Spatial Generalized Linear Mixed Models with Application to Prevalence Mapping

Xiang-Yun Huang   Zai-Xing Li

March 27, 2018

Department of Statistics, School of Science
China University of Mining and Technology,Beijing

Outline

1. Introduction (Motivations and goals)
2. Literature reviews
3. Geostatistical model (SGLMM)
4. Computing details and simulations
5. Real data analysis (Applications)
6. Discussion

Motivations

1. Radionuclide concentrations on Rongelap Island
2. Childhood malaria in the gambia
3. Loa loa prevalence in Cameroon and surrounding areas

Goals

1. parameter estimation and spatial prediction as Diggle and Giorgi (2016)
2. thesis as Varin et al. (2005)

Multiple prevalence surveys

Sample n_i individuals,observe Y_i positives,$i = 1, 2, \cdots, m$

$$Y_i \sim \mathrm{Bin}(n_i, p_i)$$

Extra-binomial variation

Sample n_i individuals,observe Y_i positives,$i = 1, 2, \cdots, m$

$$Y_i | d_i, U_i \sim \mathrm{Bin}(n_i, p_i) \quad \log\{p_i/(1-p_i)\} = d_i'\beta + U_i \quad U_i \sim N(0, \tau^2)$$

notations: Spatial Generalized Linear Mixed Models (SGLMM)

- Latent spatially correlated process
  Stationary Gaussian Process: $S(x) \sim \mathrm{SGP}\{0, \sigma^2, \rho(u)\}$
  correlation function: e.g. $\rho(u) = \exp(-|u|/\phi)$
- Linear prediction (regression model)
  $d(x)$ = covariates at location x
  Linear prediction: $\eta(x) = d(x)'\beta + S(x)$
  Link function: logit $p(x) = \log\{\eta(x)/[1-\eta(x)]\}$
- Conditional distribution for positive proportion Y_i/n_i
  $Y_i | S(\cdot) \sim \mathrm{Bin}(n_i, p(x_i))$ (binomial sampling)

Standard geostatistical prevalence sampling model:

$$\log[p(x_i)/\{1 - p(x_i)\}] = T_i = d(x_i)'\beta + S(x_i) + U_i$$

$E[Y_i|S(x_i), U_i] = n_i p_i$

theoretical variograms:

$$\begin{aligned}
V(x, x') &= \frac{1}{2}\mathrm{Var}\{S(x) - S(x')\} \\
&= \frac{1}{2}\mathrm{Cov}(S(x) - S(x'), S(x) - S(x')) \\
&= \frac{1}{2}\{E[S(x) - S(x')][S(x) - S(x')] - [E(S(x) - S(x'))]^2\} \\
&= \sigma^2 - \mathrm{Cov}(S(x), S(x')) = \sigma^2\{1 - \rho(u)\}, u = ||x - x'||
\end{aligned}$$

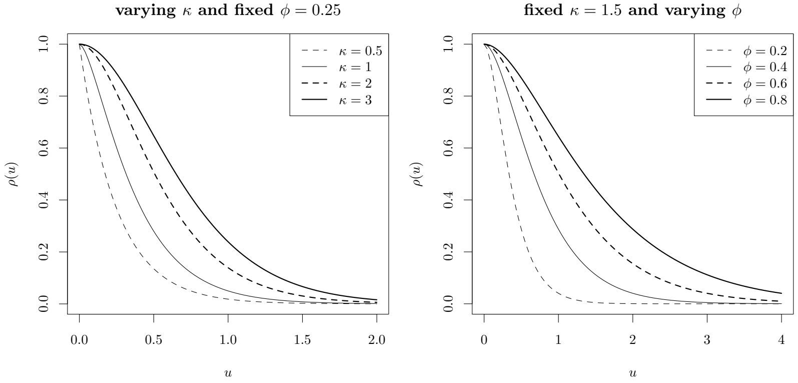
$$V_T(u_{ij}) = \frac{1}{2}\mathrm{Var}\{T_i(x) - T_j(x)\} = \frac{1}{2}E[(T_i - T_j)^2] = \tau^2 + \sigma^2(1 - \rho(u_{ij}))$$

covariance matrix:

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

Matérn class of correlation functions:

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



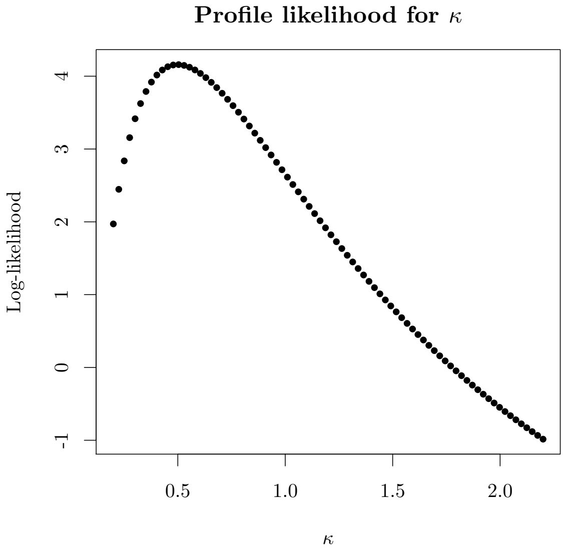


Figure 1: $\kappa = 0.4988445$,Loa loa data from Giorgi and Diggle (2016b)

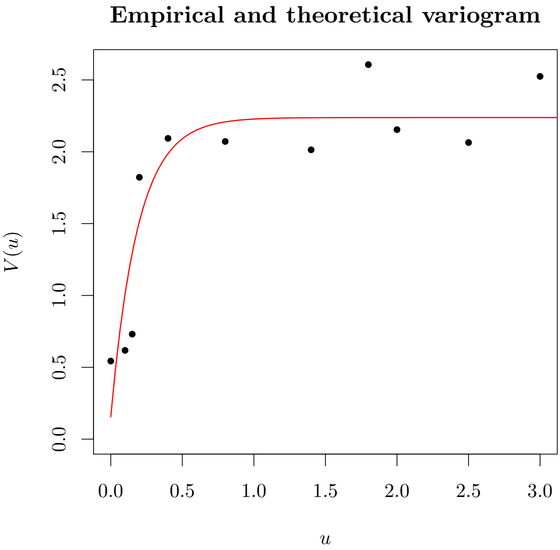


Figure 2: $\tau^2 = 0.1554, \sigma^2 = 2.0827, \phi = 0.189$ and fixed $\kappa = 0.5$

Goals:

- estimation: the coefficient vector, 95% confidence intervals
- prediction: probability of Loa loa prevalence of unknown locations

Likelihood-based methods inferences

- Monte Carlo EM gradient used by Zhang (2002)
- Monte Carlo maximum likelihood used by Christensen (2004) and Diggle and Giorgi (2016)
- Approximate Monte Carlo EM gradient used by Hosseini (2016)

Approximate Bayesian Inference

- Bayesian approach combined with MCMC methods used by Diggle et al. (1998, 2002)
- Bayesian approach combined with integrated nested Laplace approximations used by Eidsvik et al. (2009); Rue et al. (2009); Gómez-Rubio and Rue (2017)

Monte Carlo Maximum Likelihood (MCML)

let $\theta^\top = (\sigma^2, \phi, \tau^2)$, D denote the n by p matrix of covariates, $y^\top = (y_1, y_2, \cdots, y_n)$ and marginal distribution of T is $N(D\beta, \Sigma(\theta))$. The conditional distribution of $Y^\top = (Y_1, \cdots, Y_n)$ given $T^\top = t^\top = (t_1, t_2, \cdots, t_n)$ is

$$f(y|t) = \prod_{i=1}^{n} f(y_i|t_i)$$

a product of independent binomial probability functions.
The likelihood function for β and θ

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

Computing details

fixed β_0, θ_0,then we get the joint distribution of Y and T

$$f(y,t) = N(t; D\beta_0, \Sigma(\theta_0))f(y|t)$$

for pre-defined and use MCMC algorithm to obtain m samples t_i from conditional distribution of T given $Y = y$ under β_0 and θ_0, so

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

Let $\hat{\beta}_m$ and $\hat{\theta}_m$ denote MCML estimates by maximising $L_m(\beta,\theta)$ given an suitable initial values β_0 and θ_0, repeat the iterative procedure with $\beta_0 = \hat{\beta}_m$ and $\theta_0 = \hat{\theta}_m$ until convergence.
For maximization of $L_m(\beta,\theta)$, we can choose BFGS algorithm or unconstrained optimization with PORT rountines.

Case Study 1

Loa loa prevalence data from 197 village surveys in west Africa, Diggle et al. (2007)

Table 1: Loa loa prevalence data (partial)

LONGITUDE	LATITUDE	NO_EXAM	NO_INF	ELEVATION	MAX9901
8.0419	5.7367	162	0	108	0.69
8.0043	5.6803	167	1	99	0.74
8.9056	5.3472	88	5	783	0.79
8.1007	5.9174	62	5	104	0.67
8.1825	5.1045	167	3	109	0.85
8.9292	5.3556	66	3	909	0.80
11.3600	4.8850	163	11	503	0.78
8.0675	5.8978	83	0	103	0.69

- MAX9901: Maximum of all NDVI values recorded at village location, 1999-2001.
- MEAN9901, MIN9901 and STDEV9901 are as defined above.
- NDVI: normalised-difference vegetation index

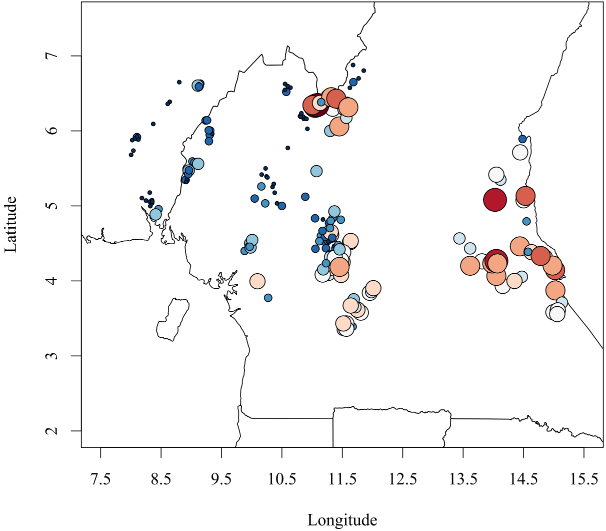


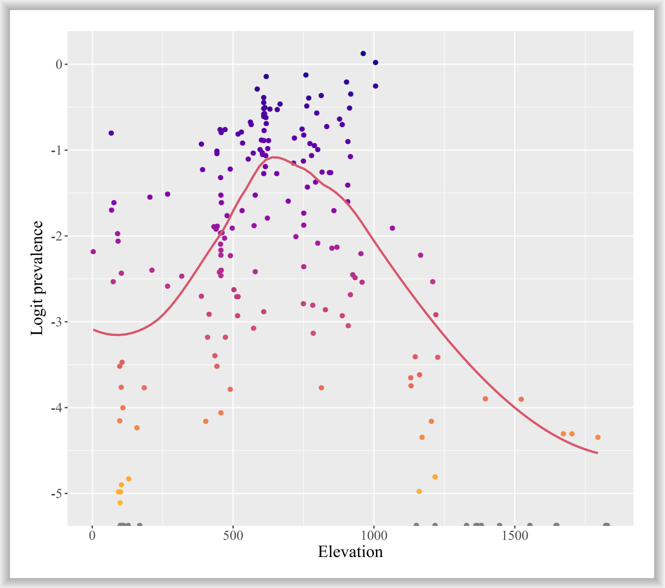
Figure 3: cex (size of circles): 0.5, 1.0, 1.5, 2.0, 2.5, 3.0 corresponds to the observed prevalence of Loa loa: [0,0.05), [0.05,0.15), [0.15,0.25), [0.25,0.35), [0.35,0.45), [0.45,0.55) and policy intervention threshold is 0.2

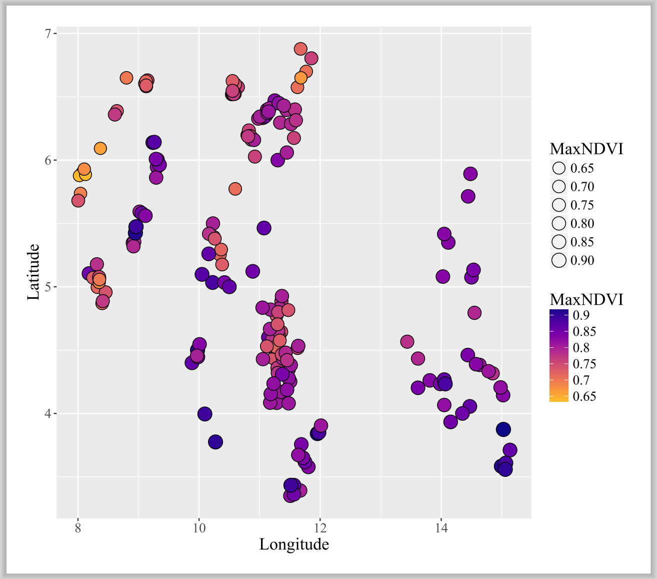
Statistical Model

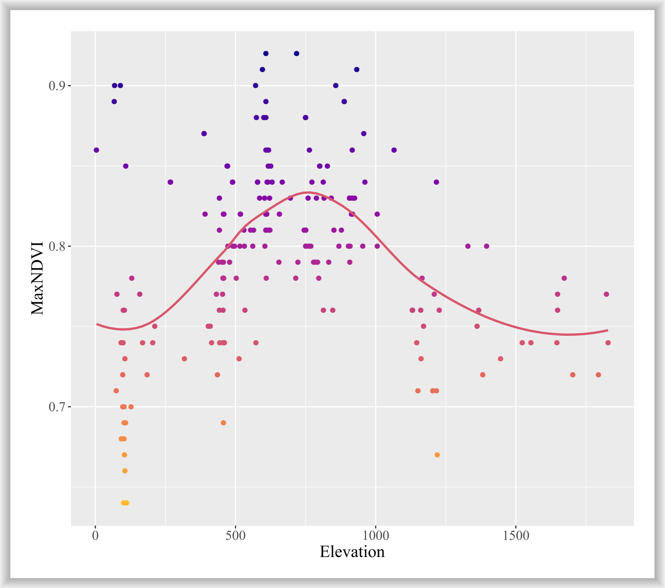
Diggle et al. (2007)
Goals:
using spatial statistical methods to address the issue of spatial correlation, and using Bayesian methods to quantify the uncertainty in the predictions from Diggle et al. (1998) to create a new map.
village level model:

$$\begin{aligned}\log\{p(x)/[1-p(x)]\} =& \alpha + f_1(\mathrm{ELEVATION}) + f_2[\max(\mathrm{NDVI})] \\ &+ f_3[\mathrm{s.d.}(\mathrm{NDVI})] + S(x)\end{aligned}$$

- $S(x)$ Gaussian process with mean zero ,variance σ^2 and correlation function $\mathrm{Corr}(S(x), S(x')) = \exp(-||x - x'||/\phi) + \tau^2/\sigma^2 \cdot \mathrm{I}_{\{x=x'\}}$
- $f_1(\cdot), f_2(\cdot)$ and $f_3(\cdot)$ are piece-wise linear functions


Logit prevalence
0
-1
-2
-3
-4
-5
0
500
1000
1500
Elevation


7
6
5
4
Latitude
8
10
12
14
Longitude
MaxNDVI
0.65
0.70
0.75
0.80
0.85
0.90
MaxNDVI
0.9
0.85
0.8
0.75
0.7
0.65


0.9
0.8
0.7
MaxNDVI
0
500
1000
1500
Elevation

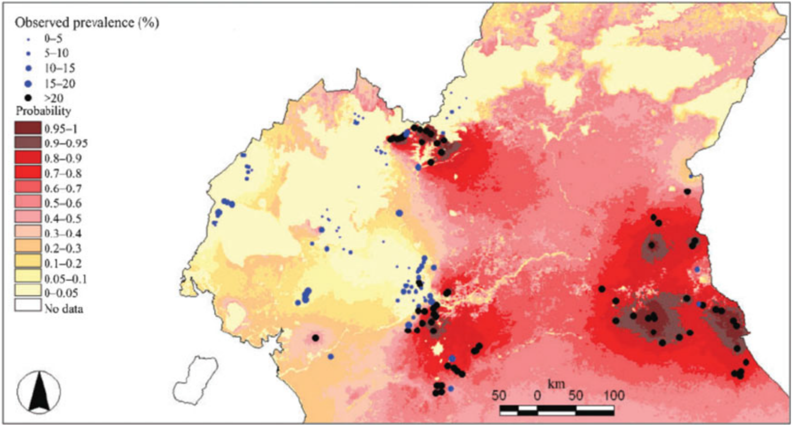


Figure 4: Predictive probability map of Loa loa prevalence in Cameroon and surrounding areas (adapted from Diggle et al. (2007)). Empirical prevalences at surveyed locations are indicated by size and color coded dots.

Case Study 2

Childhood malaria in the gambia, Diggle et al. (2002)

Table 2: Childhood malaria data (partial)

	x	y	pos	age	netuse	treated	green	phc
1850	349631.3	1458055	1	1783	0	0	40.85	1
1851	349631.3	1458055	0	404	1	0	40.85	1
1852	349631.3	1458055	0	452	1	0	40.85	1
1853	349631.3	1458055	1	566	1	0	40.85	1

- pos: presence (1) or absence (0) of malaria in a blood sample taken from the child
- netuse: whether (1) or not (0) the child regularly sleeps under a bed-net.
- treated: whether (1) or not (0) the bed-net is treated (coded 0 if netuse=0).
- green: satellite-derived measure of the green-ness of vegetation in the immediate vicinity of the village (arbitrary units).
- phc: presence (1) or absence (0) of a health center in the village.

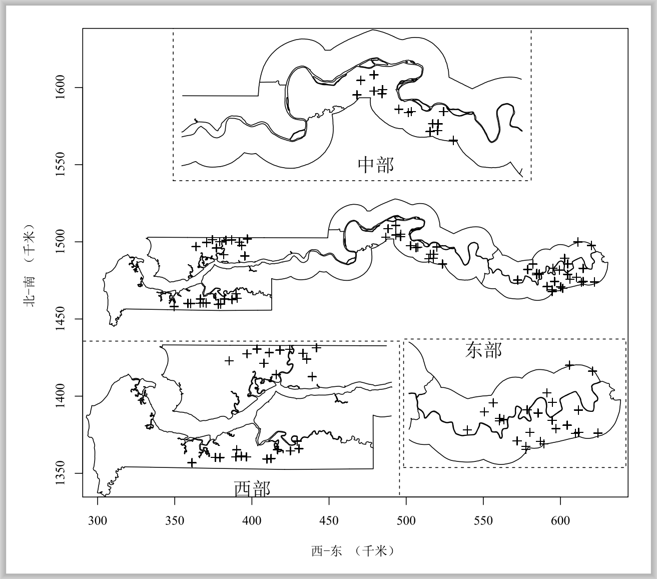
Statistical Model

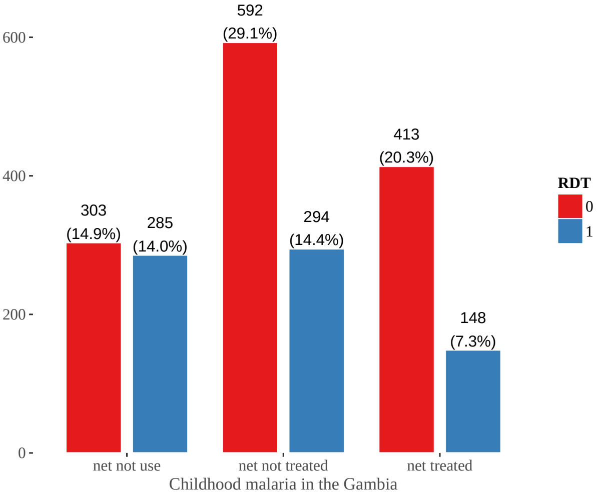
Diggle et al. (2002)

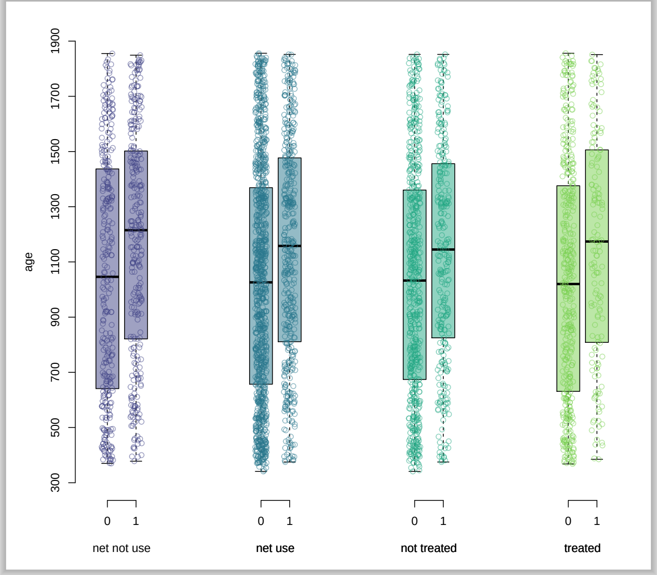
- the effects of child level covariates (age and bed net use)
- village level covariates (the primary health care and greenness of surrounding vegetation)
- separate components for residual spatial
- non-spatial extrabinomial variation

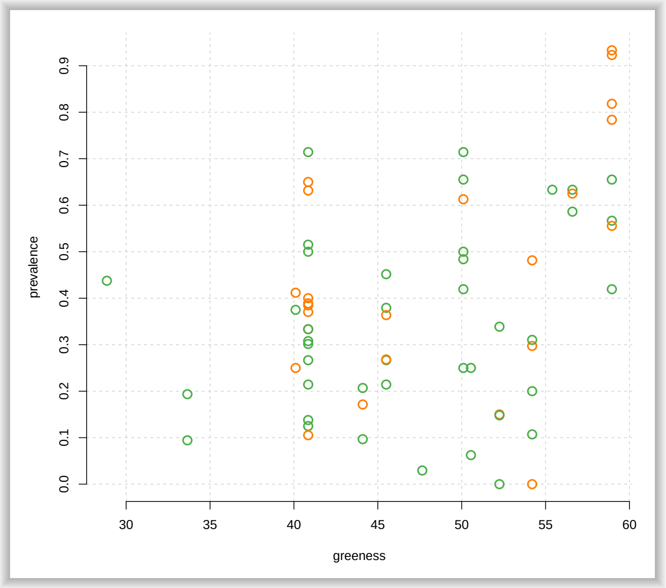
Child level model:

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


中部
东部
西部
1600
1550
1500
1450
1400
1350
300
350
400
450
500
550
600
北-南 （千米）
西-东 （千米）


600
400
200
0
592
(29.1%)
303
(14.9%)
285
(14.0%)
294
(14.4%)
413
(20.3%)
148
(7.3%)
RDT
0
1
net not use
net not treated
net treated
Childhood malaria in the Gambia


age
300
500
700
900
1100
1300
1500
1700
1900
0
1
0
1
0
1
0
1
net not use
net use
not treated
treated


prevalence
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
30
35
40
45
50
55
60
greeness

R: geoR geoRglm spatial PrevMap
Ribeiro Jr and Diggle (2016); Christensen and Ribeiro Jr (2015); Ripley (2015); Giorgi and Diggle (2016a)

Stan: Stan [1] interfaces with R (RStan) ,Python (PyStan) , MATLAB (MatlabStan) and more
Gelman et al. (2015); Bob et al. (2017)

PyMC3: Probabilistic programming in Python using PyMC3
Salvatier et al. (2016)

JAGS: **J**ust **A**nother **G**ibbs **S**ampler [2]
Bayesian hierarchical models using Markov chain Monte Carlo (MCMC)

BUGS: **B**ayesian inference **U**sing **G**ibbs **S**ampling , such as winBUGS, OpenBUGS

R-INLA: **I**ntegrated **N**ested **L**aplace **A**pproximations
Rue et al. (2009, 2016); Gómez-Rubio and Rue (2017)

[1] `http://mc-stan.org/`

[2] `https://en.wikipedia.org/wiki/Just_another_Gibbs_sampler`

Thanks

Bob, C., Andrew, G., Matthew, H., and et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

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

Christensen, O. F. and Ribeiro Jr, P. J. (2015). *geoRglm: A Package for Generalised Linear Spatial Models*. R package version 0.9-8.

Diggle, P., Moyeed, R., Rowlingson, B., and Thomson, M. (2002). Childhood malaria in the gambia: a case-study in model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4):493–506.

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

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

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

Eidsvik, J. O., Martino, S., and Rue, H. (2009). Approximate bayesian inference in spatial generalized linear mixed models. *Scandinavian Journal of Statistics*, 36(1):122.

Gelman, A., Lee, D., Guo, J., and et al. (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):837–840.

Giorgi, E. and Diggle, P. J. (2016a). Prevmap: an r package for prevalence mapping.(in press). *Journal of Statistical Software*.

Giorgi, E. and Diggle, P. J. (2016b). *PrevMap: Geostatistical Modelling of Spatially Referenced Prevalence Data*. R package version 1.4.

Gómez-Rubio, V. and Rue, H. (2017). Markov chain monte carlo with the integrated nested laplace approximation. *ArXiv e-prints*.

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

Ribeiro Jr, P. J. and Diggle, P. J. (2016). *geoR: Analysis of Geostatistical Data*. R package version 1.7-5.2.

Ripley, B. (2015). *spatial: Functions for Kriging and Point Pattern Analysis*. R package version 7.3-11.

Rue, H., Martino, S., Chopin, N., and et al. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Rue, H., Martino, S., Lindgren, F., Simpson, D., and et al. (2016). *INLA: Functions which Allow to Perform Full Bayesian Analysis of Latent Gaussian Models using Integrated Nested Laplace Approximations*. R package version 0.0-1468872408.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2(55).

Varin, C., Høst, G., and Skare, Ø. (2005). Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics and Data Analysis*, 49(4):1173–1191.

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