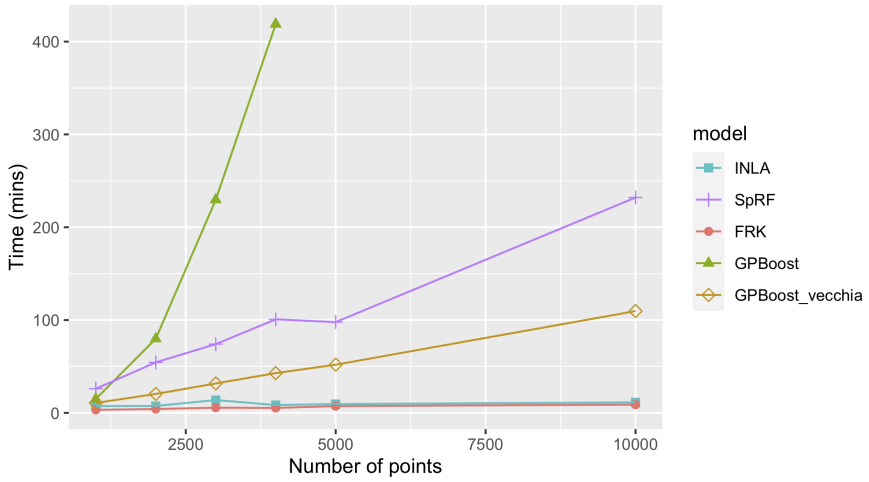


图22：对不同大小的模拟数据集应用 Vecchia近似时，GPBoost 所需的时间，并与图7中的耗时进行比较。