中国矿业大学（北京）

China University of Mining & Technology, Beijing

硕 士 学 位 论 文

空间广义线性混合效应模型及其应用

作　　者: 黄湘云

学　　院: 理学院

学　　号: TSP150701029

学科专业: 统计学

导　　师: 李再兴

2018 年 6 月

中图分类号:_____ 单位代码:_____

密　　级:_____

硕 士 学 位 论 文

中文题目:空间广义线性混合效应模型及其应用

英文题目:Spatial Generalized Linear Mixed Models and Its Applications

作　　者:黄湘云　　学　　号:TSP150701029

学科专业:统计学　　研究方向:数据分析与统计计算

导　　师:李再兴　　职　　称:教授

论文提交日期:2018 年 月 日　论文答辩日期:2018 年 月 日

学位授予日期:2018 年 月 日

中国矿业大学 (北京)

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中国矿业大学或其他教学机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

作者签名:_____ 日期:_____

关于论文使用授权的说明

本人完全了解中国矿业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅或借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

作者签名:_____ 导师签名:_____ 日期:_____

摘　要

空间广义线性混合效应模型具有广泛的应用，特别是空间统计领域，为了统计推断，相关的计算方法还没到收敛的状态。1998 年 Peter J. Diggle 实现基于马尔科夫链蒙特卡罗算法的贝叶斯估计，2002 年 Venables, W. N. 和 Ripley, B. D 实现惩罚拟似然估计，2004 年 Ole F Christensen 实现的蒙特卡罗最大似然估计，2009 年 Håvard Rue 实现近似贝叶斯推断方法—集成嵌套拉普拉斯算法。在大规模稀疏数据环境下，高效的计算方法一直是研究的重要方向。亮点在于实现了目前用以模型选择和统计推断的低秩近似、（限制）最大似然和近似贝叶斯算法，还在 Stan 框架下实现了基于贝叶斯推断的算法。并且通过模拟比较，得知低秩近似具有明显的效率优势，Stan 框架因其本身优化程度极高的计算库、并行特点和编译带来的再次优化大大加速了模拟的过程。

关键词：地质统计，空间广义线性混合效应模型，马尔科夫链蒙特卡罗

Abstract

Spatial generalized linear mixed effects model (SGLMM) has a wide range of applications, especially in the area of spatial statistics. For statistical inference, the relevant calculation methods have not yet reached the state-of-art. In 1998, Peter J. Diggle and his colleagues had bayesian estimation using Markov Chain Monte Carlo algorithms. In 2002, Venables, W. N. and Ripley, B. D fited SGLMM models via Penalized Quasi-Likelihood. In 2004, Ole F Christensen got Monte Carlo Maximum Likelihood estimations of SGLMMs. In 2009, an approximate bayesian inference — Integrated Nested Laplace Approximations was used to fit SGLMMs by Håvard Rue. In large-scale sparse settings, effective and efficient algorithms are always pursued by reseachers. Low-rank, likelihood-based and bayesian framework approaches are carried out by R language, stan library only for the latter. By comparison, the low-rank approximation has obvious efficiency advantages. The Stan framework greatly accelerates the simulation process due to its highly optimized computational library, parallel features, and re-optimization from compilation.

Key words： Geostatistics，Spatial Generalized Linear Mixed Models, Markov Chain Monte Carlo

目 录

1 绪论

1.1 研究意义

在热带地区，淋巴丝虫病和盘尾丝虫病（河盲病）是严峻的公共卫生问题，据世界卫生组织统计，在非洲撒哈拉以南、阿拉伯半岛和南美洲的 34 个国家约 2000 ～ 4000 万人感染河盲病[1]。例如，喀麦隆中部省份，Loa loa （导致河盲病的寄生虫）感染强度与疾病流行度之间存在线性关系，即 Loa loa 流行度越高感染强度越大[2]。1997 年，研究表明 Loa loa 流行度对应的高感染强度的临界值为 20%[3]。而研究个体水平的感染情况与群体水平流行度之间的关系有助于大规模给药[4]。

1.2 选题背景

空间广义线性混合效应模型（以下简称 SGLMM 模型）在地质统计中有着广泛的应用，如来自有限气象站点的污染物浓度测量，岩心样本的石油含量的评估，核污染浓度的空间分布[5]，冈比亚儿童的疟疾流行度的空间分布[6]，喀麦隆及其周边地区的热带眼线虫流行病的的空间分布[7]，以及对 LFOEP 项目（即 Lymphatic Filariasis and Onchocerciasis Elimination Programs）的决策支持[4]，Diggle 和 Giorgi 于 2016 年在 SGLMM 模型的基础上进行扩展，以适应三类新的调查数据，其一是组合随机调查数据和非随机调查数据（即潜在有偏的数据），以肯尼亚疟疾流行数据为例，组合了学校和社区的调查数据；其二是时空扩展，将时间因素考虑进模型，以马拉维 2010 年 5 月至 2013 年 6 月的疟疾流行数据为例；其三是混合分布，考虑响应变量是混合二项分布的情况[8]，检验环境和基因效应在空间相关性中的存在性[9]，流行现象的时空分析[10]。

1.3 文献综述

地质统计这个术语最初来自南非的采矿业[11]，并由 Georges Matheron 及其同事继承和发展，用以预测矿藏含量和质量。空间广义线性混合效应模型在这个地统计领域内通常又叫广义线性地统计模型。地质统计因其包含的广泛科学内容，逐渐被接受为空间统计的三大主流分支之一，其余两个是离散空间变差（discrete spatial variation）和空间点过程（spatial point processes）[12]。1994 年 Geyer 证明了蒙特卡罗最大似然（简称 MCML）积分的收敛性[13]，1998 年 Diggle 等人提出基于贝叶斯的空间统计方法应用于地质统计领域，分析了南太平洋岛上的核残留的情况，北拉纳克郡和南坎布里亚郡的弯曲杆菌感染情况[5]。随后，似然估计的统计性质和随机模拟算法的收敛性成为研究的重点：2002 年 Zhang 重点分析了空间广义线性混合效应模型的参数估计和模型预测的计算问题，应用 MCML[14] 算法求解模型；2004 年 Christensen 将 MCML 方法应用于朗格拉普岛的数据分析[15]。近年来，在大数据的背景下，寻求高效的算法成为一个新的方向，2009 年 Rue 等人提出基于近似贝叶

斯推断的集成嵌套拉普拉斯算法，简称 INLA[16]，并将其应用于空间数据建模[17]，还推广到一般的贝叶斯计算[18]。2016 年 Bonat 和 Ribeiro Jr. 综合比较了 MCML、贝叶斯 MCMC 和 INLA 方法[19]。同时，涉及空间数据分析和建模的书籍也越来越多，用于空间数据分析的分层模型[20] 和基于 R-INLA 软件的空间和时空贝叶斯模型[21]。

1.4 论文结构

第1章介绍论文相关背景，研究现状；第2章回顾了一般线性模型到广义线性混合效应模型的发展和模型结构；第3章介绍了求解空间广义线性混合效应模型的算法及相关软件实现的技术路线；第4 章展示算法模拟的结果；第5章给出基于 SGLMM 建模的案例分析；第6章给出全文总结和展望。

2 模型介绍

2.1 线性模型

线性模型的一般形式为

$$Y = X'\beta + \epsilon, \mathrm{E}(\epsilon) = 0, \mathrm{Cov}(\epsilon) = \sigma^2 I \tag{2.1}$$

其中，$Y = (y_1, y_2, \cdots, y_n)'$ 是 n 维列向量，代表对响应变量 Y 的 n 次观测；$\beta = (\beta_0, \beta_1, \cdots, \beta_{p-1})'$ 是 p 维列向量，代表模型自变量 X 的系数，β_0 是截距项；$X' = (1'_{(1\times n)}, X'_{(1)}, X'_{(2)}, \cdots, X'_{(n)})$，$1'_{(1\times n)}$ 是全 1 的 n 维列向量，而 $X'_{(i)} = (x_{1i}, x_{2i}, \cdots, x_{ni})'$ 代表对第 i 个自变量的 n 次观测；$\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)'$ 是 n 维列向量，代表模型的随机误差，$\mathrm{E}(\epsilon_i\epsilon_j) = 0, i \neq j$。求解线性模型 (2.1) 的 R 函数是 `lm`，近年来，高维乃至超高维稀疏线性模型成为热门的研究方向，相关的 R 包也越来越多，比较著名的有 `glmnet`[22] 和 `SIS`[23]。

2.2 广义线性模型

广义线性模型的一般形式

$$g(\mu) = X'\beta, \tag{2.2}$$

其中 $\mu \equiv \mathrm{E}(Y)$，g 代表联系函数，特别地，当 $Y \sim N(\mu, \sigma^2)$ 时，$g(x) = x$；当 $Y \sim \mathrm{Binomial}(n, p)$ 时，$g(x) = \ln(\frac{x}{1-x})$；当 $Y \sim \mathrm{Possion}(\lambda)$ 时，$g(x) = \ln(x)$；此处不一一列举[24]。模型 (2.2) 最早由 Nelder 和 Wedderburn[25] 提出，它弥补了模型 (2.1) 的两个重要缺点：一是因变量只能取连续值的情况，二是期望与自变量只能用线性关系联系[26]。求解广义线性模型 (2.2) 的 R 函数是 `glm`，参数估计的常用方法是拟似然法。

2.3 广义线性混合效应模型

广义线性混合模型的一般形式

$$g(\mu) = X'\boldsymbol{\beta} + Z'\mathbf{b} \tag{2.3}$$

其中，Z' 是 q 维随机效应的 $n \times q$ 的向量值矩阵，其它符号含义如前所述。混合效应模型包含线性混合效应模型、广义线性混合效应模型、广义可加混合效应模型、非线性混合效应模型等，之所以称之为混合效应，是因为模型既包含固定效应 $\boldsymbol{\beta}$ 又包含随机效应 $\mathbf{b}$。如前所述的线性和广义线性模型中的自变量就是固定效应，而随机效应是那些不能直接观察到的潜变量，但是对响应变量产生显著影响。求解模型 (2.3)的 R 包有 `nlme`[27]，`mgcv`[28] 和 `lme4`[29] 等，参数估计的方法一般有限制极大似然法。

2.4 空间广义线性混合效应模型

空间广义线性混合效应模型 (Spatial Generalized linear mixed-effects models，简写为 SGLMM)，顾名思义，它既是对模型 (2.1)、(2.2) 和 (2.3) 的延伸也是对空间数据的具体建模，属于空间分析的内容，在地质统计相关的文献中也称为广义线性地统计模型 (Generalized linear geostatistical models)[30]。

$$g(\mu_i) = T_i = d(x_i)'\beta + S(x_i) + Z_i \tag{2.4}$$

其中，$d'(x_i)$ 代表协变量对应的数据向量，$x_i \in \mathbb{R}^2$ 代表相应的空间位置，$d'(x_i)$ 即 p 个协变量在第 i 个位置的观察值。若响应变量 Y_i 服从二项分布 $\mathrm{Bin}(n_i, p_i)$ 则 $g(\mu_i) = \log[p(x_i)/\{1-p(x_i)\}]$，若响应变量 Y_i 服从泊松分布 $\mathrm{Pois}(\lambda_i)$ 则 $g(\mu_i) = \log(\lambda_i)$。此外，假定 $\mathcal{S} = \{S(x) : x \in \mathbb{R}^2\}$ 是均值为 0，方差为 σ^2，相关函数为 $\rho(x, x') = \mathrm{Cov}\{S(x), S(x')\}$ 的高斯过程；随机过程 $\mathcal{S}$ 平稳且各向同性，即 $\rho(x, x') = \mathrm{Corr}\{S(x), S(x')\} \equiv \rho(\|x, x'\|)$，$\|\cdot\|$ 表示距离，样本之间的位置间隔不大就用欧式距离，间隔很大就用球面距离；$S(x_i)$ 代表了与空间位置 x_i 相关的随机效应，简称空间效应；这里，Z_i 是相互独立且服从 $N(0, \tau^2)$ 的随机变量。

2.4.1 相关函数的选择

如前所述，模型(2.4) 包含的空间效应主要由相关函数决定，在给出相关函数之前，先计算一下空间效应的理论变差 $V(x, x')$（即空间过程的协方差函数的一半，变差源于采矿术语）和线性预测的变差 $V_T(u_{ij})$。为方便起见，记 $\rho(u) \triangleq \rho(\|x, x'\|), u \equiv \|x - x'\|$

$$\begin{aligned}
V(x, x') &= \frac{1}{2}\mathrm{Var}\{S(x) - S(x')\} \\
&= \frac{1}{2}\mathrm{Cov}(S(x) - S(x'), S(x) - S(x')) \\
&= \frac{1}{2}\{E[S(x) - S(x')][S(x) - S(x')] - [E(S(x) - S(x'))]^2\} \\
&= \sigma^2 - \mathrm{Cov}(S(x), S(x')) = \sigma^2\{1 - \rho(u)\} \\
V_T(u_{ij}) &= \frac{1}{2}\mathrm{Var}\{T_i(x) - T_j(x)\} = \frac{1}{2}\mathrm{E}[(T_i - T_j)^2] = \tau^2 + \sigma^2(1 - \rho(u_{ij}))
\end{aligned} \tag{2.5}$$

从方程 (2.5) 不难看出系数 $\frac{1}{2}$ 的化简作用，随机向量 T 的协方差矩阵如下：

$$\mathrm{Cov}(T_i(x), T_i(x)) = \sigma^2 + \tau^2, \mathrm{Cov}(T_i(x), T_j(x)) = \sigma^2\rho(u_{ij})$$

相关函数 $\rho(u)$ 的作用和地位就显而易见了，它是既决定理论变差又决定协方差矩阵的结构。常见的相关函数族有梅隆族：

$$\rho(u) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa}\mathcal{K}_{\kappa}(u/\phi), u > 0 \tag{2.6}$$

一般地，通常假定 $\rho(u)$ 单调不增，即任何两样本之间的相关性应该随距离变大而减弱，尺度参数 ϕ 控制函数 $\rho(u)$ 递减到 0 的速率，方便起见记 $\rho(u)=\rho_0(u/\phi)$，则方程(2.6)可记为

$$\rho_0(u)=\{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u)^{\kappa}\mathcal{K}_{\kappa}(u), u>0 \tag{2.7}$$

其中，$\mathcal{K}_{\kappa}(\cdot)$ 是阶数为 κ 的第二类修正的贝塞尔函数，$\kappa(>0)$ 是平滑参数，满足这些条件的空间过程 $\mathcal{S}$ 是 $\lceil\kappa\rceil-1$ 次均方可微的。

值得注意的是梅隆参数族包含指数族（这里指带有参数的指数函数），即当 $\kappa=0.5$ 时，$\rho_0(u)=\exp(-u)$，$S(x)$ 均方连续但是不可微，当 $\kappa\to\infty$ 时，$\rho_0(u)=\exp(-u^2)$，$S(x)$ 无限次均方可微。要从数据中估计 κ，为了节省计算，又不失一般性，经验做法是取离散的 κ 先验，如 $\kappa=0.5,1.5,2.5$，分别对应 $S(x)$ 均方连续、一次可微和二次可微。$S(x)$ 的性质实际代表了空间过程 $\mathcal{S}$ 的曲面平滑程度。

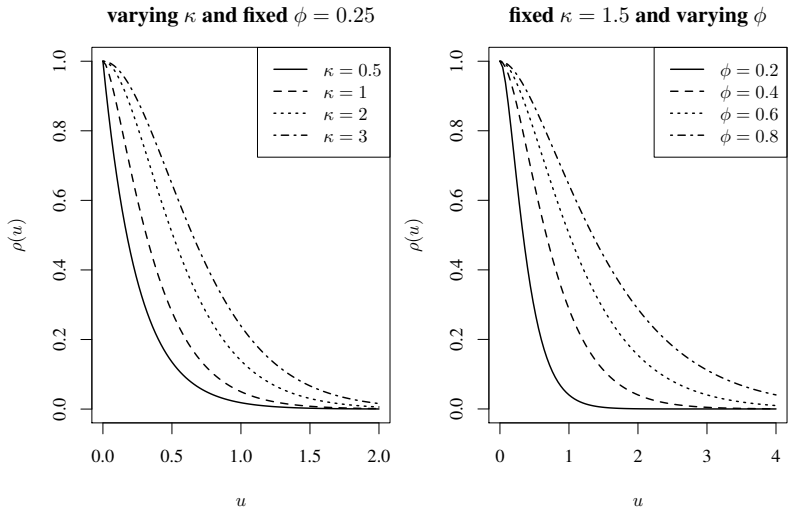

图 2.1: 固定尺度参数，相关函数随距离的变化（图左）；固定贝塞尔函数的阶，相关函数随距离的变化（图右）

从图2.1 和图 2.2 可以看出，相比于贝塞尔函数的阶 κ，尺度参数 ϕ 对相关函数的影响大些，所以在实际应用中，先固定下 κ 是可以接受的。此外，Diggle 等人于 1998 年使用幂指数族 $\rho_0(u)=\exp(-u^{\delta}), 0<\delta\leq 2$ 作为相关函数[5]，因其形式大大简化，函数图像和性质却与梅隆族相似，即当 $0<\delta<2$ 时，$S(x)$ 均方连续但不可微，当 $\delta=2$ 时，$S(x)$ 无限次可微。

2.4.2 模型识别

模型中 Z_i 与 $S(x_i)$ 项的可识别问题：向量 $T=(T_1,T_2,\cdots,T_n)$ 是协方差为矩阵 $\tau^2I+\sigma^2R$ 的多元高斯分布，其中 $R_{ij}=\rho(u_{ij};\phi)$，u_ij 是 x_i 与 x_j 之间的距离，由(2.5)知，随机过程 $T(x)$ 的相关函数在原点不连续。只要指定参数，使得 $\rho(u)$ 在原点连续，则参数 τ^2,σ^2,ϕ 就都是可识别的，显然这依赖于抽样的位置 x_i[6]。

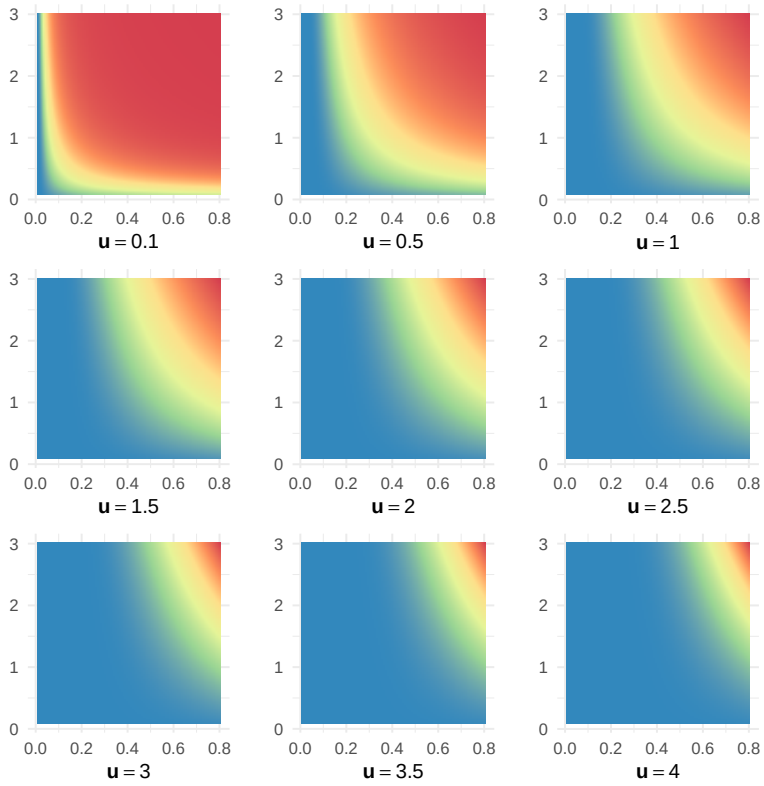

图 2.2: 相关函数随贝塞尔函数的阶和尺度参数的变化，横轴表示尺度参数，纵轴表示贝塞尔函数的阶（从蓝到红，相关性由弱变强）

2.4.3 先验分布

基于贝叶斯框架实现模型(2.4)的参数估计和预测，必然使用 MCMC 算法，自然地，需要指定模型参数 $\boldsymbol{\theta} = (\beta, \tau^2, \sigma^2, \phi)$ 的先验分布，对于 β，选择相互独立的均匀先验，而对于参数 τ^2, σ^2, ϕ，选取如下模糊先验：

$$f(\tau^2) \propto \frac{1}{\tau^2}; f(\sigma^2) \propto \frac{1}{\sigma^2}; f(\phi) \propto \frac{1}{\phi^2}$$

其中，τ^2 和 σ^2 为杰弗里斯先验，这些先验的选择是出于实用和经验的考虑（意思就是说可以取别的），如果由这些先验导出的后验不合适，则 MCMC 算法的表现就会不收敛；通常选取不同初始值，产生多条链，如果没有出现算法不收敛的情况，则这样的先验是被合适的。这些无信息的先验分布的选择对最终结果几乎没有影响，这是贝叶斯非常棒的部分，贝叶斯推断方法也得以被广泛应用[31]。

作为模型 (2.4) 求解和展示的首选工具 — R 语言在空间数据分析与可视化方面呈现越来越流行的趋势，从早些年的 `lattice` 图形[32] 到如今的 `ggplot2` 图形[33]，操作空间数据的 `sp` 对象[34] 也发展为 `sf` 对象[35]，同时整合了不少第三方软件和服务，如基于 Google Maps 的交互空间可视化[36]，基于 Google Earth 的空间可视化[37]。下面就求解模型 (2.4)的三类算法进行详细阐述，分别是贝叶斯方法、似然方法和低秩近似方法，并介绍相应的软件实现。

3 算法综述

3.1 贝叶斯

1998 年 Diggle 等人最早提出基于模型的地质统计学框架，将高斯空间随机过程和（广义）线性混合模型结合应用到空间流行病数据分析中，通过贝叶斯推断方法进行参数估计和预测[5]。2002 年 Diggle 等人使用空间广义线性混合模型分析冈比亚儿童疟疾的数据，在贝叶斯框架下，通过 Metropolis-Hastings 采样算法实现 MCMC 方法进行参数估计和模型预测[6]。

3.2 最大似然

由于贝叶斯方法构造马尔科夫链，需要很多次反复迭代，收敛速度慢，求解模型(2.4)需要花费很多时间，将最大似然和重要性采样相结合的方法出现了，称之为蒙特卡罗最大似然法，简称 MCML。1994 年 Charles J. Geyer 首先从理论证明 MCML 方法的收敛性及相关似然估计的渐进正态性，其中包括 profile 似然、近似似然和精确似然等，为后续算法的开发、改进以及应用提供了理论支持[13]。2002 年张在做模型的估计和预测的时候提出蒙特卡罗–期望极大梯度 (Monte Carlo EM Gradient) 算法，简称 MCEMG[14]。2016 年 Hosseini 在 MCEMG 的基础上提出近似蒙特卡罗–期望极大梯度 (Approximate Monte Carlo EM Gradient) 算法，简称 AMCEMG[38]。2004 年 Ole F Christensen 将 MCML 方法用于地质统计模型[15]，2016 年 Peter J. Diggle 和 Emanuele Giorgi 将 MCML 方法应用于分析西非流行病调查数据。

为描述方便起见，令 $\boldsymbol{\theta}^\top = (\sigma^2, \phi, \tau^2)$，这里和之前先验分布一节的 $\boldsymbol{\theta}$ 含义一样，是模型参数构成的一个集合，只是没有将 β 纳入进来，因为在贝叶斯框架下，要求对所有未知参数给定先验分布，包括模型系数。D 表示 $n \times p$ 的协变量矩阵, $y^\top = (y_1, y_2, \cdots, y_n)$，$T$ 的边际分布是 $N(D\beta, \Sigma(\theta))$。给定 $T^\top = t^\top = (t_1, t_2, \cdots, t_n)$ 下，$Y^\top = (Y_1, \cdots, Y_n)$ 的条件分布是独立二项概率分布函数的乘积 $f(y|t) = \prod_{i=1}^n f(y_i|t_i)$，则 β 和 $\boldsymbol{\theta}$ 的似然函数可以写成：

$$
\begin{aligned}
L(\beta, \theta) &= f(y; \beta, \boldsymbol{\theta}) \\
&= \int_{\mathbb{R}^n} N(t; D\beta, \Sigma(\theta)) f(y|t) dt \\
&= \int_{\mathbb{R}^n} \frac{N(t; D\beta, \Sigma(\theta)) f(y|t)}{N(t; D\beta_0, \Sigma(\theta_0)) f(y|t)} f(y, t) dt \\
&\propto \int_{\mathbb{R}^n} \frac{N(t; D\beta, \Sigma(\theta))}{N(t; D\beta_0, \Sigma(\theta_0))} f(t|y) dt \\
&= E_{T|y}\left[\frac{N(t; D\beta, \Sigma(\theta))}{N(t; D\beta_0, \Sigma(\theta_0))}\right]
\end{aligned}
\tag{3.1}
$$

其中，β_0, θ_0 给定，Y 和 T 的联合分布是 $f(y, t) = N(t; D\beta_0, \Sigma(\theta_0)) f(y|t)$，再使用 MCMC 算法从条件分布 $f(T|Y = y; \beta_0, \theta_0)$ 抽取 m 个样本 $t_{(i)}$，那么，可以用

如下方程近似 (3.1)

$$L_m(\beta,\theta)=\frac{1}{m}\sum_{i=1}^{n}\frac{N(t_i;D\beta,\Sigma(\theta))}{N(t_i;D\beta_0,\Sigma(\theta_0))} \tag{3.2}$$

给定合适的初始值 β_0 和 θ_0，用 $\hat{\beta}_m$ 和 $\hat{\theta}_m$ 表示最大化 $L_m(\beta,\theta)$ 获得的 MCML 估计，重复迭代 $\beta_0=\hat{\beta}_m$ 和 $\theta_0=\hat{\theta}_m$ 直到收敛。最大化 $L_m(\beta,\theta)$ 的过程中，我们可以选择 BFGS 算法。

3.3 低秩近似

模型(2.4)中 $S(x_i)$ 来自高斯过程 $\mathcal{S}=S(x),x\in\mathbb{R}^2$，可以被表示成高斯噪声的卷积形式

$$S(x)=\int_{\mathbb{R}^2}K(\|x-t\|;\phi,\kappa)\,dB(t) \tag{3.3}$$

其中，B 表示布朗运动，$\|\cdot\|$ 表示欧氏距离，$K(\cdot)$ 表示梅隆核，形如

$$K(u;\phi,\kappa)=\frac{\Gamma(\kappa+1)^{1/2}\kappa^{(\kappa+1)/4}u^{(\kappa-1)/2}}{\pi^{1/2}\Gamma((\kappa+1)/2)\Gamma(\kappa)^{1/2}(2\kappa^{1/2}\phi)^{(\kappa+1)/2}}\mathcal{K}_\kappa(u/\phi),u>0. \tag{3.4}$$

通过离散方程 (3.3)，并且让 r 充分大，可以获得低秩近似

$$S(x)\approx\sum_{i=1}^{r}K(\|x-\tilde{x}_i\|;\phi,\kappa)Z_i, \tag{3.5}$$

其中，$(\tilde{x}_1,\ldots,\tilde{x}_r)$ 表示空间网格的格点，Z_i 是独立同分布的高斯变量，均值为 0，方差 σ^2，特别地，当尺度参数 ϕ 比较大的时候，这种近似变得很有效，如图2.1所示，ϕ 越大，空间曲面越平缓，即使格点数目 r 比较小也能得到很好的效果。此外，空间格点数目 r 与样本量 n 是独立的，因此这种方法在大样本的时候，很有计算上的吸引力。对于具有复杂空间结构的模型(2.4)，保持高计算效率是一个非常有意义的方面。

3.4 相关软件

Stan 是一种概率编程语言[39]，可以替代 BUGS (**B**ayesian inference **U**sing **G**ibbs **S**ampling)[40] 作为 MCMC 的高效实现，可用于贝叶斯框架下，标准地质统计模型的参数估计，Stan 提供多种语言的接口实现，方便起见，本文采用它提供的 R 语言接口 – rstan 包[41]。基于 GPU 加速是一个不错的选择，Stan 开发者也把 GPU 加速列入开发日程。SCIKIT-CUDA[42] 和 ArrayFire[43] 等基于 CUDA 开发的通用加速框架获得越来越多的关注。

R 语言作为自由的统计计算和绘图环境，因其免费，更新快，社区庞大，扩展包更是多达 12500 个，提供了大量的前沿统计技术的代码实现。如用于一元和多元时空模型选择和预测的 spBayes 包[44]；可以对 MCMC 的输出进行诊断和分析的 coda 包[45]；MCMCvis 包提取模型参数，MCMC 算法输出的结果并可视化，产生出版级的图形，支持转化 JAGS、Stan 和 BUGS 软件输出结果[46]；基于贝叶斯方法的

空间线性混合效应模型选择和预测的 geoR 包[47]，geoRglm 包将其扩展到空间广义线性混合效应模型[48]；glmmBUGS 包提供 WinBUGS、OpenBUGS 和 JAGS 软件的统一接口，使求解 BUGS 模型的过程放在 R 环境中[49;50]；gstat 包是迁移自 S 语言的地质统计扩展包，提供了各种各样的克里金插值方法[51;52]；brms 包基于 Stan 框架拟合贝叶斯广义线性和非线性混合效应模型[53]。

4　数值模拟

模拟的空间广义线性混合效应模型分别是(4.1)和 (4.2)。RandomFields 可以模拟多元随机场[54]，geoR 包[55] 的 `grf` 函数只适合模拟少量样本点（$n < 500$），MASS 包的 `glmmPQL` 函数采用惩罚拟似然求解模型[56]。模型参数设置为 $n = 1600, \sigma^2 = 1, \phi = 25, \tau^2 = 1, \kappa = 1, \beta_0 = 1.2$，图 4.1 模拟规则网格上的采样，图4.2 模拟随机采样，图 4.4 基于 INLA 方法模拟三角网格。每个格点上重复实验 10 次，得到响应变量二项分布的概率值。REML 方法图4.6，INLA 方法图 4.5 ，低秩近似和 MCML 方法图 4.3

$$\log\left\{\frac{p(x_i)}{1-p(x_i)}\right\} = d(x_i)'\beta + S(x_i) + Z_i \tag{4.1}$$

$$\log[\lambda(x_i)] = d(x_i)'\beta + S(x_i) + Z_i \tag{4.2}$$

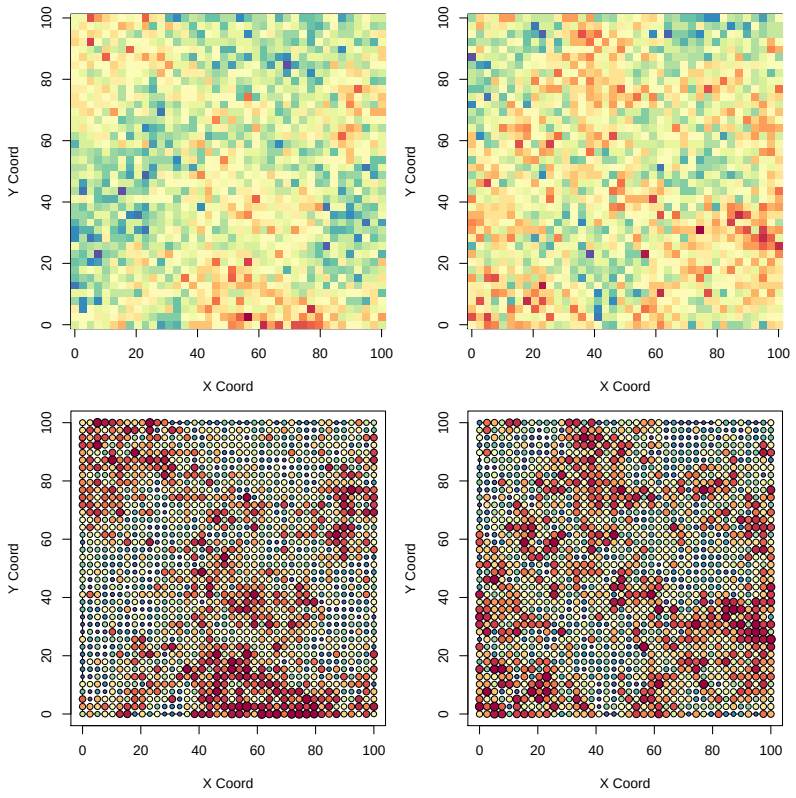

图 4.1: 模拟高斯过程：核函数分别为指数族（左图），梅隆族（右图）

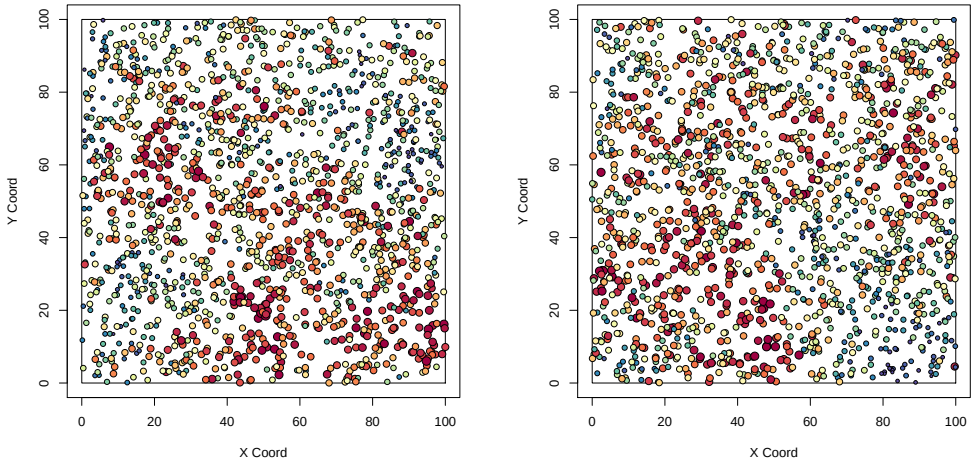

图 4.2: 模拟高斯过程：核函数分别为指数族（左图），梅隆族（右图）

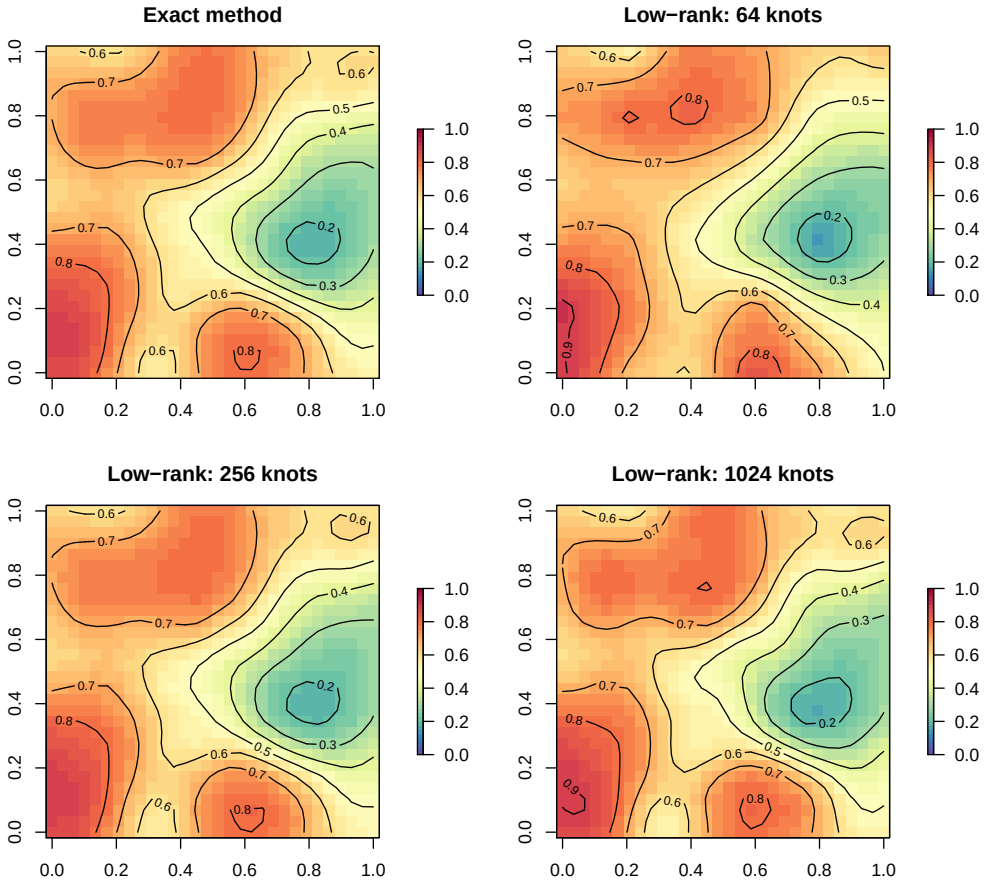

图 4.3: 低秩近似方法与精确蒙特卡罗最大似然方法

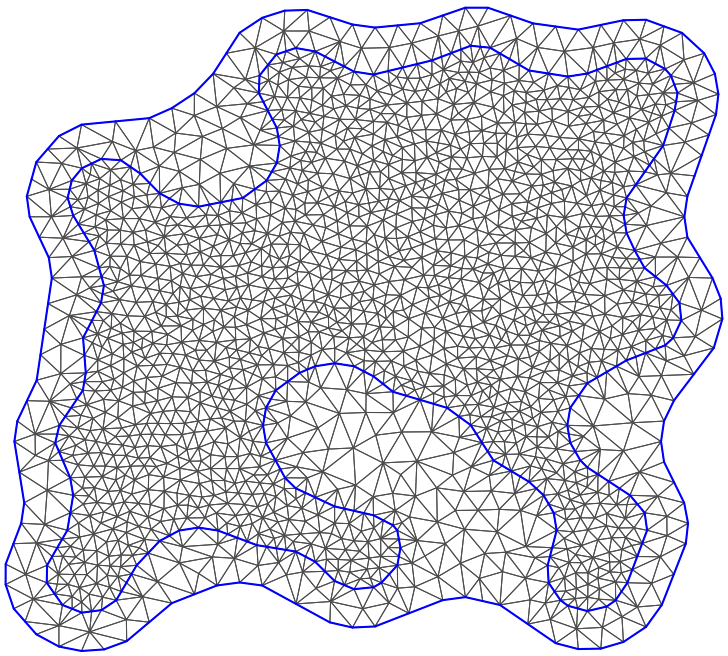

图 4.4: 基于 INLA 的三角网格划分

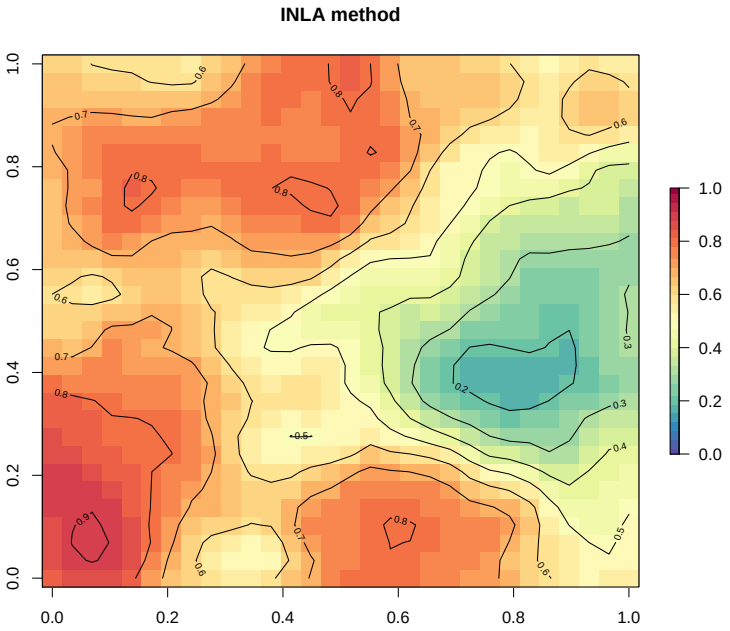

图 4.5: INLA 方法

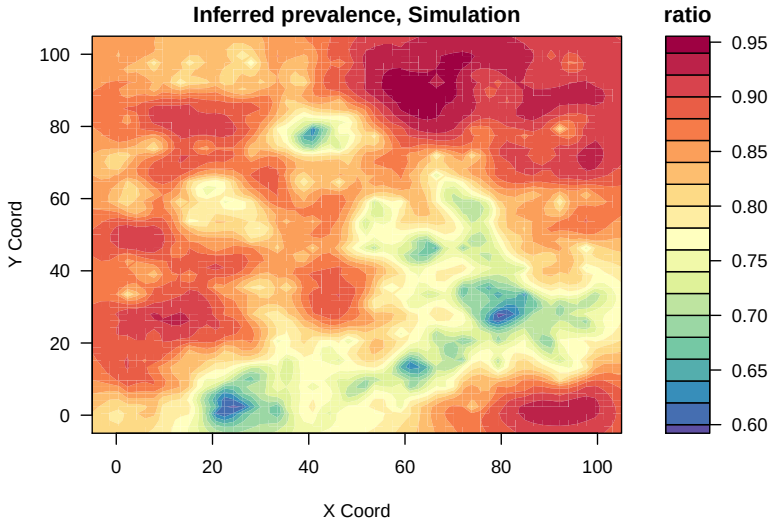

图 4.6: 规则网格上预测结果（REML 方法）

5 案例分析

5.1 喀麦隆及周边地区眼线虫病的空间分布

Loa loa (eyeworm) 是一种可致盲的热带疾病，APOC (African Programme for Onchocerciasis Control) 搜集了 168 个村庄的 21938 个样本，另外在研究区域 1 公里的范围内添加了样本周围的环境变量[57]，从美国地质调查局获得海拔信息 (https://www.usgs.gov/)，以及来自卫星数据的植被绿色度 (http://free.vgt.vito.be)。如表5.1所示

表 5.1: Loa loa 数据集（部分）

LONGITUDE	LATITUDE	NO_EXAM	NO_INF	ELEVATION
8.04	5.74	162	0	108
8.00	5.68	167	1	99
8.91	5.35	88	5	783
8.10	5.92	62	5	104
8.18	5.11	167	3	109
8.93	5.36	66	3	909

搜集的 Loa loa 数据的空间分布，如图5.1，其中圆圈的大小分六个等级：0.5, 1.0, 1.5, 2.0, 2.5, 3.0 分别对应 Loa loa 流行度的六个区间: [0,0.05), [0.05,0.15), [0.15,0.25), [0.25,0.35), [0.35,0.45), [0.45,0.55)，这里超过 0.2，就列为高感染，需要对该地区采取措施，如派遣医疗队和药品等。

建立模型 (2.4) $\log\{p_{ij}/(1-p_{ij})\} = \alpha + \beta' z_{ij} + U_i + S(x_i)$, 基于限制极大似然估计，计算得到固定效应参数如下表

参数	估计	条件标准差	t 统计量
(Intercept)	-1.009e+01	2.9790516	-3.3874
elev1	-2.825e-05	0.0006196	-0.0456
elev2	8.087e-04	0.0014786	0.5469
elev3	-1.138e-02	0.0025495	-4.4629
elev4	1.067e-02	0.0031547	3.3814
maxNDVI1	1.072e+01	2.7334338	3.9221
seNDVI	-2.906e+00	4.4210991	-0.6574

随机效应的参数 $\nu = 0.24326, \phi = 0.01345, \sigma^2 = 6.236$，相应的空间预测结果

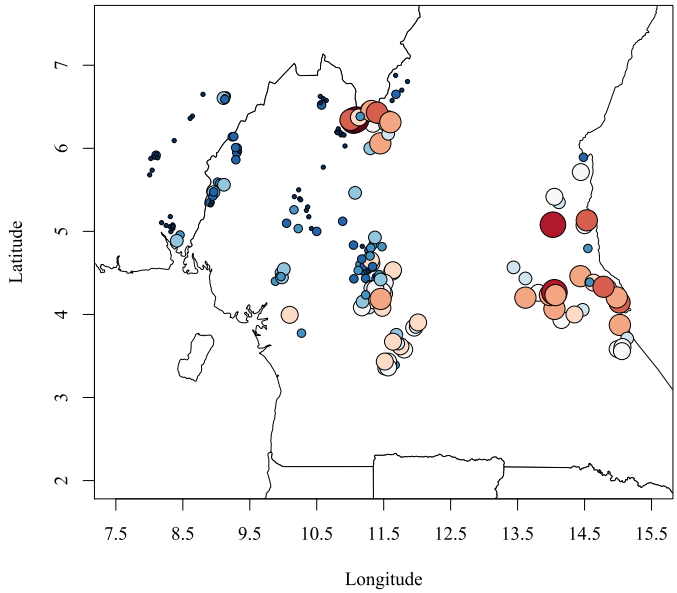

图 5.1: Loa loa 流行度观测结果，黑点是采样点

如图 5.2所示

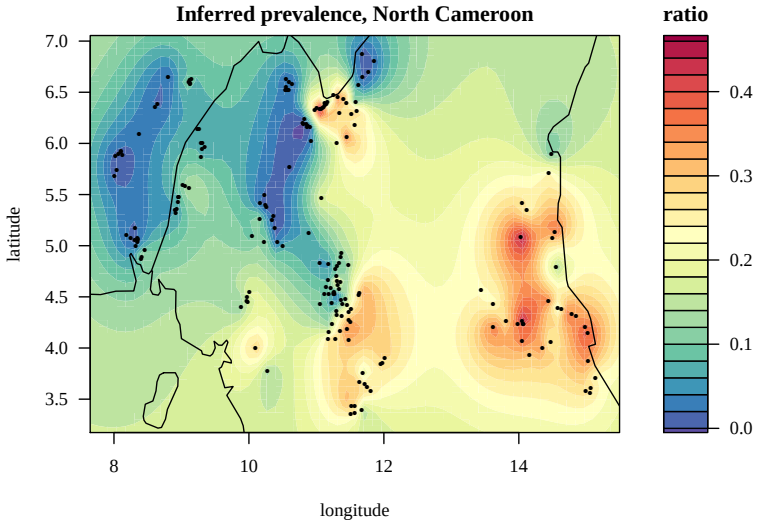

图 5.2: Loa loa 数据集上的预测结果

5.2 冈比亚儿童疟疾的空间分布

2002 年，Peter Diggle 等人分析过冈比亚儿童疟疾数据，该数据采集自冈比亚的 5 个地域，65 个村庄，2035 个 5 岁以下儿童的血液样本，如图5.3。

并记录了他们的年龄、村庄的位置（GPS 坐标）、血液中是否含有疟疾寄生虫、蚊帐是否使用、蚊帐是否杀虫、村庄周围绿色植物的覆盖度（RS 测量）、村庄是否有医疗中心[6]。调查所得的数据如表5.4 所示（篇幅所限展示部分）。数据各指标说

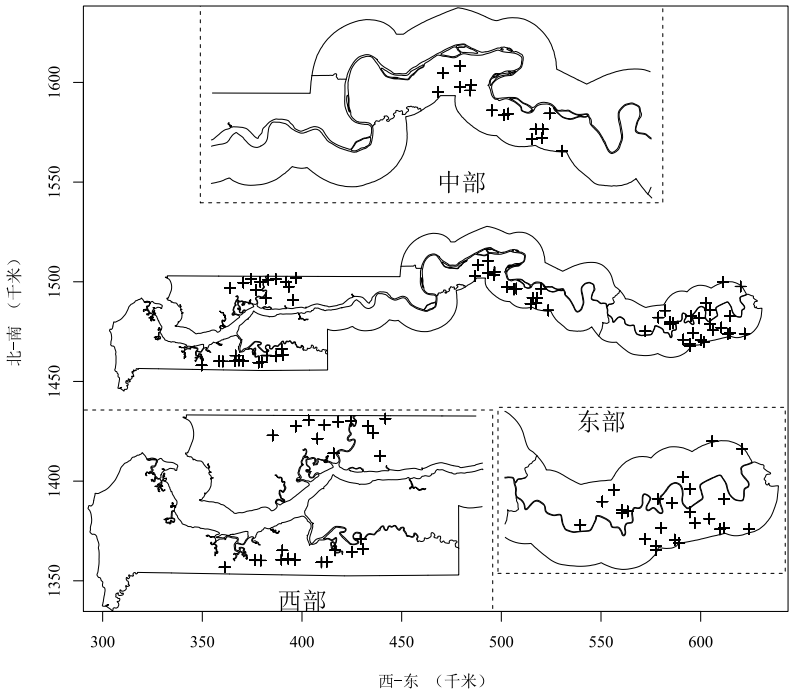

图 5.3: 采样的村庄

明如下：

变量	含义
(x, y)	村庄的坐标
pos	血样中是否出现寄生虫（1 表示是，0 表示否）
age	儿童的年龄（按天计算）
netuse	儿童是否睡在蚊帐中（1 表示是，0 表示否）
treated	蚊帐是否杀虫（1 表示是，0 表示否）
green	村庄附近的绿色植物的覆盖度
phi	村庄里是否有医疗中心（1 表示有，0 表示没有）

表 5.4: 冈比亚儿童疟疾数据（部分）

x	y	pos	age	netuse	treated	green	phc
349631	1458055	1	1783	0	0	40.9	1
349631	1458055	0	404	1	0	40.9	1
349631	1458055	0	452	1	0	40.9	1
349631	1458055	1	566	1	0	40.9	1
349631	1458055	0	598	1	0	40.9	1
349631	1458055	1	590	1	0	40.9	1

在建模之前，这些搜集的数据对疟疾产生的影响可以通过探索性分析获得直观的认识，植被覆盖度、蚊帐以及杀虫对疟疾流行度的影响，分别如图5.4和图5.5所示。总体上，植被越茂盛，疟疾流行度越大，医疗中心对疟疾的控制作用比较有限，主要原因是冈比亚医疗卫生条件太差。蚊帐对疟疾的预防效果很好，没有使用蚊帐的人群中感染比例接近 50%，而使用了蚊帐的人群中比例降至 30.5%，此外，对蚊帐杀虫也有很好的保护效果。

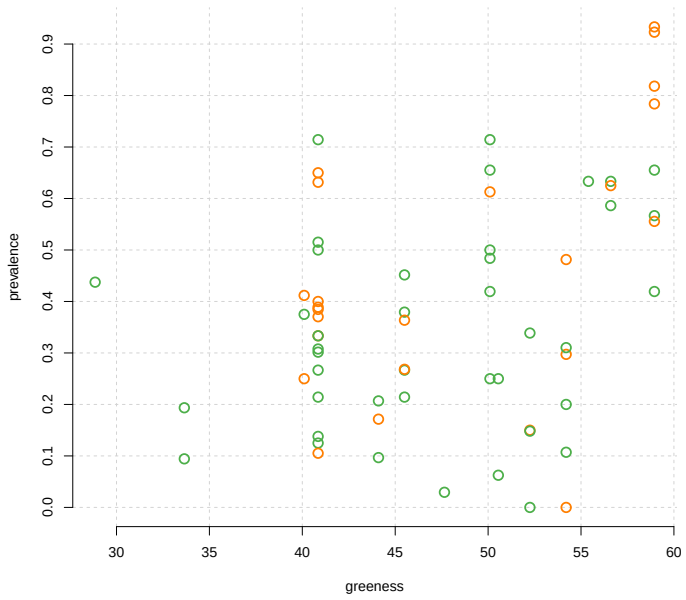

图 5.4: 疟疾流行度与村庄周围的植被覆盖度：绿色表示有医疗中心，橘黄色表示没有

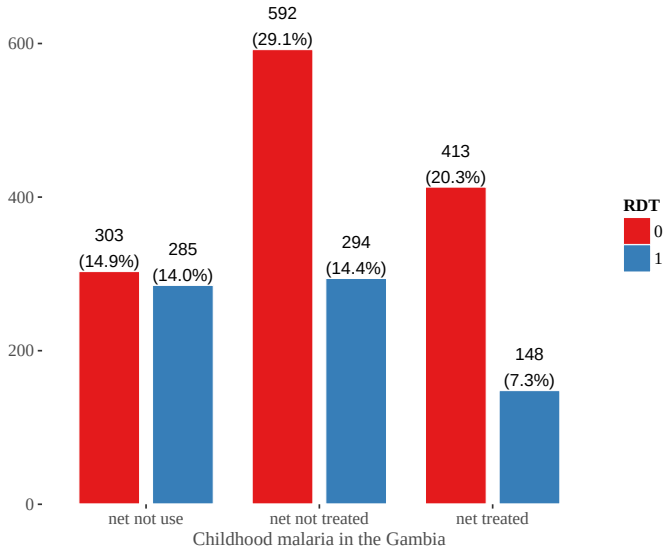

图 5.5: 蚊帐与杀虫对疟疾流行度的影响（是否有蚊帐，蚊帐是否杀虫，RDT 表示快速诊断结果：0 表示没有感染疟疾，1 表示感染疟疾）

为进一步作出定量分析，建立模型如下：

$$\log\{p_{ij}/(1-p_{ij})\} = \alpha + \beta' z_{ij} + U_i + S(x_i) \tag{5.1}$$

其中，z_{ij} 表示对第 i 个村庄的第 j 个儿童的观测值，如前所述的年龄、蚊帐使用情况等固定效应，相应地，p_{ij} 表示感染疟疾的概率。$S(x_i)$ 表示空间随机效应，U_i 表示除空间效应以外的村庄水平上的变化，也是随机效应。固定效应 β 和截距项 α 结果如下

参数	估计	条件标准差	t 统计量
(Intercept)	-1.182665	2.491648	-0.4747
avg_age	0.002613	0.001544	1.6930
netuse	-0.023868	0.009401	-2.5390
treated	-0.002662	0.009053	-0.2940
green	-0.026757	0.028247	-0.9473
phc	-0.376871	0.247394	-1.5234

此外，$\nu = 0.16465, \phi = 0.000367, \sigma^2 = 2.705$，相应的空间预测如图 5.6 所示

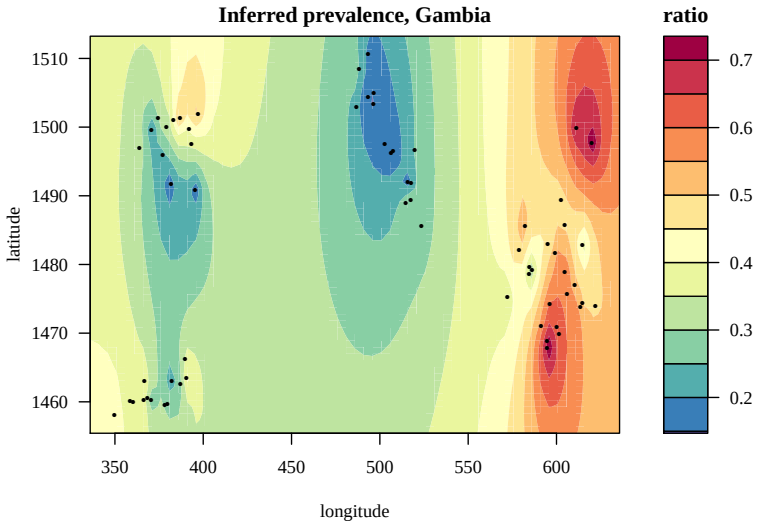

图 5.6: 冈比亚儿童疟疾空间分布预测

6 结论与展望

近年来，近似贝叶斯推断受到越来越多的关注，因其高效的计算性能，快速发展的 INLA 社区[1]，R-INLA 软件的广泛使用和日益成熟的理论。可以将近似贝叶斯推断用于空间数据建模和分析[17]，集成嵌套拉普拉斯和蒙特卡罗算法的结合也是值得研究的方向[58]，Stan 程序库在 GPU 上的并行也是提高计算效率的可行方向[2]。

数据模拟和案例分析的部分，还可以增加响应变量服从指数族其它分布的情形，如泊松分布。算法性能的比较可以同时考虑时间和计算平台，记录多次运行同一个算法的时间数据，比较它们所耗时间的分布差异，可以获得更加可靠的结果。计算平台如多核，多线程，甚至集群环境的实现和比较，可以获得算法扩展性方面的结论。此外，不能单纯看算法实现的语言方式，从文中的计算结果来看，R 语言的性能是弱于 C++ （Stan 是基于 C++ 的计算库，需要先编译源码和加载动态链接库）的，但是利用 R 编程可以快速实现算法原型。INLA 算法和软件非常高效的表现，得益于随机偏微分方程已有的实现算法和近似手段，如三角网格划分，迭代格式让算法有更快的收敛速度，近似效果没有 MCML 和 Low-Rank 好。由此可知，算法的选择需要去做效果和效率的平衡，Stan 更新迭代的速度很快，可在不久的将来进入应用界，但是却也要求更多的学习成本和优化技巧。

[1] http://www.r-inla.org/

[2] https://github.com/stan-dev/stan/wiki/Longer-Term-To-Do-List

参考文献

[1] Takougang I, Meremikwu M, Wandji S, et al. Rapid assessment method for prevalence and intensity of loa loa infection.[J]. Bulletin of the World Health Organization, 2002, 80(11): 852–858.

[2] Boussinesq M, Gardon J, Kamgno J, et al. Relationships between the prevalence and intensity of loa loa infection in the central province of cameroon.[J]. Annals of Tropical Medicine and Parasitology, 2001, 95(5): 495–507.

[3] Gardon J, Gardon-Wendel N, Demanga-Ngangue, et al. Serious reactions after mass treatment of onchocerciasis with ivermectin in an area endemic for loa loa infection.[J]. Lancet, 1997, 350(9070): 18–22.

[4] Schlüter D K, Ndeffombah M L, Takougang I, et al. Using community-level prevalence of loa loa infection to predict the proportion of highly-infected individuals: Statistical modelling to support lymphatic filariasis and onchocerciasis elimination programs[J]. Plos Neglected Tropical Diseases, 2016, 10(12): 1–15.

[5] Diggle P J, Tawn J A, Moyeed R A. Model-based geostatistics[J]. Journal of the Royal Statistical Society, Series C, 1998, 47(3): 299–350.

[6] Diggle P, Moyeed R, Rowlingson B, et al. Childhood malaria in the gambia: a case-study in model-based geostatistics[J]. Journal of the Royal Statistical Society, Series C, 2002, 51(4): 493–506.

[7] Diggle P J, Thomson M C, Christensen O F, et al. Spatial modelling and the prediction of loa loa risk: decision making under uncertainty[J]. Annals of Tropical Medicine and Parasitology, 2007, 101(6): 499–509.

[8] Diggle P J, Giorgi E. Model-based geostatistics for prevalence mapping in low-resource settings[J]. Journal of the American Statistical Association, 2016, 111(515): 1096–1120.

[9] Rousset F, Ferdy J B. Testing environmental and genetic effects in the presence of spatial autocorrelation[J]. Ecography, 2014, 37(8): 781–790.

[10] Meyer S, Held L, Höhle M. Spatio-temporal analysis of epidemic phenomena using the R package surveillance[J]. Journal of Statistical Software, 2017, 77(11): 1–55.

[11] Krige D G. A statistical approach to some basic mine valuation problems on the witwatersrand[J]. Journal of the Chemical, Metallurgical and Mining Society of South Africa, 1951, 52: 119–139.

[12] Cressie N A C. Statistics for spatial data[M]. Rev. ed. London: John Wiley and Sons, Inc., 1993: 27–104

[13] Geyer C J. On the convergence of monte carlo maximum likelihood calculations[J]. Journal of the Royal Statistical Society, Series B, 1994, 56(1): 261–274.

[14] Zhang H. On estimation and prediction for spatial generalized linear mixed models[J]. Biometrics, 2002, 58(1): 129–36.

[15] Christensen O F. Monte carlo maximum likelihood in model-based geostatistics[J]. Journal of Computational and Graphical Statistics, 2004, 13(3): 702–718.

[16] Rue H, Martino S, Chopin N. Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations (with discussion).[J]. Journal of the Royal Statistical Society, Series B, 2009, 71(2): 319–392.

[17] Lindgren F, Rue H. Bayesian spatial modelling with R-INLA[J]. Journal of Statistical Software, 2015, 63(19): 1–25.

[18] Rue H, Riebler A I, Sørbye S H, et al. Bayesian computing with INLA: A review[J]. Annual Reviews of Statistics and Its Applications, 2017, 4(1): 395–421.

[19] Bonat W H, Ribeiro Jr. P J. Practical likelihood analysis for spatial generalized linear mixed models[J]. Environmetrics, 2016, 27(2): 83–89.

[20] Banerjee S, Carlin B P, Gelfand A E. Hierarchical modeling and analysis for spatial data[M]. Second ed. Boca Raton, Florida: Chapman and Hall/CRC, 2015

[21] Marta Blangiardo M C. Spatial and spatio-temporal bayesian models with R-INLA[M]. Chichester, UK: John Wiley and Sons, 2015

[22] Simon N, Friedman J, Hastie T, et al. Regularization paths for cox's proportional hazards model via coordinate descent[J]. Journal of Statistical Software, 2011, 39(5): 1–13.

[23] Saldana D F, Feng Y. SIS: An R package for sure independence screening in ultrahigh dimensional statistical models[J]. Journal of Statistical Software, 2018, 83(2): 1–25.

[24] McCullagh P, Nelder J. Generalized linear models[M]. Second ed. Boca Raton, Florida: Chapman and Hall/CRC, 1989

[25] Nelder J A, Wedderburn R W M. Generalized linear models[J]. Journal of the Royal Statistical Society, Series A, 1972, 135(3): 370–384.

[26] 陈希孺. 广义线性模型的拟似然法[M]. 合肥: 中国科学技术大学出版社, 2011: 002–004

[27] Pinheiro J, Bates D, DebRoy S, et al. nlme: Linear and nonlinear mixed effects models [EB/OL]. 2018. https://CRAN.R-project.org/package=nlme.

[28] Wood S N. Generalized additive models: An introduction with R[M]. Second ed. Boca Raton, Florida: Chapman and Hall/CRC, 2017

[29] Bates D, Mächler M, Bolker B, et al. Fitting linear mixed-effects models using lme4 [J]. Journal of Statistical Software, 2015, 67(1): 1–48.

[30] Diggle P J, Ribeiro Jr. P J. Model-based geostatistics[M]. New York: Springer-Verlag, 2007

[31] 茆诗松, 王静龙, 濮晓龙. 高等数理统计[M]. 第二版. 北京: 高等教育出版社, 2006: 370–372

[32] Sarkar D. Lattice: Multivariate data visualization with R[M]. New York: Springer-Verlag, 2008

[33] Wickham H. ggplot2: Elegant graphics for data analysis[M]. Second ed. New York: Springer-Verlag, 2016

[34] Pebesma E J, Bivand R S. Classes and methods for spatial data in R[J]. R News, 2005, 5(2): 9–13.

[35] Pebesma E. sf: Simple features for R[EB/OL]. 2018. https://CRAN.R-project.org/package=sf.

[36] Kilibarda M, Bajat B. plotGoogleMaps: the r-based web-mapping tool for thematic spatial data[J]. Geomatica, 2012, 66(1): 37–49.

[37] Hengl T, Roudier P, Beaudette D, et al. plotKML: Scientific visualization of spatio-temporal data[J]. Journal of Statistical Software, 2015, 63(5): 1–25.

[38] Hosseini F. A new algorithm for estimating the parameters of the spatial generalized linear mixed models[J]. Environmental and Ecological Statistics, 2016, 23(2): 205–217.

[39] Carpenter B, Gelman A, Hoffman M, et al. Stan: A probabilistic programming language [J]. Journal of Statistical Software, 2017, 76(1): 1–32.

[40] Lunn D, Spiegelhalter D, Thomas A, et al. The BUGS project: Evolution, critique and future directions[J]. Statistics in Medicine, 2009, 28(25): 3049–3067.

[41] Stan Development Team. RStan: the R interface to Stan[EB/OL]. 2018. http://mc-stan.org/.

[42] Givon L E, Unterthiner T, Erichson N B, et al. SCIKIT-CUDA: A Python interface to GPU-powered libraries[EB/OL]. 2015. http://dx.doi.org/10.5281/zenodo.40565.

[43] Yalamanchili P, Arshad U, Mohammed Z, et al. ArrayFire: A high performance software library for parallel computing with an easy-to-use api[EB/OL]. Atlanta: AccelerEyes, 2015. https://github.com/arrayfire/arrayfire.

[44] Finley A O, Banerjee S, E.Gelfand A. spBayes for large univariate and multivariate point-referenced spatio-temporal data models[J]. Journal of Statistical Software, 2015, 63(13): 1–28.

[45] Plummer M, Best N, Cowles K, et al. Coda: Convergence diagnosis and output analysis for mcmc[J]. R News, 2006, 6(1): 7–11.

[46] Youngflesh C. MCMCvis: Tools to visualize, manipulate, and summarize mcmc output [EB/OL]. 2018. https://CRAN.R-project.org/package=MCMCvis.

[47] Ribeiro Jr. P J, Diggle P J. geoR: A package for geostatistical analysis[J]. R News, 2001, 1(2): 14–18.

[48] Christensen O, Ribeiro Jr. P. geoRglm: a package for generalised linear spatial models [J]. R News, 2002, 2(2): 26–28.

[49] Brown P E, Zhou L. glmmBUGS: Generalised linear mixed models and spatial models with WinBUGS, JAGS, and OpenBUGS[EB/OL]. 2018. https://CRAN.R-project.org/package=glmmBUGS.

[50] Brown P E, Zhou L. MCMC for generalized linear mixed models with glmmBUGS[J]. R Journal, 2010, 2(1): 13–17.

[51] Pebesma E J. Multivariable geostatistics in S: the gstat package[J]. Computers and Geosciences, 2004, 30(7): 683–691.

[52] Gräler B, Pebesma E, Heuvelink G. Spatio-temporal interpolation using gstat[J]. The R Journal, 2016, 8(1): 204–218.

[53] Bürkner P C. brms: An R package for bayesian multilevel models using Stan[J]. Journal of Statistical Software, 2017, 80(1): 1–28.

[54] Schlather M, Malinowski A, Menck P J, et al. Analysis, simulation and prediction of multivariate random fields with package RandomFields[J]. Journal of Statistical Software, 2015, 63(8): 1–25.

[55] Ribeiro Jr. P J, Diggle P J. geoR: Analysis of geostatistical data[EB/OL]. 2016. https://CRAN.R-project.org/package=geoR.

[56] Venables W N, Ripley B D. Modern applied statistics with S[M]. Fourth ed. New York: Springer-Verlag, 2002

[57] Thomson M C, Obsomer V, Kamgno J, et al. Mapping the distribution of loa loa in cameroon in support of the african programme for onchocerciasis control[J]. Filaria Journal, 2004, 3(1): 1–13.

[58] Gómez-Rubio V, Rue H. Markov chain monte carlo with the integrated nested laplace approximation[J]. ArXiv e-prints, 2017.

致　谢

三年时间说短不短，说长不长，但是对我却是意义重大的三年，无论是学习还是生活，学校对我的影响都是终生难忘的。首先，我要感谢父母一如既往的默默支持，没有他们就没有我的今天，虽然远隔千山万里，也照顾不到我的学习和生活，但只要想到，不管我做怎样的决定，他们都会全力支持，我很感动；然后，我要感谢我的导师，从他那里我学到严谨治学的态度，他也给予了我最大的自由，这得以让我去一些技术公司实习，接触到最前沿的正在发生深刻变革的人工智能领域，这段实习经历除了让我开阔眼界，接触了深度学习技术和计算框架，更重要的是结识了老师木（一流科技 CEO）和一些志同道合的同事，如深度学习算法研究者陈新鹏，计算框架开发者王笑舒等；此外，还要感谢新浪的总监高鹏，实习期间，除了基本业务外，让我做了很多我感兴趣的事，如学习 R 语言绘图系统和 R Markdown 生态系统，最得益的莫过于见识了大数据平台的系统架构；最后我要感谢统计之都，特别是创始人谢益辉，除了使用他开发的工具打造毕业论文模板，使得论文排版工作量直接降低了几个量级，一年多以来，还一直对我的问题有问必答。三年来，帮助过我的老师，同学，同事，朋友太多，他们当中很多都直接或间接地帮助了我的毕业论文，人生最大的幸运莫过于结识你们。

作者简介

黄湘云，男（1992-），2015 年毕业于中国矿业大学（北京），获理学学位；2018 年毕业于中国矿业大学（北京），攻读硕士学位，专业为统计学，研究方向为数据分析与统计计算。

在学期间参加科研项目

1. 国家自然科学基金项目“混合模型的方差元素检验及函数型混合模型研究”项目组成员。项目编号：11671398。2017 年 01 月-2020 年 12 月

主要获奖

1. 2015-2016 年度获研究生优秀学生一等奖学金
2. 2016-2017 年度获研究生优秀学生奖学金