

Spatio-temporal Design and Analysis

Peter J Diggle

Lancaster University

December 2022

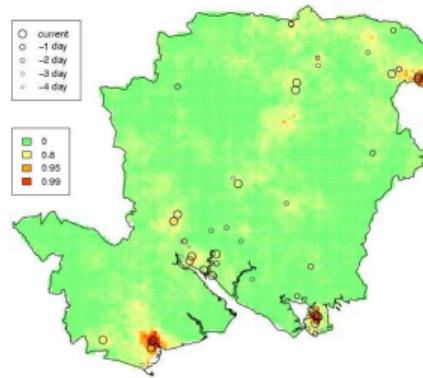


A short introduction to a big topic

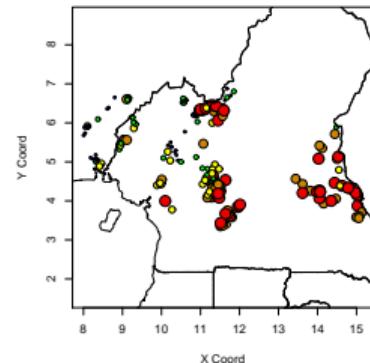
1. Spatial data-formats
2. Statistics and scientific method
3. Geospatial survey design
4. Geospatial survey analysis
5. From spatial to spatio-temporal
 - 5.1 longitudinal designs
 - 5.2 repeated cross-sectional designs
 - 5.3 adaptive designs
6. Reading list

Geospatial data-formats

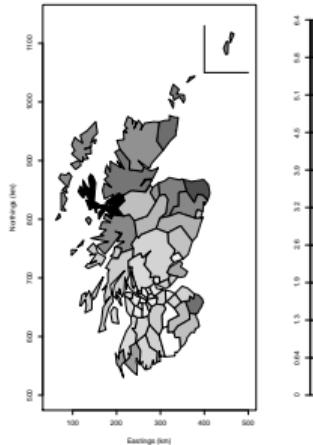
Point process



Survey



Registry



In each example, the quantity of scientific interest is a **spatially varying risk surface**,
 $\rho(x) : x \in A$

Statistical modelling principles

A statistical model 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

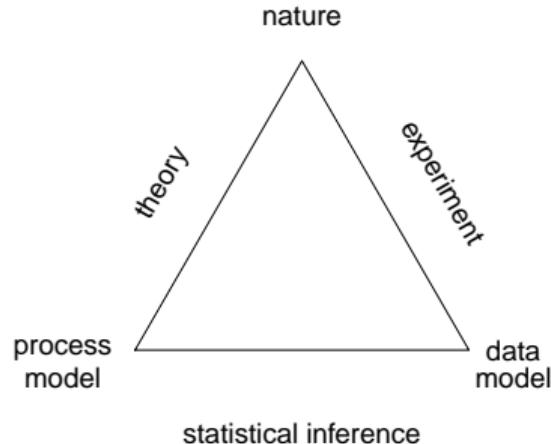
“Too many notes, Mozart”

Emperor Joseph II

“Only as many as there needed to be”

Mozart (apochryphal?)

Statistics and scientific method



A statistical model is:

- a **device** to answer a question
- a **bridge** between theoretical and applied science
- a **framework** to enable principled inference in the presence of uncertainty

Scientific purpose is more important than data-format

Analyse problems, not data

The essence of statistical modelling and inference

S = state of nature
 Y = all relevant data

Model: $[S, Y] = [S][Y|S]$

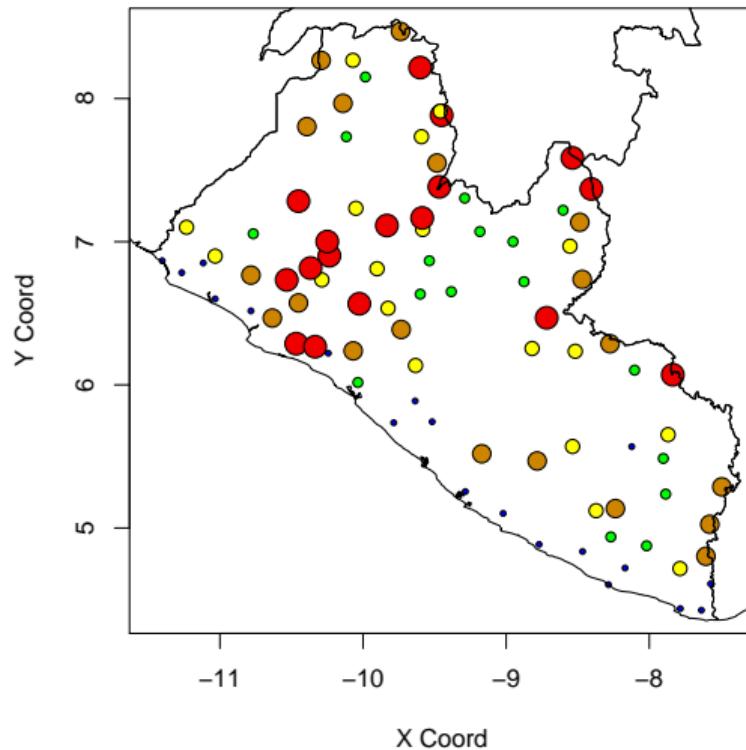
Estimation: what can you say about the parameters that define $[S]$ and $[Y|S]$?

Testing: are the data compatible with hypothesised values for one or more of the parameters?

Prediction: What can you say about the realisation of S ?

Putting it into practice: disease prevalence surveys

A prevalence survey data-set: onchocerciasis in Liberia



Prevalence survey design

- What's the question?
- Where to sample?
- How many locations?
- How many individuals per location?

Neglected Tropical Disease Control Programmes



- Ivermectin (Mectizan): annual dose clears microfilarial infections of the blood
- generally considered safe, with no serious side-effects
- mass drug administration (MDA) made possible by donation programme (Merck)
- used in multi-national programmes to combat onchocerciasis and lymphatic filariasis

What's the question?

WHO guideline: if the prevalence of microfilaraemia among adults in an Evaluation Unit is less than 1%, MDA can be stopped

Performance requirement for a survey

- $\text{PPV} \geq 0.95$ (positive predictive value)
probability that an EU which declares elimination has prevalence less than 1%
- $\text{NPV} \geq 0.75$ (negative predictive value)
probability that an EU which does not declare elimination has prevalence greater than 1%

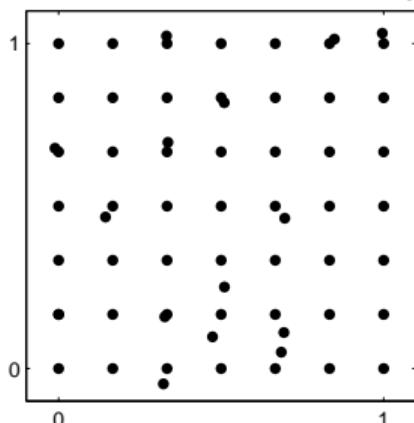
Kenya EUs



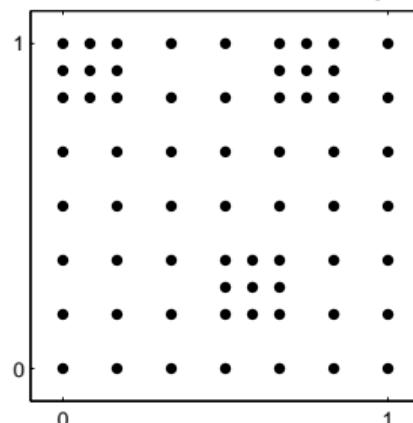
Where to sample?

Examples of lattice-based designs

A) Lattice plus close pairs design



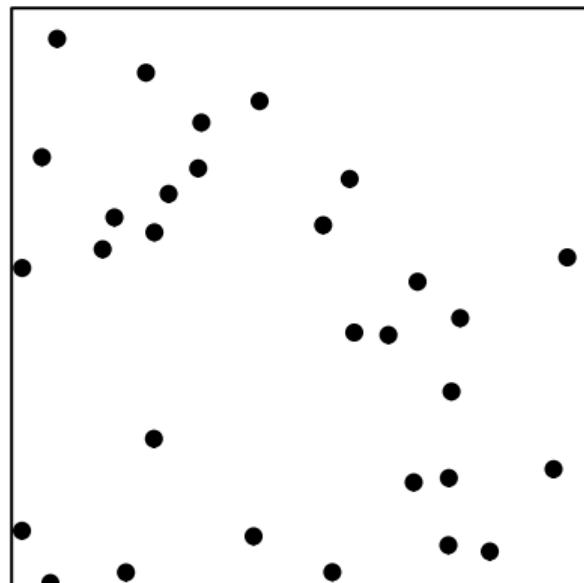
B) Lattice plus in-fill design



- Why include close pairs?
- Absence of a probability sampling framework may be problematic

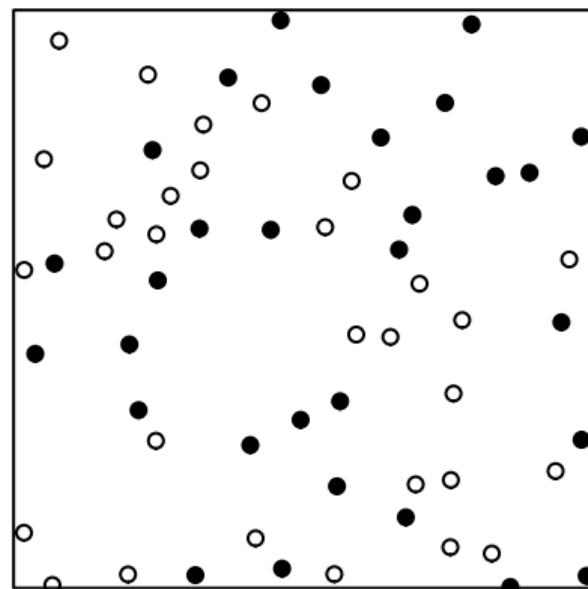
Lattice-free spatially regulated sampling designs

Sample at random subject to a minimum distance constraint



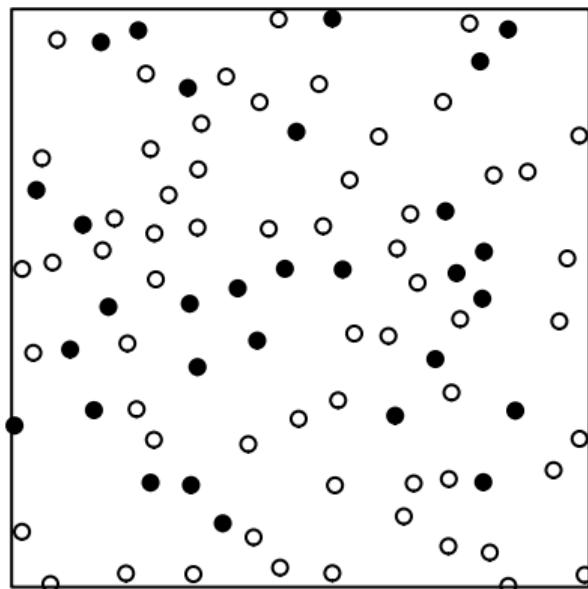
Lattice-free spatially regulated sampling designs

Sample at random subject to a minimum distance constraint



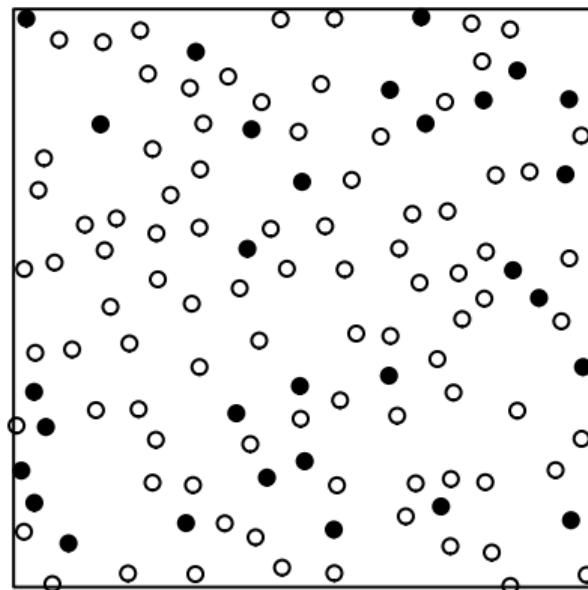
Lattice-free spatially regulated sampling designs

Sample at random subject to a minimum distance constraint



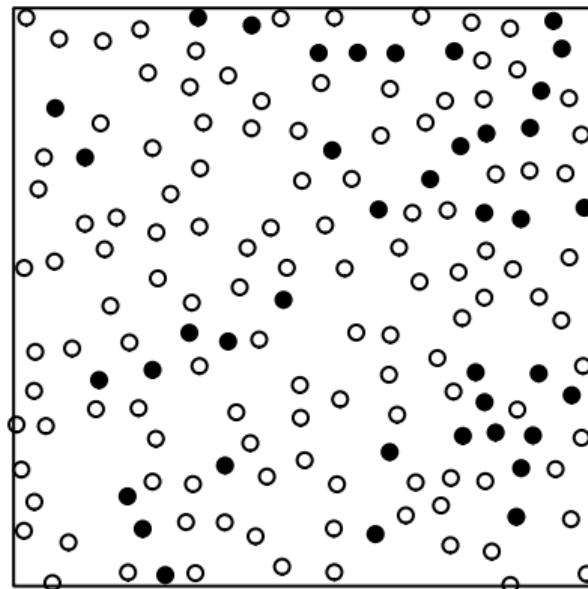
Lattice-free spatially regulated sampling designs

Sample at random subject to a minimum distance constraint



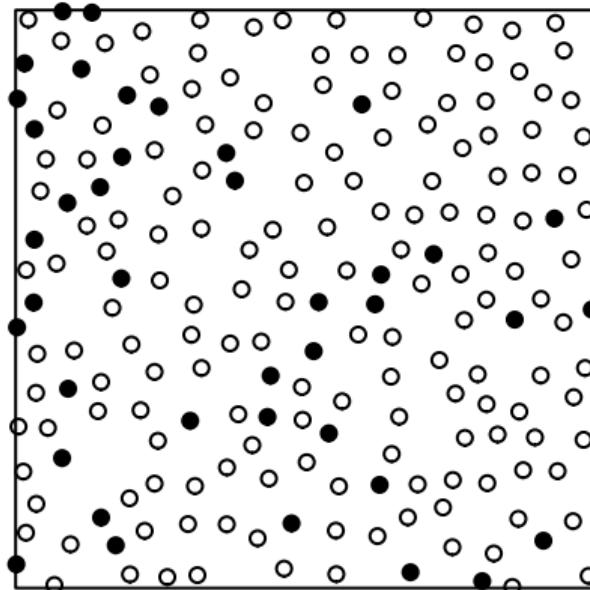
Lattice-free spatially regulated sampling designs

Sample at random subject to a minimum distance constraint

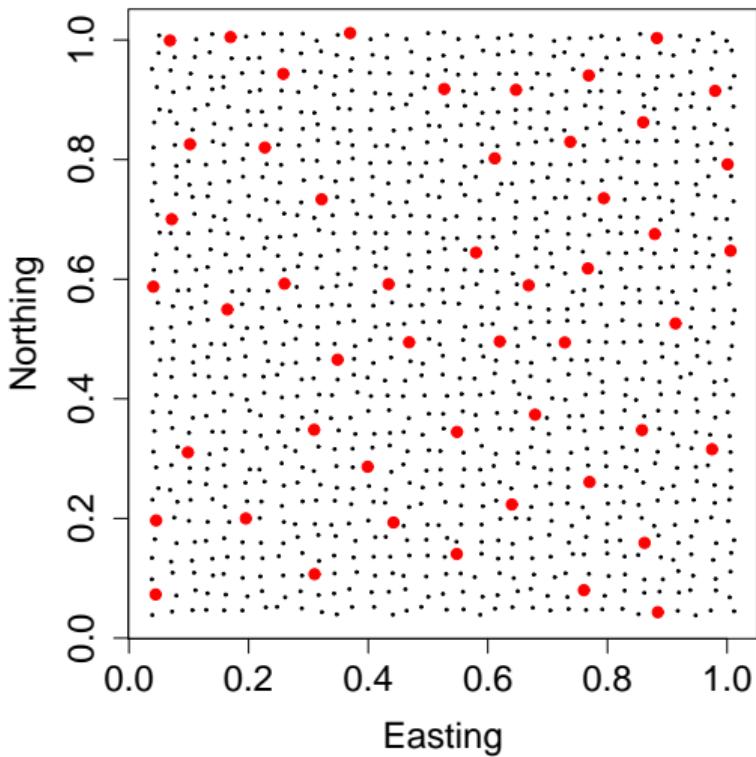
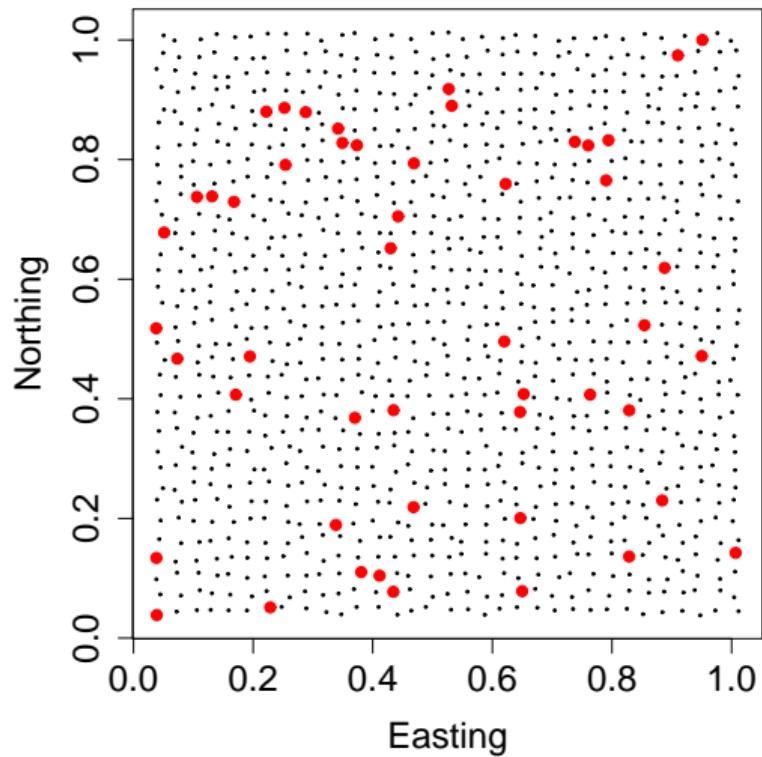


Lattice-free spatially regulated sampling designs

Sample at random subject to a minimum distance constraint



Spatially regulated sampling from a pre-specified set of locations



Analysing prevalence data 1: logistic regression

Data: $(x_i, d(x_i), n_i, Y_i) : i = 1, \dots, n$

- x : location
- $d(x)$: covariates
- n : number tested
- Y : number positive

Model:

- $\log[p(x_i)/\{1 - p(x_i)\}] = d(x_i)' \beta$
- $Y_i \sim \text{Binomial}\{n_i, p(x_i)\}$

Analysing prevalence data 2: extra-binomial variation

Data: $(x_i, d(x_i), n_i, Y_i) : i = 1, \dots, n$

Model:

- $\log[p(x_i)/\{1 - p(x)_i\}] = d(x)'_i \beta + U_i$
- $U_i \sim N(0, \nu^2)$, mutually independent
- $Y_i|U_i \sim \text{Binomial}\{n_i, p(x_i)\}$

Analysing prevalence data 3: geostatistics

Data: $(x_i, d(x_i), n_i, Y_i) : i = 1, \dots, n$

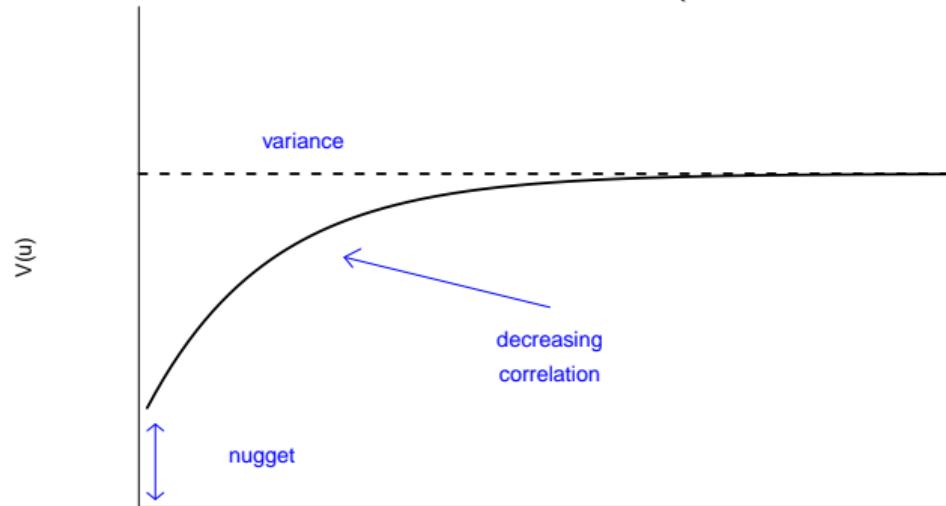
Model:

- $\log[p(x_i)/\{1 - p(x)_i\}] = d(x)'_i \beta + U_i + S(x_i)$
- $U_i \sim N(0, \nu^2)$, mutually independent
- $S(x) \sim$ Gaussian process, spatially correlated
- $Y_i | U_i, S(x_i) \sim \text{Binomial}\{n_i, p(x_i)\}$

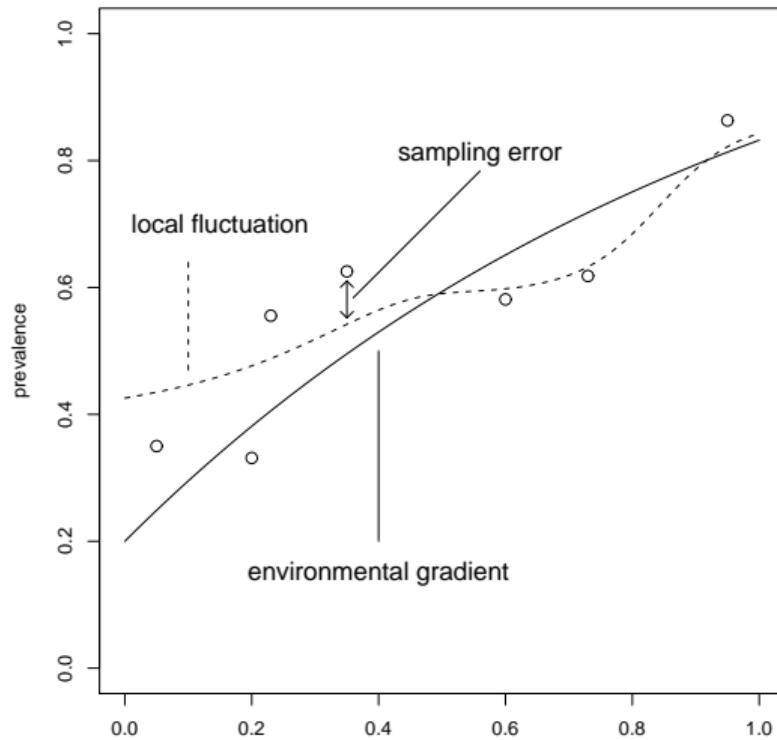
Diggle and Giorgi, 2019

Spatial covariance structure: the variogram

$$\begin{aligned} V(u) &= \frac{1}{2} \text{Var}\{Y(x) - Y(x-u)\} \\ &= \text{nugget} + \text{variance} \times (1 - \text{correlation}) \end{aligned}$$



Schematic representation of a geostatistical model



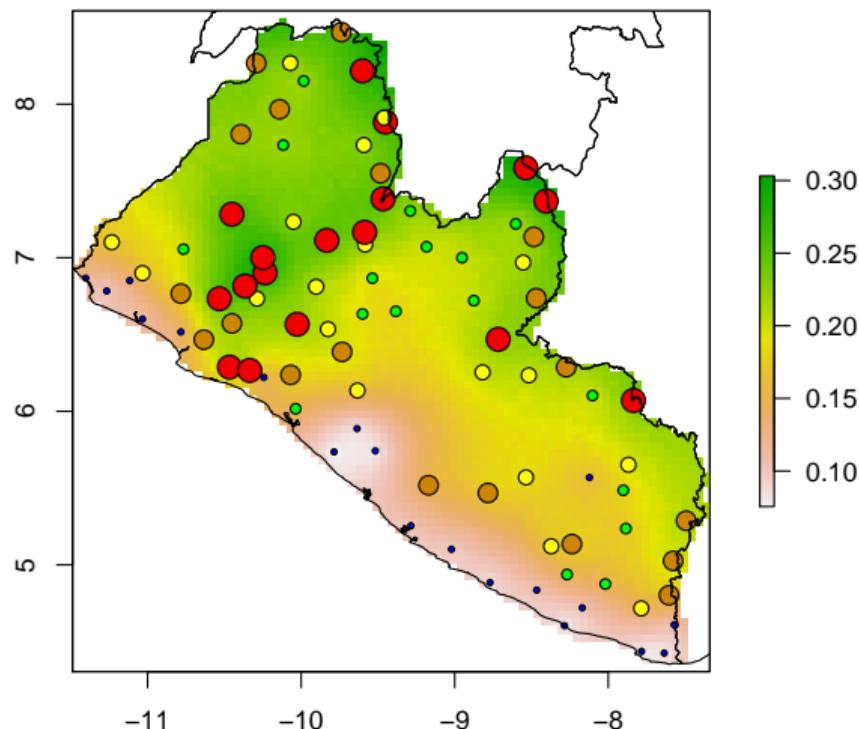
Model-based Geostatistics

- the application of general principles of statistical modelling and inference to geostatistical problems
- paradigm:
 - formulate a model for the data
 - use likelihood-based methods of inference
 - answer the scientific question

Example. What can we say about the variation in onchocerciasis prevalence, $P(x)$, throughout Liberia?

Diggle, Moyeed and Tawn, 1998.

A predicted prevalence surface: onchocerciasis in Liberia



Geostatistical prediction

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

Peter McCullagh

S = state of nature

Y = all relevant data

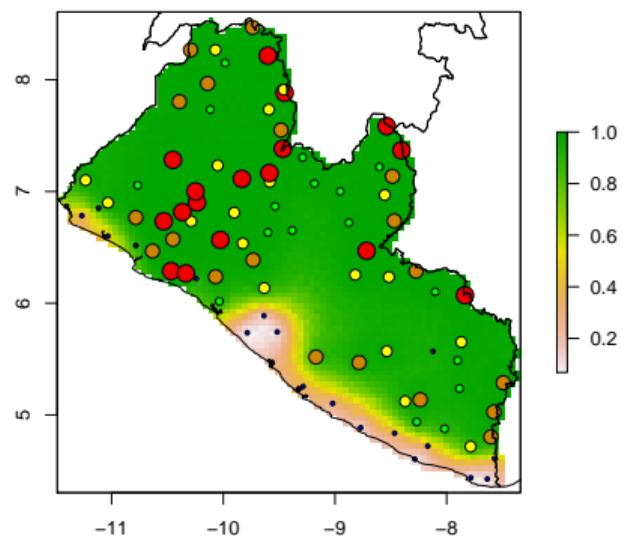
T = $\mathcal{F}(S)$ = target for prediction

Model: $[S, Y] = [S][Y|S]$

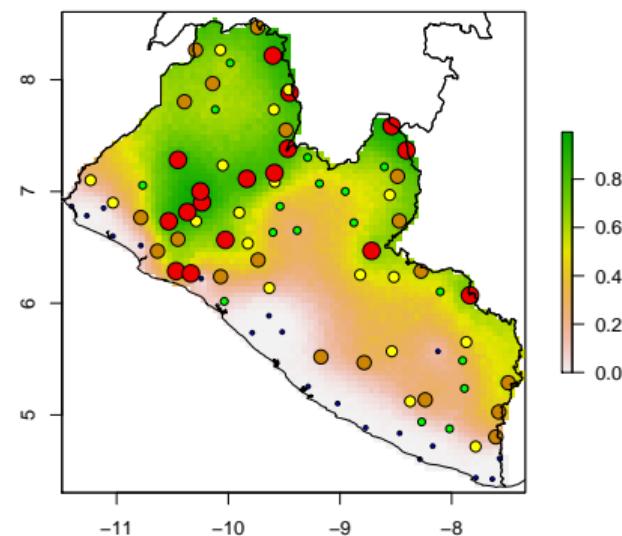
Prediction: $[S, Y] \Rightarrow [S|Y] \Rightarrow [T|Y]$

Liberia: onchocerciasis exceedance maps

$P(\text{prevalence} > 10\%)$



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



Sample size?

Answer often has two dimensions:

- n = how many locations;
- m = how many samples per location
- Typically, sampling cost is proportional to

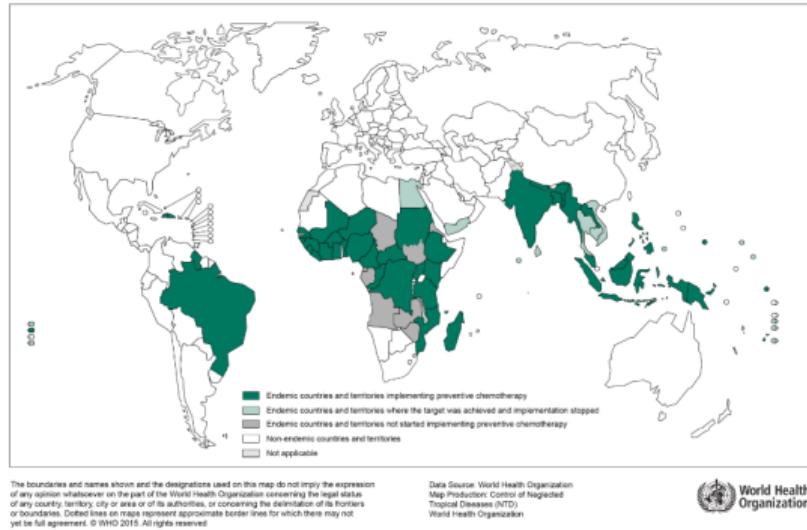
$$n + c * n * m$$

- Designing to a cost-constraint or to a performance requirement?

Case-study: Lymphatic Filariasis in Ghana



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



Elimination as a public health problem if antigen prevalence (current infection) is less than 2%

Current WHO guideline

- Consider each district as an evaluation unit (EU)
- Use tables provided by WHO to calculate for each EU
 - number of villages to sample
 - number of children to test per village
 - critical number of positive test results
- Classify each EU as **elimination indicated** or **not indicated** according to whether total observed number of positives does or does not exceed the tabulated critical number
- No account taken of **where** positives are located

Elimination target for each EU

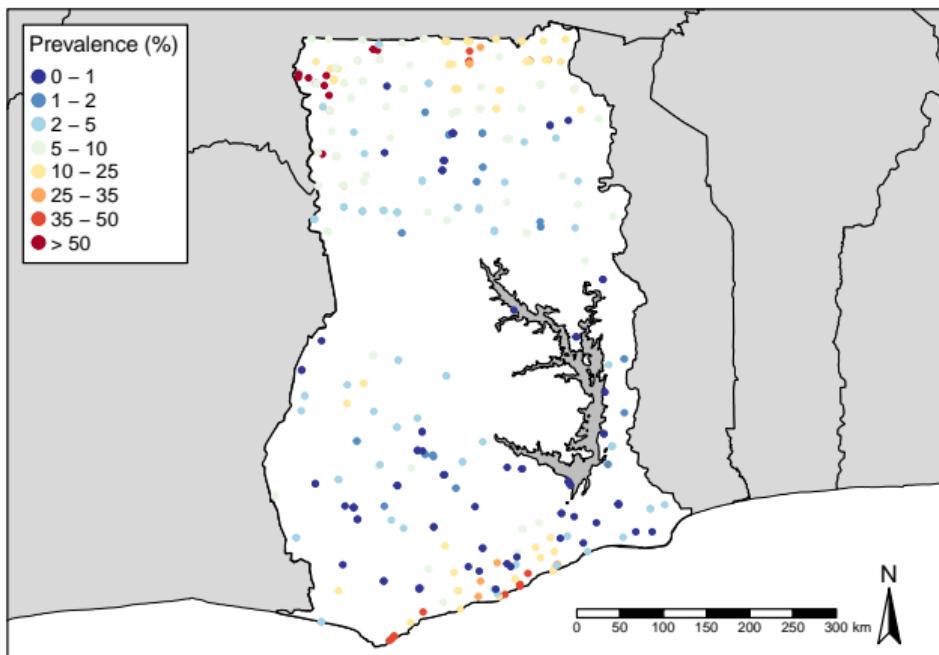
Communities $j = 1, \dots, N$ of size n_j at locations x_j , prevalence surface $P(x)$

$$T = \frac{\sum n_j P(x_j)}{\sum n_j} \quad \text{Elimination} \Leftrightarrow T < 0.02$$

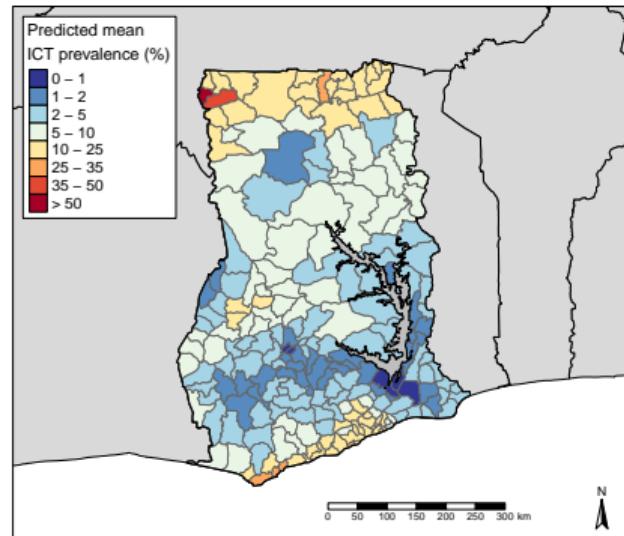
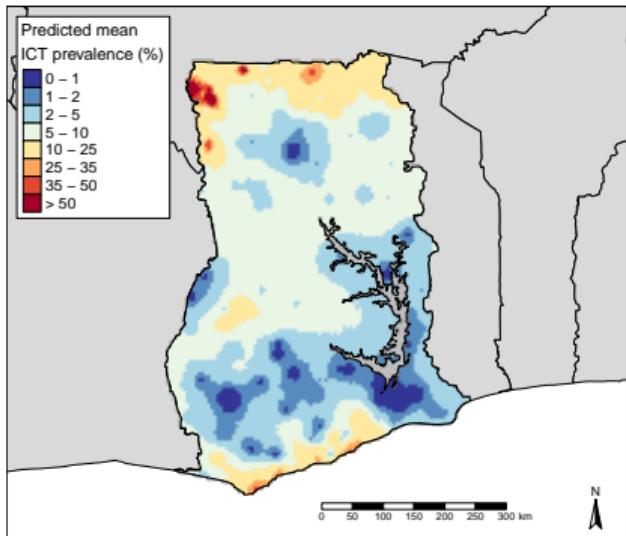
Probabilistic prediction

- Draw samples from predictive distribution of T
- Choose probability threshold q
- Elimination indicated $\Leftrightarrow \text{Prob}(T < 0.02 | \text{data}) > q$

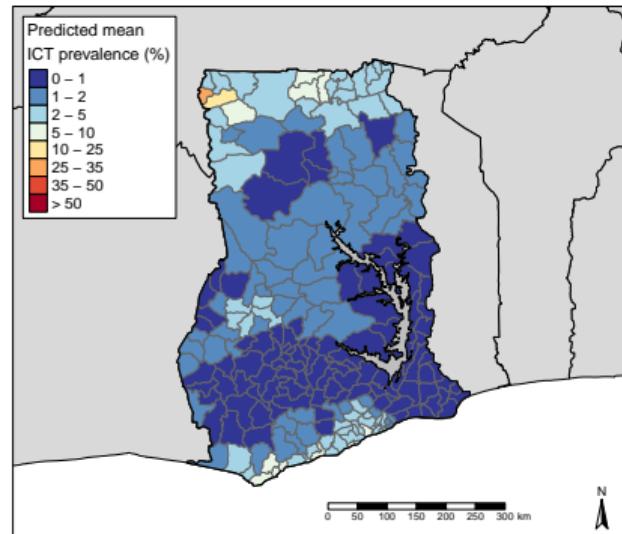
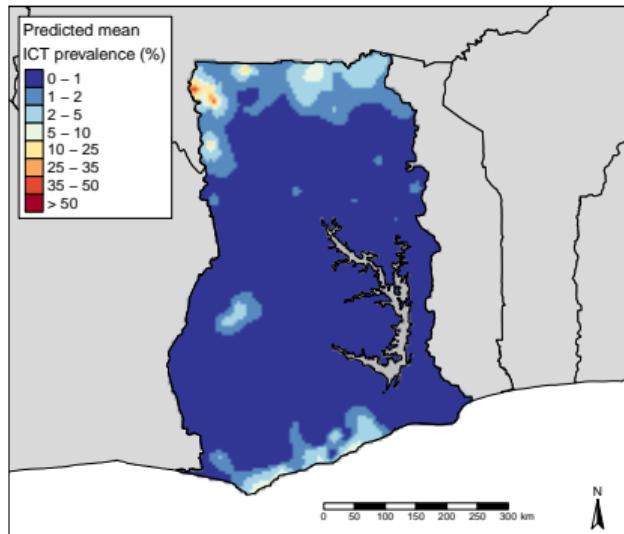
Ghana: baseline LF prevalence at sampled locations



Baseline prevalence maps



Projected prevalence maps



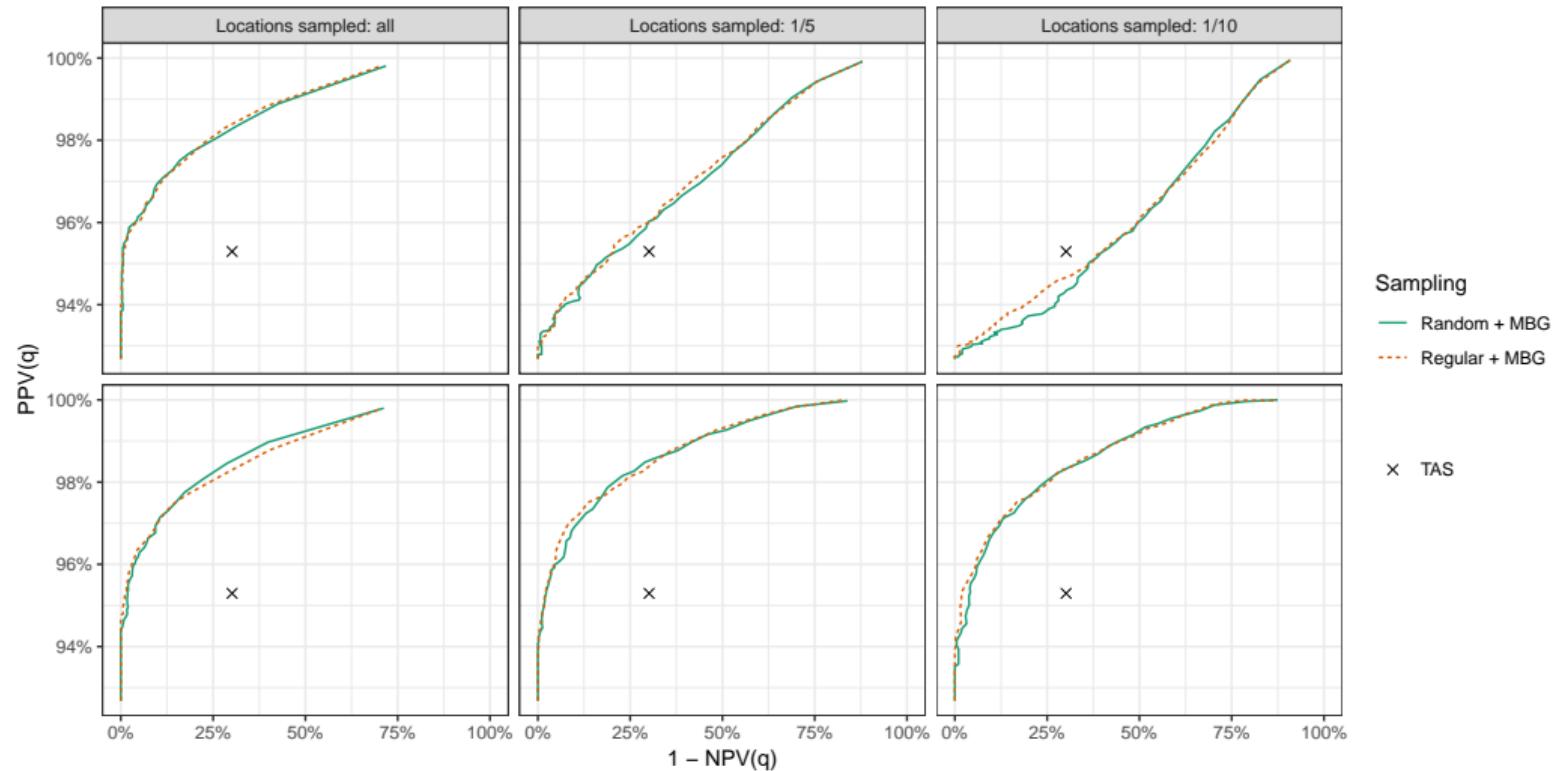
Evaluation of predictive performance

- simulate test results over region of interest
- for each EU
 - apply WHO guideline to simulated data
 - apply model-based geostatistics to simulated data
 - compare actual and indicated elimination status
- construct tables of true/false positive/negative indications
- calculate positive and negative predictive values:

$PPV = \text{Prob}(\text{elimination achieved} \mid \text{elimination indicated})$

$NPV = \text{Prob}(\text{elimination not achieved} \mid \text{elimination not indicated})$

Results: positive and negative predictive values



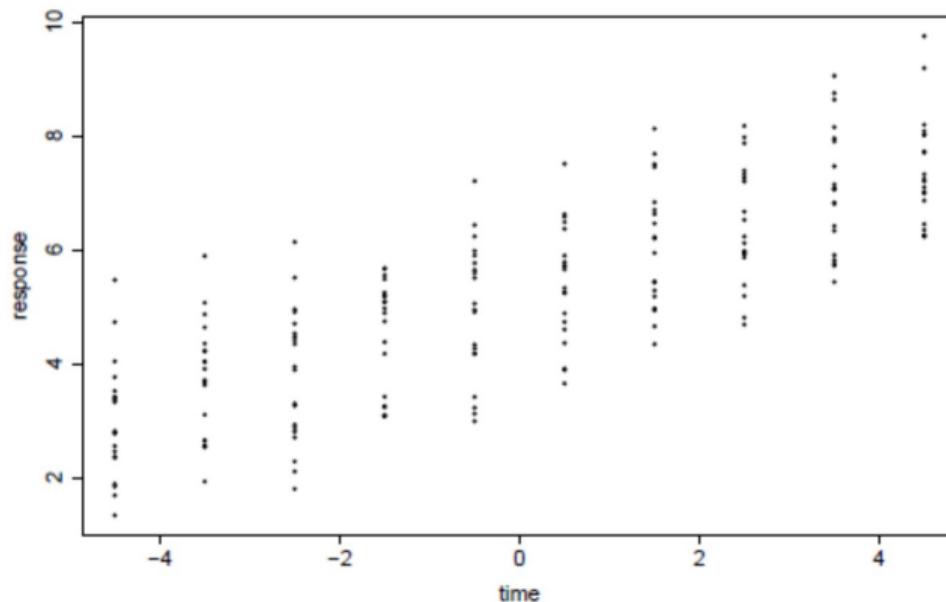
From spatial to spatio-temporal

Problem statement: how best to sample a target population at a sequence of times $t = 1, 2, \dots, m$

- longitudinal sampling
 - $t = 1$: choose a set of sampling locations
 - $t = 2, 3, \dots, m$: continue to sample at the same set of locations
 - **Example:** when primary objective is to detect changes over time
- repeated cross-sectional sampling
 - choose a new set of sampling locations on each of $t = 1, 2, \dots, m$
 - **Example:** when primary objective is to accumulate information on as much of the target population as possible over time
- adaptive sampling
 - $t = 1$: choose an initial set of sampling locations
 - $t = 2, 3, \dots, m$: choose a new set of sampling locations informed by analysis of data collected at times $s < t$
 - **Example:** when primary objective is to map achievement of a policy requirement

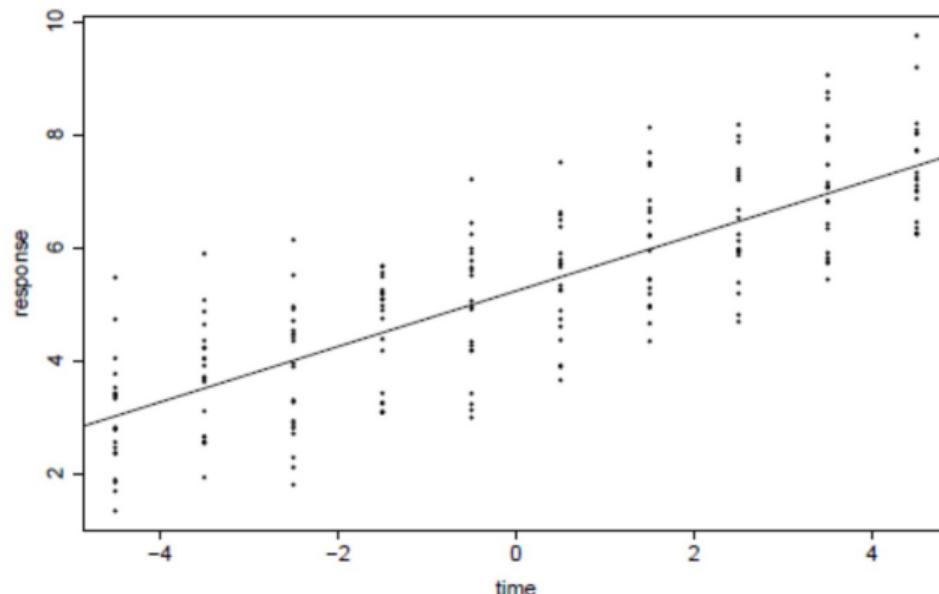
Longitudinal vs repeated cross-sectional: a non-spatial example

$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \quad i = 1, \dots, m \quad t = 1, \dots, n$$



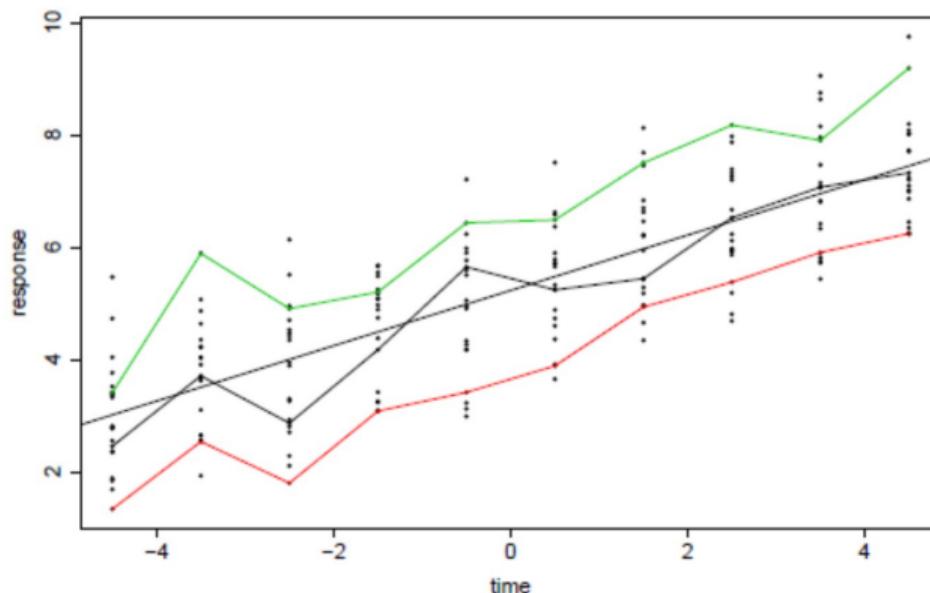
Longitudinal vs repeated cross-sectional: a non-spatial example

$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \quad i = 1, \dots, m \quad t = 1, \dots, n$$



Longitudinal vs repeated cross-sectional: a non-spatial example

$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \quad i = 1, \dots, m \quad t = 1, \dots, n$$



Longitudinal vs repeated cross-sectional: a non-spatial example

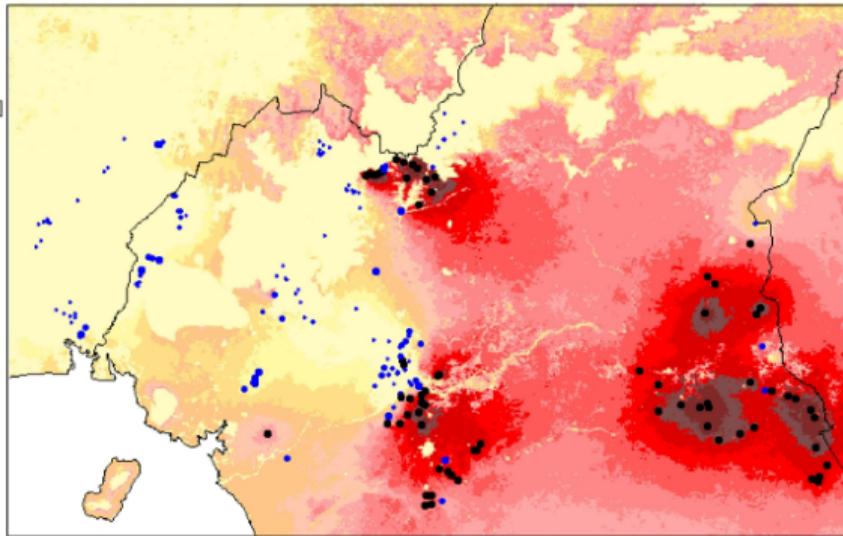
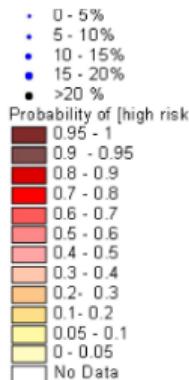
$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \quad i = 1, \dots, m \quad t = 1, \dots, n$$

Parameter estimates and standard errors:

	ignoring correlation		recognising correlation	
	estimate	standard error	estimate	standard error
α	5.234	0.074	5.234	0.202
β	0.493	0.026	0.493	0.011

Adaptive sampling

- The WHO considers areas **safe** for protective mass drug administration where the prevalence of Loa loa (eyeworm) is less than 20%
- the map shows **prevalence data**, as blue and black dots, and the **probability that prevalence is greater than 20% given the data**, as a coloured image.



Where would you sample next?