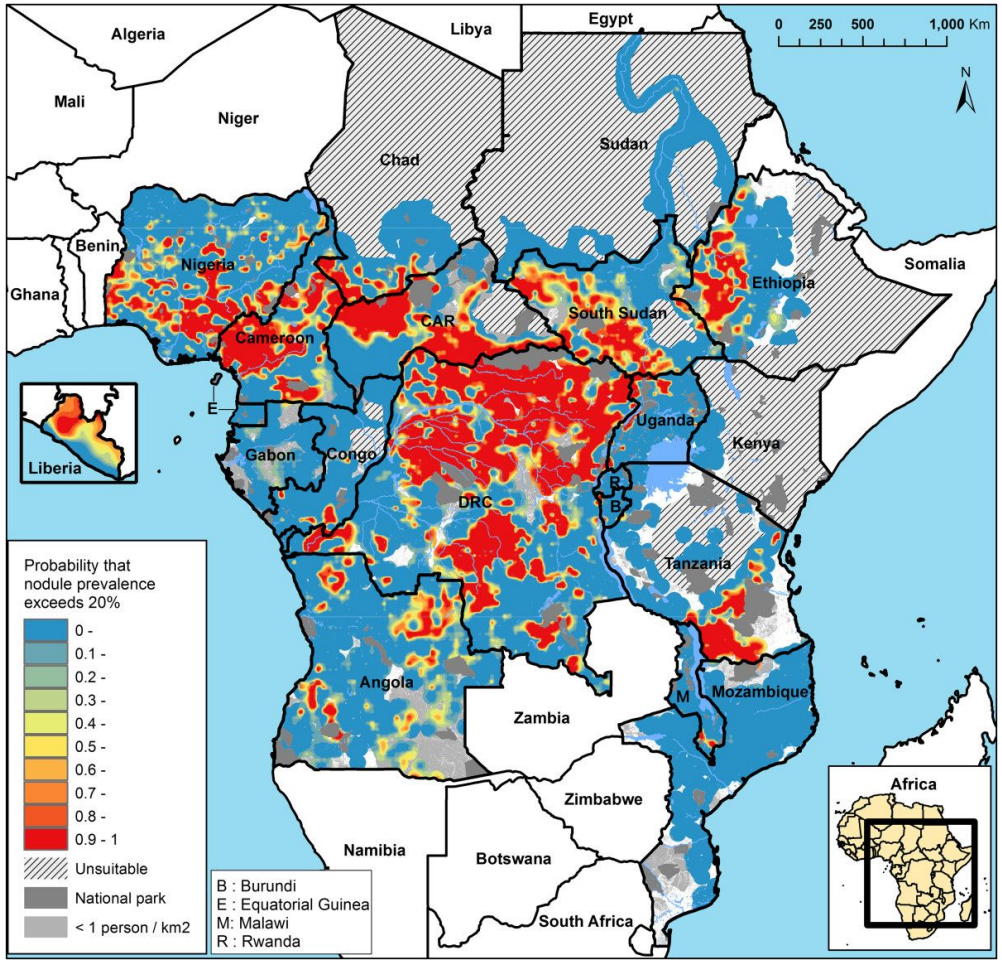
A geostatistical framework for combining spatially referenced prevalence data from multiple diagnostics

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

Toowoomba 21-25 October 2019

What is geostatistics?

- **Geostatistical problems:** making inference on a spatially continuous surface using noisy spatially discrete data.



Prevalence mapping

Prevalence mapping

- Data: (x_i, y_i, n_i), $x_i \in A$.

Prevalence mapping

- Data: (x_i, y_i, n_i), $x_i \in A$.
- Prevalence: $p(x_i)$

Prevalence mapping

- Data: (x_i, y_i, n_i), $x_i \in A$.
- Prevalence: $p(x_i)$
- Modelling spatial variation

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

Prevalence mapping

- Data: (x_i, y_i, n_i), $x_i \in A$.
- Prevalence: $p(x_i)$
- Modelling spatial variation

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

- $S(x)$ is a stationary and isotropic Gaussian process with correlation function

$$\mathrm{corr}\{S(x), S(x')\} = \rho(u; \phi).$$

Prevalence mapping

- Data: (x_i, y_i, n_i), $x_i \in A$.
- Prevalence: $p(x_i)$
- Modelling spatial variation

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

- $S(x)$ is a stationary and isotropic Gaussian process with correlation function

$$\mathrm{corr}\{S(x), S(x')\} = \rho(u; \phi).$$

- Objective: to predict $p(x)$ for all $x \in A$

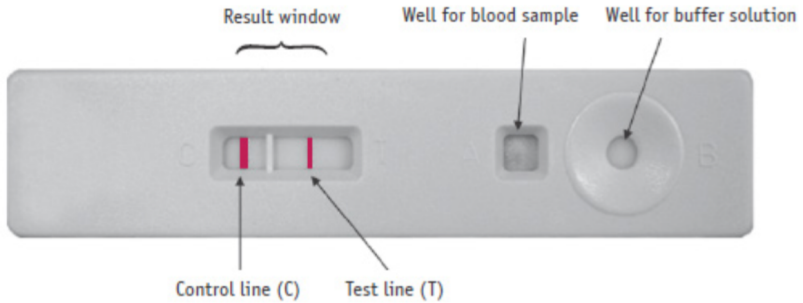
Combining multiple diagnostics

Case I: Predicting a gold-standard diagnostic using a more economic but biased alternative

Combining multiple diagnostics

Case II: Joint predictions of complementary diagnostics

Figure 1. Typical RDT for Malaria Diagnosis



Source: Good practices for selecting and procuring rapid diagnostic tests for malaria. Geneva: WHO; 2011.

A joint modelling framework

- The data:

$$\mathcal{D} = \{(x_{ik}, n_{ik}, y_{ik}) : i = 1, \ldots, n; k = 1, \ldots, K\} \tag{1}$$

A joint modelling framework

- The data:

$$\mathcal{D} = \{(x_{ik}, n_{ik}, y_{ik}) : i = 1, \ldots, n; k = 1, \ldots, K\} \tag{1}$$

- How should we jointly model the $p_k(x)$?

A joint modelling framework

- The data:

$$\mathcal{D} = \{(x_{ik}, n_{ik}, y_{ik}) : i = 1, \ldots, n; k = 1, \ldots, K\} \tag{1}$$

- How should we jointly model the $p_k(x)$?
- Crainiceanu, Diggle and Rowlingson (2008), or CDRM

$$\begin{cases} \text{logit}\{p_1(x_i)\} = d^\top(x_i)\beta + S(x_i) \\ \text{logit}\{p_2(x_i)\} = \alpha_0 + \alpha_1 \text{logit}\{p_1(x_i)\} + Z_i, \end{cases} \tag{2}$$

- Limitation: use of a single spatial process in (2) to capture spatial variation in both diagnostics.

Two classes of bivariate geostatistical models

Two classes of bivariate geostatistical models

- Case I

$$\begin{cases} f_1\{p_1(x_i)\} = d^\top(x_i)\beta_1 + S_1(x_i) + Z_{i1} \\ f_2\{p_2(x_i)\} = d^\top(x_i)\beta_2 + S_2(x_i) + Z_{i2} + \alpha f_1\{p_1(x_i)\}. \end{cases} \tag{3}$$

Two classes of bivariate geostatistical models

- Case I

$$\begin{cases} f_1\{p_1(x_i)\} = d^\top(x_i)\beta_1 + S_1(x_i) + Z_{i1} \\ f_2\{p_2(x_i)\} = d^\top(x_i)\beta_2 + S_2(x_i) + Z_{i2} + \alpha f_1\{p_1(x_i)\}. \end{cases} \tag{3}$$

- Case II

$$f_k\{p_{jk}(x_i)\} = d_{ij}^\top \beta_k + \nu_k \left[S_k(x_i) + T(x_i)\right] + Z_{ik}. \tag{4}$$

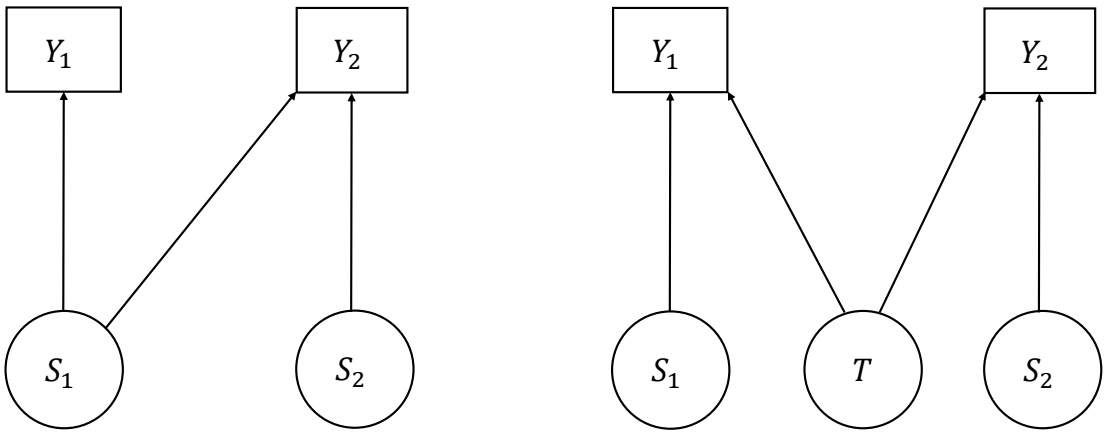
Two classes of bivariate geostatistical models

- Case I

$$\begin{cases} f_1\{p_1(x_i)\} = d^\top(x_i)\beta_1 + S_1(x_i) + Z_{i1} \\ f_2\{p_2(x_i)\} = d^\top(x_i)\beta_2 + S_2(x_i) + Z_{i2} + \alpha f_1\{p_1(x_i)\}. \end{cases} \tag{3}$$

- Case II

$$f_k\{p_{jk}(x_i)\} = d_{ij}^\top \beta_k + \nu_k \left[S_k(x_i) + T(x_i)\right] + Z_{ik}. \tag{4}$$



(a) (b)

Inference

Inference

- $W_k = \{W_k(x_i); i = 1, \ldots, N\}$ (e.g. in Case II, $W_k(x_i) = \nu_k[S_k(x_i) + T(x_i)]$)

Inference

- $W_k = \{W_k(x_i); i = 1, \ldots, N\}$ (e.g. in Case II, $W_k(x_i) = \nu_k[S_k(x_i) + T(x_i)]$)
- Let $h(\cdot)$ be "the density function of $\cdot$".

Inference

- $W_k = \{W_k(x_i); i = 1, \ldots, N\}$ (e.g. in Case II, $W_k(x_i) = \nu_k[S_k(x_i) + T(x_i)]$)
- Let $h(\cdot)$ be "the density function of $\cdot$".
- The likelihood function for θ

$$L(\theta) = \int h(y|w)h(w)dw \tag{5}$$

Inference

- $W_k = \{W_k(x_i); i = 1, \ldots, N\}$ (e.g. in Case II, $W_k(x_i) = \nu_k[S_k(x_i) + T(x_i)]$)
- Let $h(\cdot)$ be "the density function of $\cdot$".
- The likelihood function for θ

$$L(\theta) = \int h(y|w)h(w)dw \tag{5}$$

- We approximate (5) using MCMC to give

$$L(\theta) \approx L_B(\theta) = \frac{1}{B}\sum_{b=1}^{B} \frac{h(w_{(b)})h(y|w_{(b)})}{h_0(w_{(b)})h_0(y|w_{(b)})}, \tag{6}$$

Model validation

- The empirical variogram for W_k

$$\hat{\gamma}_k(u) = \frac{1}{2|N(u)|} \sum_{(i,j)\in N(u)} \left\{ \hat{W}_k(x_i) - \hat{W}_k(x_j) \right\}^2, \tag{7}$$

- A variogram-based validation algorithm.

Model validation

- The empirical variogram for W_k

$$\hat{\gamma}_k(u) = \frac{1}{2|N(u)|} \sum_{(i,j)\in N(u)} \left\{ \hat{W}_k(x_i) - \hat{W}_k(x_j) \right\}^2, \tag{7}$$

- A variogram-based validation algorithm.
  1. Obtain $\hat{W}_k(x_i)$ from two separate standard geostatistical models and compute the empirical variogram $\hat{\gamma}_k$, for $k = 1, 2$.

Model validation

- The empirical variogram for W_k

$$\hat{\gamma}_k(u) = \frac{1}{2|N(u)|} \sum_{(i,j)\in N(u)} \left\{ \hat{W}_k(x_i) - \hat{W}_k(x_j) \right\}^2, \tag{7}$$

- A variogram-based validation algorithm.
  1. Obtain $\hat{W}_k(x_i)$ from two separate standard geostatistical models and compute the empirical variogram $\hat{\gamma}_k$, for $k = 1, 2$.
  2. Simulate prevalence data as in (1) under the fitted model.

Model validation

- The empirical variogram for W_k

$$\hat{\gamma}_k(u) = \frac{1}{2|N(u)|} \sum_{(i,j)\in N(u)} \left\{ \hat{W}_k(x_i) - \hat{W}_k(x_j) \right\}^2, \qquad (7)$$

- A variogram-based validation algorithm.
  1. Obtain $\hat{W}_k(x_i)$ from two separate standard geostatistical models and compute the empirical variogram $\hat{\gamma}_k$, for $k = 1, 2$.
  2. Simulate prevalence data as in (1) under the fitted model.
  3. Fit separate standard geostatistical models as in 1 and compute the empirical variogram for the simulated dataset.

Model validation

- The empirical variogram for W_k

$$\hat{\gamma}_k(u) = \frac{1}{2|N(u)|} \sum_{(i,j)\in N(u)} \left\{ \hat{W}_k(x_i) - \hat{W}_k(x_j) \right\}^2, \qquad (7)$$

- A variogram-based validation algorithm.
  1. Obtain $\hat{W}_k(x_i)$ from two separate standard geostatistical models and compute the empirical variogram $\hat{\gamma}_k$, for $k = 1, 2$.
  2. Simulate prevalence data as in (1) under the fitted model.
  3. Fit separate standard geostatistical models as in 1 and compute the empirical variogram for the simulated dataset.
  4. Repeat 2-3 a large enough number of times, say M.

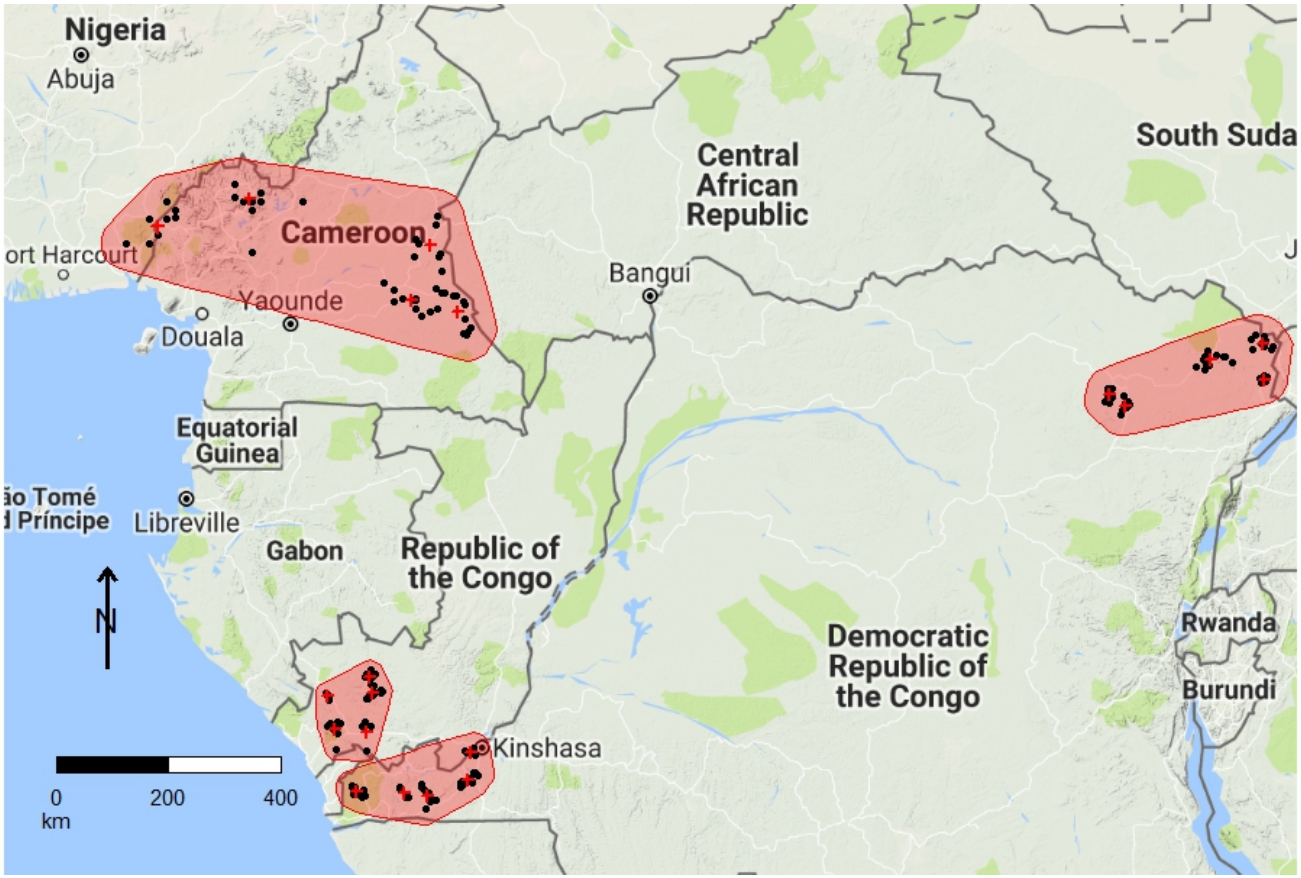
Model validation

- The empirical variogram for W_k

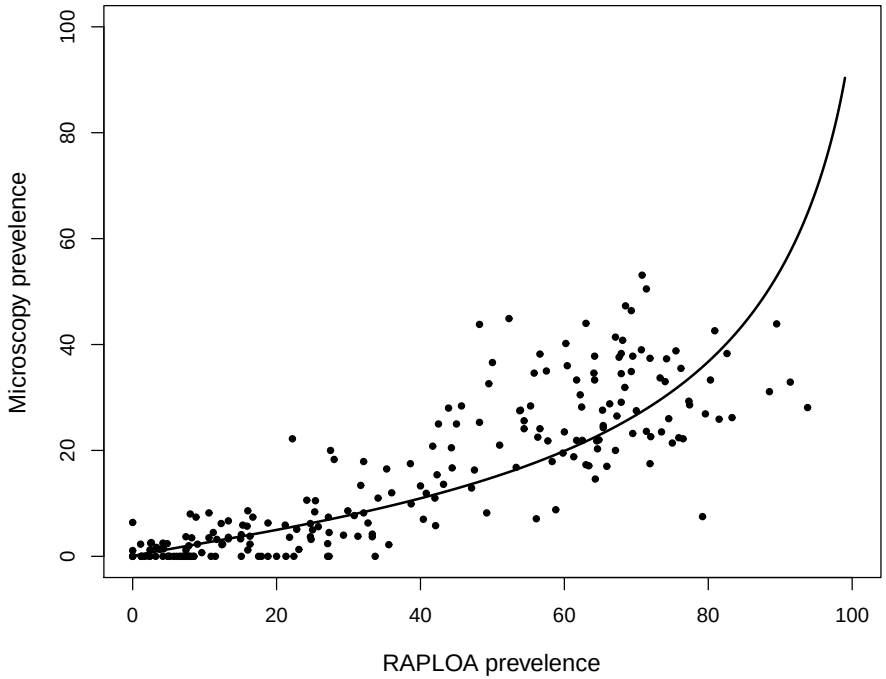
$$\hat{\gamma}_k(u) = \frac{1}{2|N(u)|} \sum_{(i,j)\in N(u)} \left\{ \hat{W}_k(x_i) - \hat{W}_k(x_j) \right\}^2, \qquad (7)$$

- A variogram-based validation algorithm.
  1. Obtain $\hat{W}_k(x_i)$ from two separate standard geostatistical models and compute the empirical variogram $\hat{\gamma}_k$, for $k = 1, 2$.
  2. Simulate prevalence data as in (1) under the fitted model.
  3. Fit separate standard geostatistical models as in 1 and compute the empirical variogram for the simulated dataset.
  4. Repeat 2-3 a large enough number of times, say M.
  5. Use the resulting M empirical variograms to generate 95% confidence intervals.

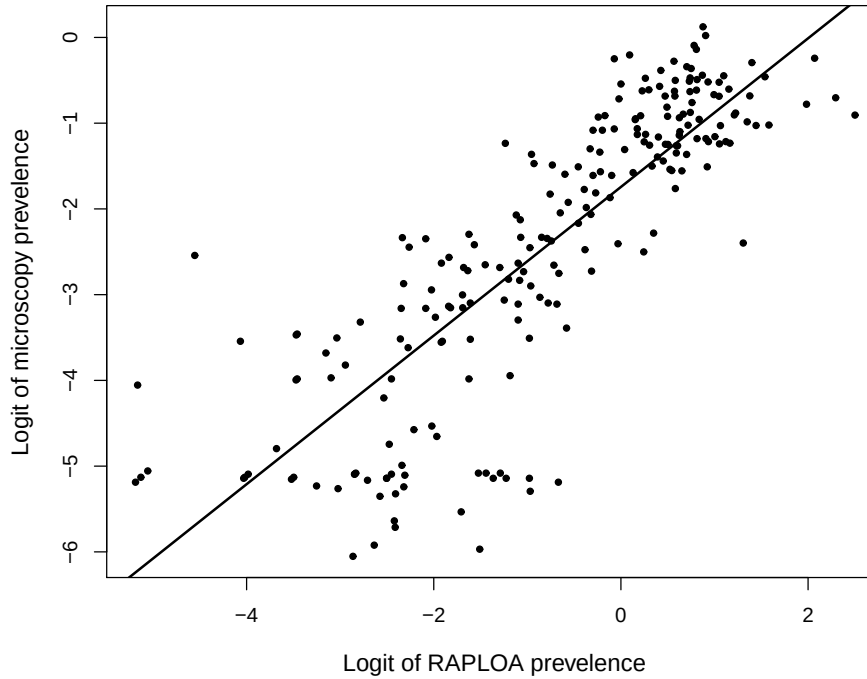
Application: re-analysis of the *Loa loa* data



Application: re-analysis of the *Loa loa* data

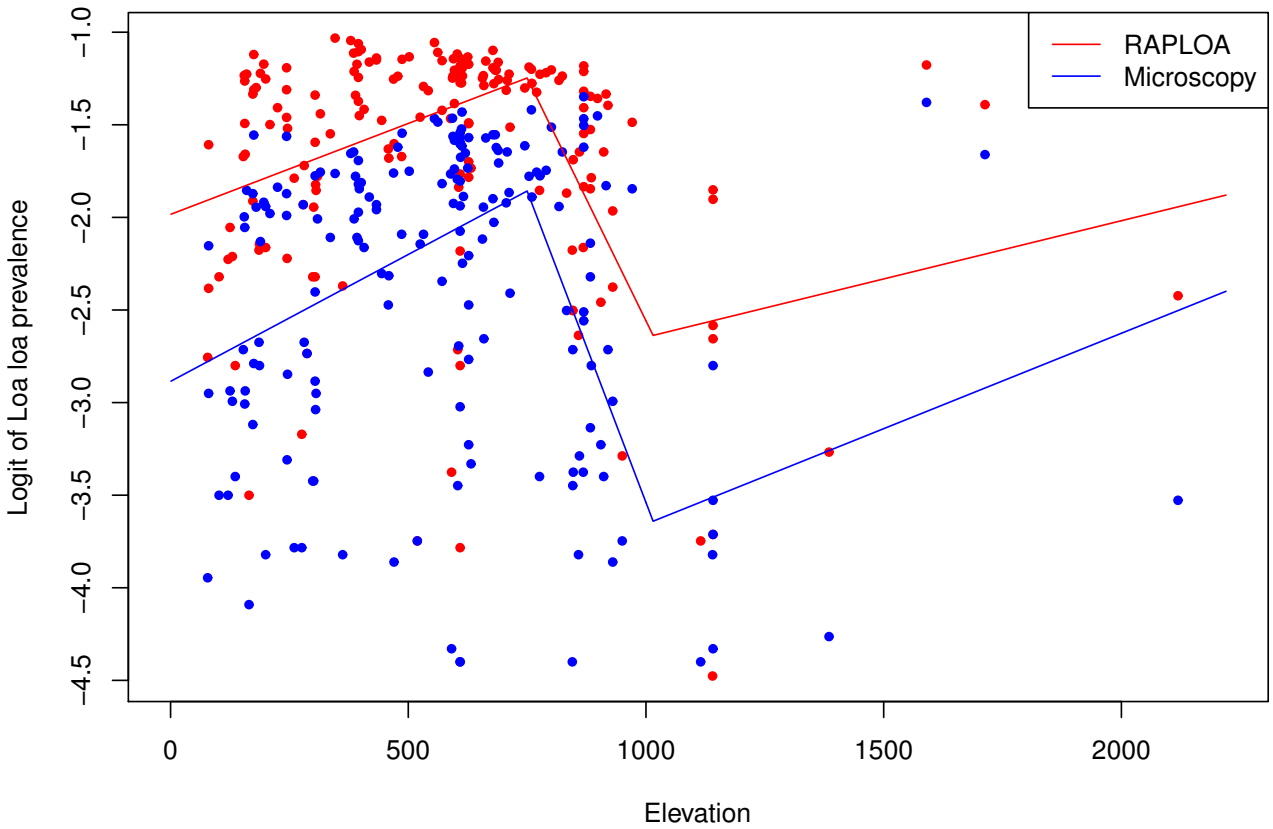


(a)



(b)

Application: re-analysis of the *Loa loa* data



Application: re-analysis of the *Loa loa* data

Application: re-analysis of the *Loa loa* data

- *Model 1*

$$\begin{cases} \text{logit}\{p_1(x_i)\} = \mu_1(x_i) + S_1(x_i) + Z_{i1} \\ \text{logit}\{p_2(x_i)\} = \mu_2(x_i) + \alpha\text{logit}\{p_1(x_i)\} + Z_{i2} \end{cases} \tag{8}$$

Application: re-analysis of the *Loa loa* data

- *Model 1*

$$\begin{cases} \text{logit}\{p_1(x_i)\} = \mu_1(x_i) + S_1(x_i) + Z_{i1} \\ \text{logit}\{p_2(x_i)\} = \mu_2(x_i) + \alpha\text{logit}\{p_1(x_i)\} + Z_{i2} \end{cases} \tag{8}$$

- *Model 2*

$$\begin{cases} \text{logit}\{p_1(x_i)\} = \mu_1(x_i) + S_1(x_i) + Z_{i1} \\ \text{logit}\{p_2(x_i)\} = \mu_2(x_i) + \alpha\text{logit}\{p_1(x_i)\} + S_2(x_i) + Z_{i2} \end{cases} \tag{9}$$

- Spatial correlation

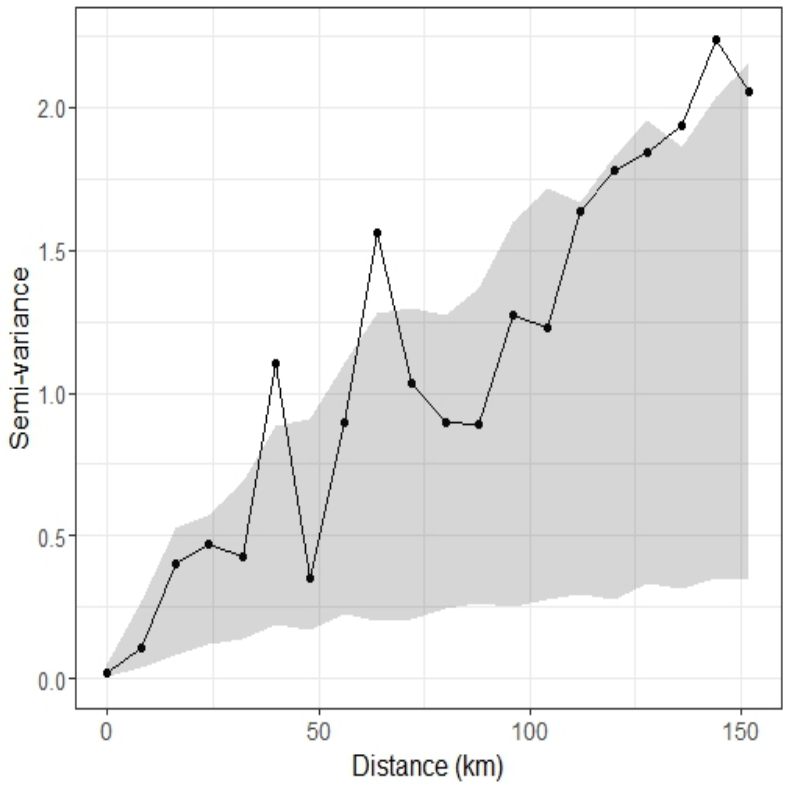
$$\text{corr}\{S_k(x), S_k(x')\} = \exp\{-\|x - x'\|/\phi_k\}$$

Application: re-analysis of the *Loa loa* data

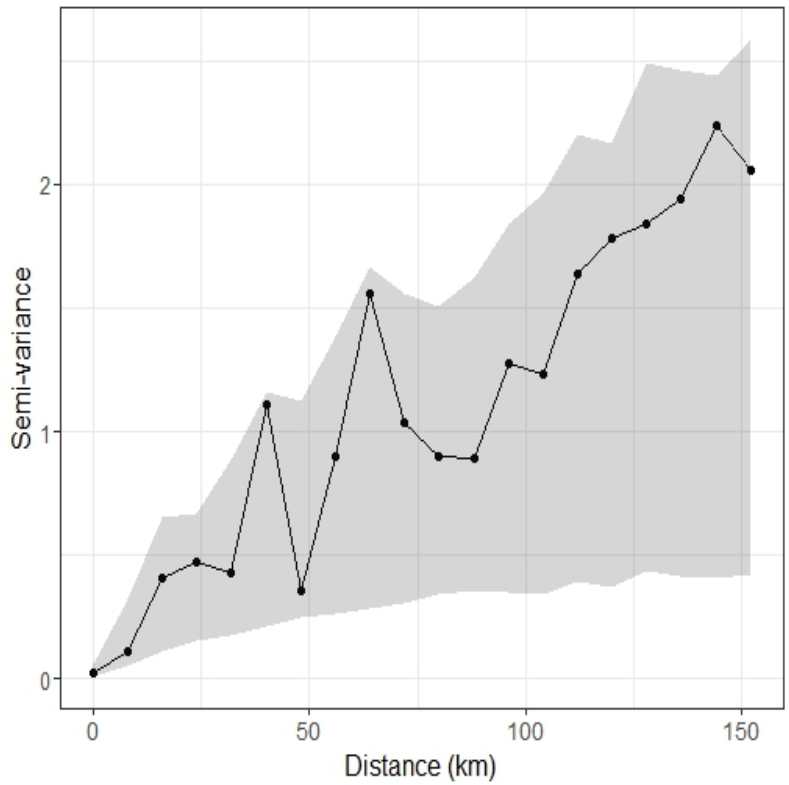
Parameter	Model 1	Model 2
σ_1^2	1.581	1.617
	(0.669, 3.738)	(0.679, 3.851)
σ_2^2	—	0.216
		(0.111, 0.419)
ϕ_1	182.037	187.388
	(64.657, 512.512)	(65.171, 538.807)
ϕ_2	—	23.686
		(6.150, 91.220)
τ_1^2	0.205	0.324
	(0.081, 0.521)	(0.052, 6.229)
τ_2^2	0.324	0.104
	(0.055, 5.873)	(0.018, 5.797)
α	1.005	1.017
	(0.902, 1.107)	(0.939, 1.095)

Application: re-analysis of the *Loa loa* data

Validation of Model 1 (a) and Model 2 (b) Validation *Model 1*



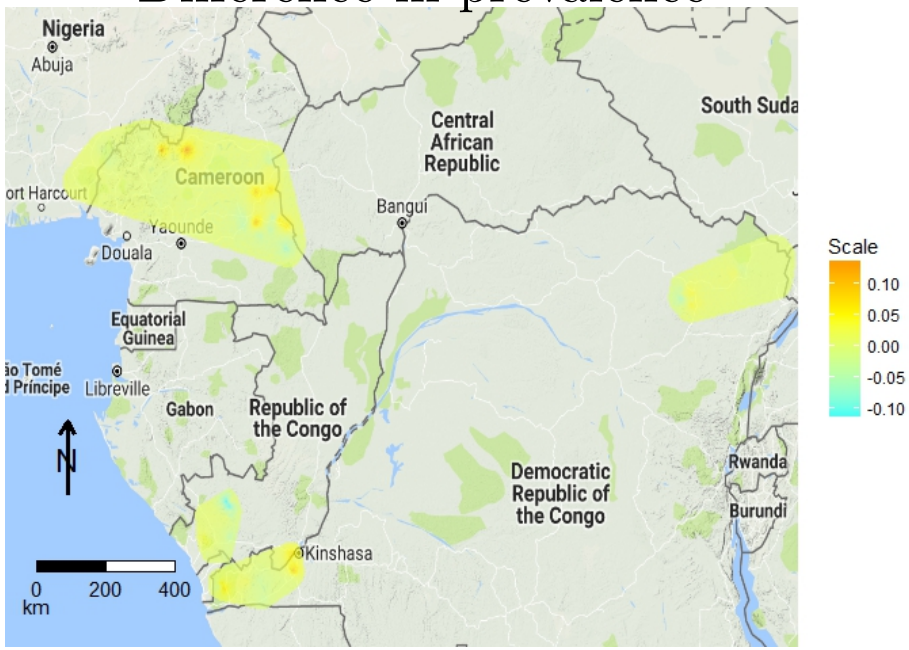
(a)



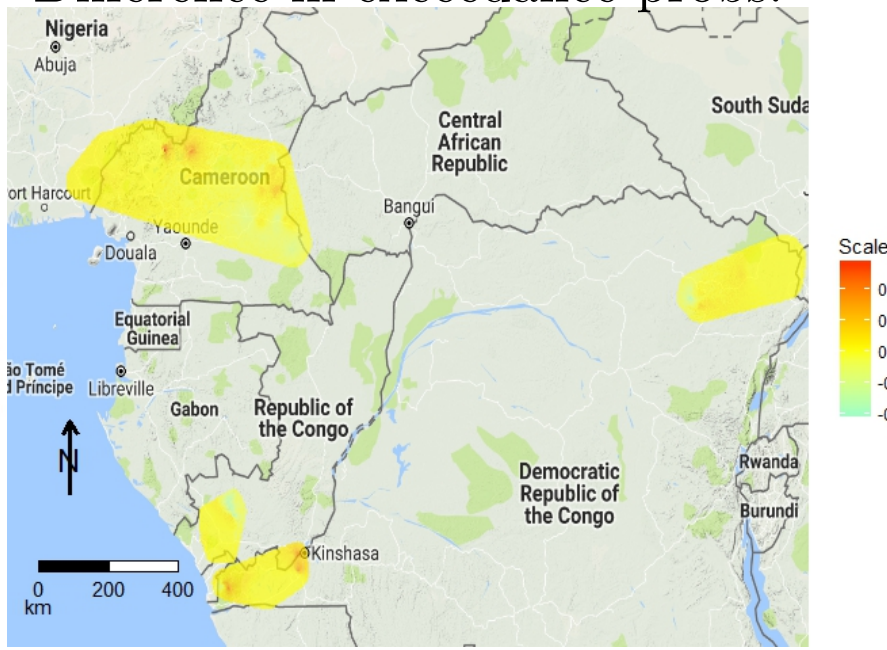
(b)

Application: re-analysis of the *Loa loa* data

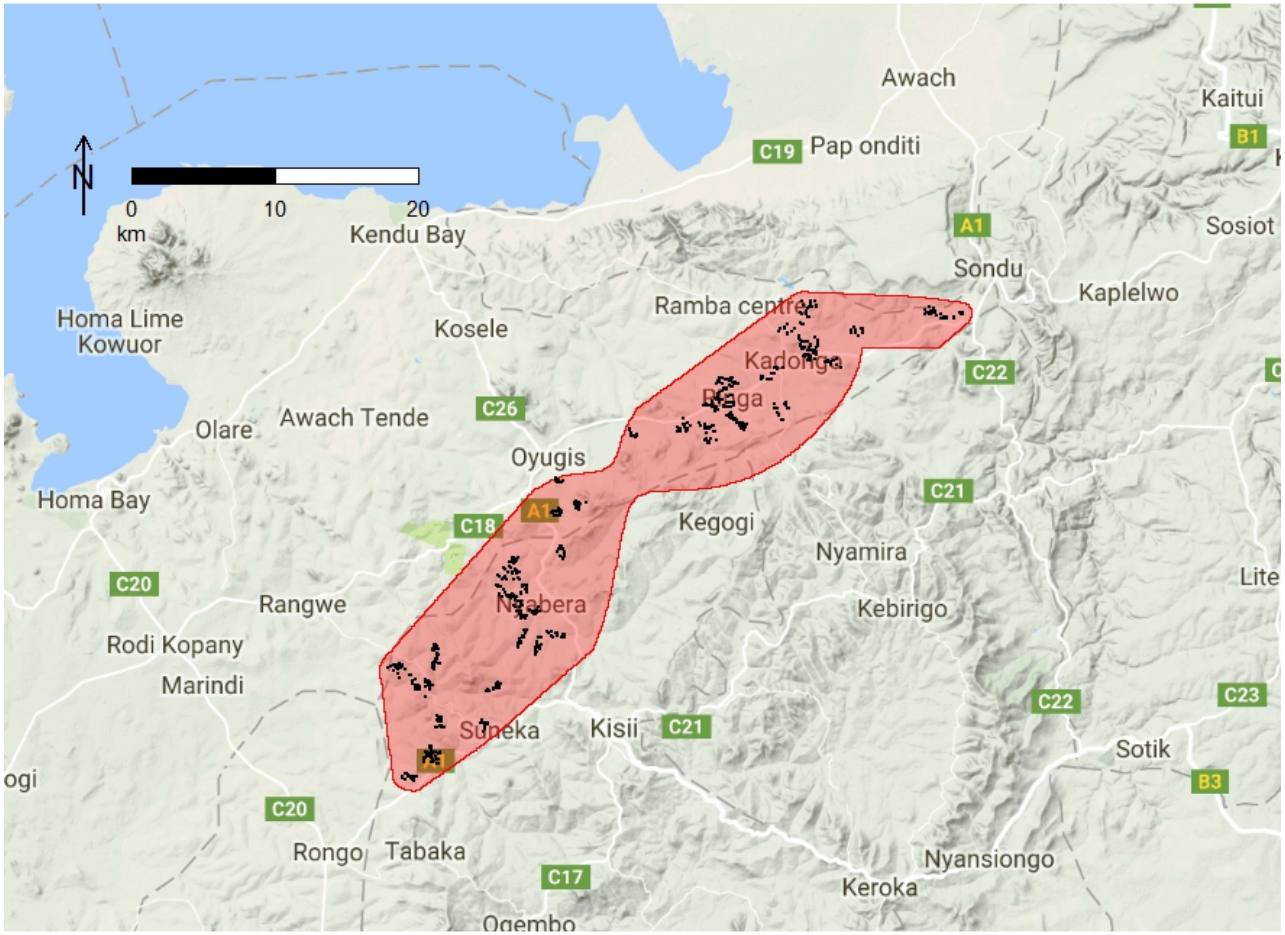
Difference in prevalence



Difference in exceedance probs.



Application: Joint prediction of malaria prevalence using RDT and PCR



Application: Joint prediction of malaria prevalence using RDT and PCR

- Our model for the data

$$\log\left\{\frac{p_{jk}(x_i)}{1-p_{jk}(x_i)}\right\} = \beta_{k,0} + \sum_{l=1}^{3}\beta_{k,l}d_{ij,l} + \nu_k T(x_i) \qquad (10)$$

Application: Joint prediction of malaria prevalence using RDT and PCR

- Our model for the data

$$\log\left\{\frac{p_{jk}(x_i)}{1-p_{jk}(x_i)}\right\} = \beta_{k,0} + \sum_{l=1}^{3}\beta_{k,l}d_{ij,l} + \nu_k T(x_i) \qquad (10)$$

- Question 1: what is the most parsimonious and empirically compatible model?

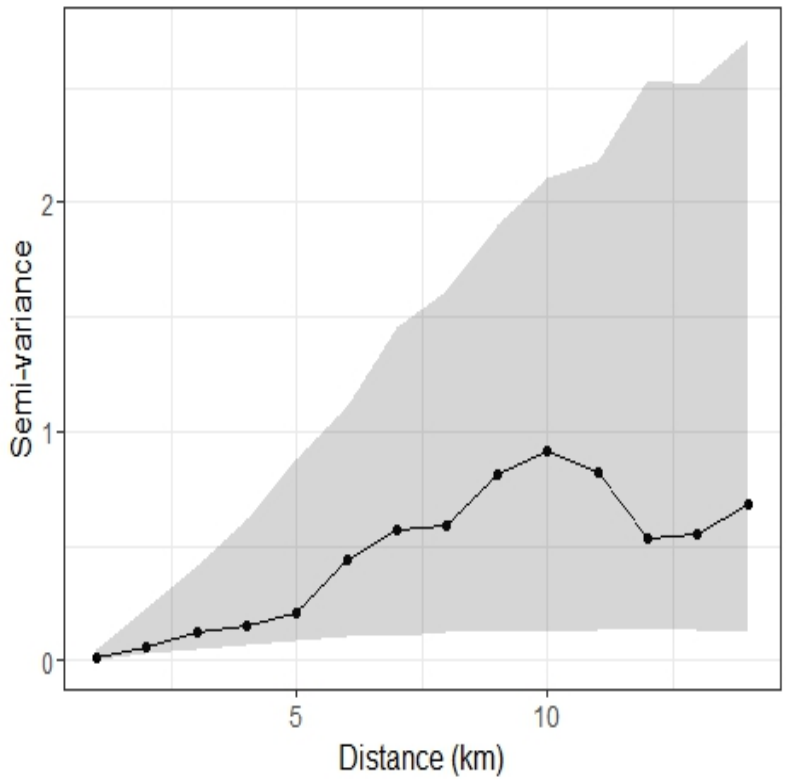
Application: Joint prediction of malaria prevalence using RDT and PCR

- Our model for the data

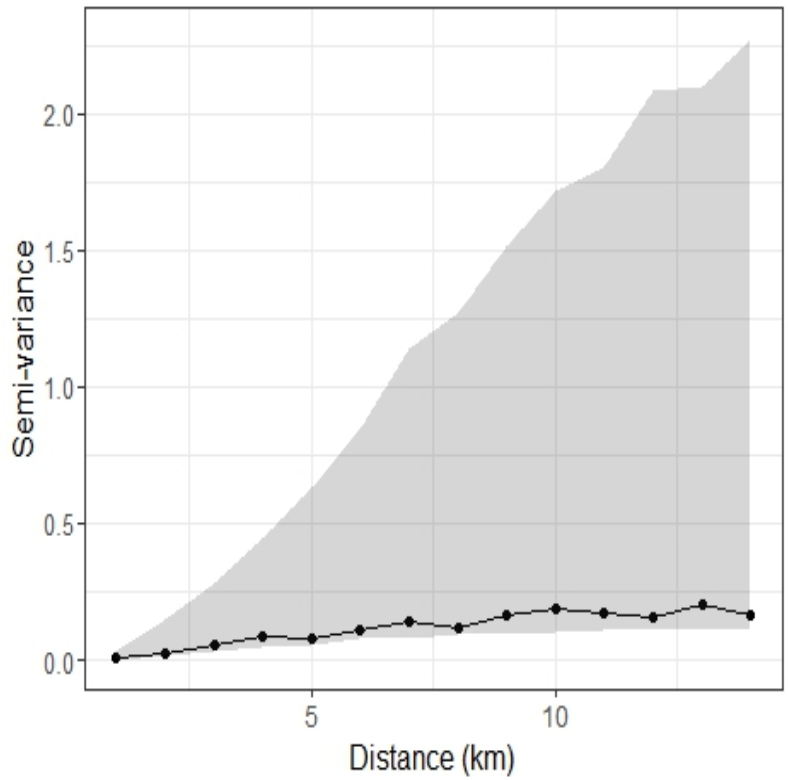
$$\log\left\{\frac{p_{jk}(x_i)}{1-p_{jk}(x_i)}\right\} = \beta_{k,0} + \sum_{l=1}^{3}\beta_{k,l}d_{ij,l} + \nu_k T(x_i) \qquad (10)$$

- Question 1: what is the most parsimonious and empirically compatible model?
- Question 2: is a joint model an *improvement* over two separate models?

Application: Joint prediction of malaria prevalence using RDT and PCR



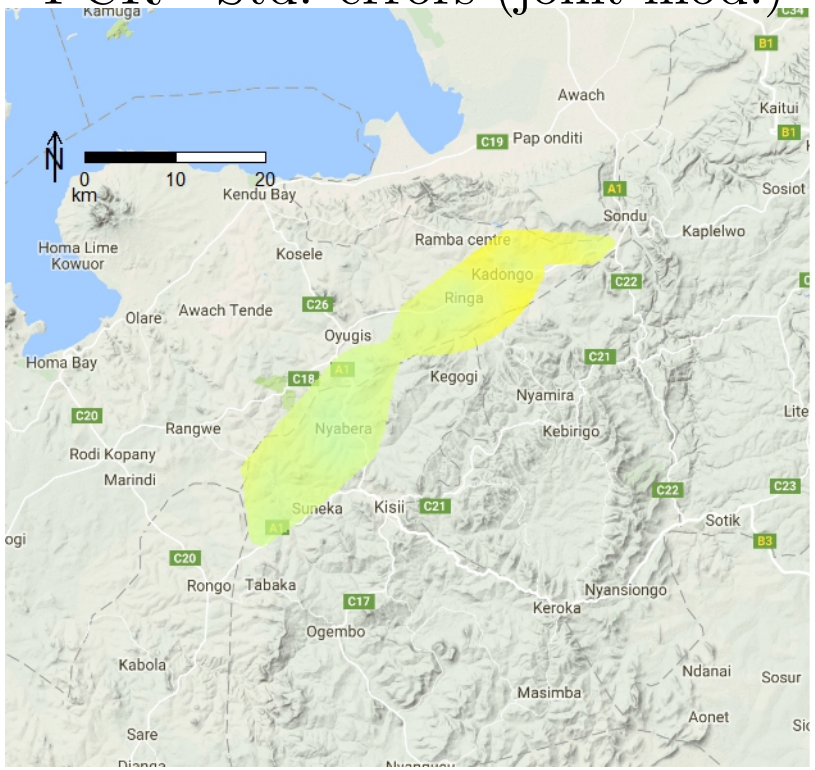
(a)



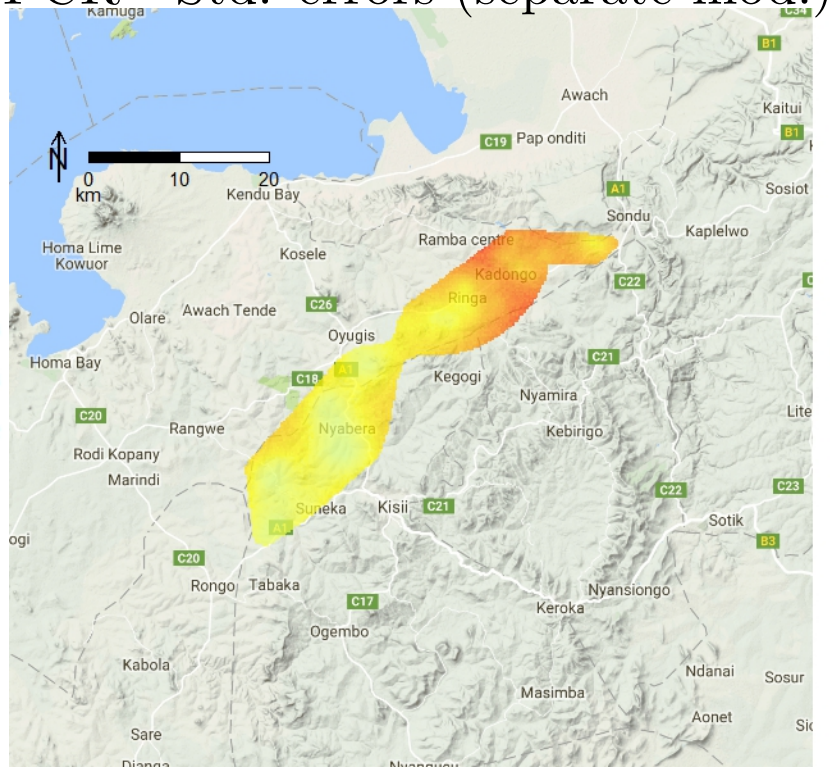
(b)

Application: Joint prediction of malaria prevalence using RDT and PCR

PCR - Std. errors (joint mod.)

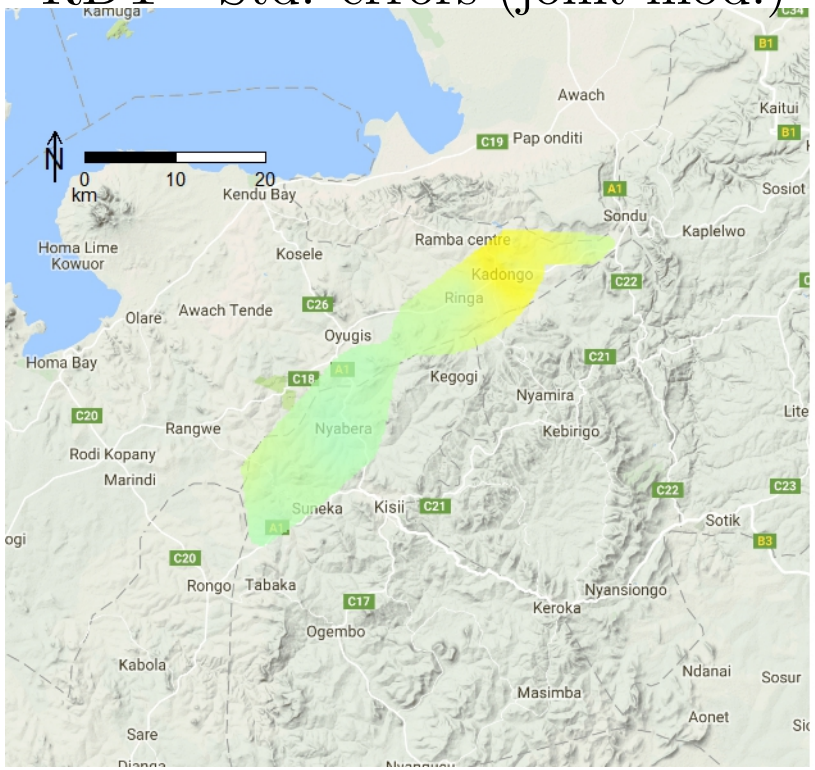


PCR - Std. errors (separate mod.)

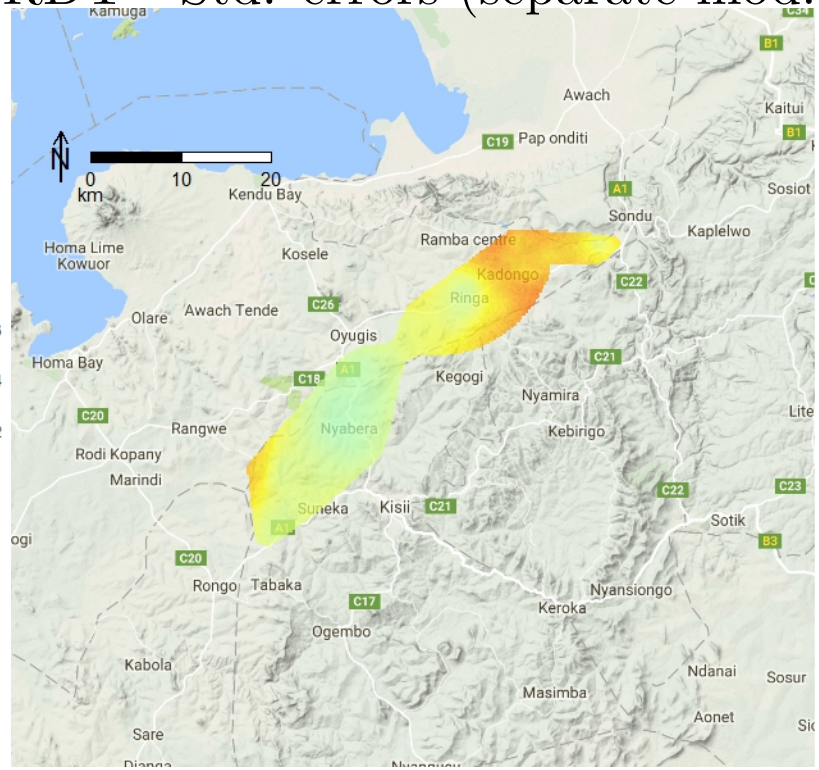


Application: Joint prediction of malaria prevalence using RDT and PCR

RDT - Std. errors (joint mod.)



RDT - Std. errors (separate mod.)



Conclusions

- The problem should inform the modelling approach.
- It is important to develop methods of inference that allow the borrowing of strength of information across multiple diagnostics.
- Ignoring this diagnostic specific spatial variation can lead to unreliably narrow prediction intervals for prevalence.
- Methodology can be easily extended to more than two diagnostics.