Statistical Methods for real-time monitoring of health outcomes

Peter J Diggle

CHICAS, Faculty of Health and Medicine, Lancaster University

January 2018

Context of this talk

- **increasing availability of electronically recorded health outcome data**
- **at community and/or individual level**
- **accruing in "real-time"**
- **often spatially referenced**
- **to be used for prediction and/or explanation**
- **case-studies:**
  - **hospital-acquired MRSA**
  - **monitoring progression towards end-stage renal failure**
  - **statistical modelling to support lymphatic filariasis control**

References

Hospital-acquired MRSA

Deng, L., Diggle, P.J. and Cheesbrough, J. (2012). Estimating incidence rates using exact or interval-censored data, with an application to hospital-acquired infections. *Statistics in Medicine* **31**, 963–977.

Predicting renal failure

Diggle, P.J., Sousa, I. and Asar,Ö. (2015). Real-time monitoring of progression towards renal failure in primary care patients. *Biostatistics*, **16**, 522–536.

Asar,Ö, Bolin, D., Diggle, P.J. and Wallin, J. (2017). Linear mixed Eefects modelling for non-Gaussian Repeated measurement data (submitted)

Lymphatic filariasis control

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*, **10**, 12, e0005157. doi:10.1371/journal.pntd.0005157

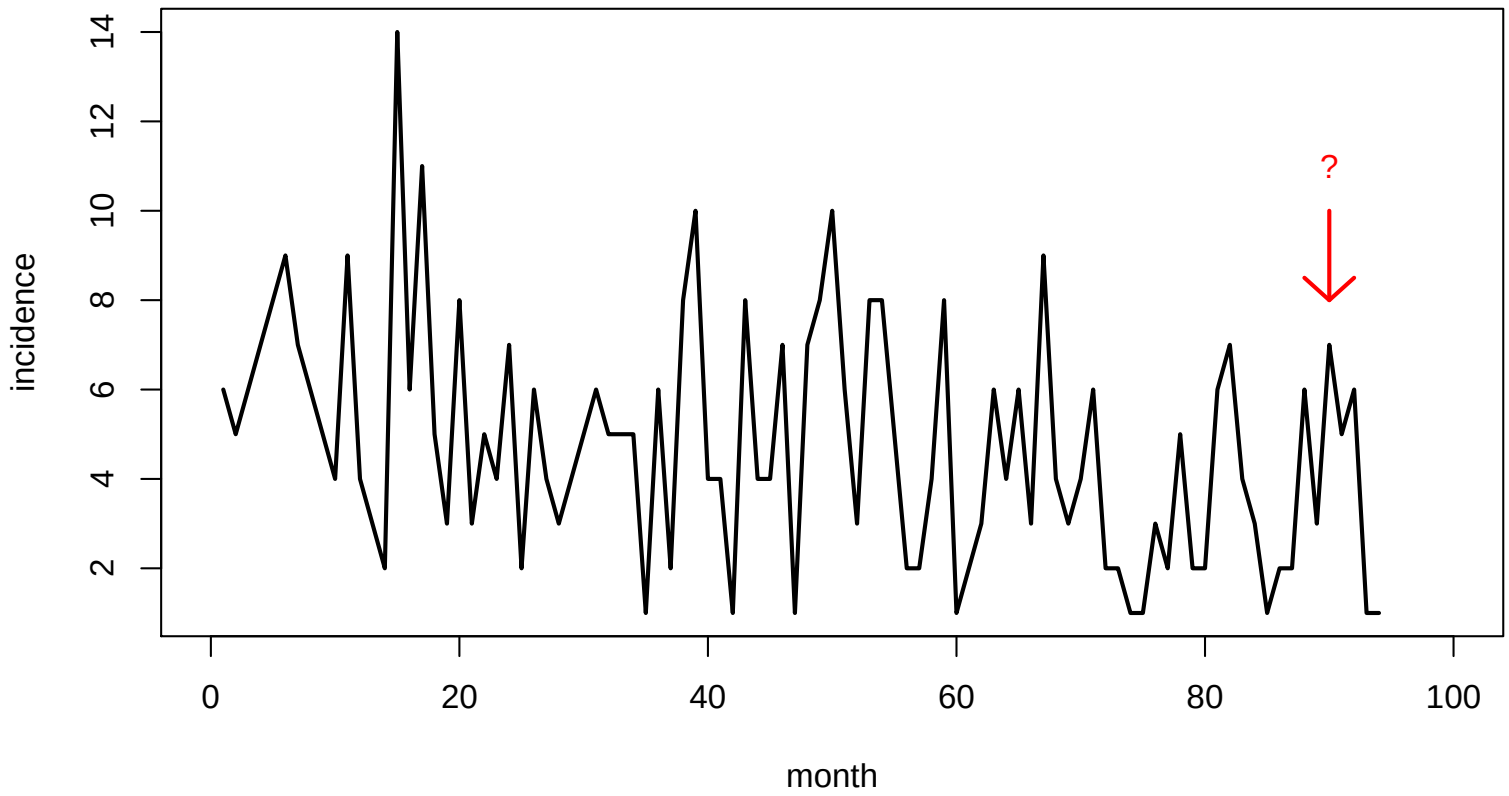
Giorgi, E., Schlüter, D.K. and Diggle, P.J. (2017). Bivariate geostatistical modelling of the relationship between Loa loa prevalence and intensity of infection. *Environmetrics*, **17**, DOI: 10.1002/env.2447

MRSA

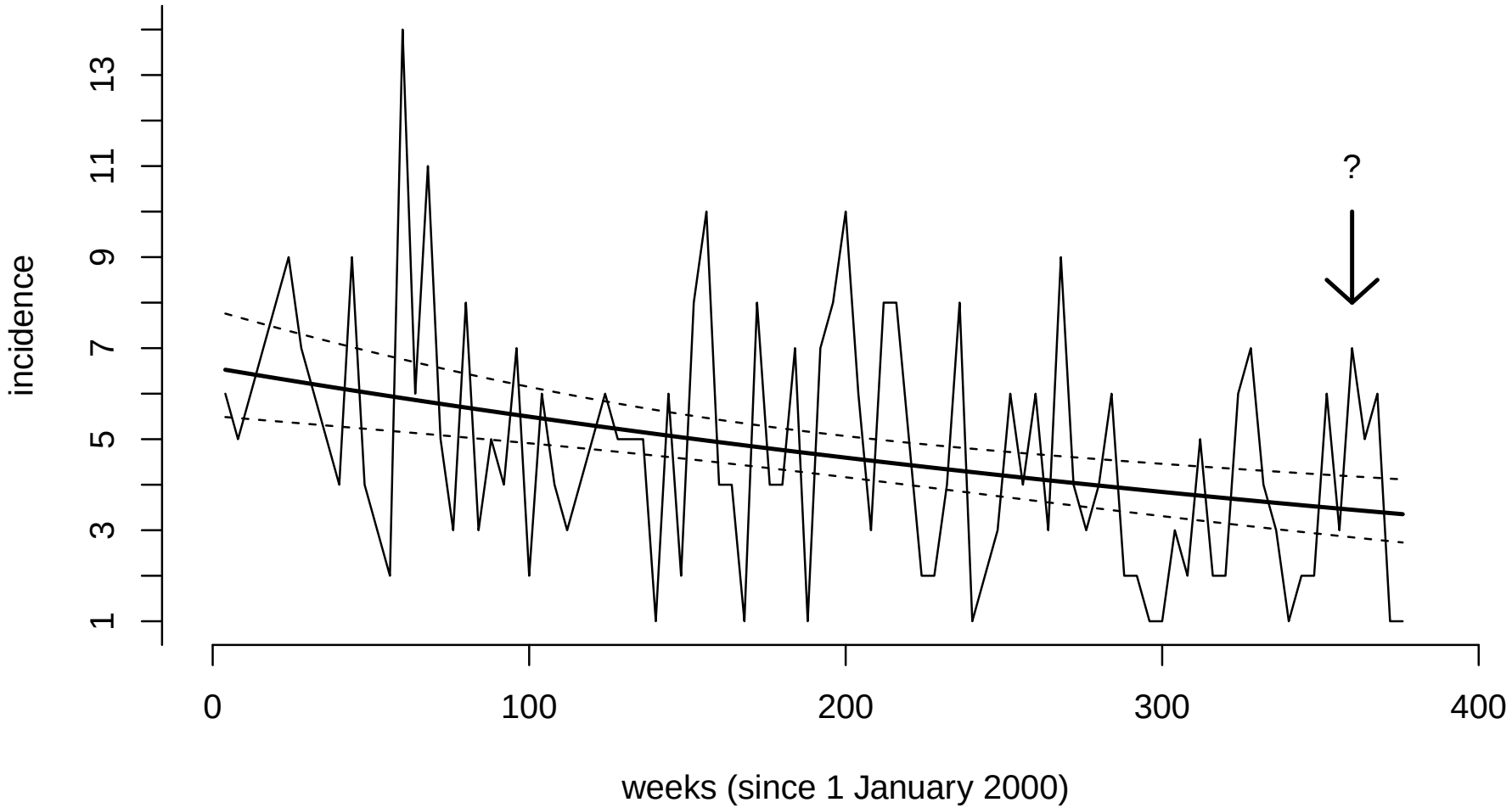
Hospital-acquired infections

1. **Health Services Journal, June 2009**
   In 2000 ... at least 100,000 cases ... annually.
   More than one in 10 NHS trusts in England have seen an increase in cases of MRSA.
   "threatening all those who use our healthcare system."
   (Edward Leigh, Conservative MP)
2. **Health Promotion Agency (2009)**
   MRSA rates ... between 1.6 and 1.8 cases per 100,000 occupied-bed-days ... by September 2008 had reduced by 59% compared to base-year (2002)

MRSA: in a Lancashire Hospital



Fitted Poisson process model



Kidney failure

Kidney failure: diagnosis, treatment and survival

Diagnosis

- Serum creatinine $\Rightarrow$ estimated glomerular filtration rate

$$\text{eGFR} = 186 \times \left(\frac{\text{SCr}}{88.4}\right)^{-1.154} \times \text{age}^{-0.203} (\times 0.742 \text{ if female})$$

- progression can be asymptomatic for many years
- SCr easy to measure from blood-sample (but noisy)

Treatment and survival

- aggressive control of blood-pressure
- renal replacement therapy: dialysis and transplantation
- early diagnosis and intervention can slow rate of progression

	Survival rate (%) to year			
	1	2	5	10
Dialysis	79.3	64.7	33.6	10.2
Transplant (living)	98.4	96.5	90.0	76.0

Royal Salford Hospital, NW England

Clinical guideline

Loss of $> 5\%$ eGFR per year $\Rightarrow$ refer to secondary care

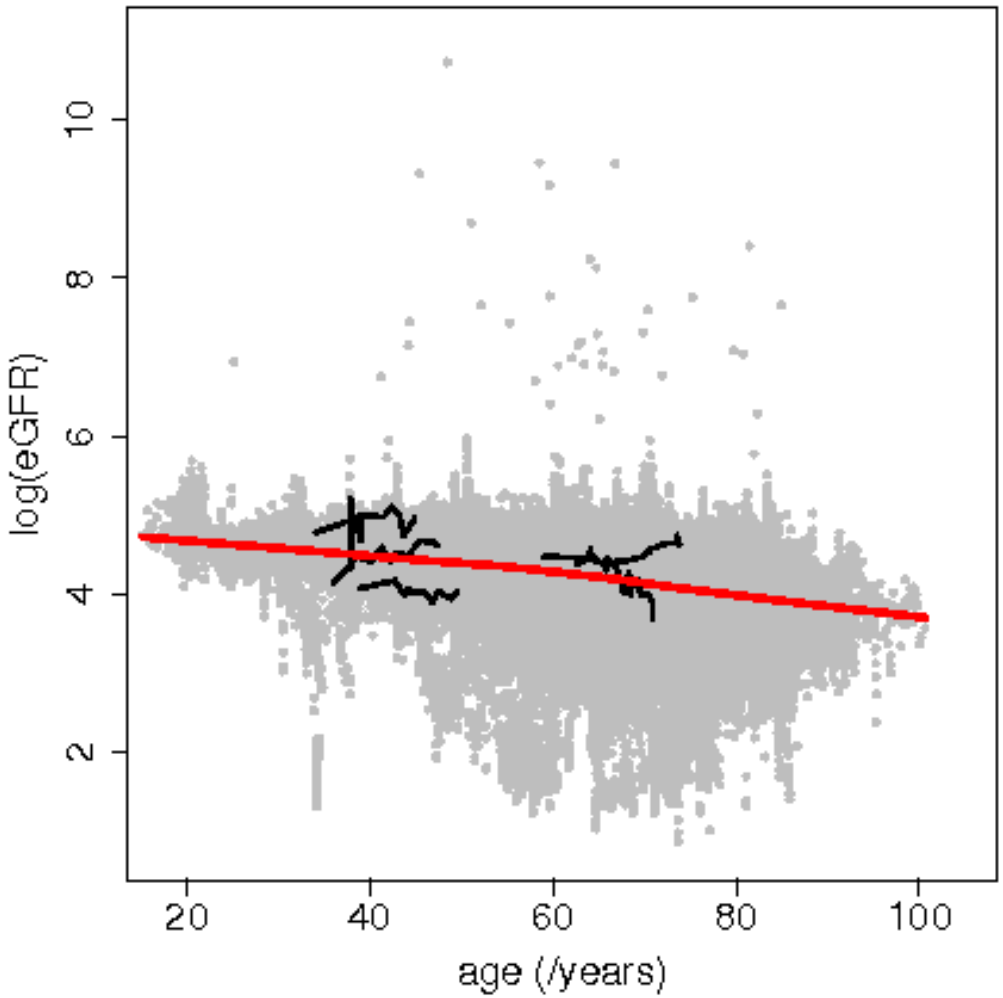
Data

- **measurements $Y_{ij} = \log \text{eGFR}$ at times t_{ij}, explanatory variables x_i (age, sex)**
  - **$i = 1, ..., m = 22,910$ "at-risk" primary care patients**
  - **$j = 1, ..., n_i \leq 305$ (median $n_i = 12$)**
  - **$0 \leq 10.02$ years follow-up (median 4.46)**
- $\mathcal{H}_i(t) = \{x_i, (t_{ij}, y_{ij}) : t_{ij} \leq t\}$

Statistical objective

$$\mathrm{P}\left(\frac{d}{dt}\log \text{GFR} < -0.05 | \mathcal{H}_i(t)\right) = \quad ?$$

Data: all cross-sectional and selected longitudinal



Dynamic regression model

- **subjects** $i = 1, \ldots, n$ **observed at times** $t_{ij}, j = 1, \ldots n_i$

$$Y_{ij} = \log(\text{eGFR})$$

- **expected value of** Y_{ij} **linear in initial age and time since recruitment**
- **rate of progression varies randomly:**
  - **between subjects:** random effect U_i
  - **within subject:** stochastic process $W_i(t_{ij})$

Dynamic Regression Model

$$
\begin{aligned}
Y_{ij} &= \alpha_0 + \alpha_1 \times I(\text{female}) \\
&+ \beta_1 \times \text{age}_{i1} + \beta_2 \times (\text{age}_{ij} - \text{age}_{i1}) + \beta_3 \times \max(0, \text{age}_{ij} - 56.5) \\
&+ U_i + W_i(t_{ij}) + Z_{ij}
\end{aligned}
$$

- Z_{ij}: measurement error, $\mathrm{N}(0, \tau^2)$
- U_i: between-subject random intercept, $\mathrm{N}(0, \omega^2)$
- $W_i(t)$: within-subject stochastic process

Model $W_i(t)$ as integrated Brownian motion

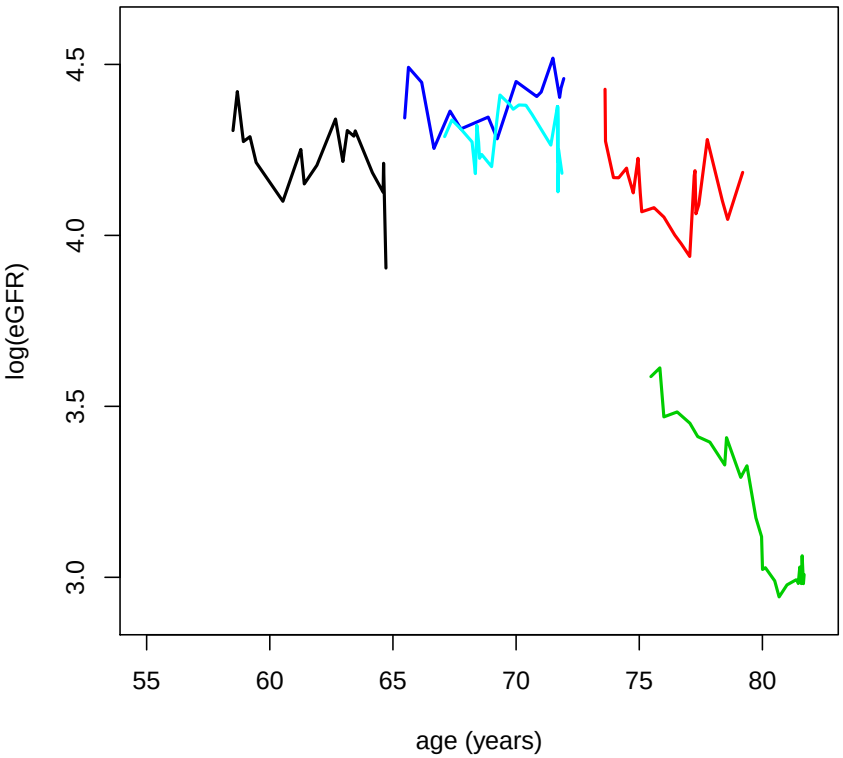
$$
\begin{aligned}
W_i(t) &= \int_0^t B_i(u)du \\
B_i(u)|B_i(s) &\sim \mathrm{N}\left(B_i(u), (u-s)\sigma^2\right)
\end{aligned}
$$

$B_i(u)$ is rate of progression for subject i at time t

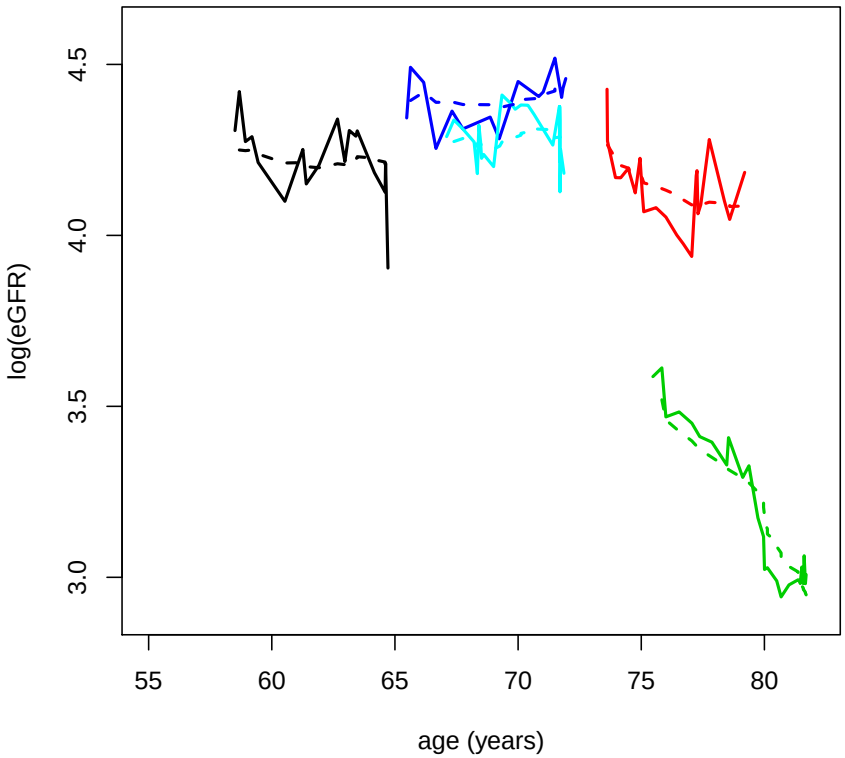
Maximum likelihood estimates of model parameters

Parameter		Estimate	SE
α_0	intercept	4.6006	0.0203
α_1	female	-0.0877	0.0048
β_1	age on entry	-0.0048	0.0004
β_2	follow-up	-0.0232	0.0011
β_3	age>56.5	-0.0075	0.0006
ω^2	intercept	0.1111	0.0012
σ^2	signal	0.0141	0.0002
τ^2	noise	0.0469	0.0001

Modelling progression



Modelling progression



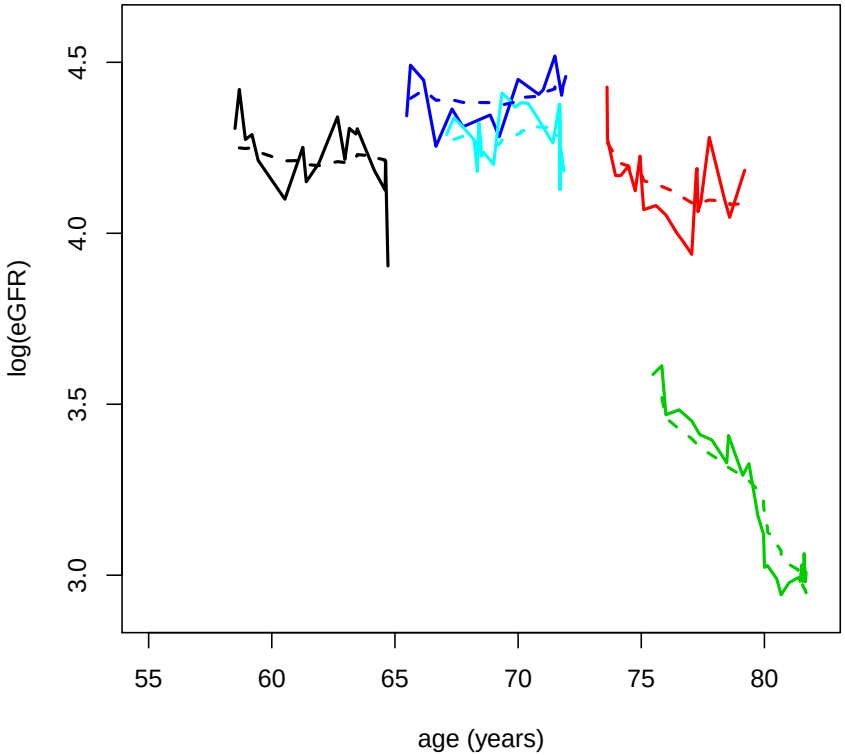
Prediction of subject-specific rate of change

Goal: calculate predictive distributions,

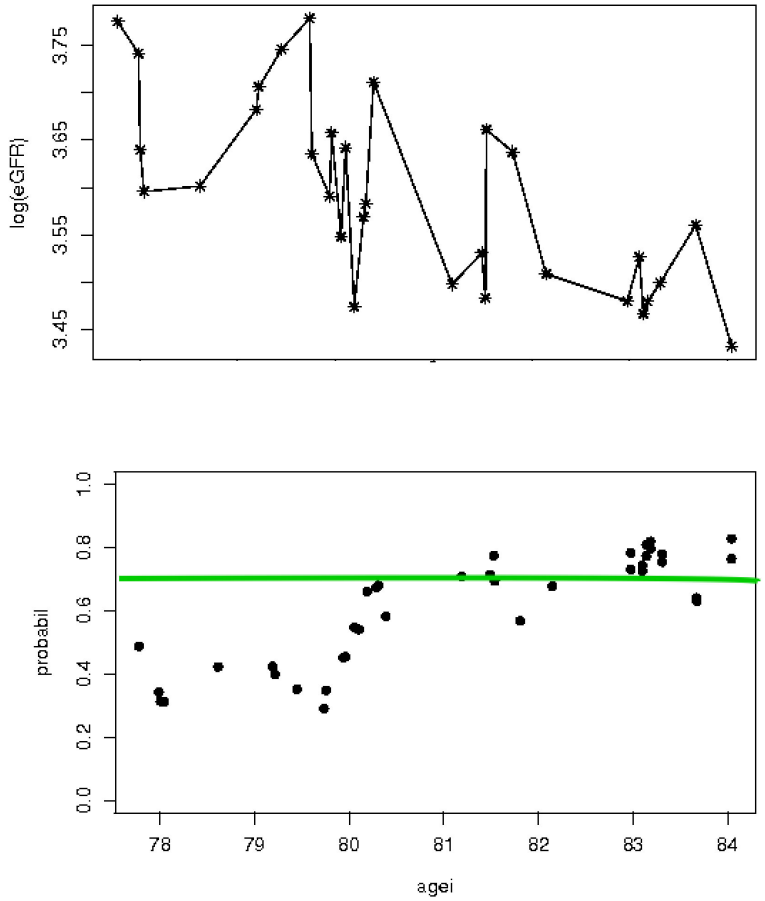
$$[B_i(t_{ij}) | Y_{i1}, ..., Y_{ij}; \hat{\theta}]$$

- $B_i(t_{ij})$ is current rate of change of eGFR

Modelling progression



Predicting rate of change in GFR



Is the Gaussian assumption critical?

For estimating mean response profiles	**Probably not**
For predicting individual response profiles	**Probably**
For spotting extreme behaviour	**Almost certainly**

Extending the Gaussian linear model

$$Y_{ij} = x'_{ij}\beta + d'_{ij}U_i + W_i(t_{ij}) + Z_{ij}$$

- any or all of the stochastic terms non-Gaussian
- continuous-time interpretation for $W_i(t)$

Distributional family

$$Y = \mu + \sqrt{T}Z,$$

- $\mu = \mathrm{E}[Y]$
- $\Sigma = \mathrm{var}(Y)$
- $T \sim$ generalized inverse Gaussian distribution (GIG)
- $Z \sim \mathrm{N}(0, V)$.

Non-Gaussian noise

$$Y_{ij} = x'_{ij}\beta + d'_{ij}U_i + W_i(t_{ij}) + Z_{ij}$$

$$Z_{ij} = \sqrt{T_{ij}}Z^*_{ij}$$

- $T_{ij} \sim$ **iid GIG**
- $Z^*_{ij} \sim \mathrm{iidN}(0, \tau^2)$.

Non-Gaussian random effects

$$Y_{ij} = x'_{ij}\beta + d'_{ij}U_i + W_i(t_{ij}) + Z_{ij}$$

$$U_i = \sqrt{T_i}U_i^*$$

- $T_i \sim$ **iid GIG**
- $U_i^* \sim \mathrm{iidN}(\mathbf{0}, V)$.

Non-Gaussian stochastic process

$$Y_{ij} = x'_{ij}\beta + d'_{ij}U_i + W_i(t_{ij}) + Z_{ij}$$

$$\mathcal{D}W_i(t) = dL_i(t),$$

- $\mathcal{D}$ = differential operator
- $dL_i \sim$ continuous-time white noise ($\Rightarrow W_i(t)$ at least continuous)

Integrated random walk: $\mathcal{D} = \frac{\partial^2}{\partial t^2}$

Matérn: $\mathcal{D} = (\frac{\partial^2}{\partial t^2} - \kappa)^{\alpha/2}$

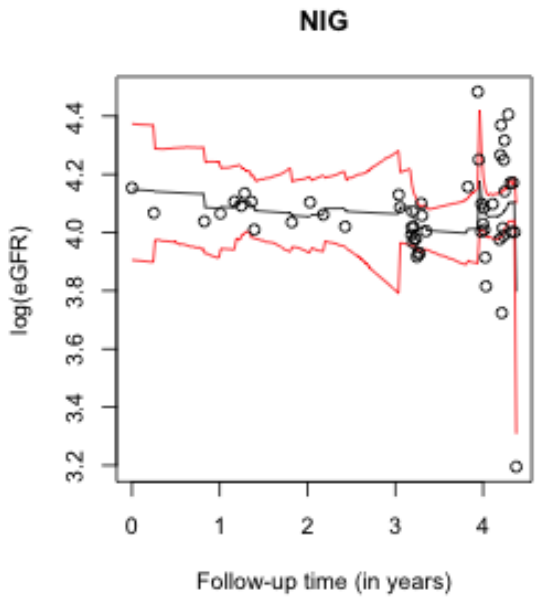
Low-rank approximation for fast computation

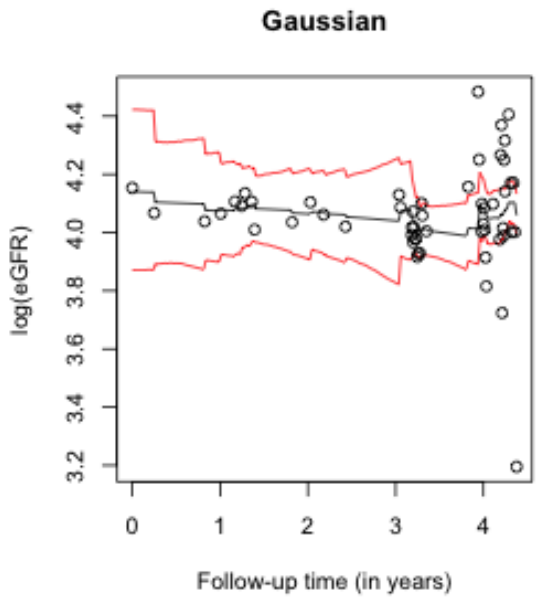
Kidney failure data: Gaussian model OK?

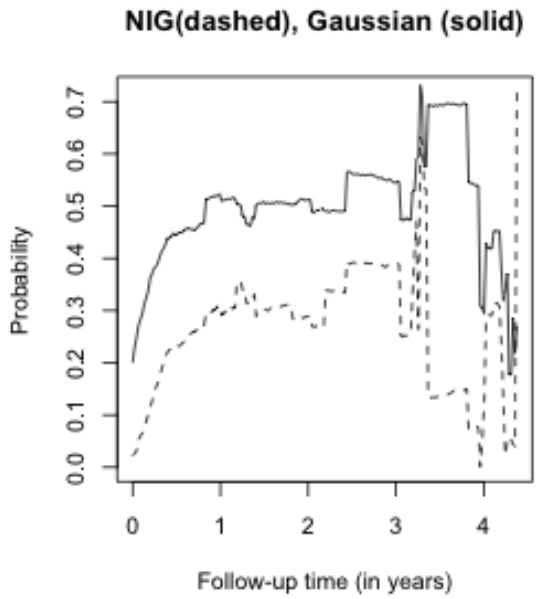
Comparing predictive inferences for current value

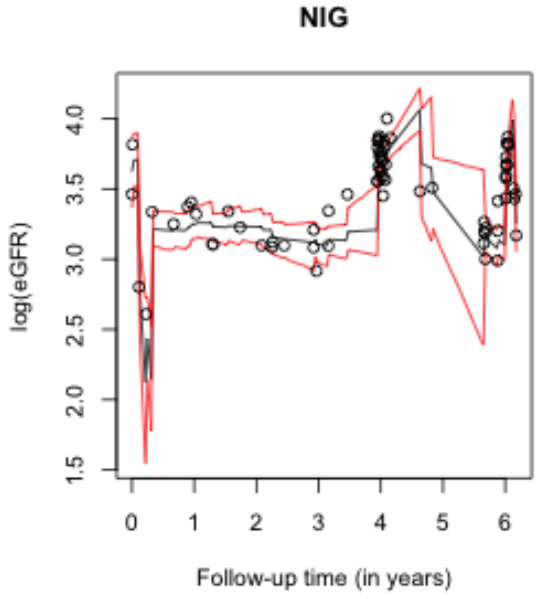
Distribution			Results		
Process	Random Effects	Noise	MAE	Coverage	Width
None	Normal	Normal	0.175	93.80	1.020
None	GIG	Normal	0.178	93.86	1.014
None	GIG	GIG	0.182	92.79	0.971
Normal	Normal	Normal	0.168	93.82	0.990
Normal	Normal	GIG	0.126	94.95	0.910
GIG	Normal	Normal	0.119	96.13	0.794
GIG	GIG	Normal	0.169	92.37	0.847
GIG	GIG	GIG	0.115	95.63	0.801

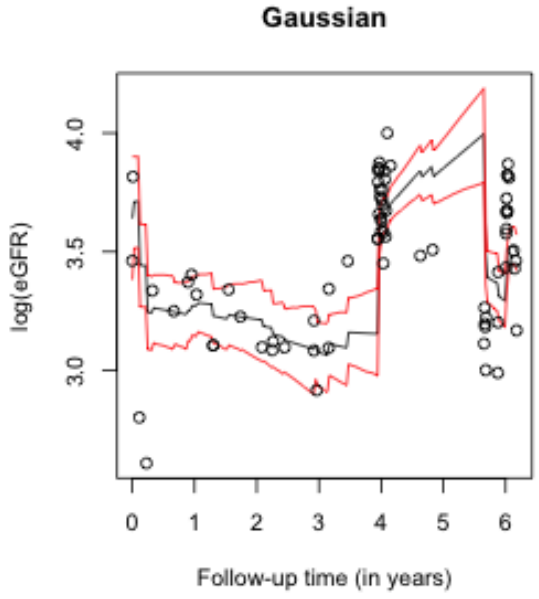
Comparing predictions of rate of change (two patients)

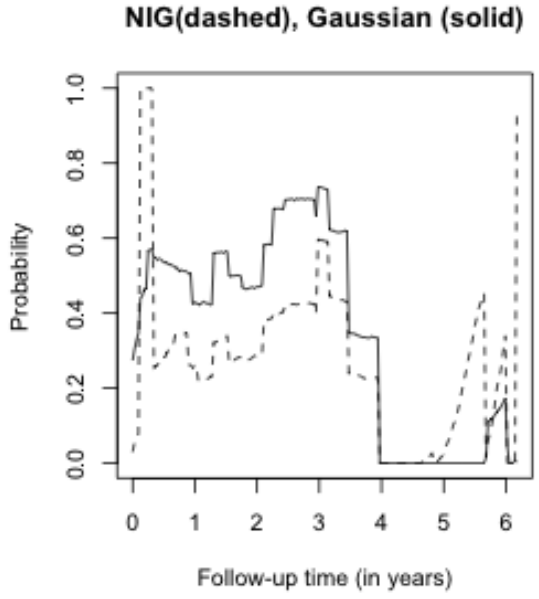












Work-in-prospect

Field-testing: comparative evaluation against current methods

- eye-balling
- OLS fit to three most recent values

Informative follow-up: eGFR more likely to be measured when subject is in poor health

$\Rightarrow$ joint modelling of eGFR measurements and follow-up times

Feedback: prediction algorithm needs to know about inerventions

Implementation: in clinical practice...needs informatics expertise

Lymphatic filariasis

Lymphatic filariasis control

Lymphatic filariasis: a vector-borne disease

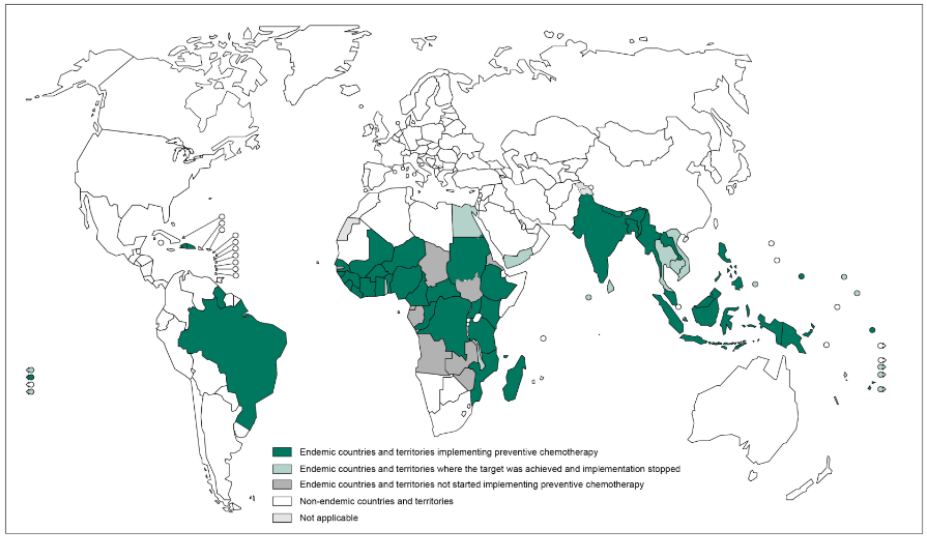
- **impairs lymphatic system, can lead to abnormal enlargement of body parts, causes pain, severe disability, social stigma**
- **856 million people in 52 countries require preventive chemotherapy**

http://www.who.int/mediacentre/factsheets/fs102/en/

Global Programme for Elimination of Filariasis

- **launched in 2000**
- **target is to achieve elimination by 2020.**
- **treatment with preventive chemotherapy:**
  - **mass drug administration (MDA) annually**
  - **albendazole (400 mg) plus ivermectin (150-200 mcg/kg)**

Distribution and status of preventive chemotherapy for lymphatic filariasis, worldwide, 2014



Data Source: World Health Organization
Map Production: Control of Neglected Tropical Diseases (NTD)
World Health Organization

World Health Organization

http://www.who.int/mediacentre/factsheets/fs102/en/

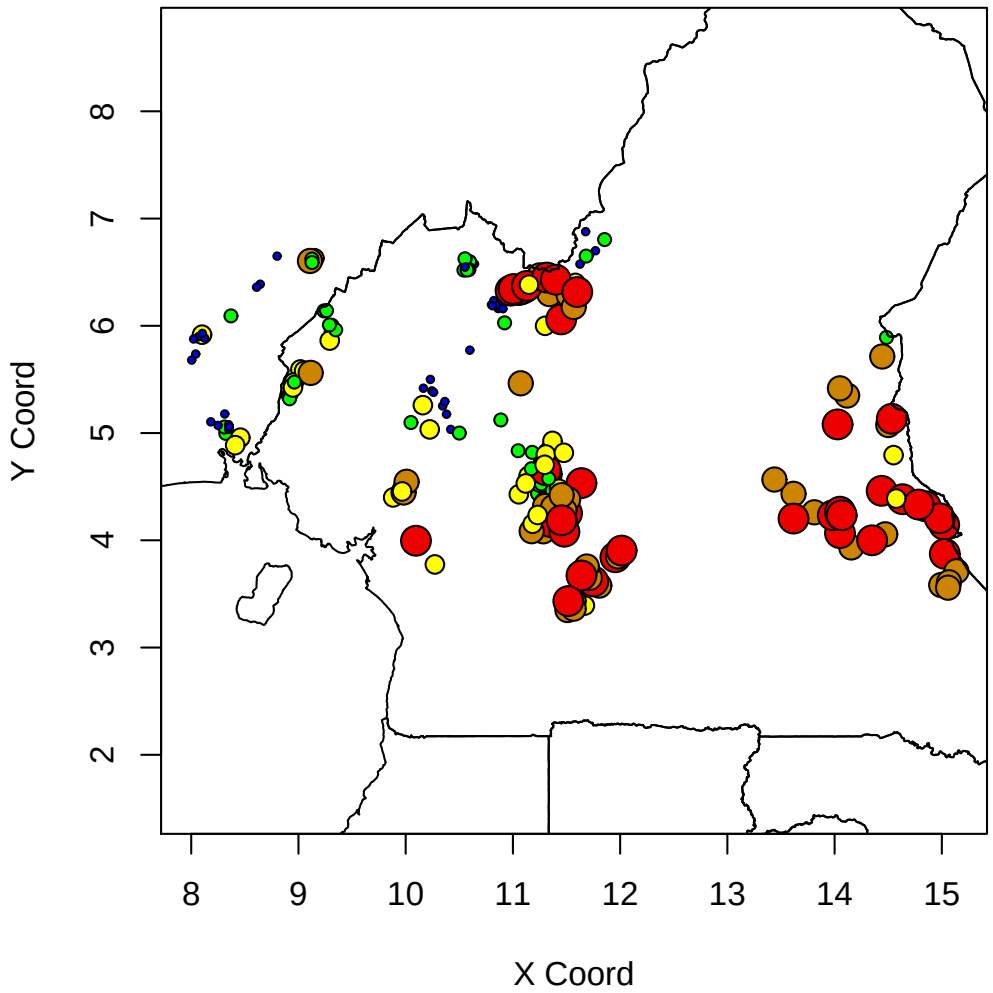
The Loa loa problem.

People who are heavily co-infected with *Loa loa* parasites can experience serious (occasionally fatal) adverse reactions to Mectizan

Loa loa young

...and old

A geostatistical data-set: Loa loa prevalence surveys



Canonical geostatistical problem: **predict prevalence throughout mapped region**

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

- **the application of general principles of statistical modelling and inference to geostatistical problems**
- **paradigm:**
  - **specify the scientific question**
  - **design the study and collect/collate data**
  - **formulate the statistical model**
  - **fit the model using likelihood-based methods**
  - **answer the scientific question**

Addressing the Loa loa problem (1)

The Loa loa problem

People who are heavily co-infected with *Loa loa* parasites can experience serious (occasionally fatal) adverse reactions to ivermectin

Current strategy

- **heavily co-infected people are more likely to be found in high prevalence areas**
- **areas with prevalence greater than 20% declared high-risk**
- **map *Loa loa* prevalence using model-based geostatistics**
- **identify areas with high predictive probability of exceeding 20% prevalence threshold**

Mapping Loa loa prevalence

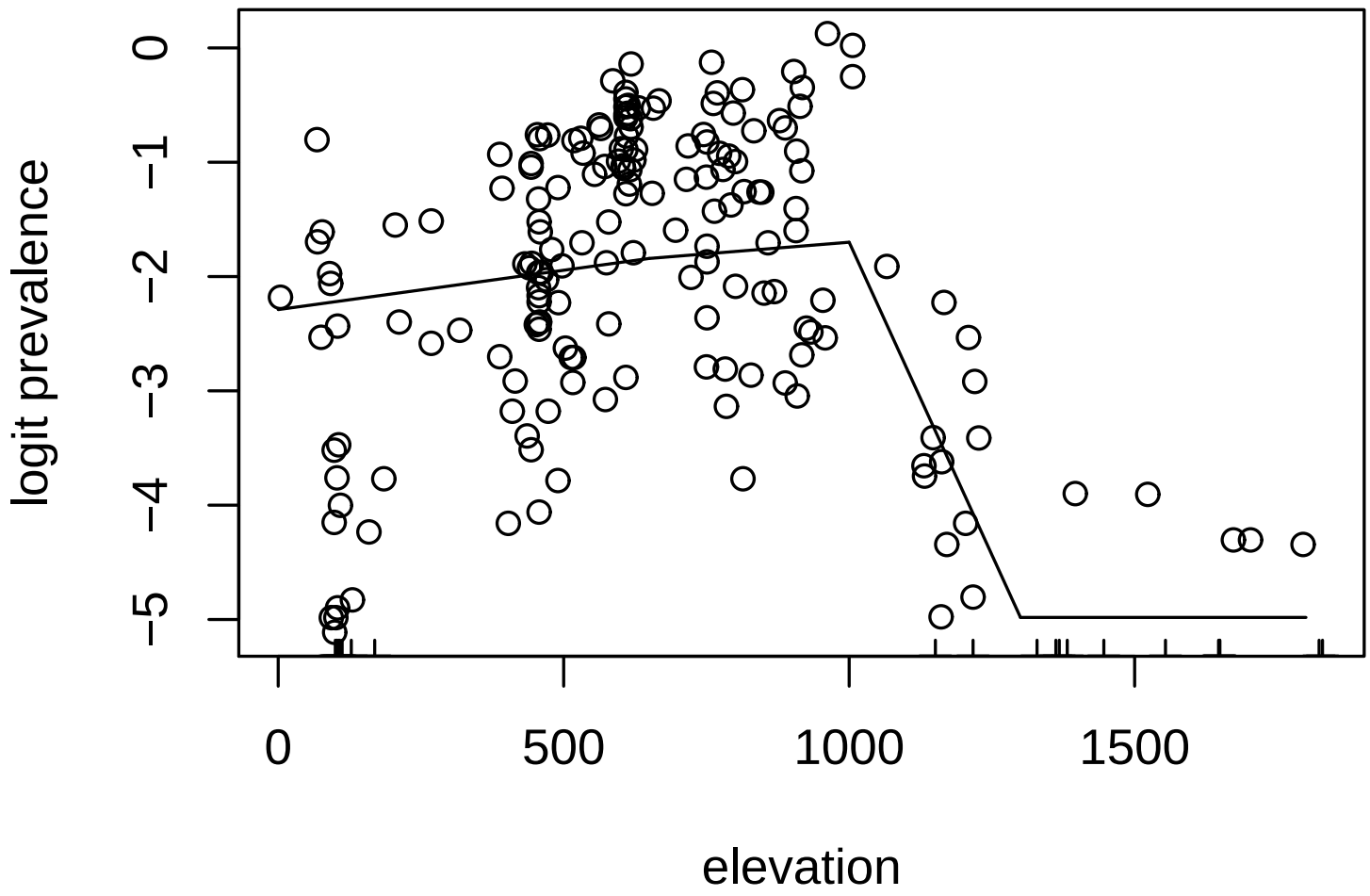
Parasitological survey data

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

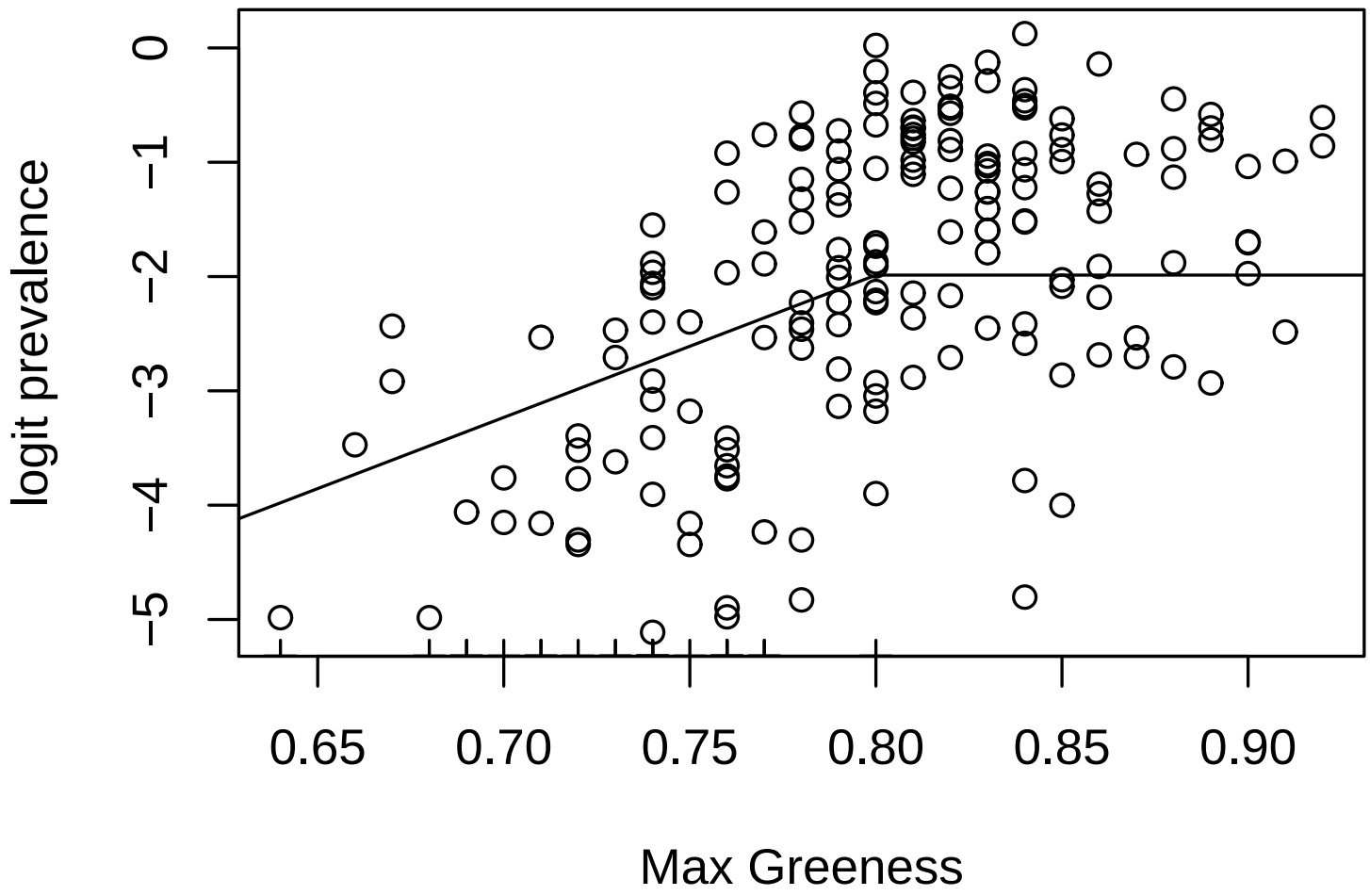
Environmental data (satellite images)

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

logit prevalence vs elevation



logit prevalence vs max NDVI



Loa loa mapping: geostatistical model

- **Latent spatially correlated process**

  $S(x) \sim \mathrm{SGP}\{\mu, \sigma^2, \rho(u)\}$

  Matérn correlation, $\rho(u; \phi, \kappa)$, fix $\kappa = 0.5$

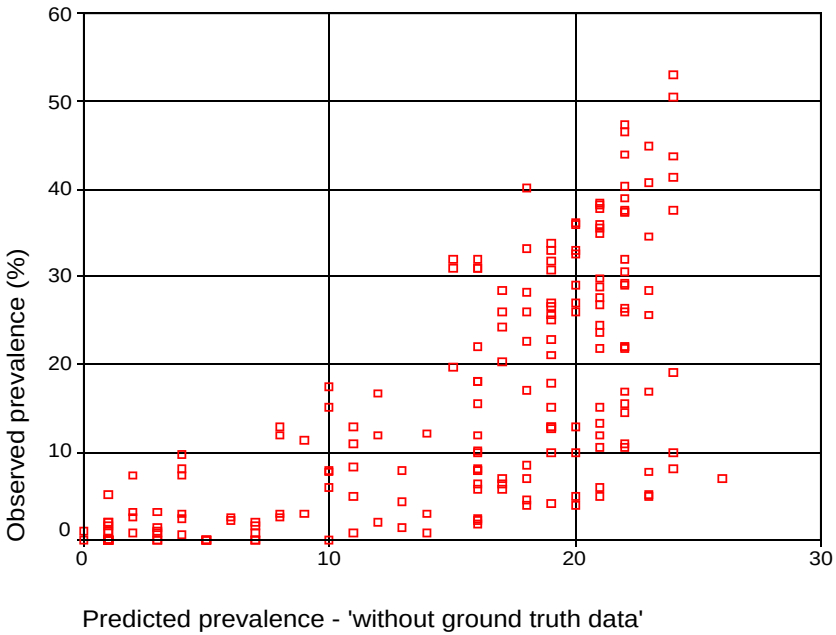
- **Prevalence**

  $p(x) = \exp\{d(x)'\beta + S(x)\}/[1 + \exp\{d(x)'\beta + S(x)\}]$

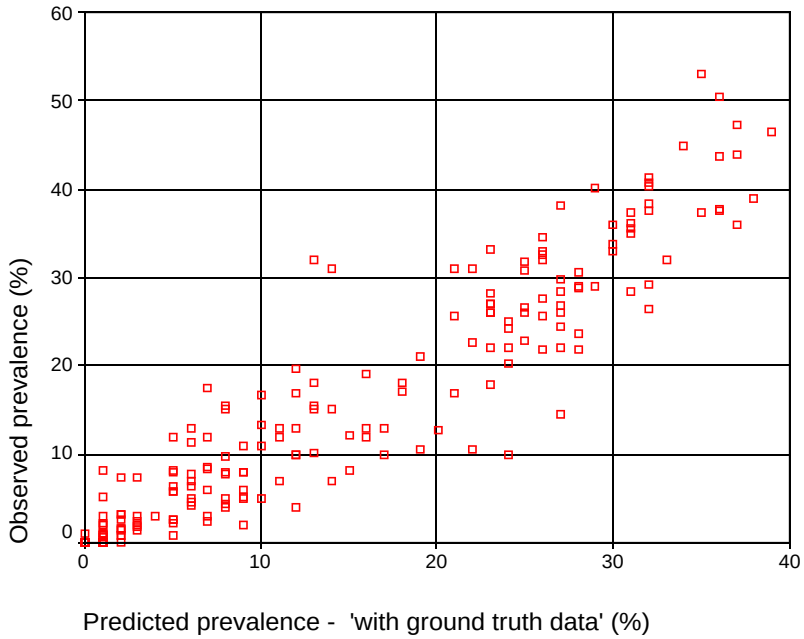
- **Conditional distribution of empirical prevalence Y_i/n_i**

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

Observed vs fitted prevalence



Logistic regression



Model-based geostatistics

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

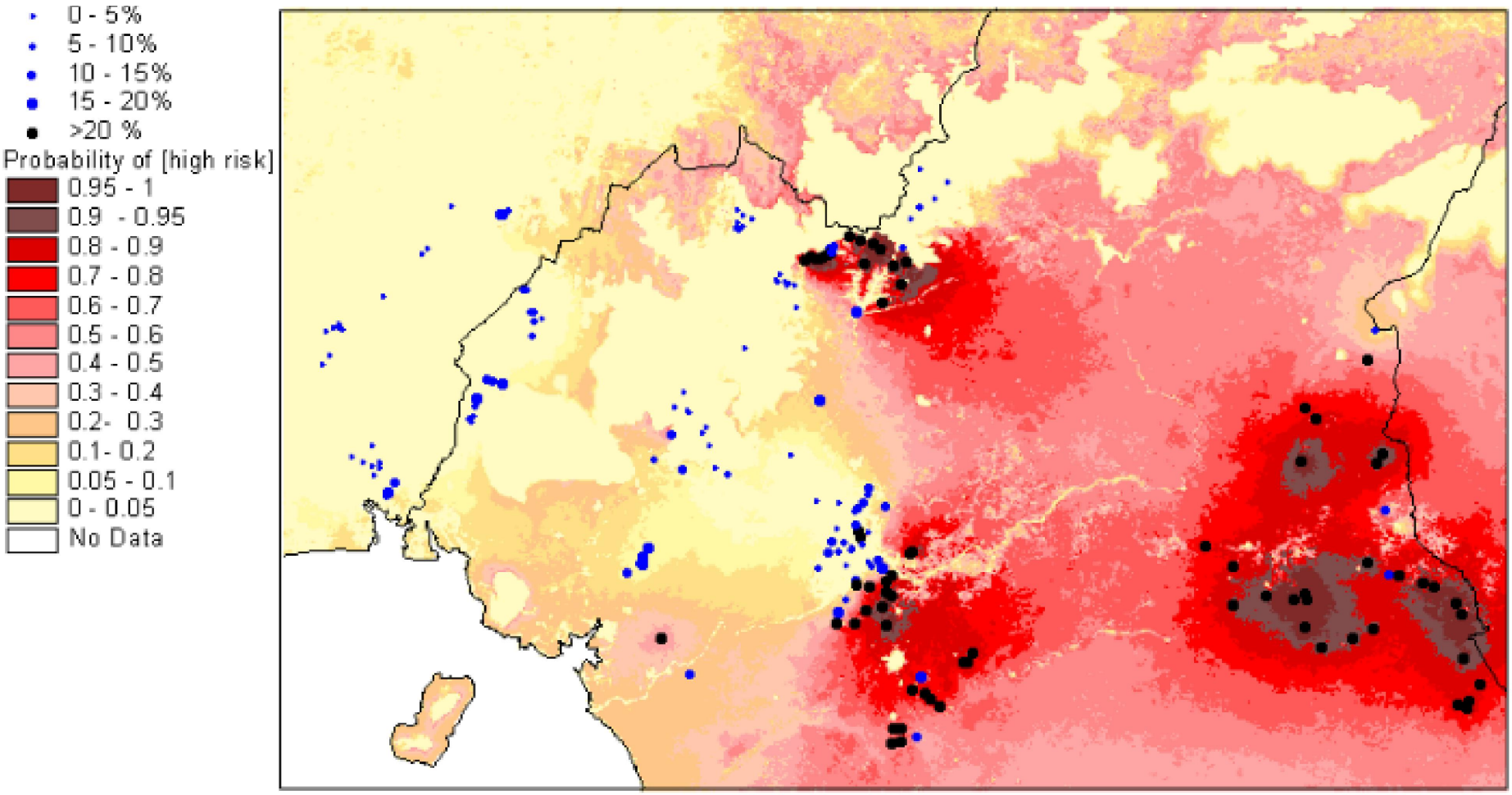


Figure 6: PCM for [high risk] in Cameroon based on ERMr with ground truth data.

Addressing the Loa loa problem (2)

Prevalence is only a proxy outcome, albeit a convenient one

A better strategy?

- **model prevalence and levels of infection**
- **estimate community-level prevalence**
- **predict number of highly infected individuals**

Formulating the question

- **Level of infection:** Y **(parasites per ml of blood)**
- **Prevalence:** $\mathrm{P}(Y > 0)$
- **High-risk individual:** $Y > 8000$

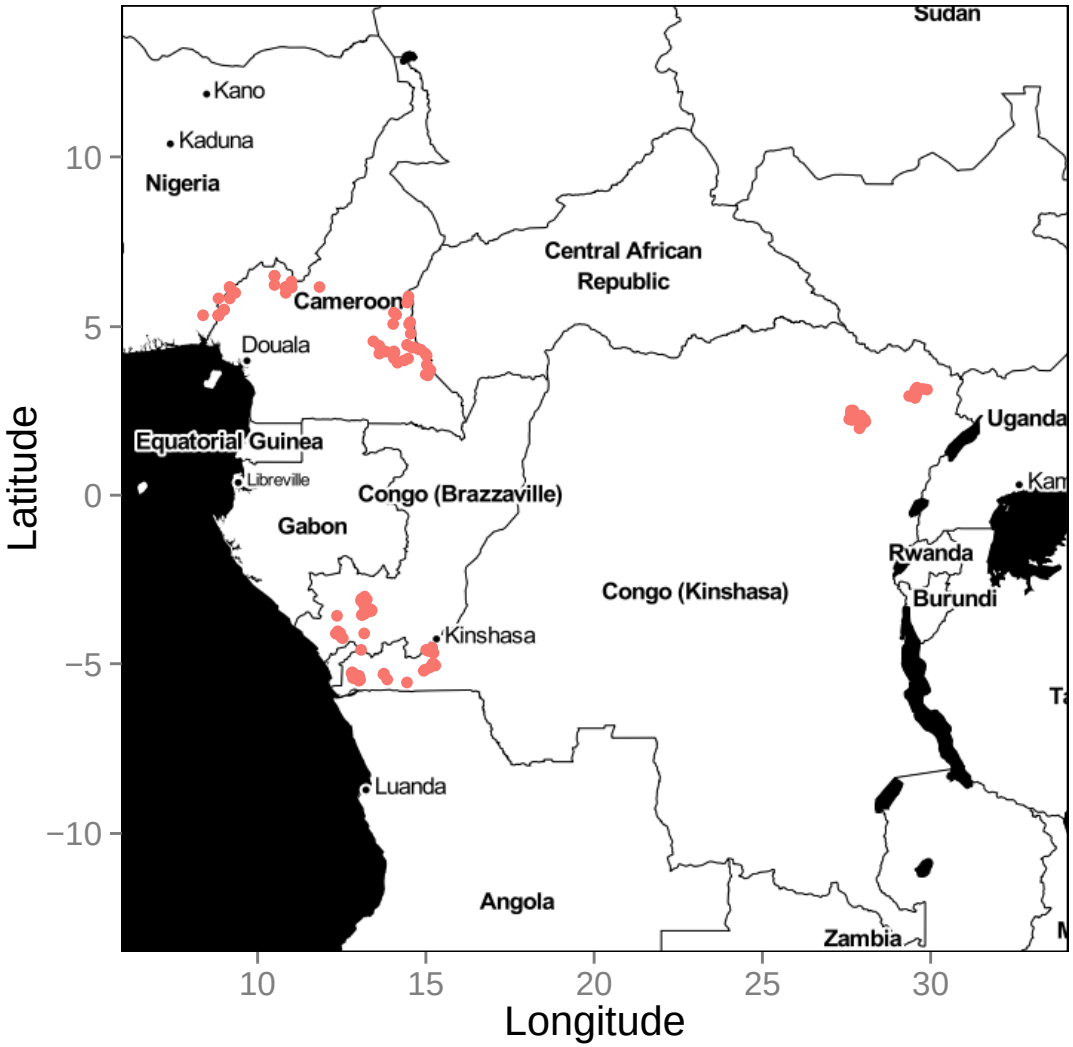
Target for prediction: **proportion ($\Rightarrow$ number) of highly infected individuals in a community**

Data: **from a single community**

n : number of individuals tested
Z : number testing positive
d : covariates

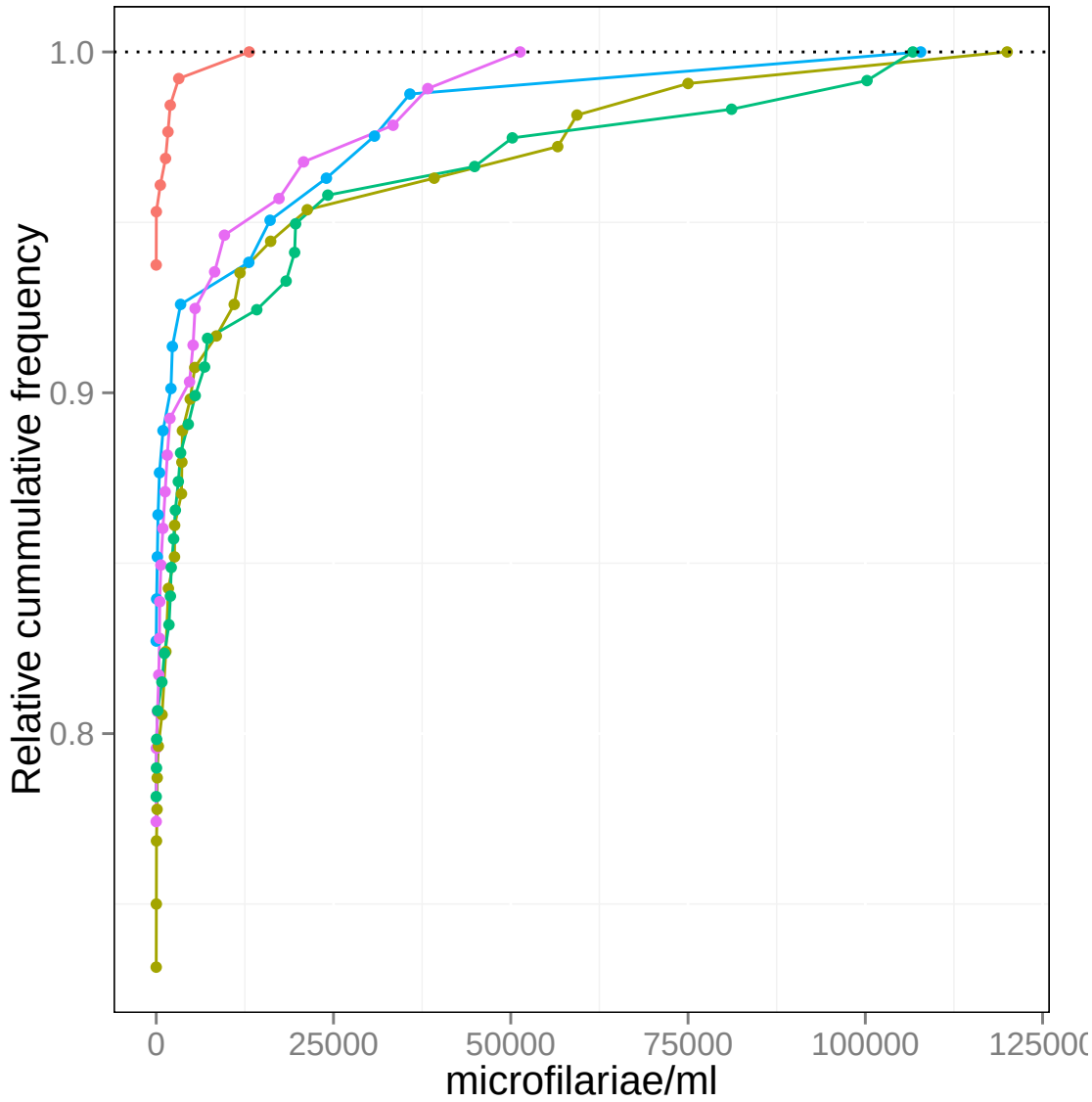
Required: $\mathrm{P}(Y > 8000|Z; n, d)$

Data

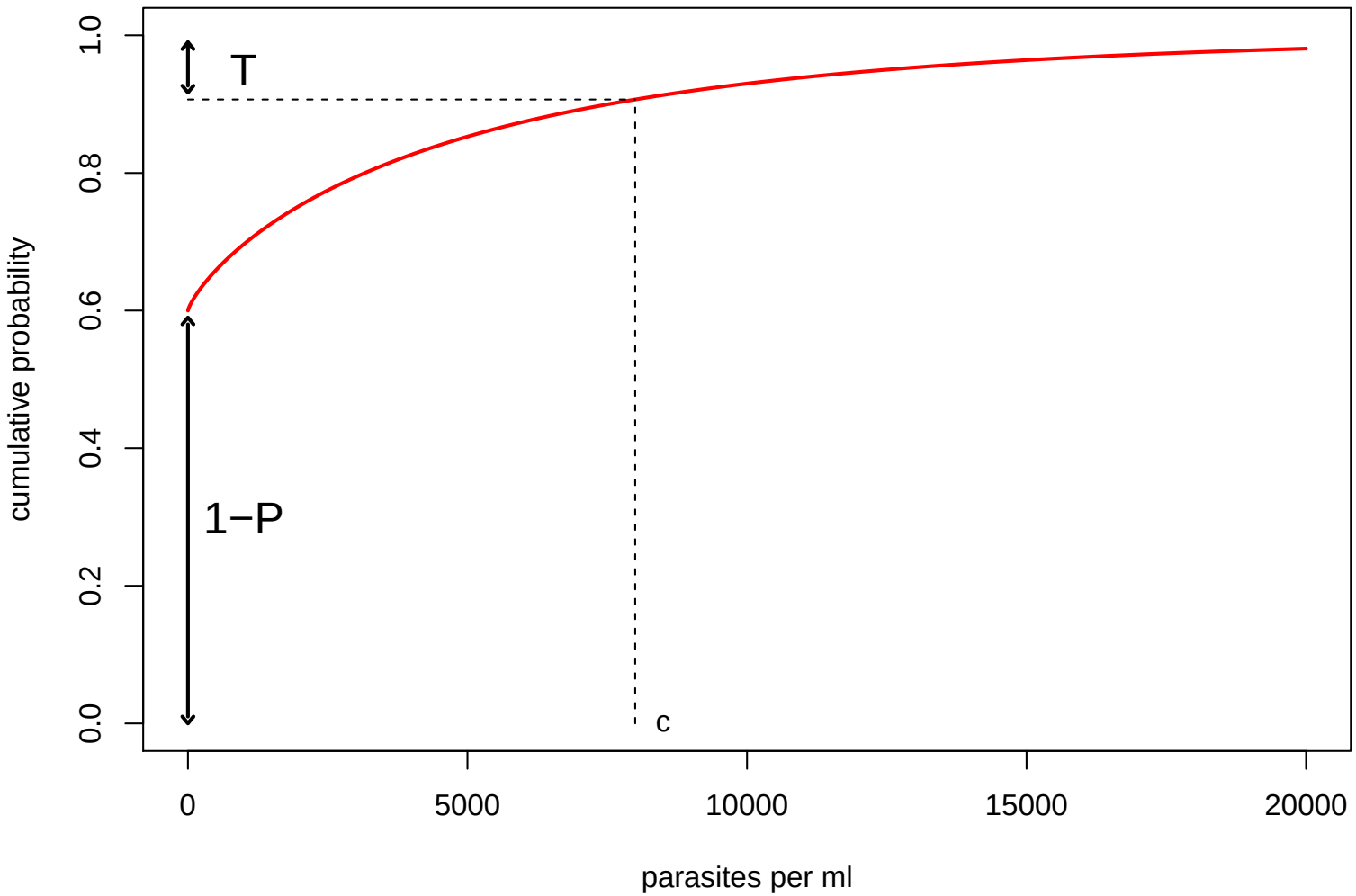


- **223 villages**
- **24 to 229 individuals per village, total 19,128**

Cumulative distribution of infection levels (5 villages)



Schematic: P=prevalence; T=proportion highly infected

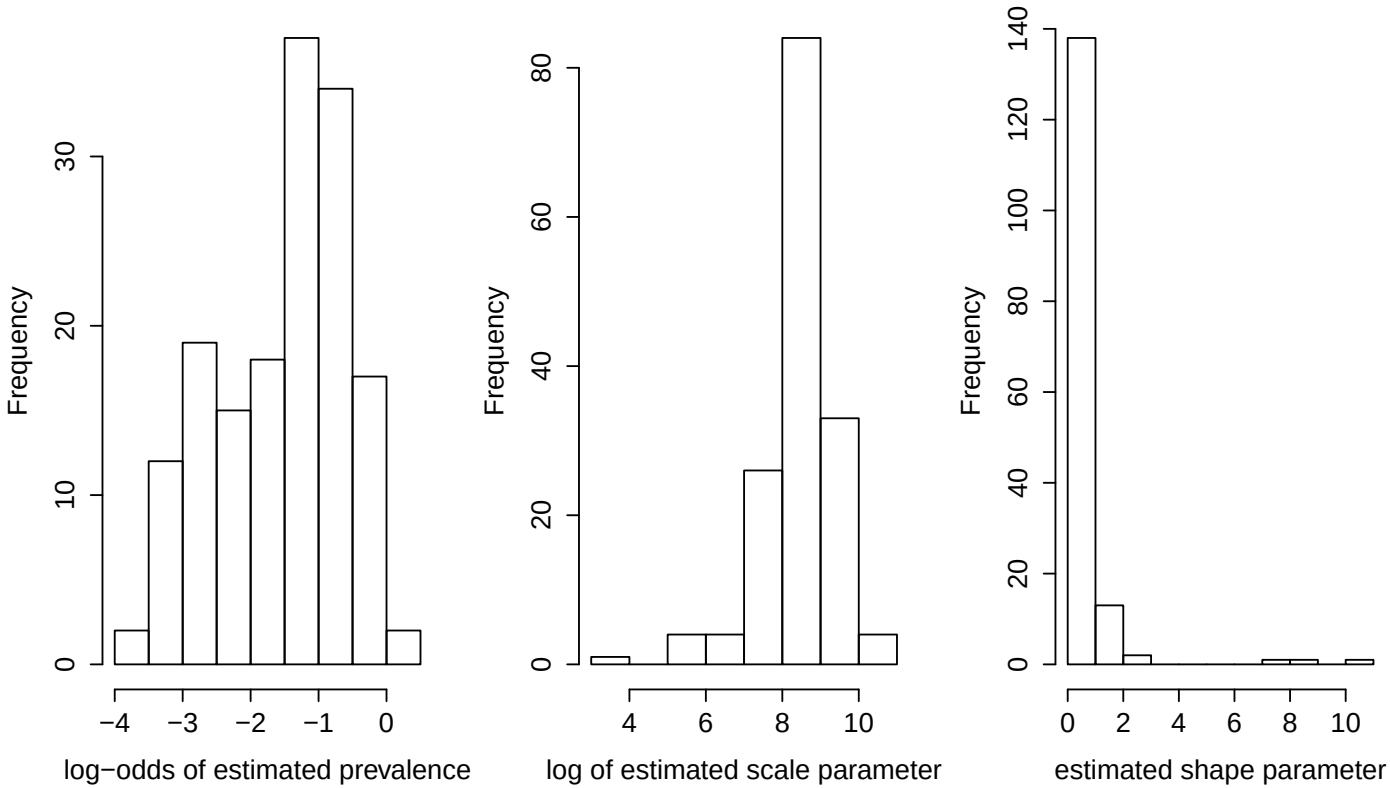


Statistical model

Family of distributions for $Y \geq 0$, positive probability at $Y = 0$,

$$F(y) = (1 - \rho) + \rho G(y; \lambda, \kappa)$$

Parameter estimates from 156 villages



Statistical model (continued)

Family of distributions for $Y \geq 0$, positive probability at $Y = 0$, parameterised through community-level covariates and random effects (unexplained community-level heterogeneity)

$$F(y) = (1 - \rho) + \rho G(y; \lambda, \kappa)$$

- $G(\cdot)$: continuous distribution function on $\mathbb{R}^+$ ((Weibull)
- κ : shape parameter
- $\log\{\rho/(1-\rho)\} = d'\alpha + U$
- $\log \lambda = d'\beta + V$
- $(U, V) \sim \mathrm{BVN}(0, \Sigma)$

Maximum likelihood estimates

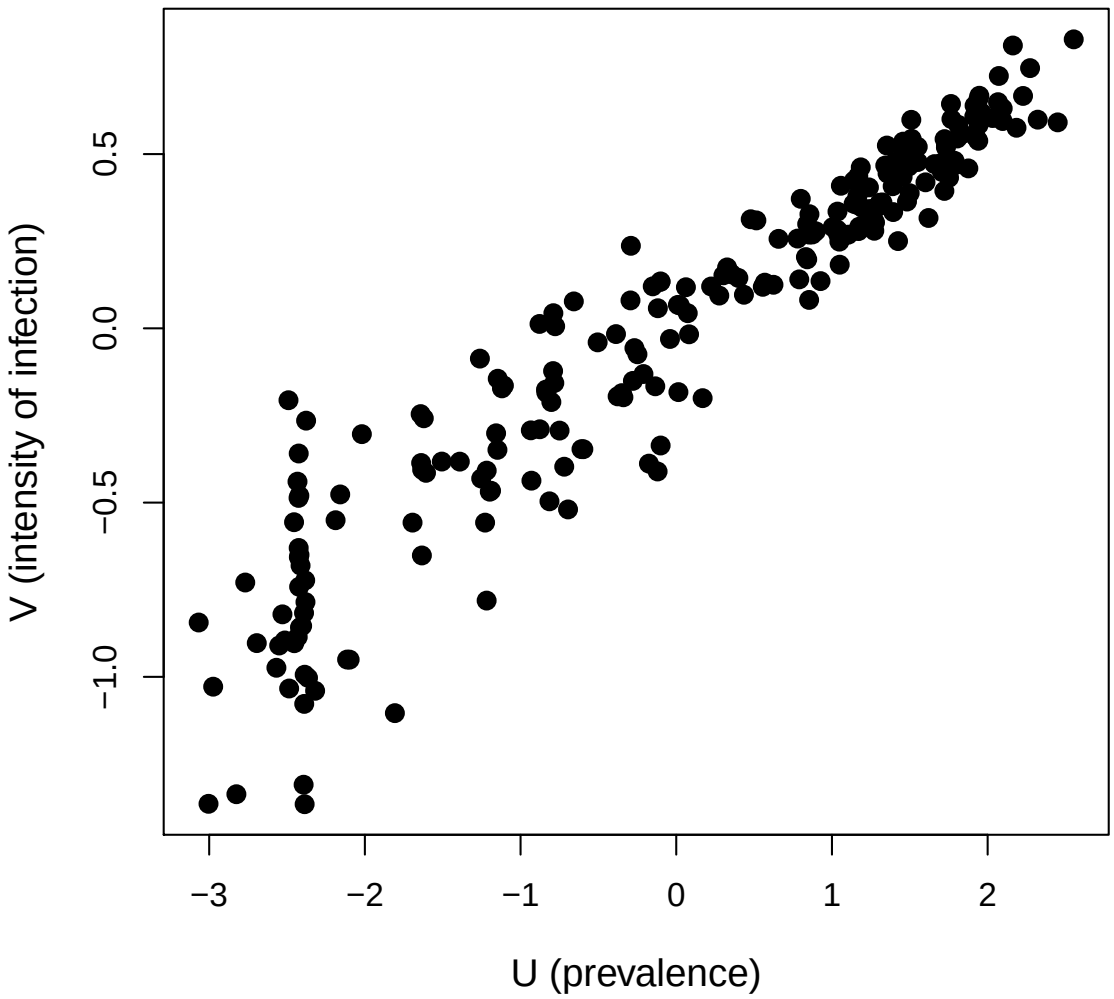
Weibull shape parameter

$\hat{\kappa} = 0.555 \qquad 95\%\ \text{CI} = (0.539, 0.572)$

Random effects

	Estimate	95% CI	
σ^2_U	2.069	1.637	2.616
σ^2_V	0.380	0.231	0.625
ρ_{UV}	0.680	0.454	0.824

Predicted random effects



Prediction

- **Target for prediction:** $T = \rho(U) \times \{1 - G(8000; \lambda(V), \kappa)\}$
- **Plug-in prediction:** substitute parameter estimates for unknown true values
- **Predictive distribution:** $[U, V|\text{data}] \Rightarrow [T|\text{data}]$

Sampling from the predictive distribution

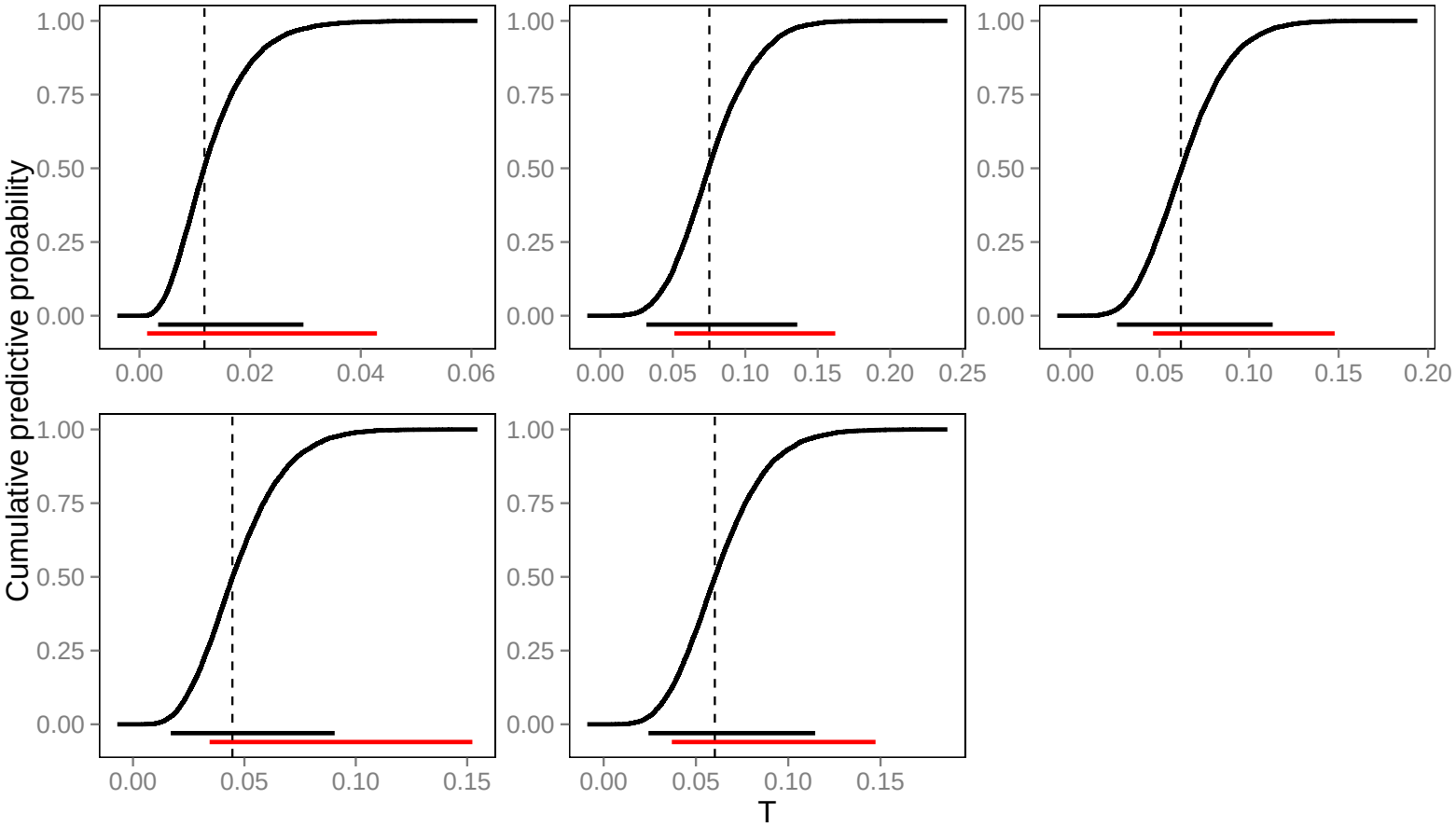
$$[U, V|Z] = [U|Z][V|U, Z] = [U|Z][V|U]$$

- $[U|Z] = [Z|U][U]/[Z]$

  Gaussian quadrature for $[Z]$

- $[V|U] = \mathrm{N}\{\rho U \sigma_U/\sigma_V, (1 - \rho^2)\sigma_V^2\}$

Prediction: model-based vs empirical



Black lines: model-based 95% predictive intervals

Red lines: 95% confidence intervals based on binomial sampling distribution of observed numbers with parasite count $>$ 8000/ml

Finite population correction

Community size N, sample size n, of whom h 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 the predictive distribution of Q;**
2. **Sample a value M from a binomial distribution with number of trials $N - n$ and probability of success Q;**
3. **Repeat 1 and 2 many times to give probability distribution of M, and hence of $H = h + M$**

Geostatistical extension

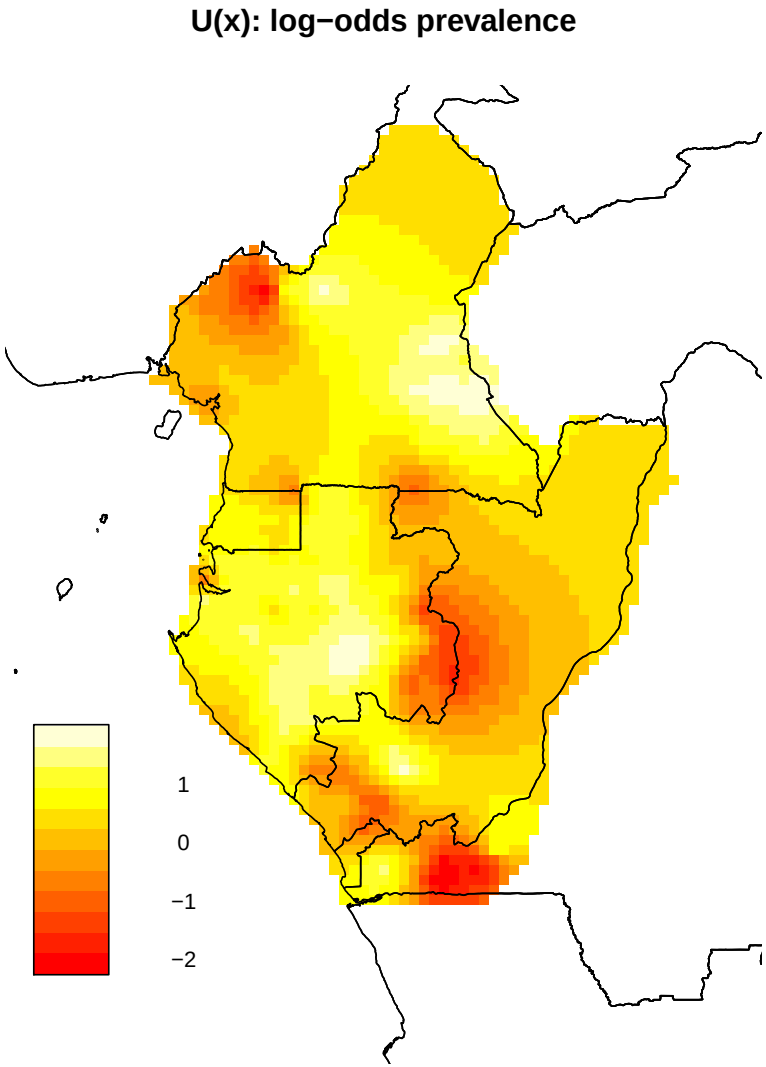
Current model

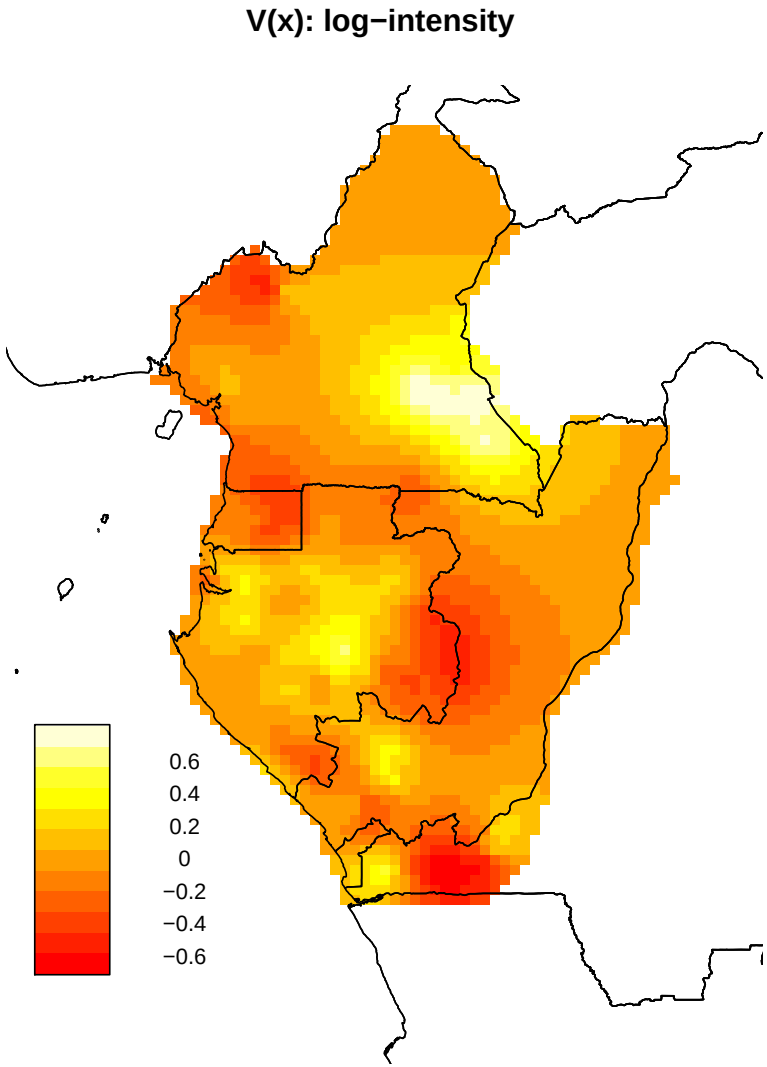
Independent $(U_i, V_i) : i = 1, \ldots, m \Rightarrow$ only village-specific information is helpful

Borrowing strength: use information on 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 $\{(U(x), V(x)) : x \in \mathbb{R}^2\}$**

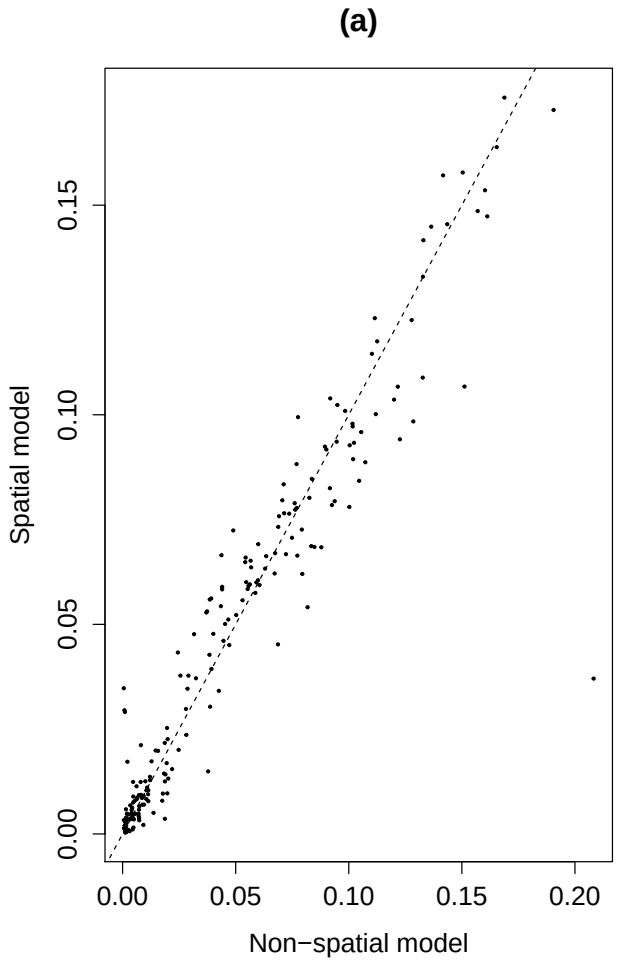
Predicted random effects $U(x)$ and $V(x)$



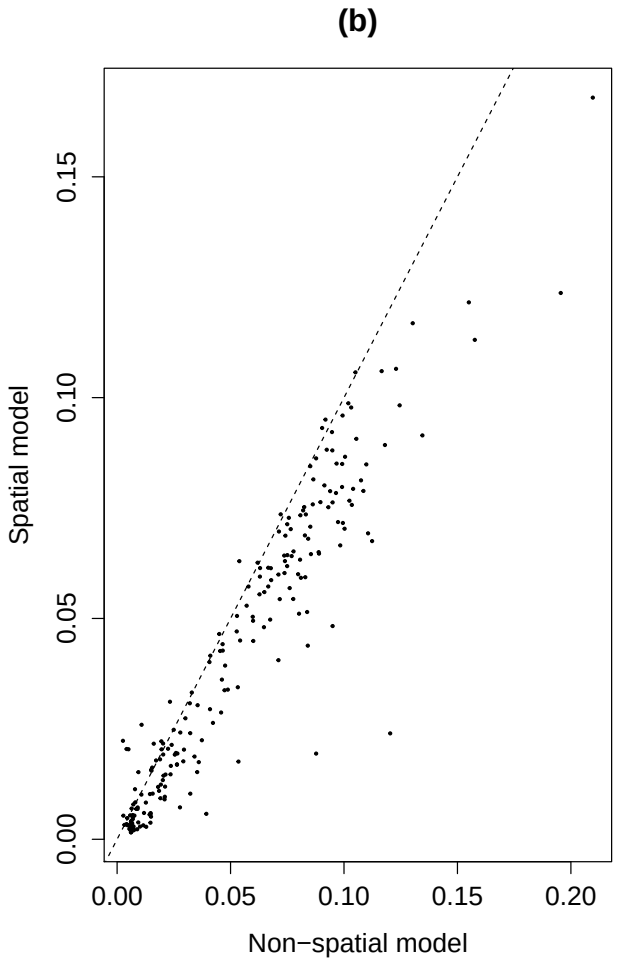


Comparison between non-spatial and spatial model predictions

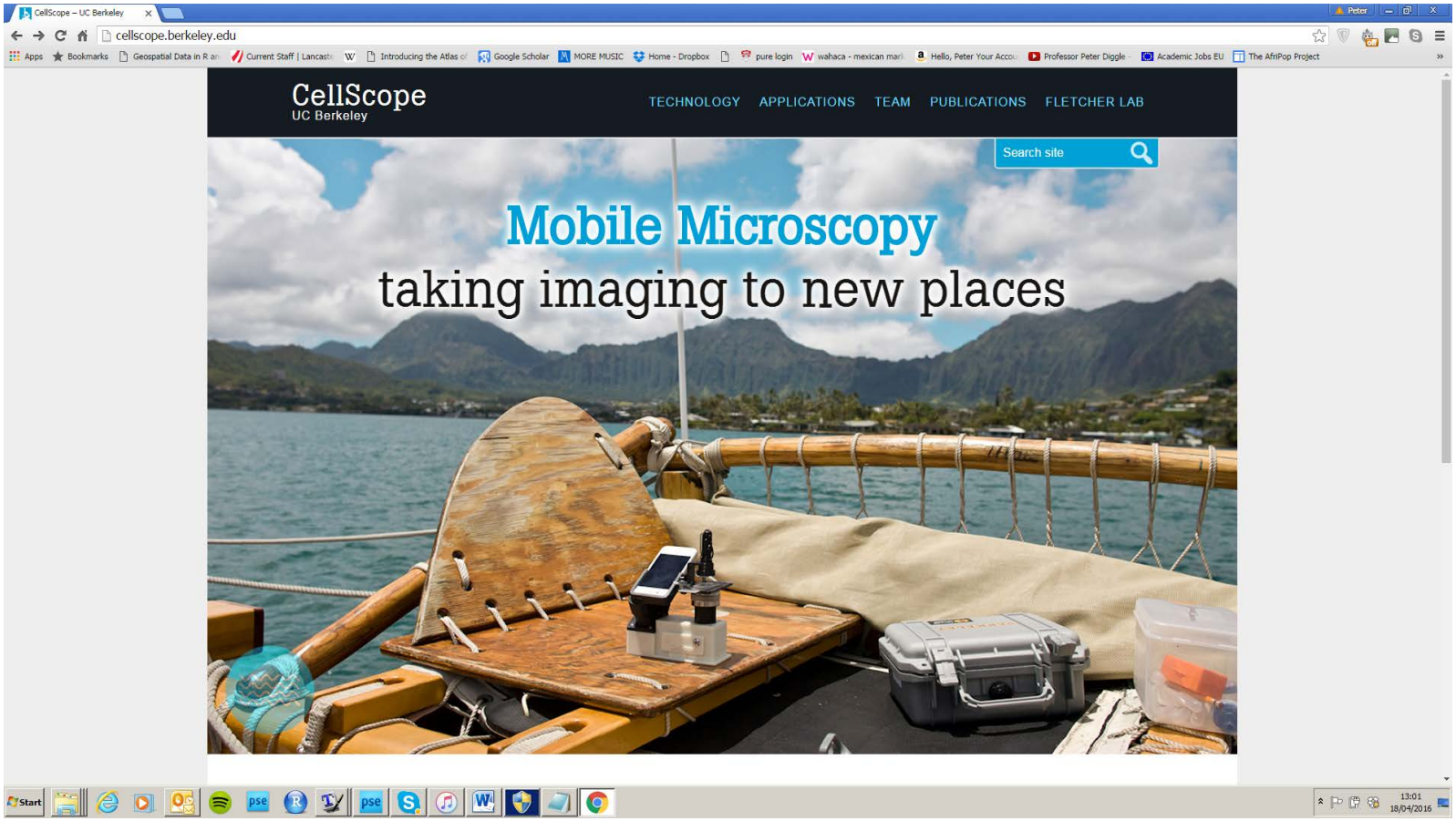
(a) point



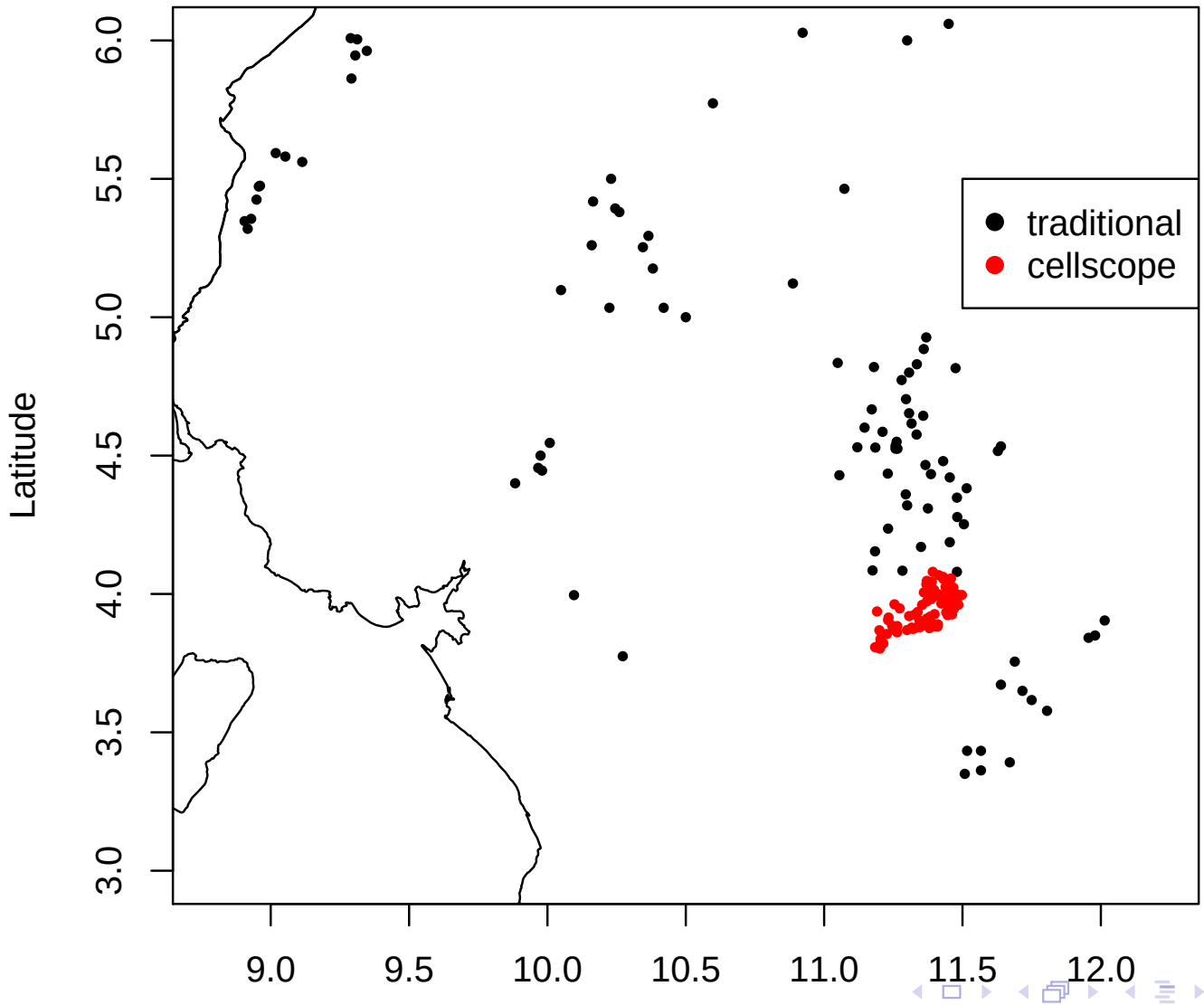
(b) length of 95% interval



Exploiting new technology ... cellscope



Field-testing cellscope vs MF: sampling locations



Infection status: cellscope vs MF

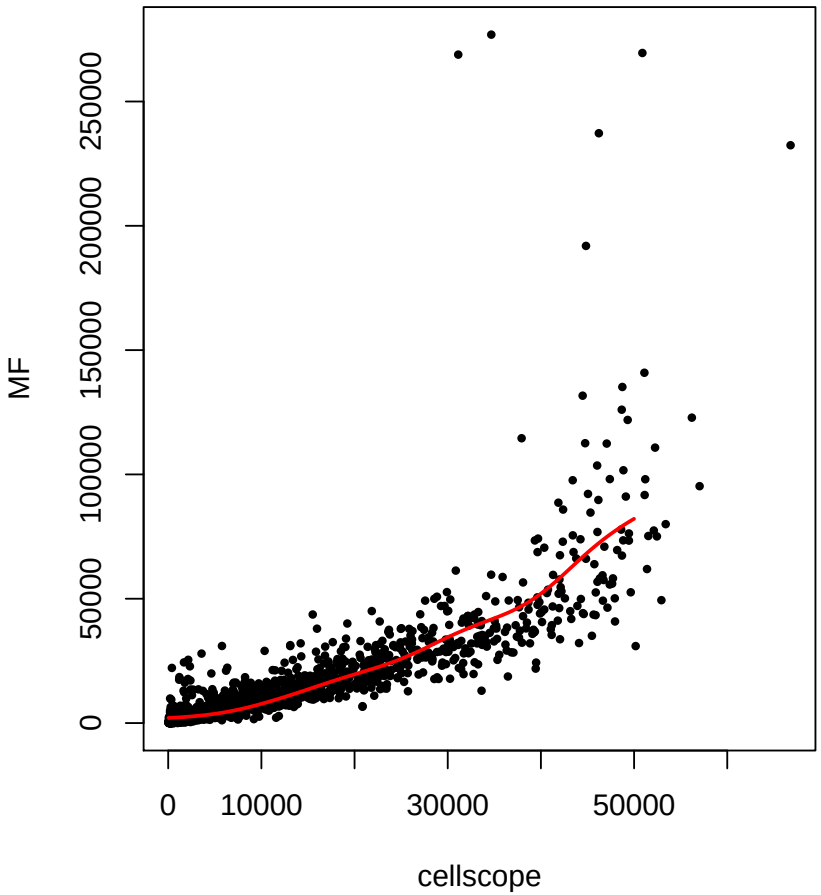
Binary classification: 0/1=uninfected/infected

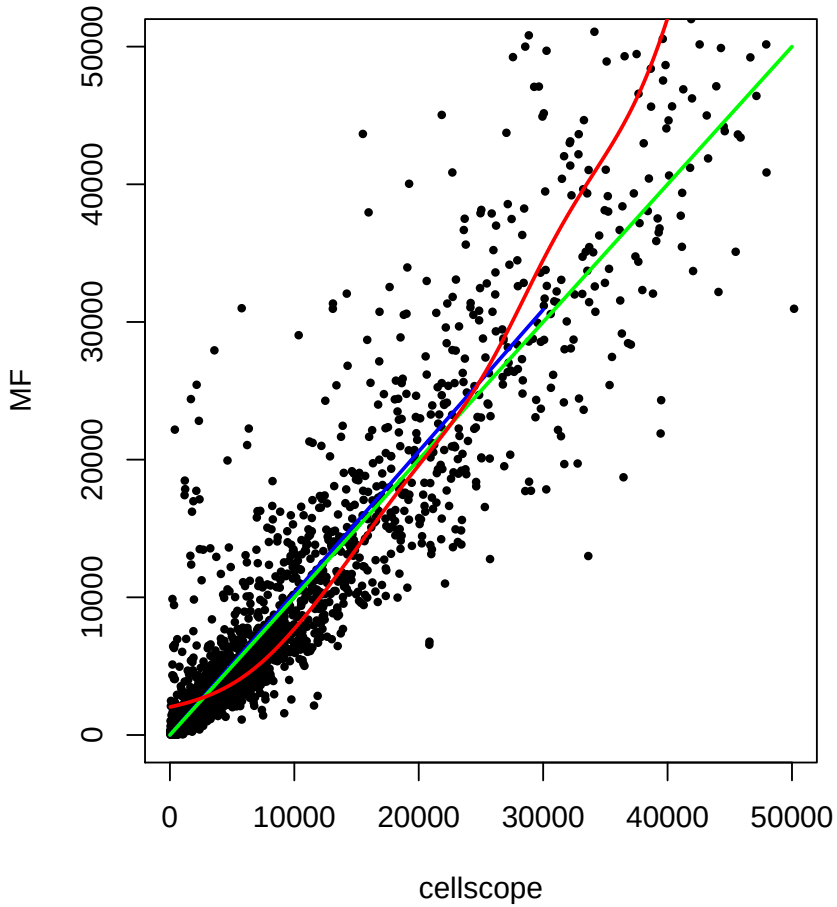
		cellscope		
		0	1	Total
MF	0	12135	357	12492
	1	268	2421	2689
	Total	12413	2779	15181

Sensitivity/specificity of cellscope considered as proxy for MF status, and *vice versa*

gold standard	SE	SP
MF	0.900	0.971
cellscope	0.871	0.971

Infection levels: cellscope vs MF





Maximum likelihood parameter estimates: cellscope vs MF

	Schlüter et al (2016)		88 villages			
	MF		MF		Cellscope	
	Estimate	SE	Estimate	SE	Estimate	SE
α	-2.470	0.125	-1.477	0.075	-1.553	0.022
β	8.20	0.097	8.702	0.056	8.660	0.033
σ^2_U	2.99	0.365	0.129	0.044	0.146	0.023
σ^2_V	0.545	0.131	0.068	0.035	0.072	0.011
ρ	0.699	0.082	0.631	0.105	0.516	0.079
κ	0.556	0.008	0.604	0.009	0.678	0.010

Current work

- **presence of spatial correlation suggests environmental effects**
- **but available covariates from remotely sensed images had little impact on predictive inferences**
- **social/genetic effects have also been hypothesised, but no candidate covariates yet available**
- **current debate:**

  predictive inference or test-and-treat?

  (a false dichotomy?)

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