Model-Based Geostatistics for Prevalence Mapping in Low-Resource Settings

Peter J Diggle, Emanuele Giorgi and Daniela Schlüter

Lancaster University

References

Diggle, P.J., Thomson, M.C., Christensen, O.F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, H., Boussinesq, M. and Molyneux, D.H. (2007). Spatial modelling and prediction of Loa loa risk: decision making under uncertainty. *Annals of Tropical Medicine and Parasitology*, **101**, 499–509.

Zoure, H.G.M., Noma, M., Tekle, A.H., Amazigo, U.V., Diggle, P.J., Giorgi, E. and Remme, J.H.F. (2014). The geographic distribution of onchocerciasis in the 20 participating countries of the African Programme for Onchocerciasis Control: 2. Pre-control endemicity Levels and estimated number infected. *Parasites and Vectors*, **7**, 326

Diggle, P.J. and Giorgi, E. (2016). Model-based geostatistics for prevalence Mapping in low-resource settings (with Discussion). *Journal of the American Statistical Association* (to appear)

Schlüter, D.K., Ndeffo-Mbah, M.L., Takougang, I., Ukety, T., Wanji, S., Galvani, A.P. and Diggle, P.J. (2016). Using community-level prevalence of Loa loa infection to predict the proportion of highly-infected individuals: statistical modelling to support lymphatic filariasis elimination programs. *PLoS Neglected Tropical Diseases* (submitted).

R packages: `geoR`, `PrevMap`

Acknowledgements

MLW, Blantyre, Malawi Sanie Sesay, Anja Terlouw

APOC, Ouagadougou: Hans Remme, Honorat Zoure, Sam Wanji

IRI, Columbia University: Madeleine Thomson

...and many others

Low resource settings

Prevalence mapping

Single prevalence survey

Sample n individuals, observe Y positives

$$Y \sim \text{Bin}(n, p)$$

Multiple prevalence surveys

Sample n_i individuals, observe Y_i positives, $i = 1, ..., m$

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

Extra-binomial variation

Sample n_i individuals, observe Y_i positives, $i = 1, ..., m$

$$Y_i | d_i, U_i \sim \text{Bin}(n_i, p_i) \quad \log\{p_i/(1-p_i)\} = d_i'\beta + U_i$$

This talk

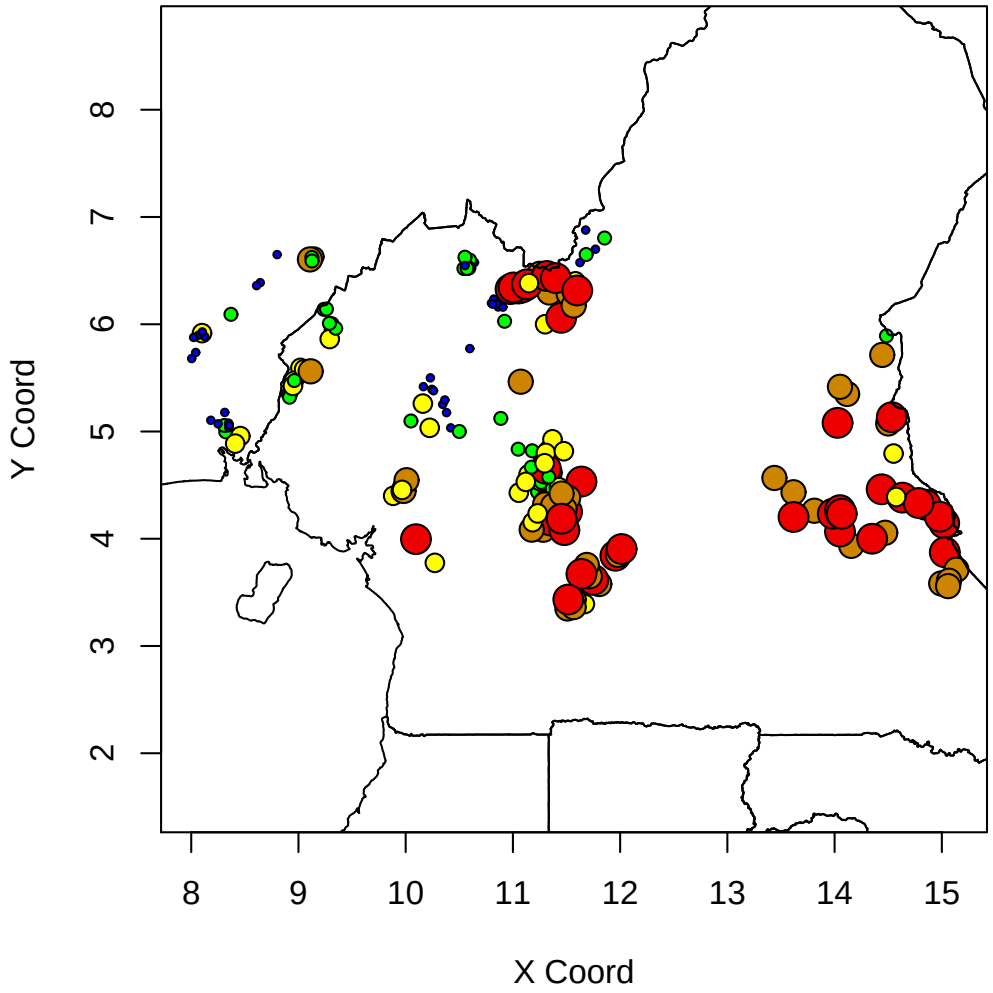
What to do if the d_i and/or the U_i are spatially structured

Geostatistics

- **Traditionally** a self-contained methodology for spatial prediction, developed at École des Mines, Fontainebleau, France
- **Nowadays** that part of spatial statistics that is concerned with data obtained by spatially discrete sampling of a spatially continuous process
- **Geostatistical prevalence data**

$$(n_i, y_i, d_i, x_i) : i = 1, ..., n$$

Loa loa prevalence surveys in West Africa



Model-based (Diggle, Moyeed and Tawn, 1998)

- **The application of general principles of statistical modelling and inference to geostatistical problems**
  - **formulate a model for the data**
  - **use likelihood-based methods of inference**
  - **answer the scientific question**
- **Design is also important, but not considered in DM&T (1998).**

Statistical modelling principles

- models are devices to answer questions
- models should:
  - be not demonstrably inconsistent with the data;
  - incorporate the underlying science, where this is well understood
  - be as simple as possible, within the above constraints

"Too many notes, Mozart"

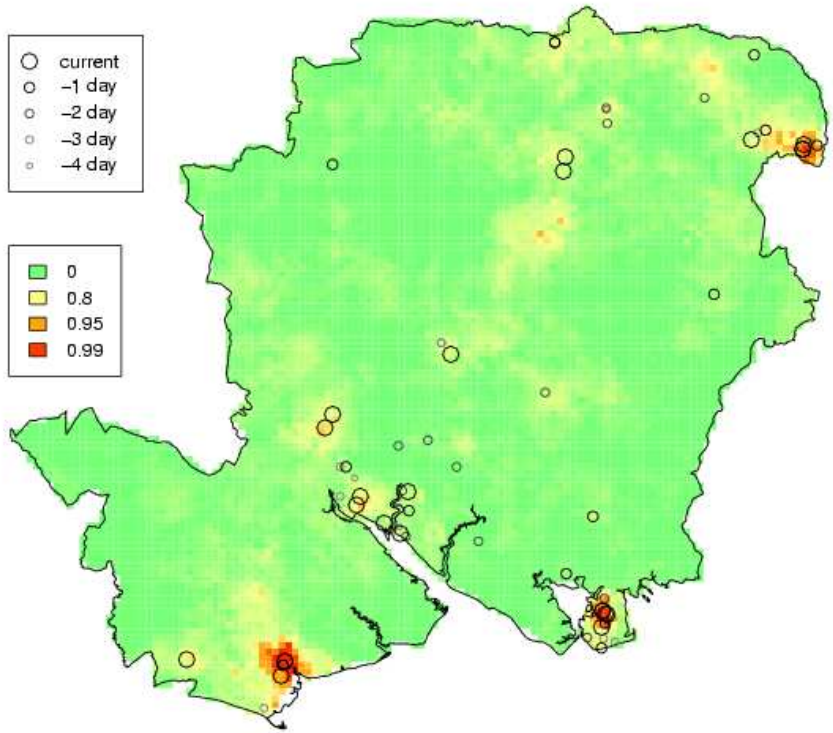
Emperor Joseph II

"Only as many as there needed to be"

Mozart (apochryphal?)

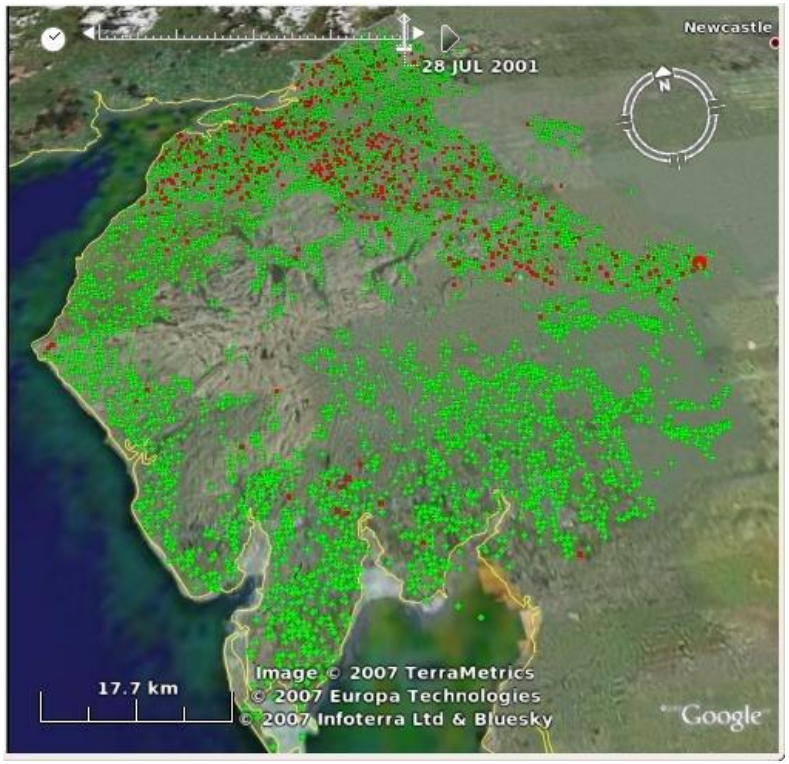
Empirical modelling: The AEGISS project (Diggle, Rowlingson and Su, 2005)

- **early detection of anomalies in local incidence**
- **data on 3374 consecutive reports of non-specific gastro-intestinal illness**
- **log-Gaussian Cox process, space-time correlation $\rho(u, v)$**



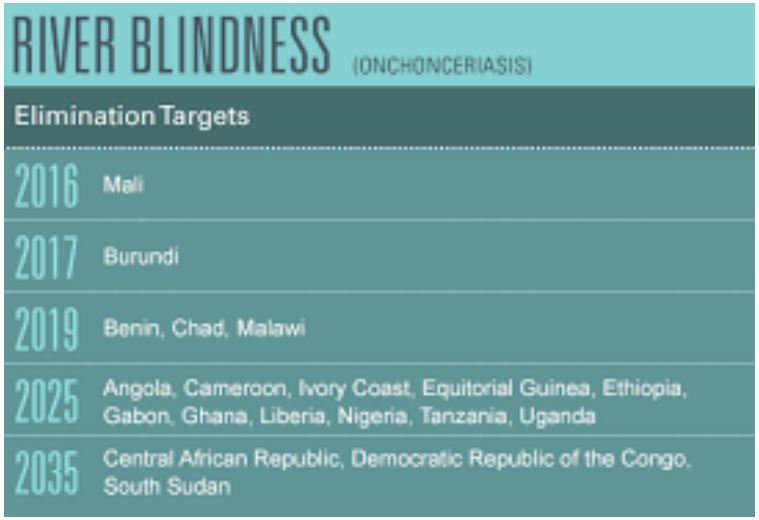
Mechanistic modelling: the 2001 UK FMD epidemic (Diggle, 2006)

- **Predominantly a classic epidemic pattern of spread from an initial source**
- **Occasional apparently spontaneous outbreaks remote from prevalent cases**
- **$\lambda(x, t|\mathcal{H}_t)$ =conditional intensity, given history $\mathcal{H}_t$**



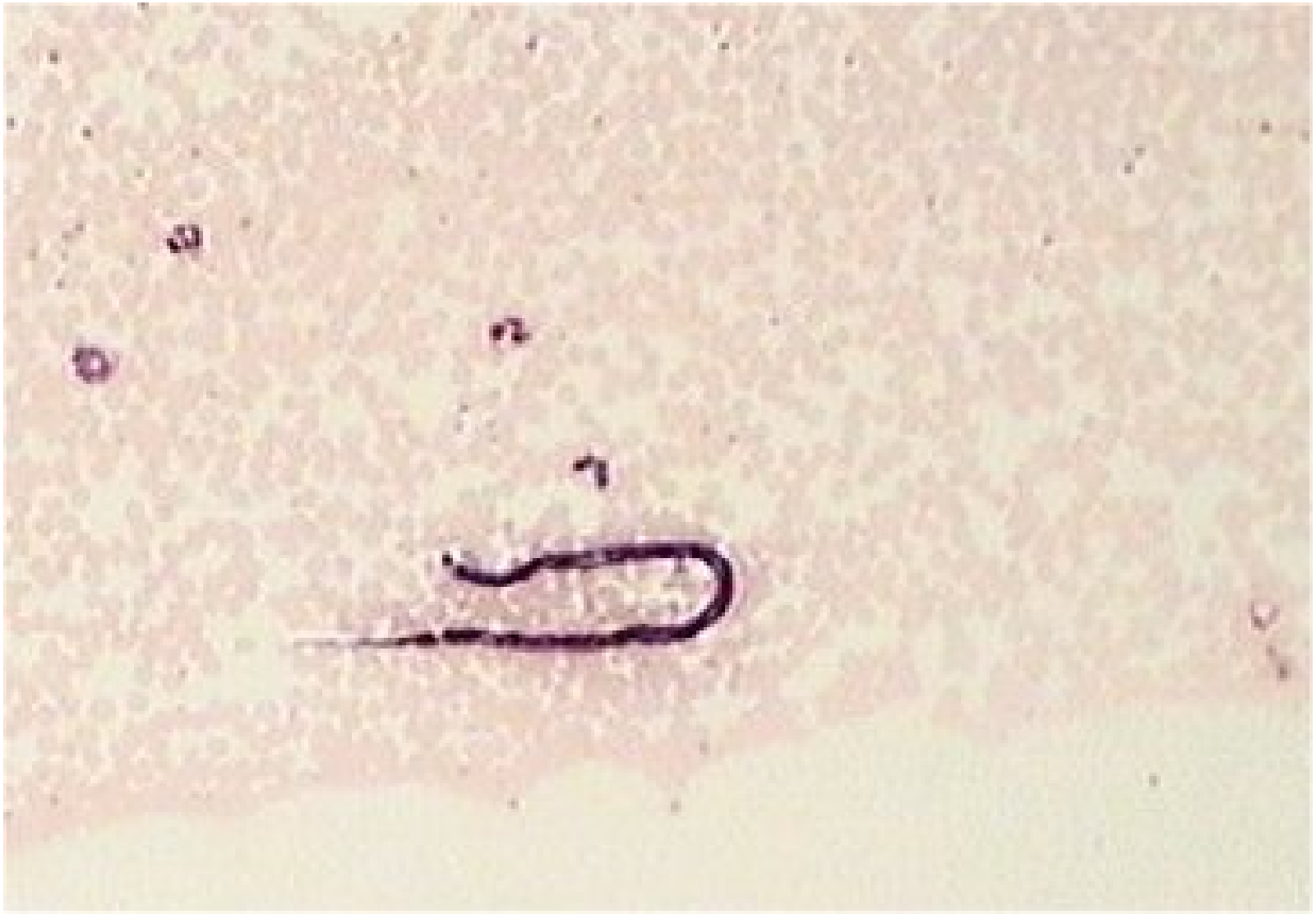
Onchocerciasis (River Blindness)

African Programme for Onchocerciasis Control (APOC)



- **Ivermectin (Mectizan): provides long-term protection if taken annually**
- **generally considered safe, with no serious side-effects**
- **mass distribution made possible by donation programme (Merck)**
- **multi-national programme coordinated by WHO**
- **recent decision to raise ambition from control to elimination**
- **Loa loa: a spanner in the works**

Loa loa young



...and old

The Loa loa prediction problem

Ground-truth survey data

- **random sample of subjects in each of a number of villages**
- **blood-samples test positive/negative for *Loa loa***

Environmental data (satellite images)

- **measured on regular grid to cover region of interest**
- **elevation, green-ness of vegetation**

Objectives

- **predict local prevalence throughout study-region (Cameroon)**
- **compute local exceedance probabilities,**

$$\mathrm{P}(\text{prevalence} > 0.2 | \text{data})$$

Statistical prediction: Bayes' Theorem

"The answer to any prediction problem is a probability distribution"

Peter McCullagh

S = state of nature
Y = all relevant data
T = $\mathcal{F}(S)$ = target for prediction

Model: $[S, Y] = [S][Y|S]$
Prediction: $[S, Y] \Rightarrow [S|Y] \Rightarrow [T|Y]$

The Loa loa modelling strategy

- use relationship between environmental variables and ground-truth prevalence to construct preliminary predictions via logistic regression
- use local deviations from regression model to estimate smooth residual spatial variation
- model-based approach acknowledges uncertainty in predictions

Loa loa: a generalised linear model

- **Latent spatially correlated process**

  $S(x) \sim \mathrm{SGP}\{0, \sigma^2, \rho(u))\}$
  $\rho(u) = \exp(-|u|/\phi)$

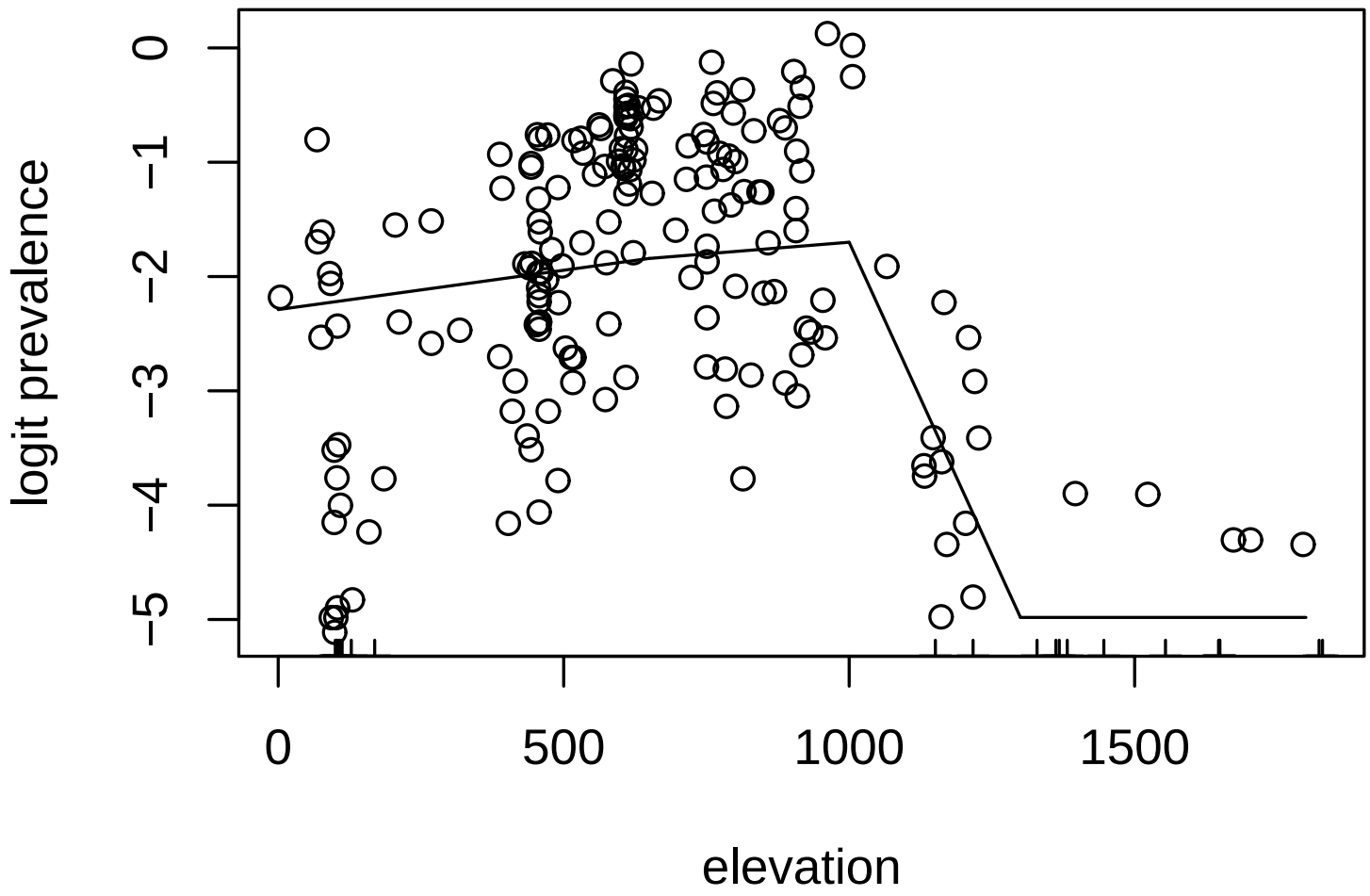
- **Linear predictor (regression model)**

  $d(x) =$ environmental variables at location x
  $\eta(x) = d(x)'\beta + S(x)$
  $p(x) = \log[\eta(x)/\{1 - \eta(x)\}]$

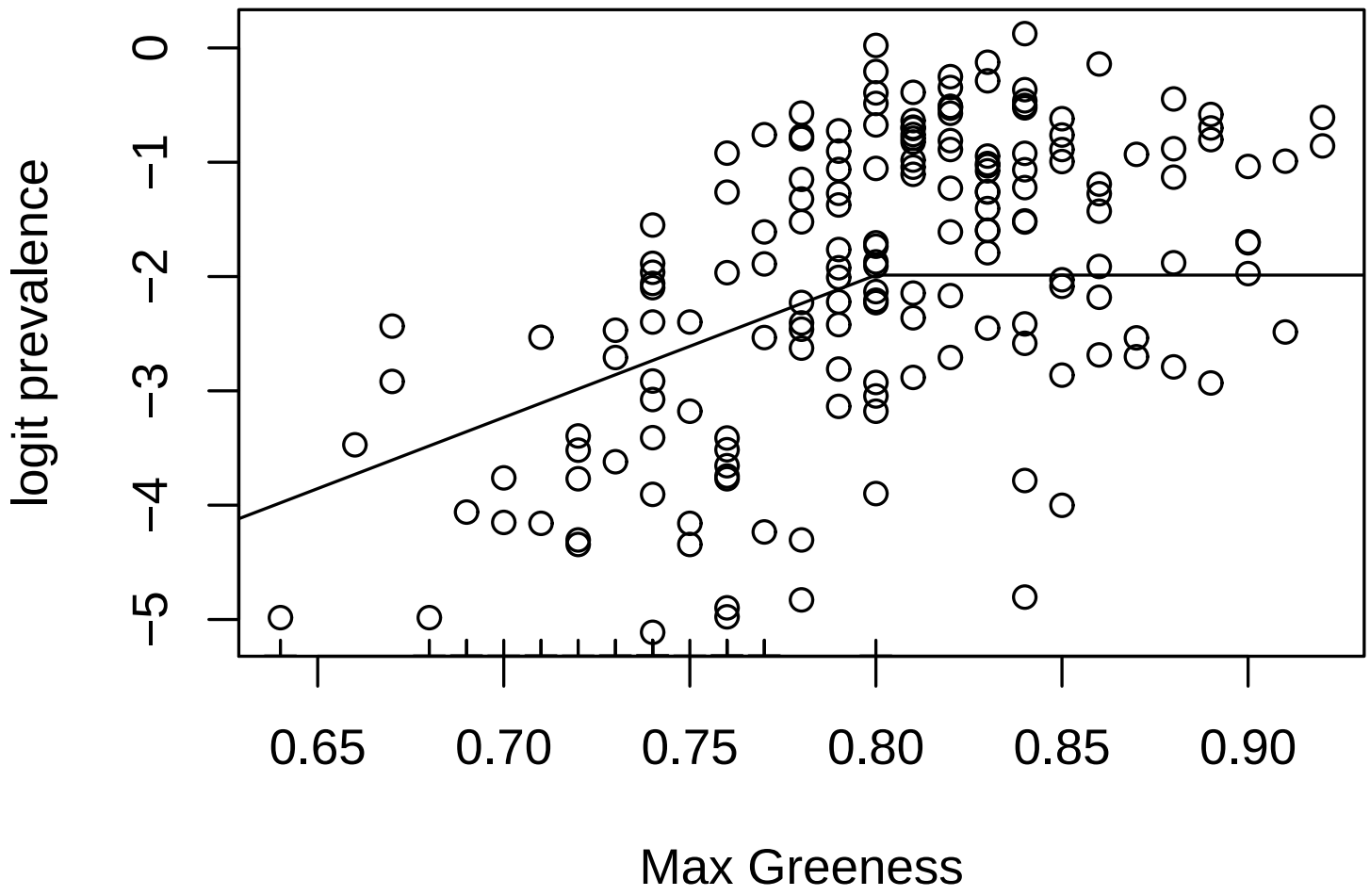
- **Conditional distribution for positive proportion Y_i/n_i**

  $Y_i|S(\cdot) \sim \mathrm{Bin}\{n_i, p(x_i)\}$ (binomial sampling)

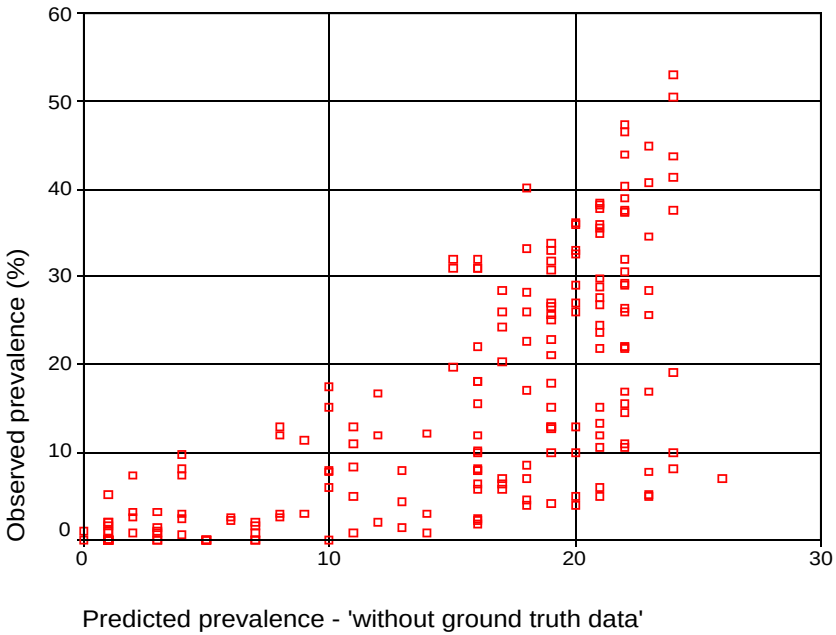
logit prevalence vs elevation



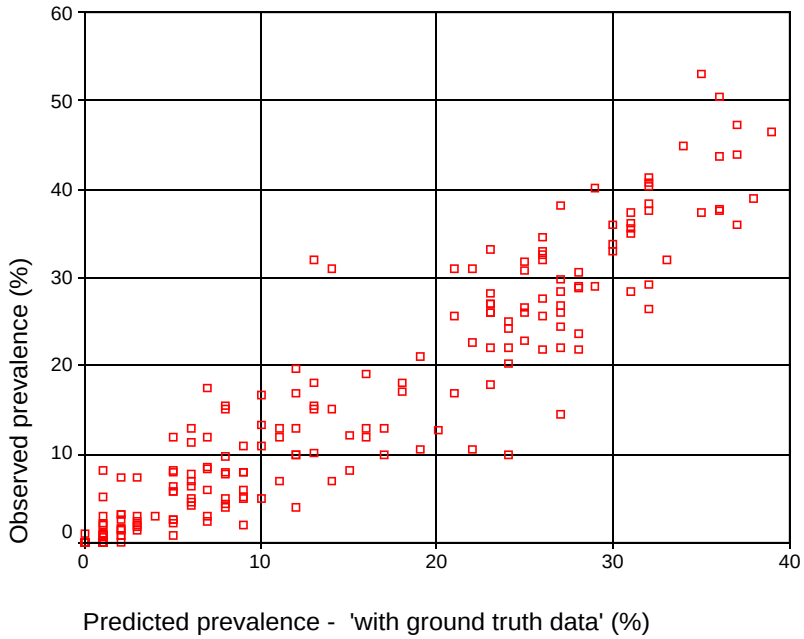
logit prevalence vs max NDVI



How useful is the geostatistical modelling?



Logistic regression



Model-based geostatistics

Probabilistic exceedance map for Cameroon (Diggle et al, 2007)

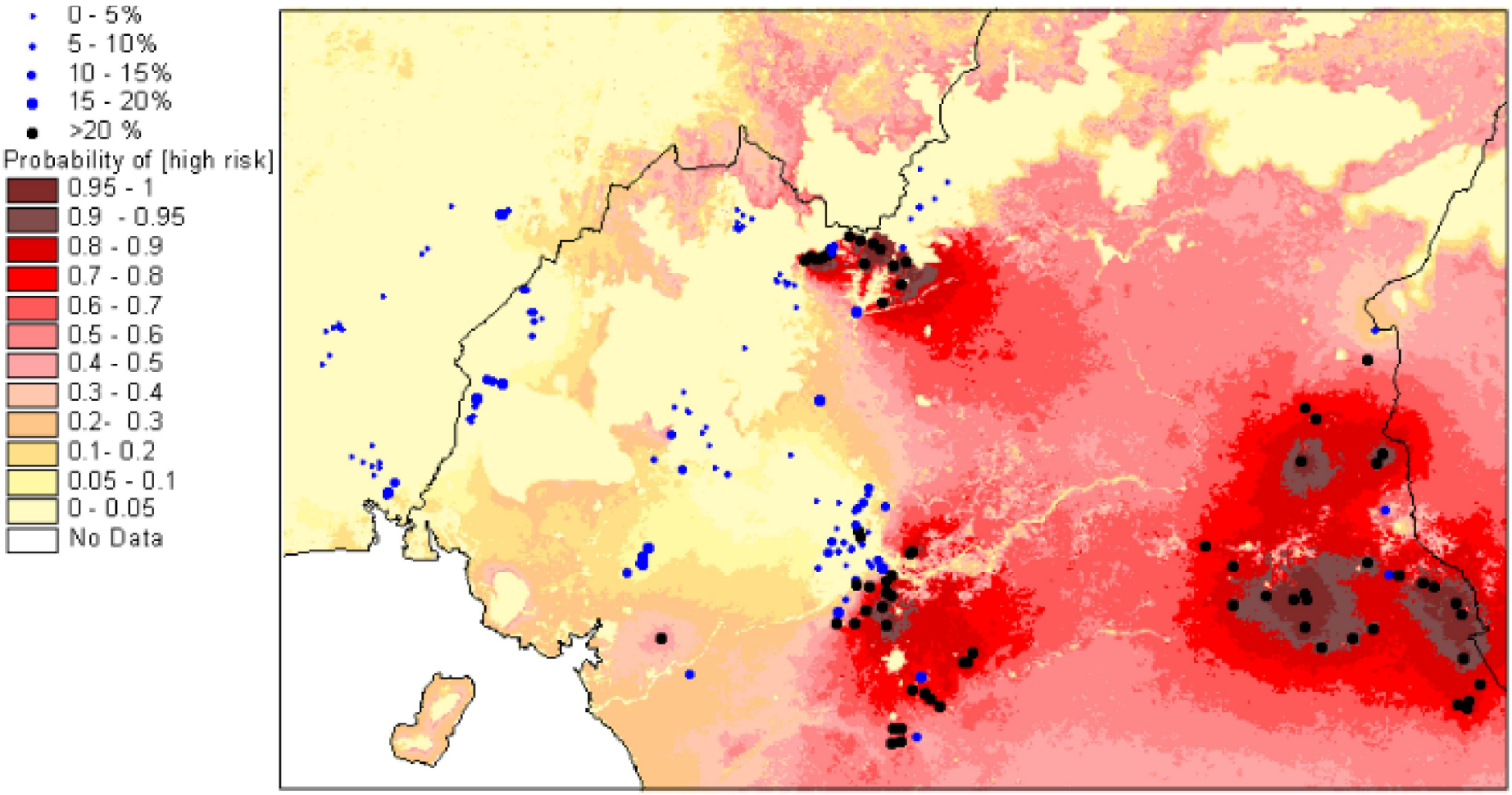


Figure 6: PCM for [high risk] in Cameroon based on ERMr with ground truth data.

Extending the model

- **Non-spatial extra-binomial variation**
- **Low-rank approximations**
- **Zero-inflation**
- **Spatio-temporal variation**
- **Multivariate spatial variation**

Spatially structured zero-inflation

- public health experts have strong sense that some areas are fundamentally unsuitable for onchocerciasis transmission
- hence need to incorporate mix of structural and chance zeros

Non-spatial model

$$Y_i \sim \begin{cases} 0 & : \quad \text{wp } q_i \\ \text{Bin}(n_i, p_i) & : \quad \text{wp } 1 - q_i \end{cases}$$

Spatial model

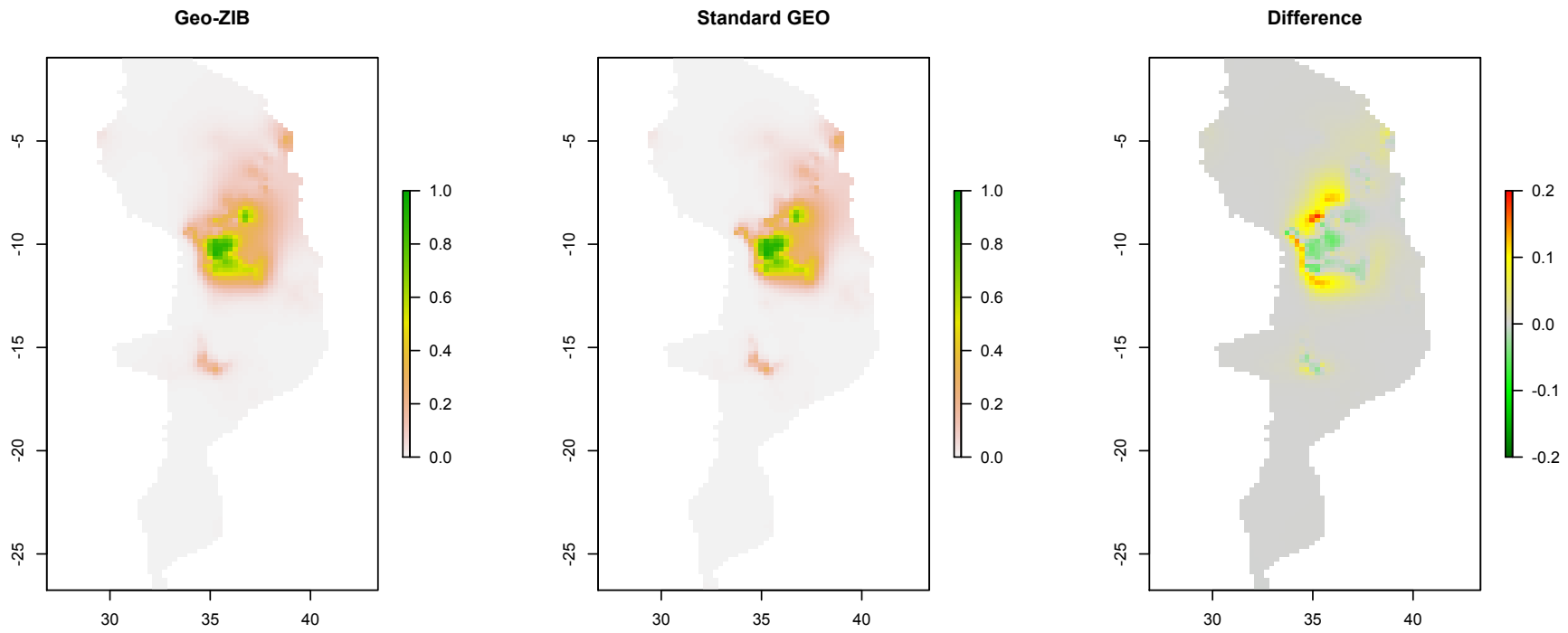
$\{q_i, p_i\} \rightarrow \{Q(x), P(x)\} : x \in \mathbb{R}^2 \sim$ bivariate stochastic process

Double logistic Gaussian process

$$P(Y = y|S(x)) = \begin{cases} Q(x) + (1 - Q(x)) \times \text{Bin}(0; n, P(x)) & : y = 0 \\ (1 - Q(x)) \times \text{Bin}(y; n, P(x)) & : y > 0 \end{cases}$$

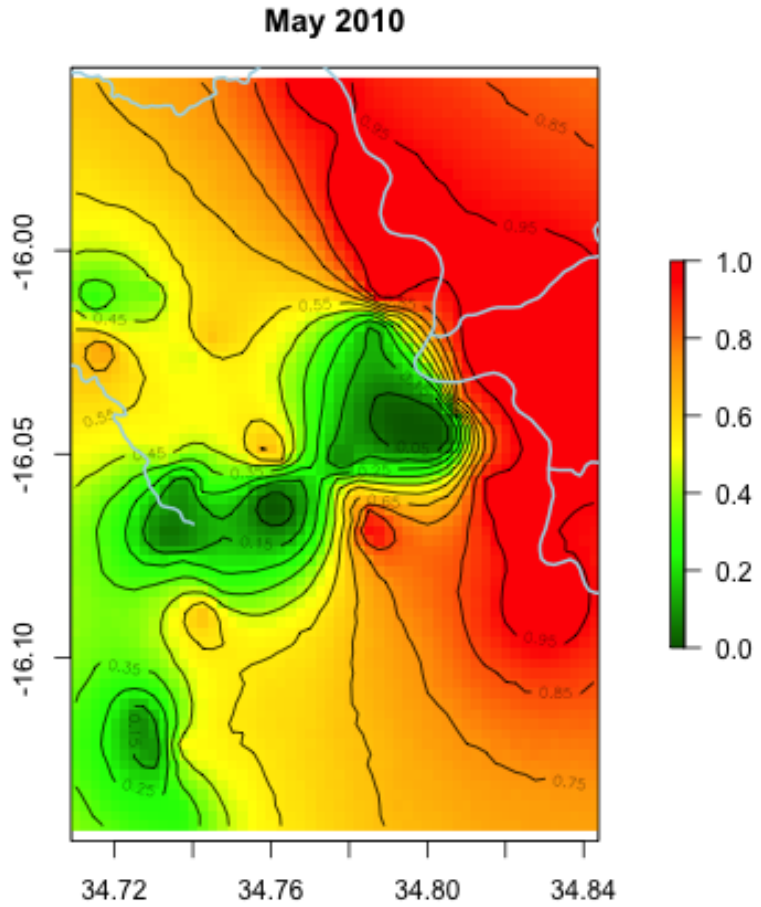
- $S(x) = \{S_1(x), S_2(x)\} \sim$ bivariate Gaussian process
- $\text{logit}(Q(x)) = \mu_1 + S_1(x)$
- $\text{logit}(P(x)) = \mu_2 + S_2(x)$

Mozambique/Malawi/Tanzania: probability exceedance map



Spatio-temporal mapping: rolling malaria indicator surveys

Hotspots: P(prevalence $>$ 20%)



Loa loa revisited: identifying "safe" communities

- **People who are highly infected with *Loa loa* parasites are at risk of serious adverse reactions to Mectizan**
- **Measuring individual parasite load in the field is difficult**
- **Can we predict proportion of highly infected individuals given only an estimate of prevalence?**

Identifying "safe" communities: formulating the question

- **Individual-level infection:** Y **(parasites per ml of blood)**
- **Community-level prevalence:** $\mathrm{P}(Y > 0)$
- **High-risk individual:** $Y > c$ $\quad$ $c = 8000, 20000, 30000?$

Target for prediction: proportion ($\Rightarrow$ number) of highly infected individuals in a community

Data from a single community:

n : number of individuals tested
Z : number testing positive

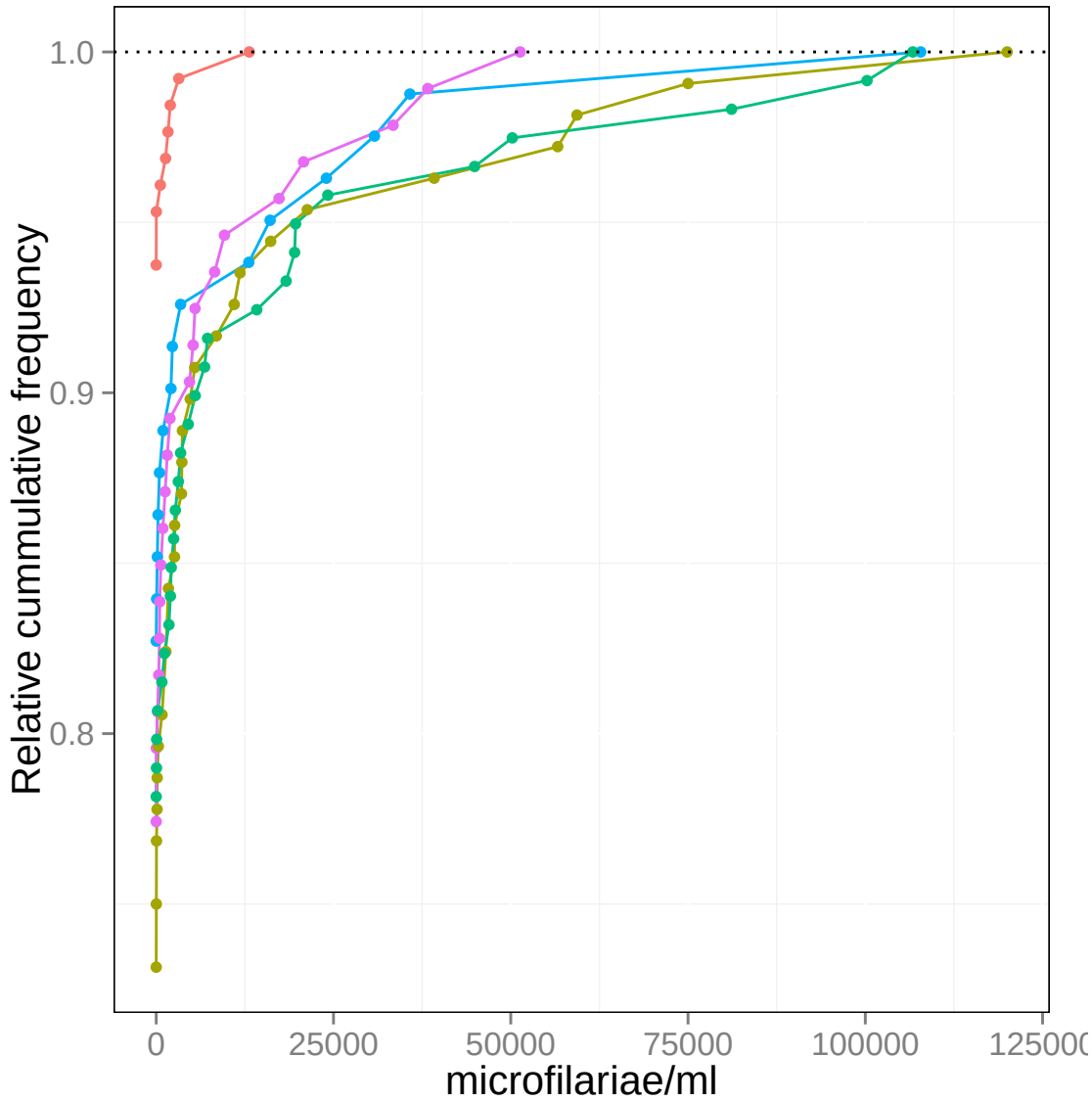
Required: $\mathrm{P}(Y > c|Z; n)$

Data for model-building

Provided by Task Force for Global Health in two stages (with thanks to original sources):

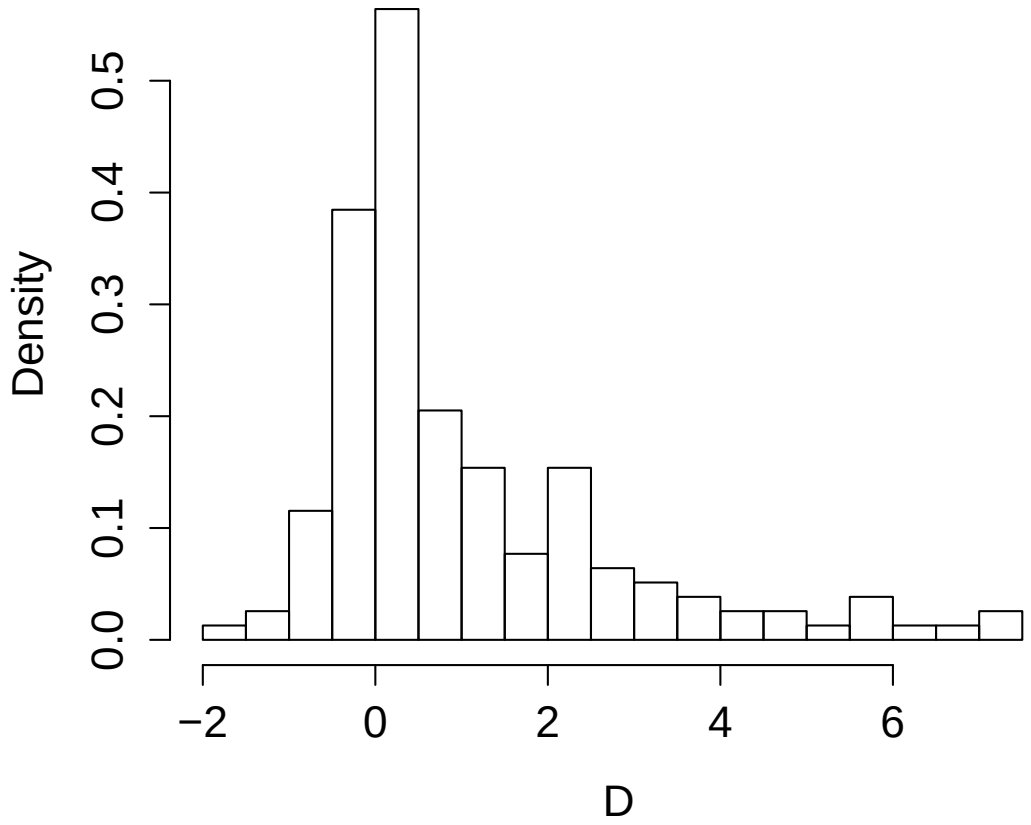
1. **development data:** 222 communities in Cameroon, Congo and DRC
2. **validation data:** 245 communities in Equatorial Guinea, Gabon and Cameroon

Cumulative distribution of infection levels (5 villages)

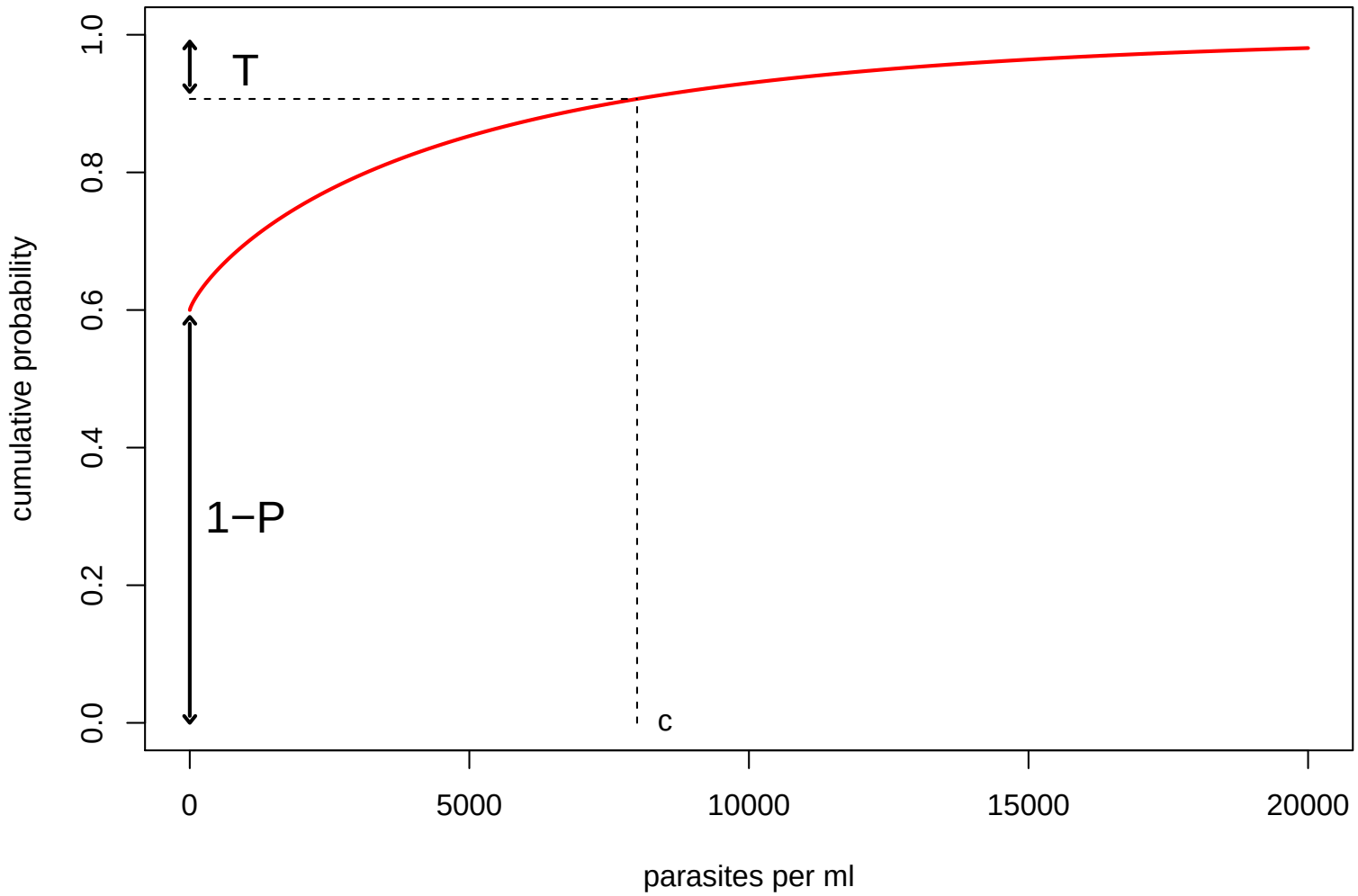


Gamma or Weibull?

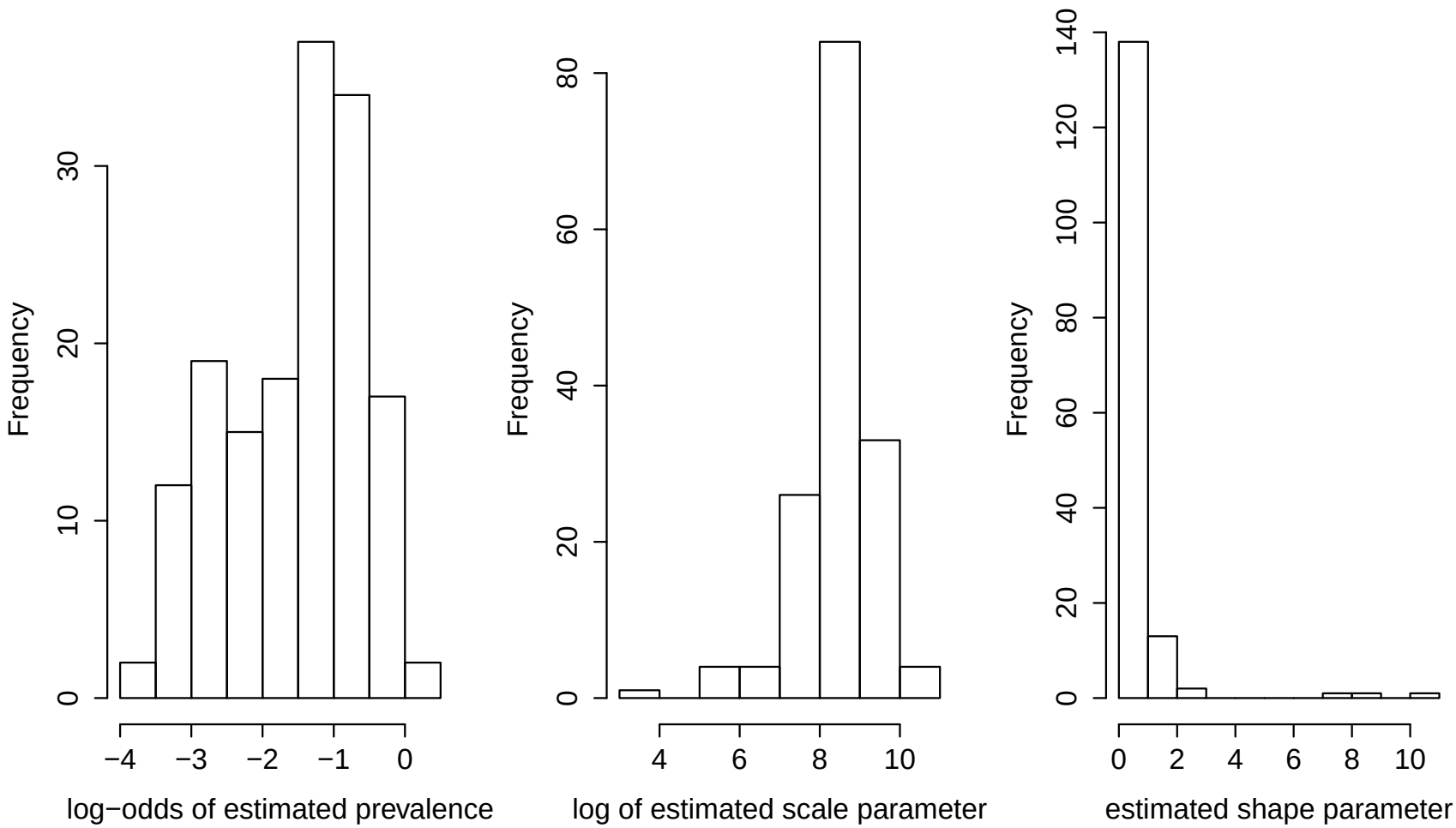
$$D = 2(\hat{L}_W - \hat{L}_G)$$



The statistical model
P=prevalence; T=proportion highly infected



Village-specific parameter estimates: 156 villages in development set



Model-fitting: maximum likelihood estimation

$$\theta = (\alpha, \beta, \kappa, \mathbf{\Sigma}) \quad \rightarrow \quad \rho = \rho(\alpha, U),\ \lambda = \lambda(\beta, V)$$

log-likelihood contribution from single village, *i*

- n_i = number sampled, z_i = number positive
- $y_{ij} : j = 1, ..., n_i; \quad (y_{ij} > 0 : j = 1, ...z_i \leq n_i)$
- $L_i(\theta|U, V) = (n_i - z_i)\log(1-\rho) + z_i \log \rho + \sum_{j=1}^{z_i} \log G'(y_{ij}; \rho, \lambda)$
- $L_i(\theta) = \int \int L_i(\theta|U, V)\mathrm{BVN}(0, \mathbf{\Sigma})dUdV$

log-likelihood from *m* villages

- $L(\theta) = \sum_{i=1}^{m} L_i(\theta)$
- integration by quasi Monte Carlo (Gaussian quadrature) or MCMC (Metropolis)

Model-fitting: results

- **Within a community:**

  probability that individual infection level is greater than x:

$$G(x) = P \exp\{-(x/L)^{\kappa}\}$$

$$\log\{P/(1-P)\} = \alpha + U \qquad \log L = \beta + V$$

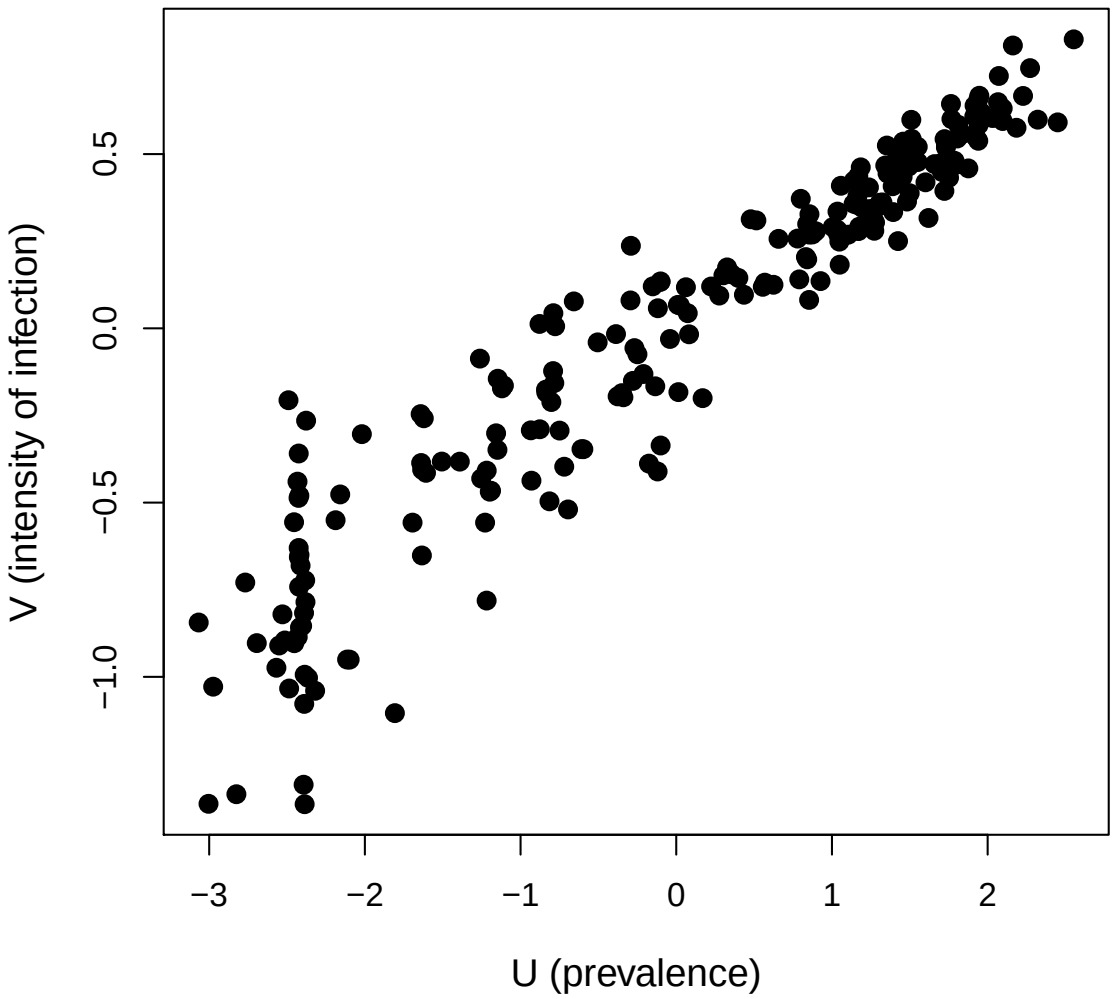
$$\hat{\alpha} = -2.47 \quad \hat{\beta} = 8.20 \quad \hat{\kappa} = 0.56$$

- **Between communities:**

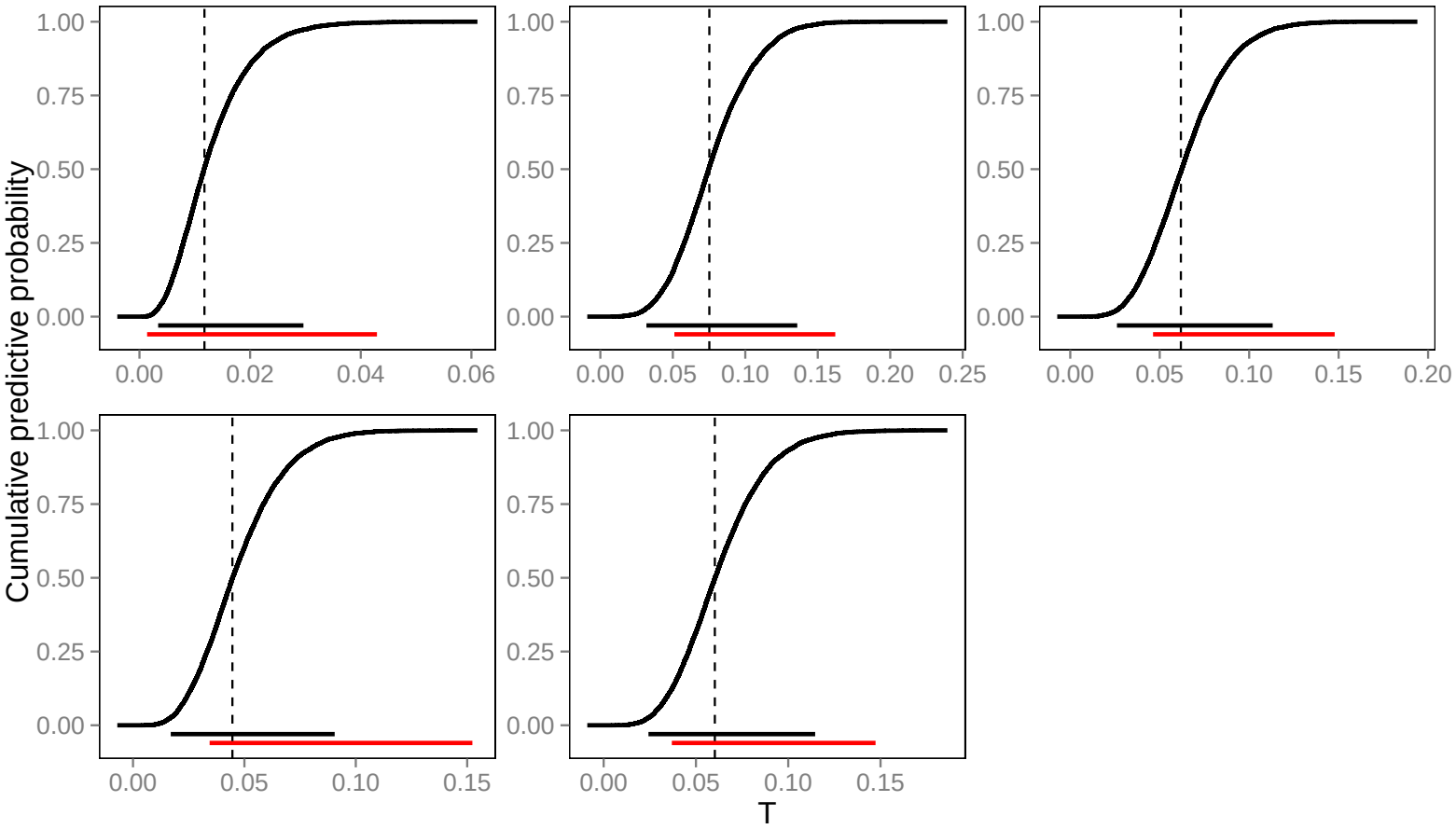
  $(U, V) \sim$ **zero-mean bivariate Normal**

$$\sigma_U^2 = 2.89 \quad \sigma_V^2 = 0.48 \quad \rho = 0.74$$

Predicted random effects (conditional expectations)



Results: predictive distributions and 95% intervals



———: parametric **———: empirical**

But we can now do better

The Loa loa problem re-formulated

People who are highly infected with *Loa loa* parasites are at risk of serious adverse reactions to Mectizan

- **Define a safe community as one for which the proportion of individuals carrying at least c parasites/per ml blood is at most q**

  Example: $c = 8000$ 20000? 30000? $q = 0.01$? 0.005?

- **New technology (cellscope) allows routine collection of data on (approximate) individual levels of infection (parasites/per ml blood)**

- **Given such data on a sample of individuals:**
  - **calculate the probability that the community is safe**
  - **set an upper limit for the probable number of highly infected people in the community**

Selected results from validation data: P(safe)

ID	n	npos	n above 20k	n above 30k	$c = 20k, q =$ 0.005	$c = 20k, q =$ 0.01	$c = 30k, q =$ 0.005	$c = 30k, q =$ 0.01	μ^+
Equatorial Guinea									
4844	44	13	0	0	0.017	0.1	0.12	0.39	3549.2
4864	44	10	0	0	0.37	0.66	0.69	0.89	760.0
Gabon									
6270	37	7	0	0	0.13	0.39	0.39	0.7	2857.1
9068	37	1	0	0	0.9	0.97	0.96	0.99	40.0
Cameroon									
4403	140	2	0	0	0.99	1.00	1.00	1.00	2050.0
NA	140	27	3	2	0.0012	0.047	0.05	0.37	6785.9

The finite population correction

Community size N, sample size n, of whom n^+ are highly infected

Predictive target thus far is:

$Q =$ **probability that a randomly sampled individual is highly infected**

To predict actual number, H, of highly infected individuals:

1. **Sample a value q from predictive distribution of Q**
2. **Sample a value m from binomial distribution,**

$$M \sim \mathrm{Bin}(N - n, q)$$

3. **Repeat 1 and 2 many times to give sample from predictive distribution of M, and hence of $H = n^+ + M$**

Selected results from validation data: 95% upper limit on number at risk

ID	n	$npos$	m_{30k}	number of highly infected individuals N=500	N=1000	N=5000
Equatorial Guinea						
4844	44	13	0	5 (1, 17)	12 (2, 34)	61 (16, 171)
4864	44	10	0	1 (0, 7)	3 (0, 14)	15 (1, 70)
Gabon						
6270	37	7	0	3 (0,12)	6 (0, 23)	32 (5, 116)
9068	37	1	0	0 (0, 3)	0 (0, 5)	2 (0, 21)
Cameroon						
4403	140	2	0	0 (0, 1)	0 (0, 2)	1 (0, 8)
NA	140	27	2	6 (3, 12)	12 (5, 25)	59 (25, 123)

Extensions

Current model

Independent $(U_i, V_i) : i = 1, \ldots, m \Rightarrow$ only village-specific information is helpful

Borrowing strength

Use information from neighbouring communities

- **data from communities $i = 1, \ldots, m$ at locations x_i**
- **spatially correlated random effects: $(U_i, V_i) \rightarrow (U(x_i), V(x_i))$**
- **bivariate Gaussian process model for $\{(U(x), V(x)) : x \in \mathbb{R}^2\}$**
- **which takes us back to model-based geostatistics!**

Closing remarks

- **principled statistical methods**
  - **make assumptions explicit**
  - **deliver optimal estimation within the declared model**
  - **make proper allowance for predictive uncertainty**
- **but there is no such thing as a free lunch**

"We buy information with assumptions"

C H Coombs

- **which is why statistics is at its most effective when conducted as a dialogue with substantive science**
- **and this should guide the way we teach statistics**

Diggle, P.J. (2015). Statistics: a data science for the 21st century. *Journal of the Royal Statistical Society* **A 178 793–813.**

Non-spatial extra-binomial variation

- **Latent spatially correlated process**

  $S(x) \sim \mathrm{SGP}\{0, \sigma^2, \rho(u))\} \quad \rho(u) = \exp(-|u|/\phi)$

- **Latent spatially independent random effects**

  $U_i \sim \mathrm{iidN}(0, \nu^2)$

- **Linear predictor (regression model)**

  $d(x) = \text{environmental variables at location } x$
  $\eta(x_i) = d(x_i)'\beta + S(x_i) + U_i$
  $p(x_i) = \log[\eta(x_i)/\{1 - \eta(x_i)\}]$

- **Conditional distribution for positive proportion Y_i/n_i**

  $Y_i|S(\cdot) \sim \mathrm{Bin}\{n_i, p(x_i)\}$ (binomial sampling)

Low-rank approximations (Rodrigues and Diggle, 2010)

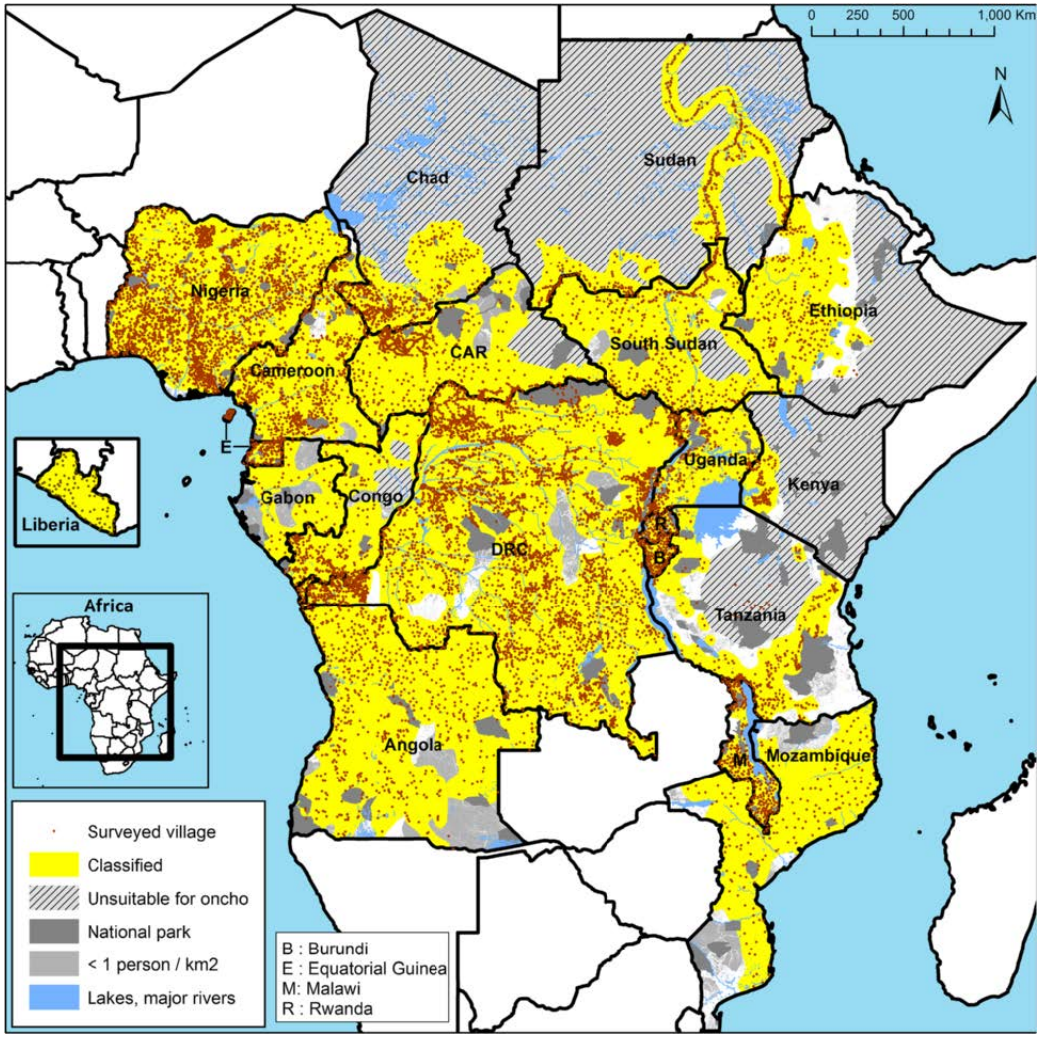
$$S(x) \approx \mu + \sum_{j=1}^{M} \phi^{-2} w\{(x - k_j)/\phi\} Z_j$$

- $w(u)$: kernel function
- $Z_j \sim$ iid $\mathrm{N}(0, \nu^2)$
- $k_j \in A \subset \mathbb{R}^2$: fixed set of points

Choose $w(\cdot)$ to approximate to preferred family of correlation functions

Computation linear in number of prediction points

Application: onchocerciasis mapping Africa-wide (Zoure et al, 2014): 14,473 survey locations



Application: onchocerciasis mapping Africa-wide (Zoure et al, 2014): low-rank model

- $M = 10,734$ points X_j in regular lattice at spacing 0.1 degrees
- $w(u)$ to approximate twice-differentiable Matérn correlation,

$$w(u) = \exp(-2\sqrt{2}\,u)$$

Parameter	estimate	95% confidence interval
μ	2:451	(2.469, 2.432)
ν^2	31:570	(31.038, 32.112)
ϕ	65:208	(64.993, 66.301)

Application: onchocerciasis mapping Africa-wide (Zoure et al, 2014): exceedance probabilities