Geostatistical modelling of spatially structured zero-inflation

Dr Emanuele Giorgi
Lancaster University
e.giorgi@lancaster.ac.uk

Toowoomba 21-25 October 2019

Geostatistical methods for prevalence mapping

- **Prevalence mapping**: $(y_i, n_i, x_i; i = 1, \ldots, n) \rightarrow p(x),\ x \in A$

Geostatistical methods for prevalence mapping

- **Prevalence mapping**: $(y_i, n_i, x_i; i = 1, \ldots, n) \rightarrow p(x),\ x \in A$
- Ingredients of a geostatistical model:

Geostatistical methods for prevalence mapping

- **Prevalence mapping**: $(y_i, n_i, x_i; i = 1, \ldots, n) \rightarrow p(x),\ x \in A$
- Ingredients of a geostatistical model:
  1. $S(x_i) \sim GP(0, \sigma^2, \rho(\cdot; \phi))$;

Geostatistical methods for prevalence mapping

- **Prevalence mapping**: $(y_i, n_i, x_i; i = 1, \ldots, n) \rightarrow p(x),\ x \in A$
- Ingredients of a geostatistical model:
  1. $S(x_i) \sim GP(0, \sigma^2, \rho(\cdot; \phi))$;
  2. $Z_i \sim N(0, \tau^2)$ i.i.d.;

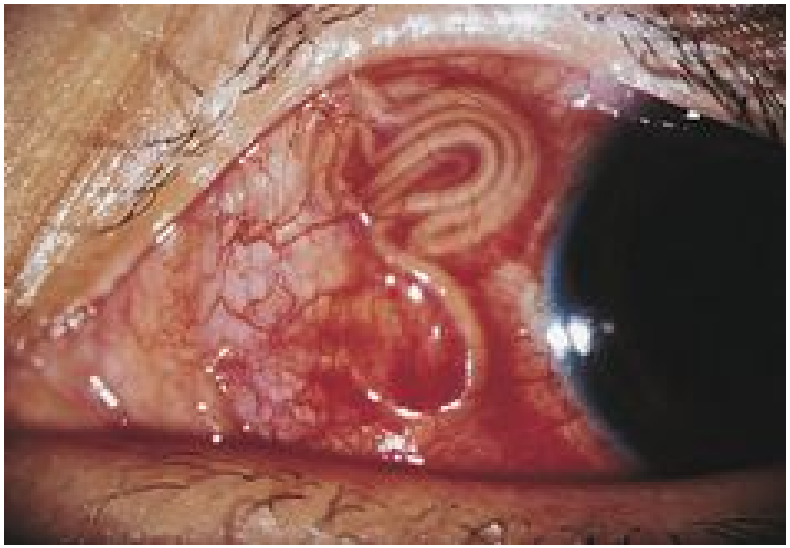
Geostatistical methods for prevalence mapping

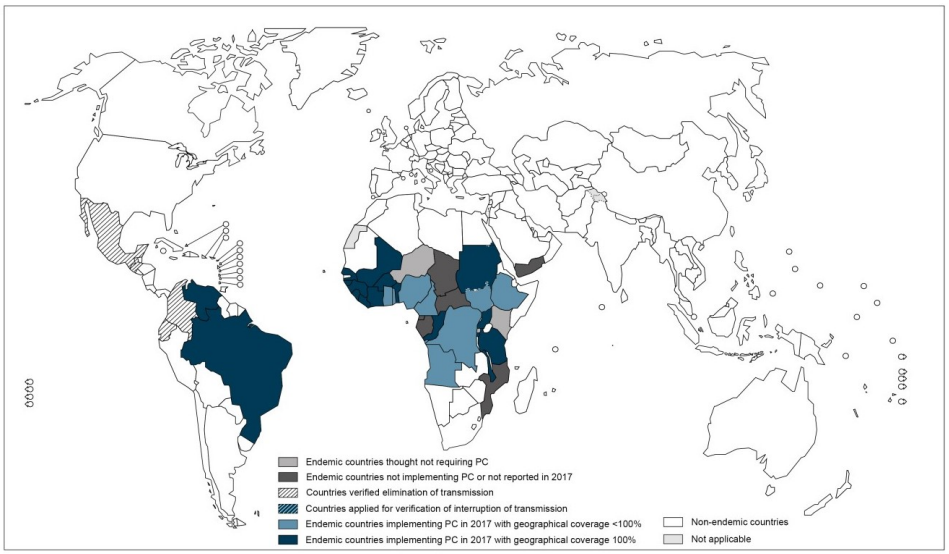
- **Prevalence mapping**: $(y_i, n_i, x_i; i = 1, \ldots, n) \rightarrow p(x),\ x \in A$
- Ingredients of a geostatistical model:
  1. $S(x_i) \sim GP(0, \sigma^2, \rho(\cdot; \phi))$;
  2. $Z_i \sim N(0, \tau^2)$ i.i.d.;
  3. $Y_i | S(x_i), Z_i \sim \mathrm{Bin}(n_i, p(x_i))$;

Geostatistical methods for prevalence mapping

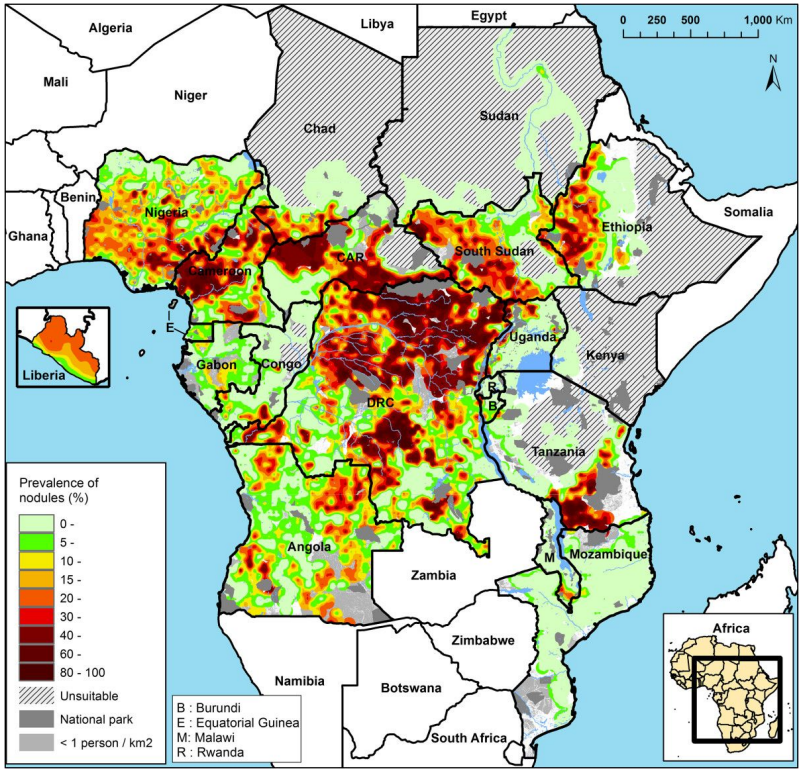
- **Prevalence mapping**: $(y_i, n_i, x_i; i = 1, \ldots, n) \rightarrow p(x),\ x \in A$
- Ingredients of a geostatistical model:
  1. $S(x_i) \sim GP(0, \sigma^2, \rho(\cdot; \phi))$;
  2. $Z_i \sim N(0, \tau^2)$ i.i.d.;
  3. $Y_i | S(x_i), Z_i \sim \mathrm{Bin}(n_i, p(x_i))$;
  4. $\log\{p(x_i)/[1 - p(x_i)]\} = d(x_i)^\top \beta + S(x_i) + Z_i$.

Onchocerciasis: the disease

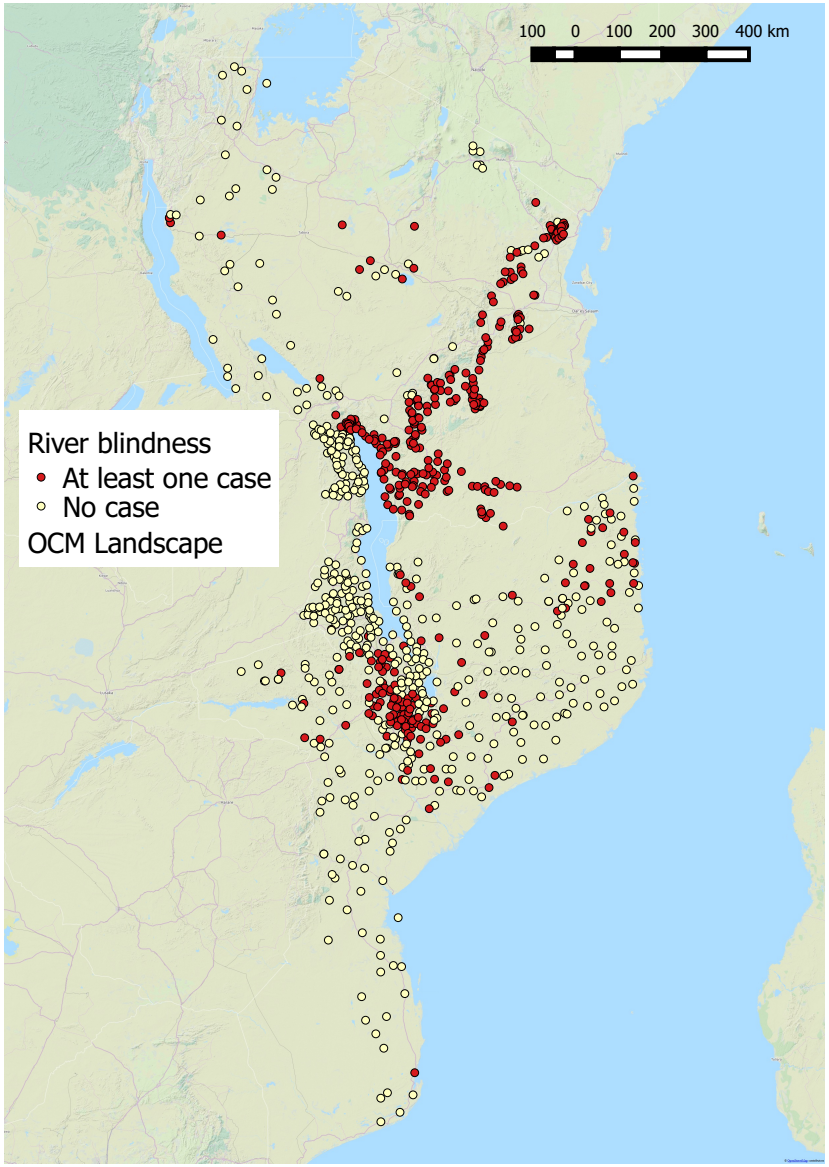




The REMO project



A closer look



Zero-inflation

Zero-inflation

$$Y_i = Y_{1,i} \times Y_{2,1}$$

Zero-inflation

$$Y_i = Y_{1,i} \times Y_{2,1}$$

- $Y_{1,i} \sim \text{Bernoulli}(\pi_\text{i})$

Zero-inflation

$$Y_i = Y_{1,i} \times Y_{2,1}$$

- $Y_{1,i} \sim \text{Bernoulli}(\pi_\text{i})$
- $Y_{2,i} \sim \text{Binomial}(\text{p}(\text{x}_\text{i}))$

Zero-inflation

$$Y_i = Y_{1,i} \times Y_{2,1}$$

- $Y_{1,i} \sim \text{Bernoulli}(\pi_\text{i})$
- $Y_{2,i} \sim \text{Binomial}(\text{p}(\text{x}_\text{i}))$
- $P(Y_i = 0) > P(Y_{2,i} = 0)$

Zero-inflation

$$Y_i = Y_{1,i} \times Y_{2,1}$$

- $Y_{1,i} \sim \text{Bernoulli}(\pi_\text{i})$
- $Y_{2,i} \sim \text{Binomial}(\text{p}(\text{x}_\text{i}))$
- $P(Y_i = 0) > P(Y_{2,i} = 0)$
- $E[Y_i] = n_i \pi_i p(x_i)$

Zero-inflation

$$Y_i = Y_{1,i} \times Y_{2,1}$$

- $Y_{1,i} \sim \text{Bernoulli}(\pi_\text{i})$
- $Y_{2,i} \sim \text{Binomial}(\text{p}(\text{x}_\text{i}))$
- $P(Y_i = 0) > P(Y_{2,i} = 0)$
- $E[Y_i] = n_i \pi_i p(x_i)$
- **Spatially structured zero-inflation**: $\pi_i = \pi(x_i)$

Zero-inflation

$$Y_i = Y_{1,i} \times Y_{2,1}$$

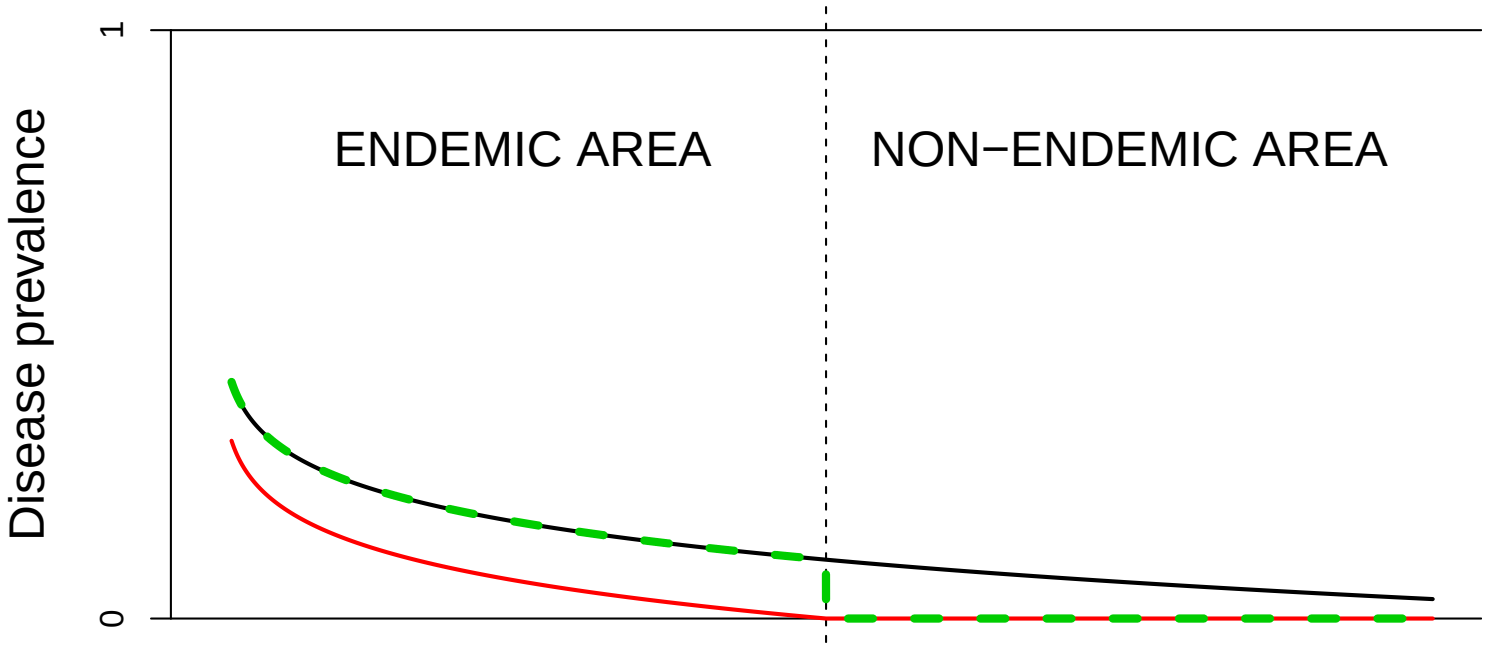
- $Y_{1,i} \sim \text{Bernoulli}(\pi_\text{i})$
- $Y_{2,i} \sim \text{Binomial}(\text{p}(\text{x}_\text{i}))$
- $P(Y_i = 0) > P(Y_{2,i} = 0)$
- $E[Y_i] = n_i \pi_i p(x_i)$
- **Spatially structured zero-inflation**: $\pi_i = \pi(x_i)$
- How should we model $\pi(x)$?

What happens at the border?

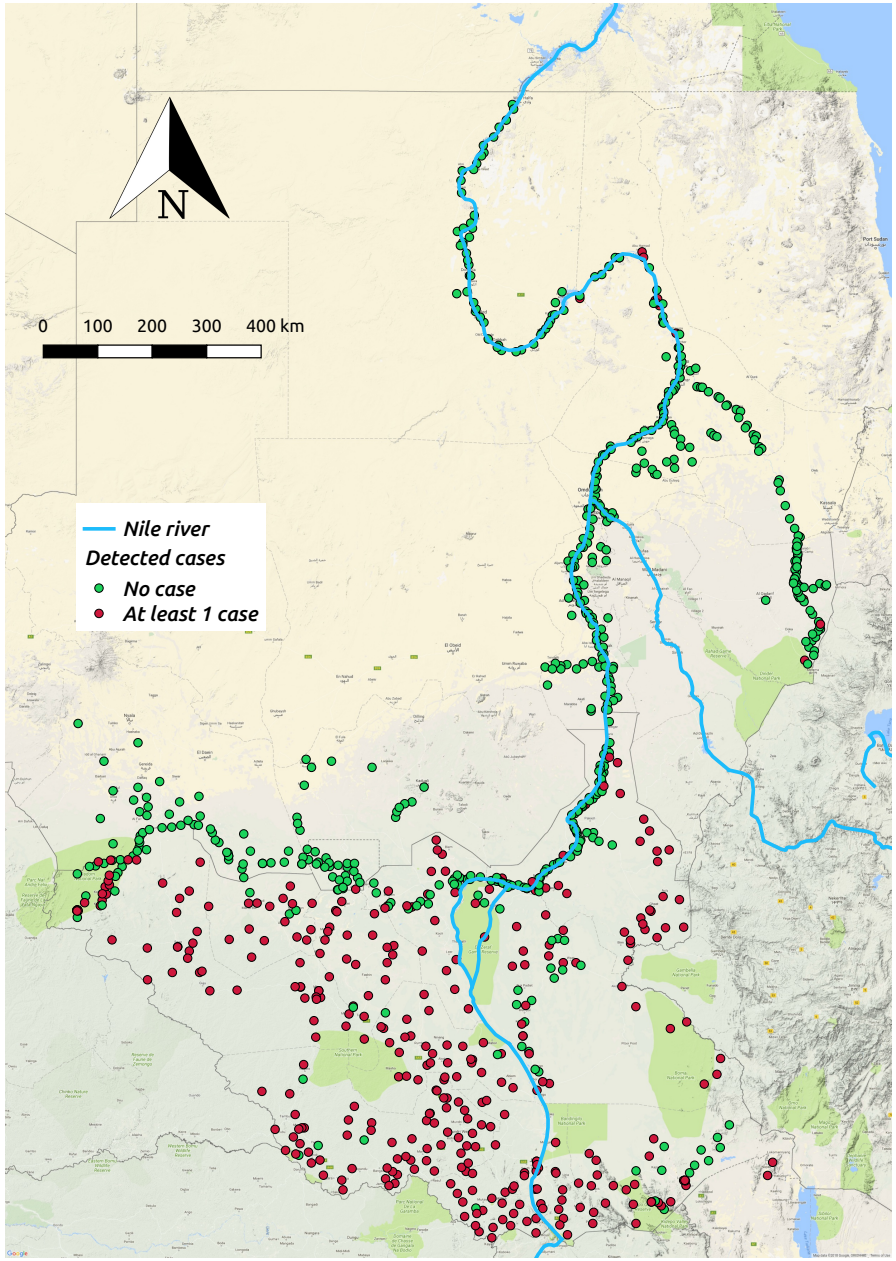
- **Endemic:** "especially of a disease or a condition, regularly found and very common among a particular group or in a particular area." (Cambridge Dictionary, 4th Edition)

What happens at the border?

- **Endemic:** "especially of a disease or a condition, regularly found and very common among a particular group or in a particular area." (Cambridge Dictionary, 4th Edition)



Onchocerciasis in Sudan and South Sudan



A double logistic GP

- $Y_i = Y_{1,i} \times Y_{2,i}$

A double logistic GP

- $Y_i = Y_{1,i} \times Y_{2,i}$
- Model for $Y_{1,i}$ ($P[Y_{1,i} = 1 | T(x_i)] = \pi(x_i)$)

$$\begin{aligned} \log\left\{\frac{\pi(x_i)}{1-\pi(x_i)}\right\} &= \gamma + T(x_i) \\ \mathrm{cov}\{T(x), T(x')\} &= \nu^2 \exp\{-\|x - x'\|/\delta\} \end{aligned}$$

A double logistic GP

- $Y_i = Y_{1,i} \times Y_{2,i}$
- Model for $Y_{1,i}$ ($P[Y_{1,i} = 1 | T(x_i)] = \pi(x_i)$)

$$\begin{aligned} \log\left\{\frac{\pi(x_i)}{1-\pi(x_i)}\right\} &= \gamma + T(x_i) \\ \text{cov}\{T(x), T(x')\} &= \nu^2 \exp\{-\|x - x'\|/\delta\} \end{aligned}$$

- Model for $Y_{2,i}$

A double logistic GP

- $Y_i = Y_{1,i} \times Y_{2,i}$
- Model for $Y_{1,i}$ ($P[Y_{1,i} = 1 | T(x_i)] = \pi(x_i)$)

$$\begin{aligned} \log\left\{\frac{\pi(x_i)}{1-\pi(x_i)}\right\} &= \gamma + T(x_i) \\ \text{cov}\{T(x), T(x')\} &= \nu^2 \exp\{-\|x - x'\|/\delta\} \end{aligned}$$

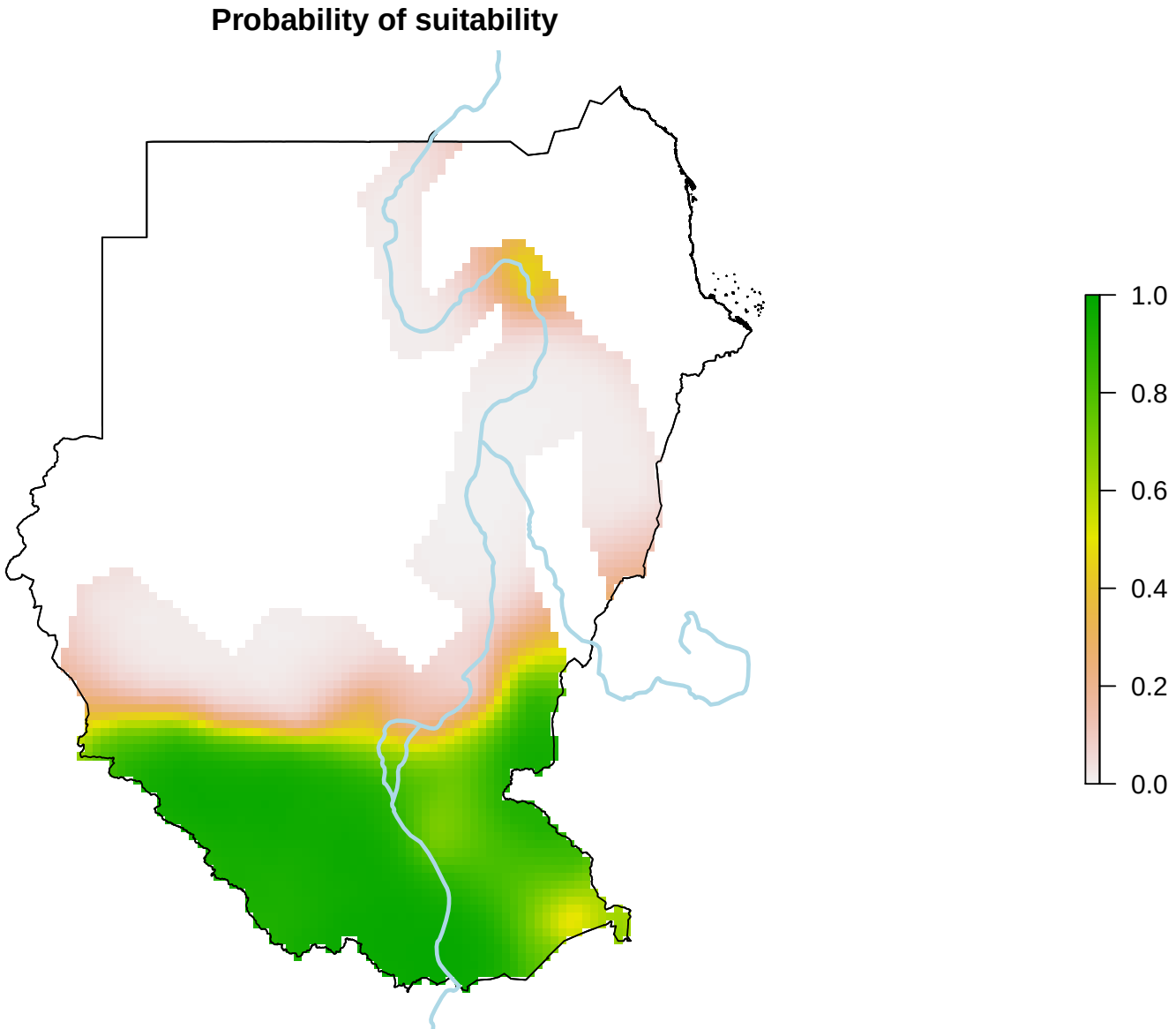
- Model for $Y_{2,i}$

$$\begin{aligned} \log\left\{\frac{p(x_i)}{1-p(x_i)}\right\} &= \beta + S(x_i) + Z_i, \\ \text{cov}\{S(x), S(x')\} &= \sigma^2 \exp\{-\|x - x'\|/\phi\} \end{aligned}$$

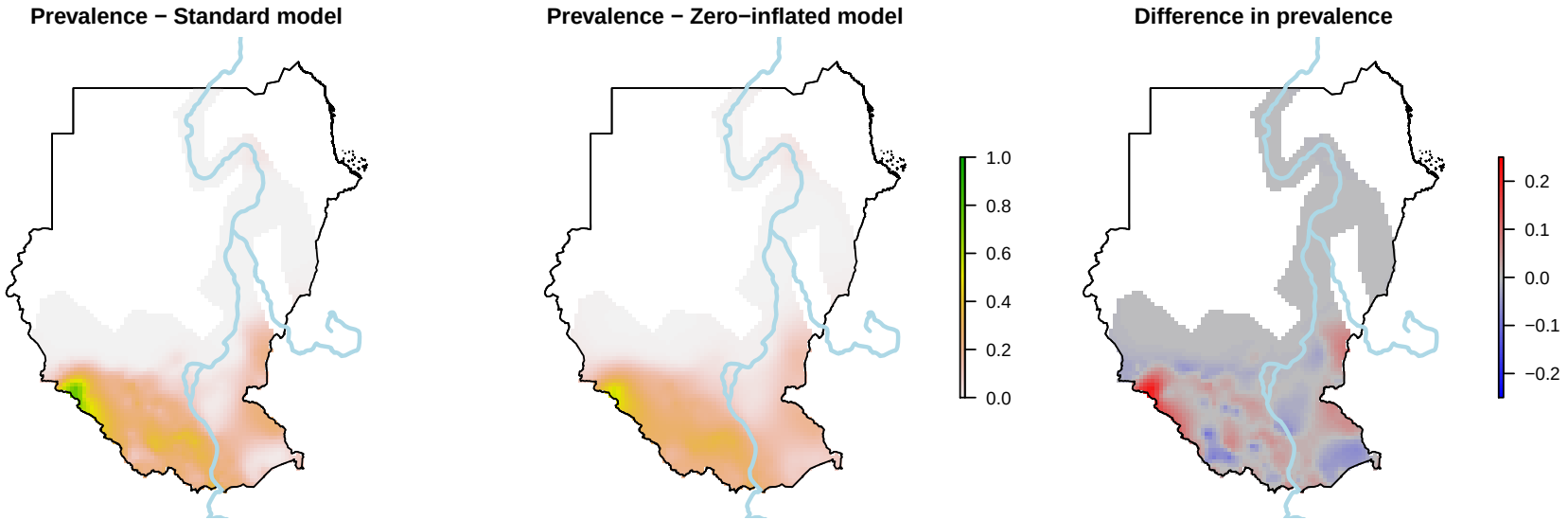
Monte Carlo maximum likelihood estimation

	Standard model		Zero-inflated model	
Parameter	Estimate	95% CI	Estimate	95% CI
β	-4.851	(-11.862, 2.160)	-1.819	(-2.618, -1.019)
$\log(\sigma^2)$	3.140	(1.413, 4.867)	-0.310	(-1.093, 0.474)
$\log(\phi)$	7.392	(5.647, 9.136)	5.769	(4.911, 6.627)
$\log(\tau^2/\sigma^2)$	-3.295	(-5.051, -1.539)	-0.812	(-1.600, -0.025)
γ			-1.561	(-9.526, 6.404)
$\log(\nu^2)$			3.217	(1.337, 5.097)
$\log(\delta)$			7.565	(5.697, 9.434)

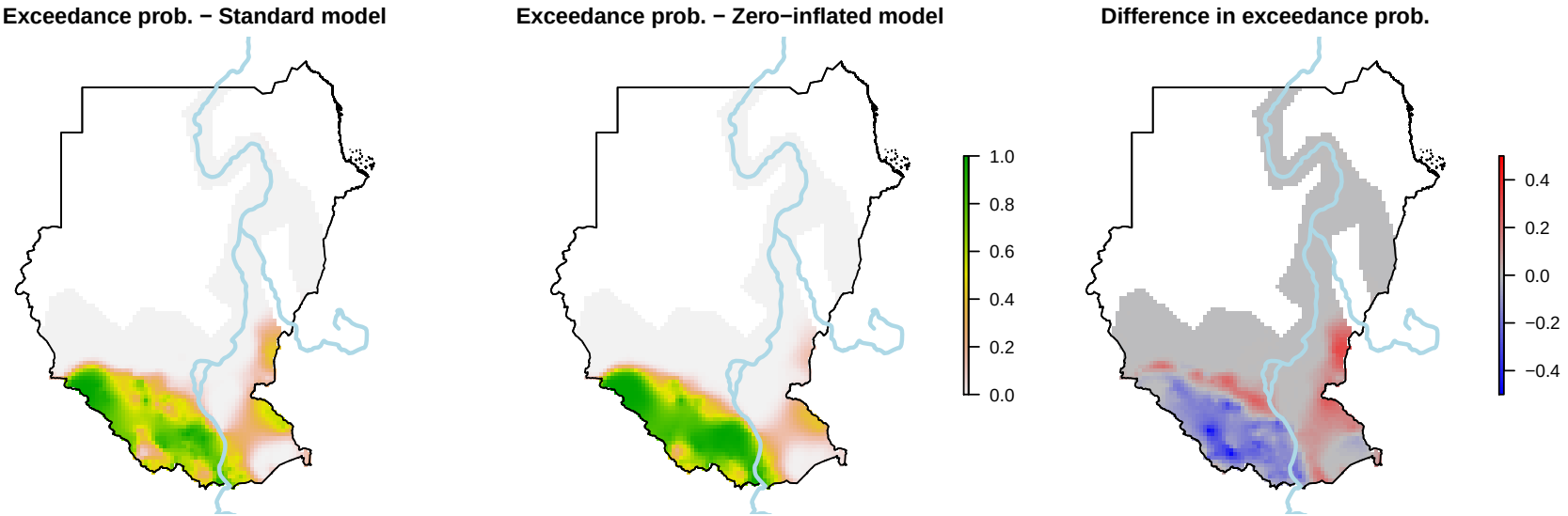
Mapping of $\pi(x)$



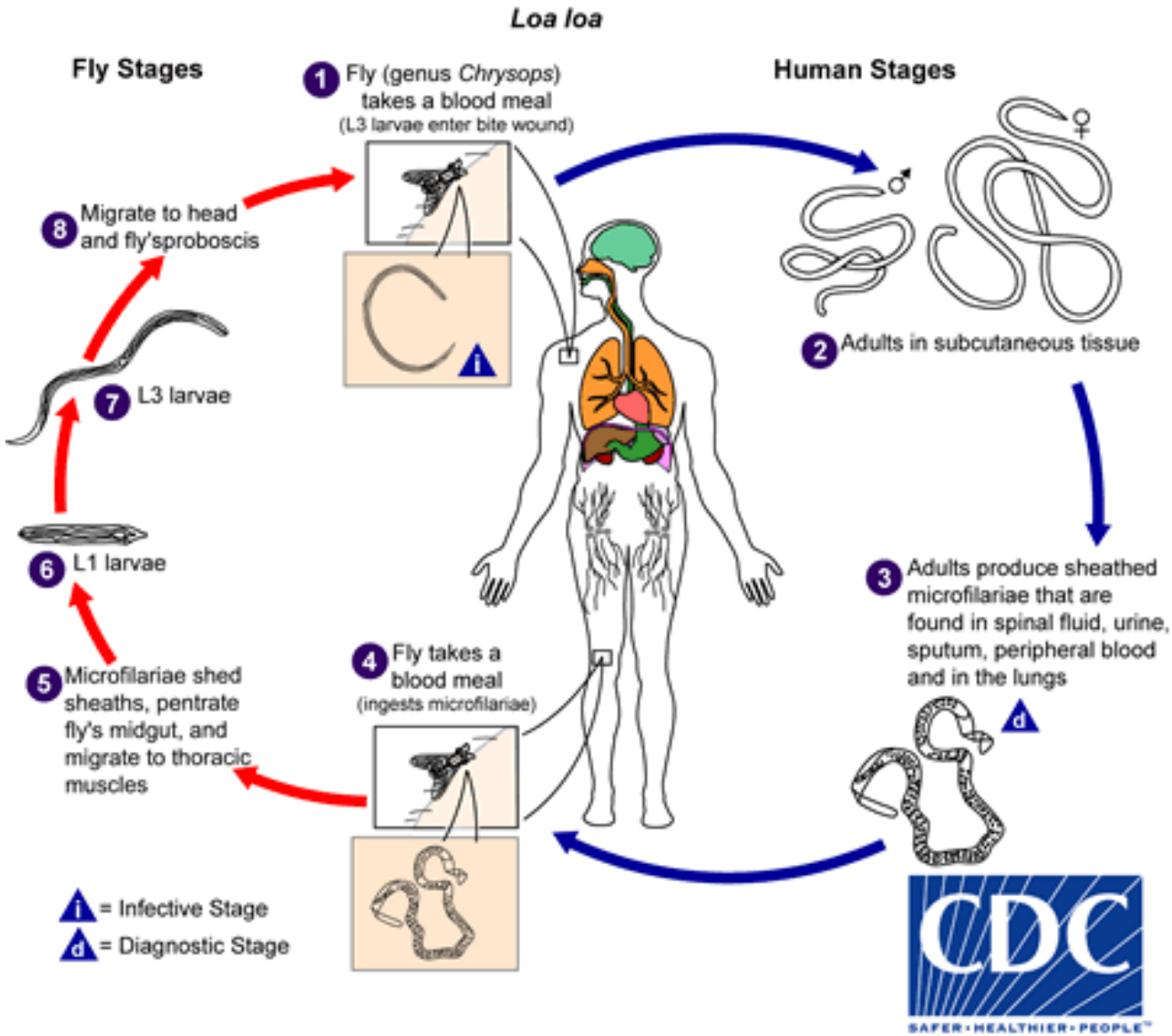
Mapping of oncho prevalence: $\pi(x)p(x)$



Mapping of exceedance probabilities: $P[\pi(x)p(x) > 0.2|y]$



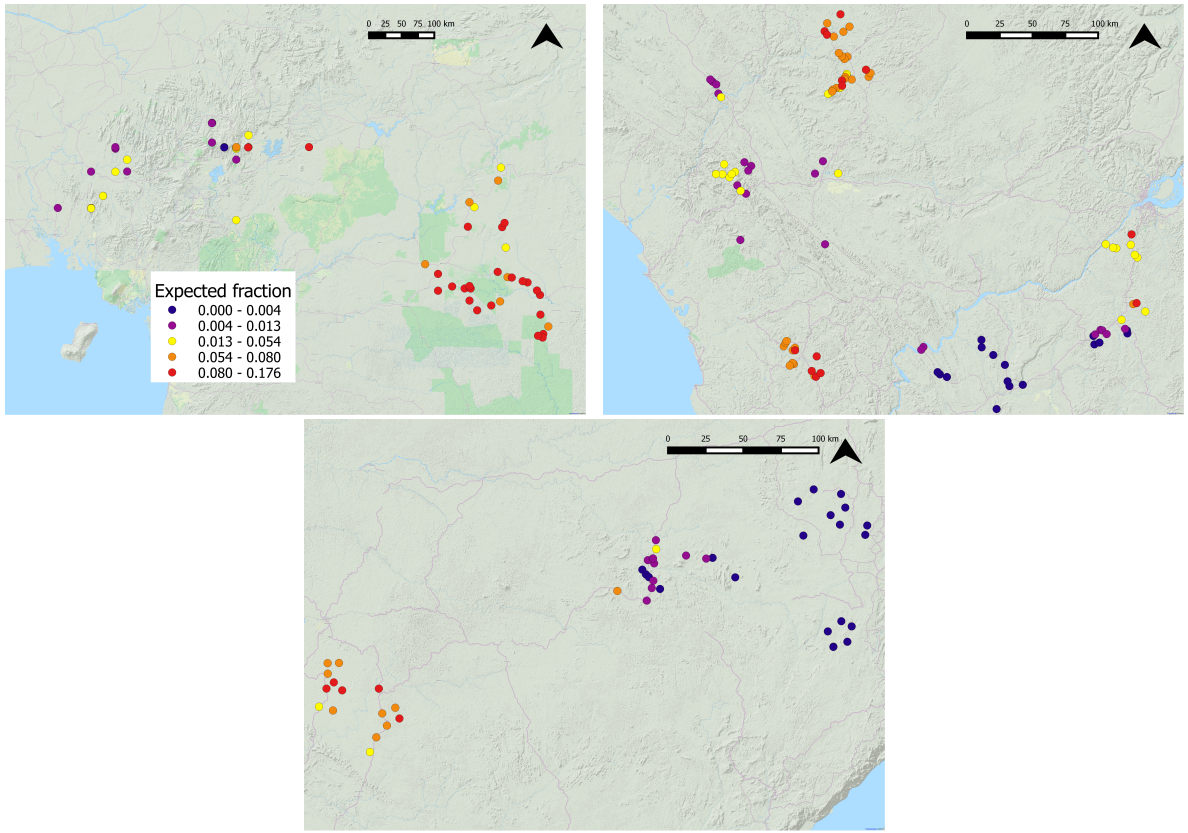
Loa loa: the disease



The data

The data

- **Outcome:** $Y_j(x_i)$, number of micro-filariae eggs per ml in a blood samples



Bivariate modelling of prevalence and intensity of infection

Bivariate modelling of prevalence and intensity of infection

- $Y_j(x_i) = Y_{i,1} \times Y_{2,i}$

Bivariate modelling of prevalence and intensity of infection

- $Y_j(x_i) = Y_{i,1} \times Y_{2,i}$
- Model for $Y_{1,i}$ ($P[Y_{1,i} = 1 | T(x_i)] = \pi(x_i)$)

$$\log\left\{\frac{\pi(x_i)}{1-\pi(x_i)}\right\} = \mu_1 + \sigma_1[S_1(x) + T(x)]$$

Bivariate modelling of prevalence and intensity of infection

- $Y_j(x_i) = Y_{i,1} \times Y_{2,i}$
- Model for $Y_{1,i}$ ($P[Y_{1,i} = 1|T(x_i)] = \pi(x_i)$)

$$\log\left\{\frac{\pi(x_i)}{1-\pi(x_i)}\right\} = \mu_1 + \sigma_1[S_1(x) + T(x)]$$

- Model for $Y_{2,i}$

$$P[Y_{2,i} < c|S_2(x), T(x)] = 1 - \exp\left\{\left[\frac{c}{\mu(x_i)}\right]^{\kappa}\right\}$$

Bivariate modelling of prevalence and intensity of infection

- $Y_j(x_i) = Y_{i,1} \times Y_{2,i}$
- Model for $Y_{1,i}$ ($P[Y_{1,i} = 1|T(x_i)] = \pi(x_i)$)

$$\log\left\{\frac{\pi(x_i)}{1-\pi(x_i)}\right\} = \mu_1 + \sigma_1[S_1(x) + T(x)]$$

- Model for $Y_{2,i}$

$$\begin{aligned} P[Y_{2,i} < c|S_2(x), T(x)] &= 1 - \exp\left\{\left[\frac{c}{\mu(x_i)}\right]^{\kappa}\right\} \\ \log\{\mu(x_i)\} &= \mu_2 + \sigma_2[S_2(x) + T(x)] \end{aligned}$$

Bivariate modelling of prevalence and intensity of infection

- $Y_j(x_i) = Y_{i,1} \times Y_{2,i}$
- Model for $Y_{1,i}$ $(P[Y_{1,i} = 1|T(x_i)] = \pi(x_i))$

$$\log\left\{\frac{\pi(x_i)}{1-\pi(x_i)}\right\} = \mu_1 + \sigma_1[S_1(x) + T(x)]$$

- Model for $Y_{2,i}$

$$\begin{aligned} P[Y_{2,i} < c|S_2(x), T(x)] &= 1 - \exp\left\{\left[\frac{c}{\mu(x_i)}\right]^{\kappa}\right\} \\ \log\{\mu(x_i)\} &= \mu_2 + \sigma_2[S_2(x) + T(x)] \end{aligned}$$

- The standardized variogram:

$$\gamma_h(u) = 1 - \frac{1}{2}\left(\exp\{-u/\phi_{S_h}\} + \exp\{-u/\phi_T\}\right), \text{ for } h = 1, 2.$$

Bivariate modelling of prevalence and intensity of infection

- $Y_j(x_i) = Y_{i,1} \times Y_{2,i}$
- Model for $Y_{1,i}$ ($P[Y_{1,i} = 1|T(x_i)] = \pi(x_i)$)

$$\log\left\{\frac{\pi(x_i)}{1-\pi(x_i)}\right\} = \mu_1 + \sigma_1[S_1(x) + T(x)]$$

- Model for $Y_{2,i}$

$$\begin{aligned} P[Y_{2,i} < c|S_2(x), T(x)] &= 1 - \exp\left\{\left[\frac{c}{\mu(x_i)}\right]^{\kappa}\right\} \\ \log\{\mu(x_i)\} &= \mu_2 + \sigma_2[S_2(x) + T(x)] \end{aligned}$$

- The standardized variogram:

$$\gamma_h(u) = 1 - \frac{1}{2}\left(\exp\{-u/\phi_{S_h}\} + \exp\{-u/\phi_T\}\right), \text{ for } h = 1, 2.$$

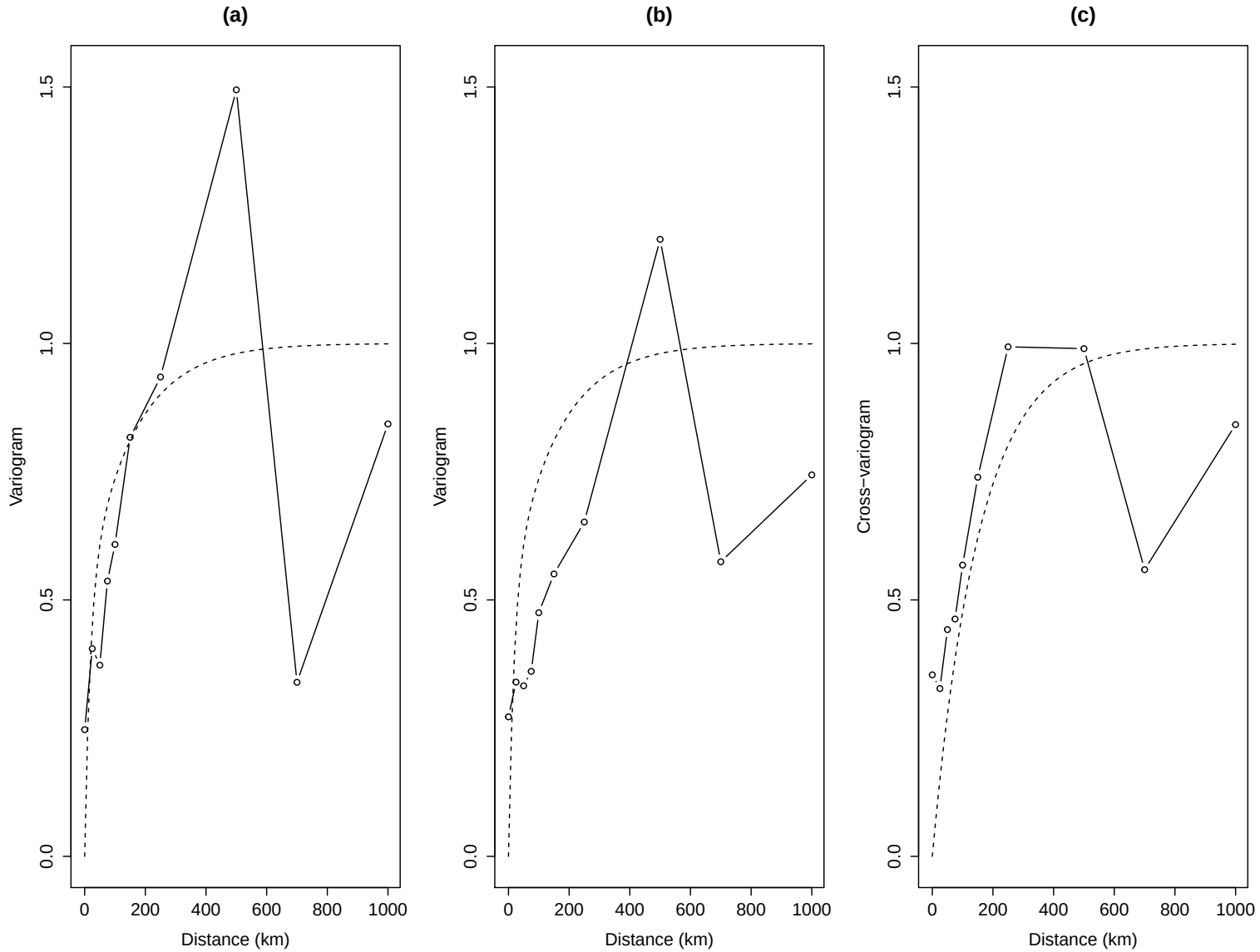
- The standardized cross-variogram:

$$\begin{aligned} \gamma_{12}(u) &= \frac{1}{2}E[(S_1(x) + T(x) - S_2(x') - T(x'))^2] \\ &= 1 - \exp\{-u/\phi_T\}. \end{aligned}$$

Monte Carlo maximum likelihood estimation

	Estimate	95% CI
μ_1	-2.187	(-2.230, -2.144)
μ_2	8.258	(8.190, 8.327)
σ_1^2	0.874	(0.663, 1.152)
σ_2^2	0.146	(0.111, 0.193)
ϕ_S	17.982	(13.012, 24.850)
ϕ_T	154.520	(72.402, 329.774)
κ	0.552	(0.537, 0.568)

Variogram check



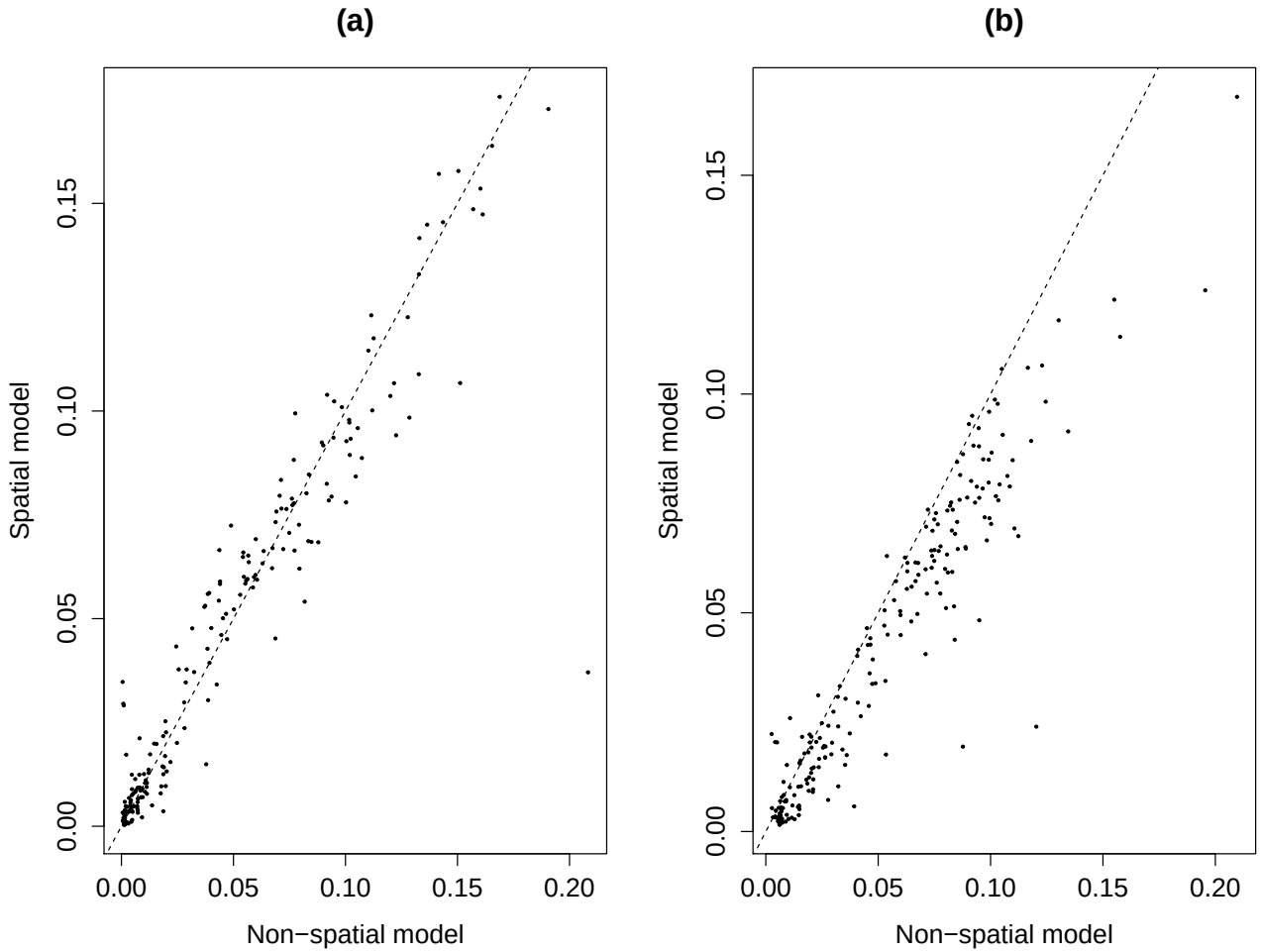
Mapping proportion of exceeding 8000 eggs per ml

Mapping proportion of exceeding 8000 eggs per ml

- **Non-spatial model**: Replace $S_1(x), S_2(x), T(x)$ with i.i.d. realisations of a bivariate Gaussian $(Z_{1,i}, Z_{2,i})$ with $\mathrm{cor}(Z_{1,i}, Z_{2,i}) = \rho$

Mapping proportion of exceeding 8000 eggs per ml

- **Non-spatial model**: Replace $S_1(x), S_2(x), T(x)$ with i.i.d. realisations of a bivariate Gaussian $(Z_{1,i}, Z_{2,i})$ with $\mathrm{cor}(Z_{1,i}, Z_{2,i}) = \rho$



Simulation study

Simulation study

- **Goal**: quantify the loss in efficiency of the non-spatial model.

Simulation study

- **Goal**: quantify the loss in efficiency of the non-spatial model.
- We simulate 1,000 data-sets under the fitted bivariate spatial model and fit two models: the spatial model (the true model); the non-spatial model (the misspecified model).

Simulation study

- **Goal**: quantify the loss in efficiency of the non-spatial model.
- We simulate 1,000 data-sets under the fitted bivariate spatial model and fit two models: the spatial model (the true model); the non-spatial model (the misspecified model).

Model	RMSE	PIL90	PIL95	PIL99	CP90	CP95	CP99
Spatial	0.013	0.033	0.040	0.053	0.895	0.947	0.989
Nonspatial	0.015	0.041	0.049	0.066	0.901	0.949	0.988

Discussion

- What about the other prevalence behaviours?

Discussion

- What about the other prevalence behaviours?
  - **Discontinuous**: Modelling $Y_{1,i}$ using an Ising process (?)

Discussion

- What about the other prevalence behaviours?
  - **Discontinuous**: Modelling $Y_{1,i}$ using an Ising process (?)
  - **Smoothly to zero**:

$$P[Y_i = y_i | S(x_i)] = \begin{cases} 0 & \text{if } S(x_i) < 0 \\ \text{Bin}(y_i; p(x_i)) & \text{if } S(x_i) \geqslant 0 \end{cases}$$

Discussion

- What about the other prevalence behaviours?
  - **Discontinuous**: Modelling $Y_{1,i}$ using an Ising process (?)
  - **Smoothly to zero**:

$$
\begin{aligned}
P[Y_i = y_i | S(x_i)] &= \begin{cases} 0 & \text{if } S(x_i) < 0 \\ \text{Bin}(y_i; p(x_i)) & \text{if } S(x_i) \geqslant 0 \end{cases} \\
p(x_i) &= 2\frac{\exp\{\alpha + S(x_i)\}}{1 + \exp\{\alpha + S(x_i)\}} - 1
\end{aligned}
$$

References

1. Diggle PJ, Giorgi E, (2019). Model-based Geostatistics for Global Public Health: Methods and Applications. Chapman & Hall/CRC.
2. Giorgi, E., Schl′uter, D. K., Diggle, P. J. (2017). Bivariate geostatistical modelling of the rela- tionship between Loa loa prevalence and intensity of infection. Environmetrics. doi:10.1002/env.2447
3. Diggle PJ, Giorgi E (2016). Model-based geostatistics for prevalence mapping in low-resource settings (with discussion). Journal of the American Statistical Association. 111:1096-1120