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

Peter J Diggle[1,2] and Emanuele Giorgi[1]

Lancaster University[1] and University of Liverpool[2]

CHICAS combining health information, computation and statistics

Lancaster Medical School | Lancaster University

UNIVERSITY OF LIVERPOOL | INSTITUTE OF INFECTION AND GLOBAL HEALTH

References

Diggle, P.J., Moyeed, R.A. and Tawn, J.A. (1998). Model-based Geostatistics (with Discussion). *Applied Statistics* **47** 299–350.

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.

Giorgi, E., Sesay, S.S., Terlouw, D.J. and Diggle, P.J. (2014). Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. *Journal of the Royal Statistical Society* A (to appear)

Rodrigues, A. and Diggle, P.J. (2010). A class of convolution-based models for spatio-temporal processes with non-separable covariance structure. *Scandinavian Journal of Statistics*, **37**, 553–567.

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

Acknowledgements

CHICAS, Lancaster: Ole Christensen, Barry Rowlingson, Michelle Stanton, Ben Taylor, Rachel Tribbick

MLW, Blantyre, Malawi Sanie Sesay, Anja Terlouw

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

IRI, Columbia University: Madeleine Thomson

...and many others

Outline

- introduction
- general remarks on statistical modelling
- the standard binomial geostatistical model: Loa loa
- low-rank approximations: river blindness
- combining data from multiple surveys: malaria
- spatially structured zero-inflation: river blindness re-visited
- implementation
- closing remarks

Low resource settings

Prevalence mapping 1

Single prevalence survey

Sample n individuals, observe Y positives

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

Multiple prevalence surveys

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

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

Prevalence mapping 2

Extra-binomial variation

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

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

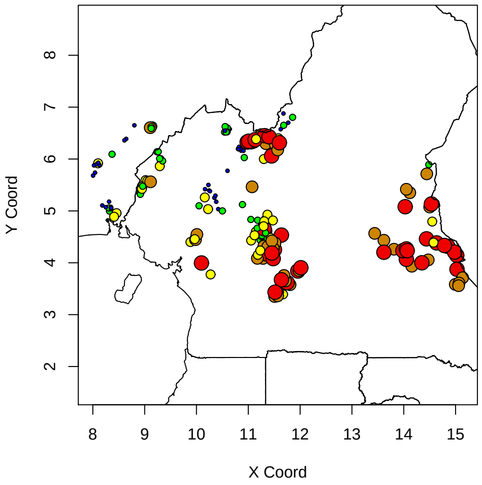
This talk

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 which is concerned with data obtained by spatially discrete sampling of a spatially continuous process**

A geostatistical data-set: Loa loa prevalence surveys



Model-based Geostatistics (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**

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”

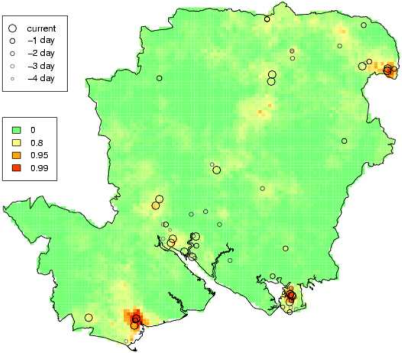
Emperor Joseph II

“Only as many as there needed to be”

Mozart (apochryphal?)

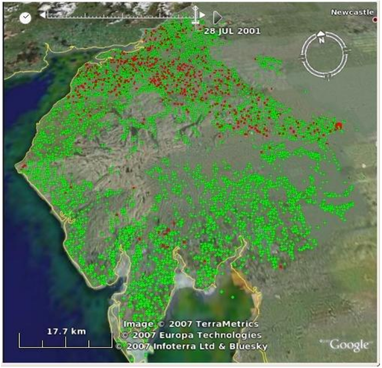
Empirical modelling: The AEGISS project (Diggle, Rowlingson and Su, 2005)

- **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 $\rho(\mathbf{u}, \mathbf{v})$**



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)

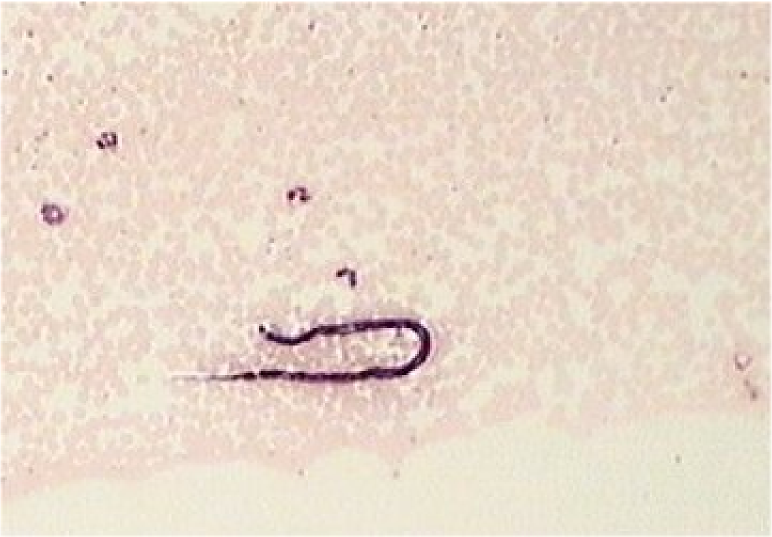
APOC

African
Programme for
Onchocerciasis
Control

- "river blindness" – endemic in wet tropical regions
- donation programme of mass treatment with ivermectin
- approximately 60 million treatments to date, in 19 countries
- serious adverse reactions experienced by some patients highly co-infected with *Loa loa* parasites
- precautionary measures put in place before mass treatment in areas of high *Loa loa* prevalence

http://www.who.int/pbd/blindness/onchocerciasis/en/

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)

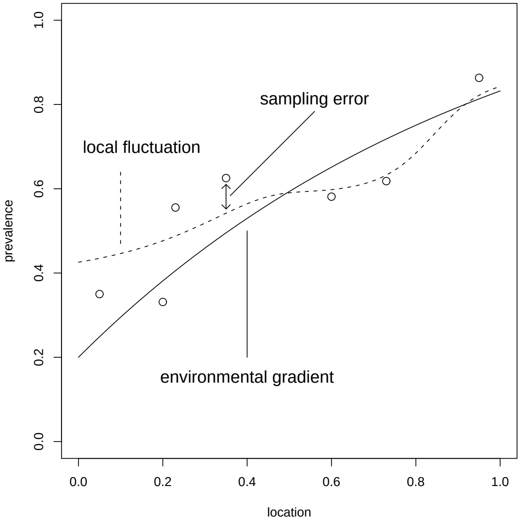
- measured on regular grid to cover region of interest
- elevation, green-ness of vegetation

Objectives

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

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

Schematic representation of Loa loa model



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**

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

Peter McCullagh

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)**

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

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

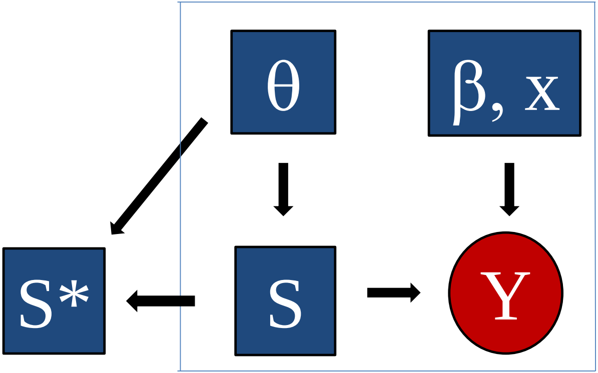
$$Y_i|S(\cdot) \sim \mathrm{Bin}\{n_i, p(x_i)\} \text{ (binomial sampling)}$$

Conditional dependence structure

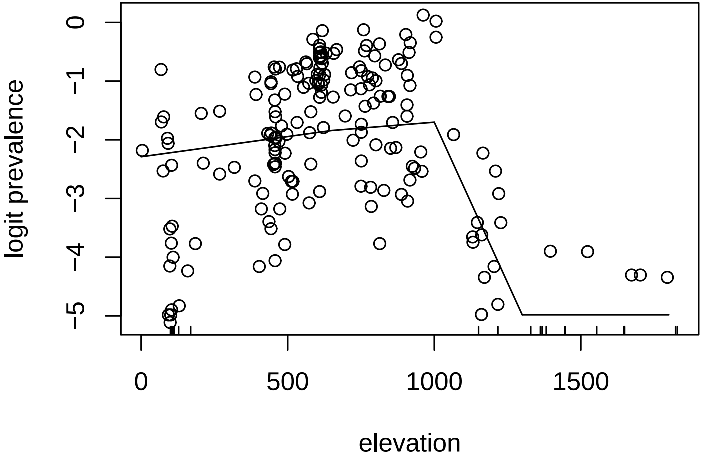
Signal: $\mathbf{S}, \mathbf{S}^*$ **(data-locations and prediction locations)**

Data: $\mathbf{Y}$ **(data-locations only)**

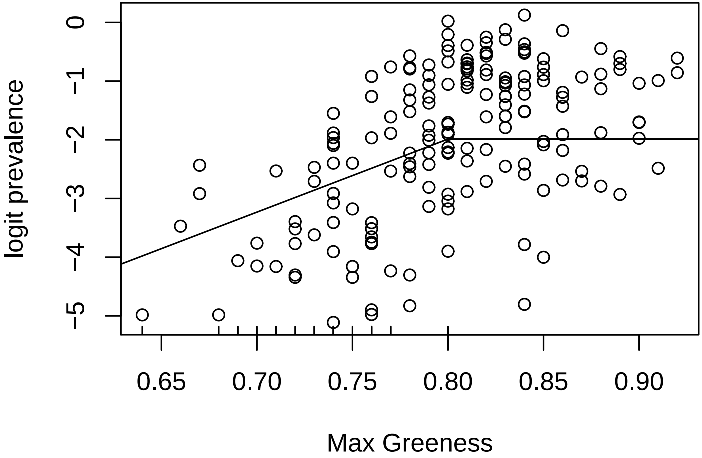
Parameters: β **(regression terms),** θ **(covariance structure)**



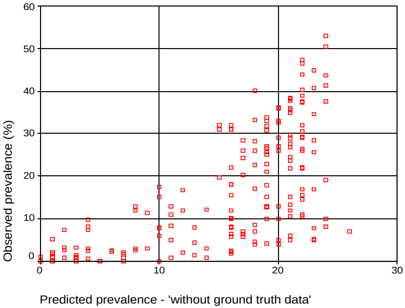
logit prevalence vs elevation

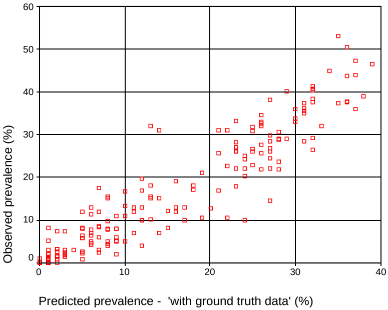


logit prevalence vs max NDVI



How useful is the geostatistical modelling?





Logistic regression

Model-based geostatistics

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

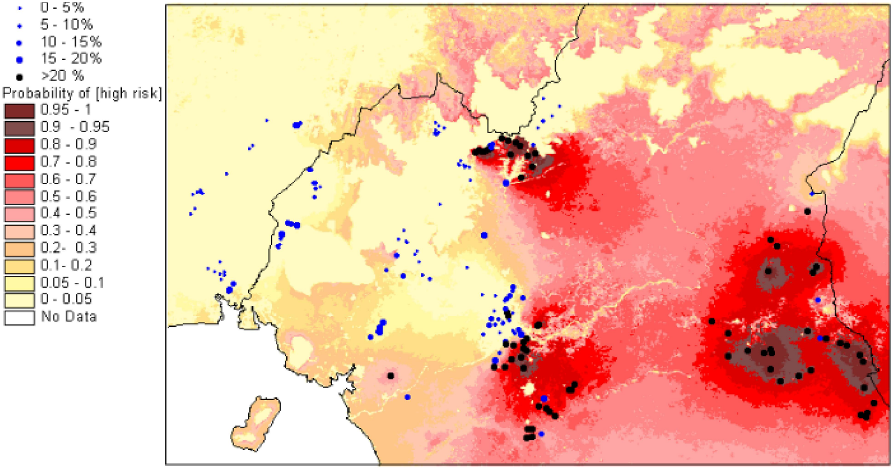


Figure 6: PCM for [high risk] in Cameroon based on ERMr with ground truth data.

Extending the model

- **non-spatial extra-binomial variation**
- **low-rank approximations;**
- **combining data from multiple surveys**
  - **randomised and non-randomised**
  - **at different times**
- **spatially structured zero-inflation.**

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**

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

- **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**

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

Low-rank approximations (Rodrigues and Diggle, 2010)

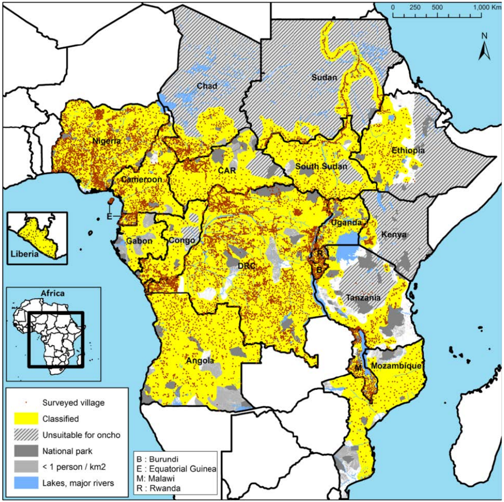
$$S(x) \approx \mu + \sum_{j=1}^{M} w(x - k_j) Z_j$$

- $w(u)$: 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



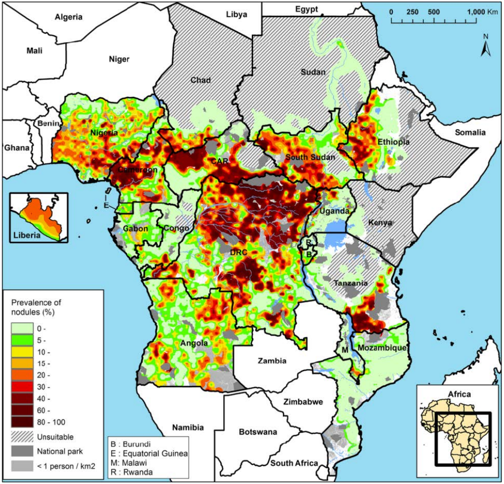
Application: onchocerciasis mapping Africa-wide (Zoure et al, 2014): low-rank model

- $M = 10734$ points X_j in regular lattice at spacing 0.1 degrees
- to approximate Matérn correlation, $M(\phi, \kappa)$, $\kappa = 2$

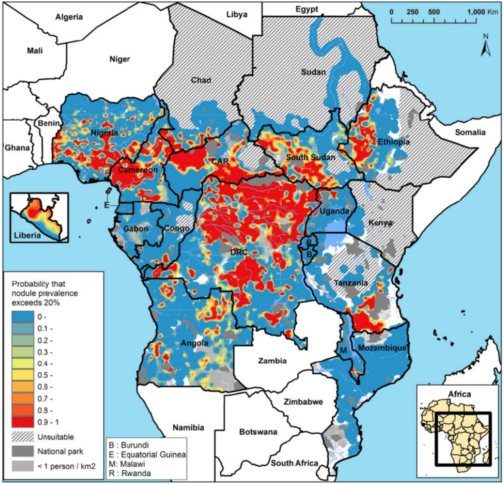
$$w(u) = \phi^{-1} \exp(-2\sqrt{2}\, u/\phi)$$

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): prevalence estimates



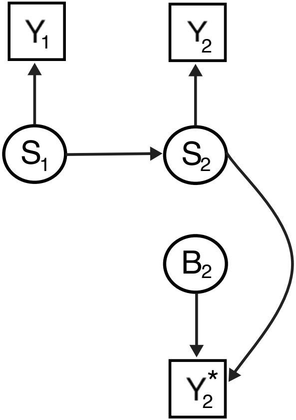
Application: onchocerciasis mapping Africa-wide (Zoure et al, 2014): exceedance probabilities



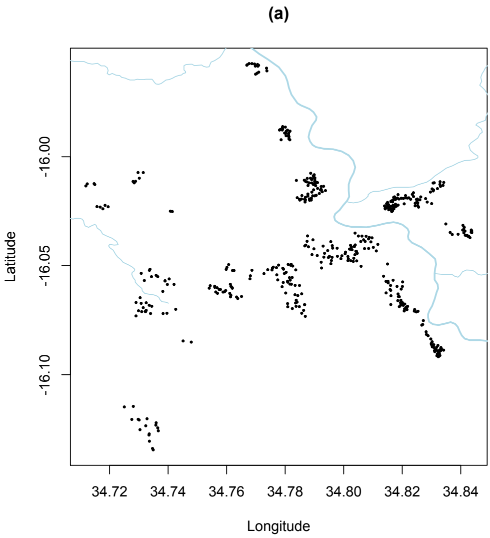
Multiple surveys (Giorgi et al, 2015)

Surveys: $i = 1, \ldots, r$ **locations** $x_{ij} : j = 1, \ldots, n_i$

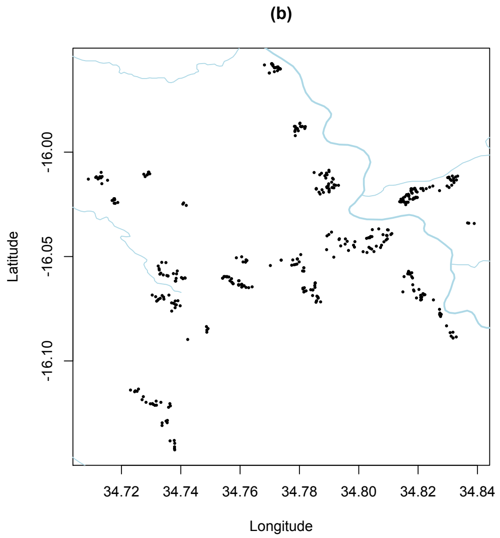
$$\eta_{ij} = d(x_{ij})^\top \beta_1 + S_i(x_{ij}) + I(i \in \mathcal{B})[B_i(x_{ij}) + d(x_{ij})'\beta_i] + U_{ij}$$



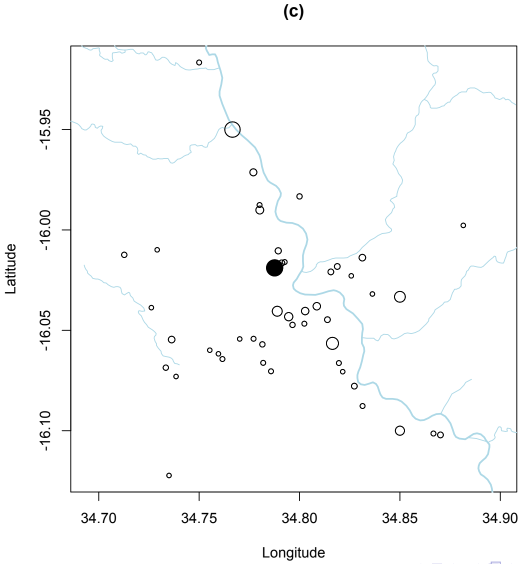
Application: malaria mapping, Chikhwawa district, Malawi (Giorgi et al, 2015): rMIS individual locations



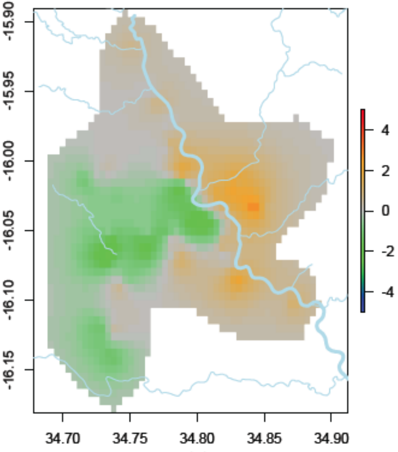
Application: malaria mapping, Chikhwawa district, Malawi (Giorgi et al, 2015): eMIS individual locations



Application: malaria mapping, Chikhwawa district, Malawi (Giorgi et al, 2015): EAG village locations and prevalences

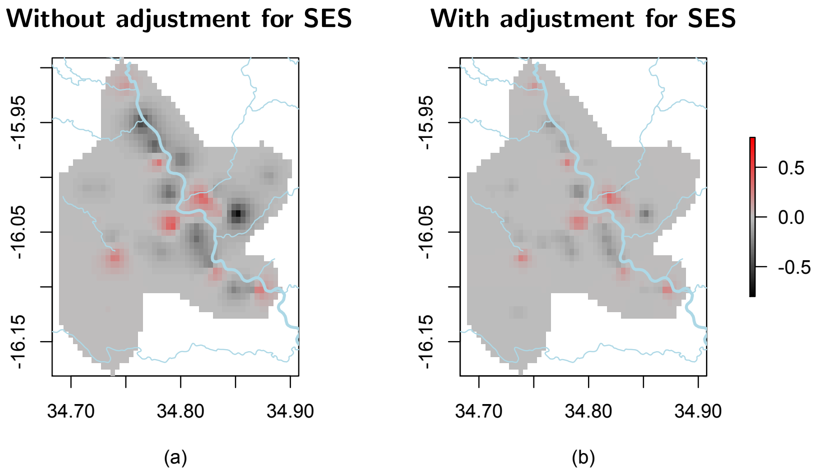


Application: malaria mapping, Chikhwawa district, Malawi (Giorgi et al, 2015): estimated prevalence map



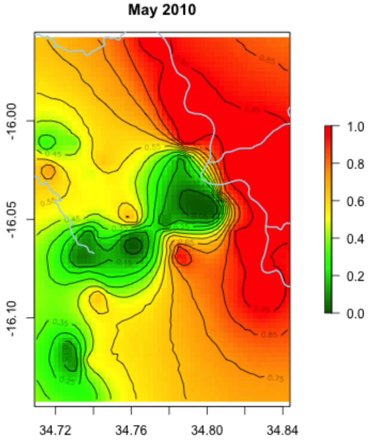
(c)

Application: malaria mapping, Chikhwawa district, Malawi (Giorgi et al, 2015): estimated bias maps



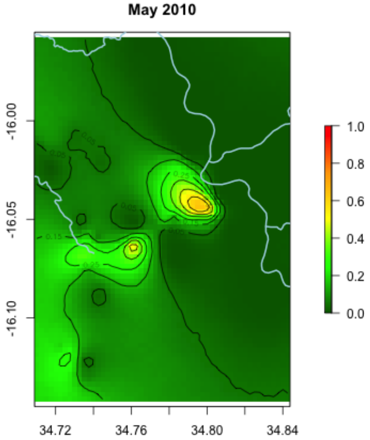
Continuous time: rolling malaria indicator surveys

Hotspots: $P(\text{prevalence} > 20\%)$



Continuous time: rolling malaria indicator surveys

Coldspots: P(prevalence $< 5\%$)



Spatially structured zero-inflation: river blindness re-visited

- **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

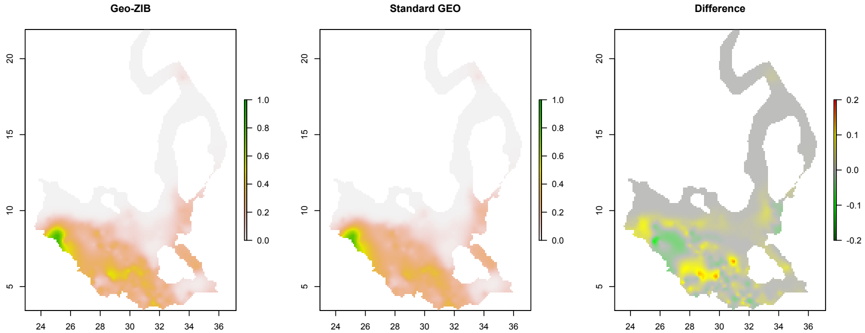
$$\{q_i, p_i\} \rightarrow \{Q(x), P(x)\} : x \in \mathbb{R}^2 \sim \text{bivariate stochastic process}$$

Double logistic Gaussian process

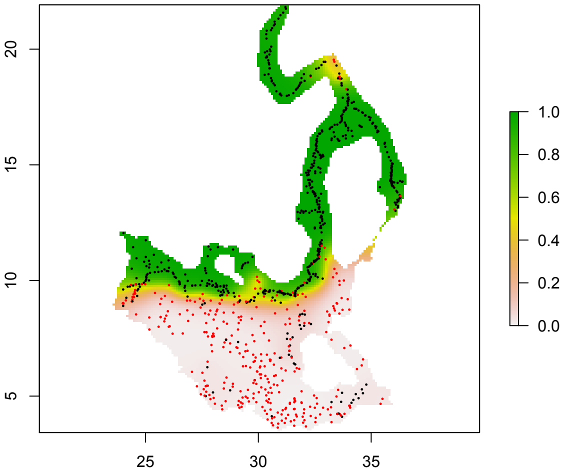
$$P(Y = y|S_1(x), S_2(x)) = \begin{cases} Q(x) + (1 - Q(x)) \times Bin(0; n, p(x)) & : y = 0 \\ (1 - Q(x)) \times Bin(y; n, p(x)) & : y > 0 \end{cases}$$

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

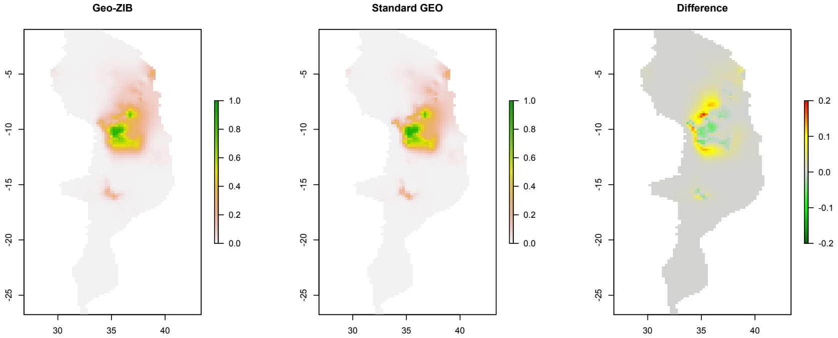
Sudan: probability exceedance map



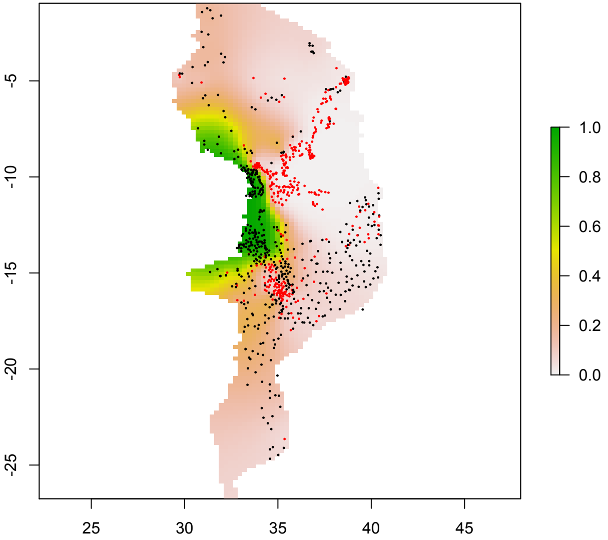
Sudan: non-transmissible probability map (**Q(x)**)



Mozambique/Malawi/Tanzania: probability exceedance map



Mozambique/Malawi/Tanzania: non-transmissible probability map (**Q(x)**)



Implementation

- **Monte Carlo maximum likelihood**
- **Plug-in prediction**
- **R package** `PrevMap`
- **Bayesian version?**

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 ...especially to science students