Spatio-temporal Design and Analysis

Peter J Diggle

Lancaster University

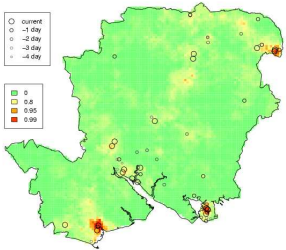
December 2022

A short introduction to a big topic

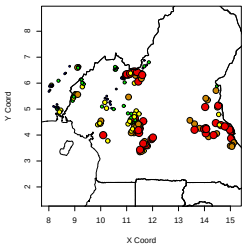
1. Spatial data-formats
2. Statistics and scientific method
3. Geospatial survey design
4. Geospatial survey analysis
5. From spatial to spatio-temporal
   - 5.1 longitudinal designs
   - 5.2 repeated cross-sectional designs
   - 5.3 adaptive designs
6. Reading list

Geospatial data-formats

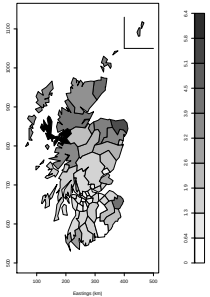
Point process



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

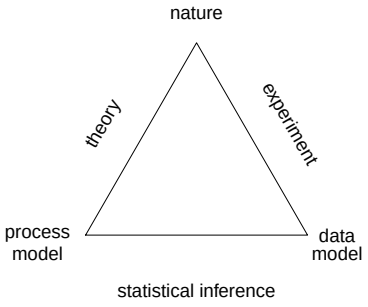
"Too many notes, Mozart"

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

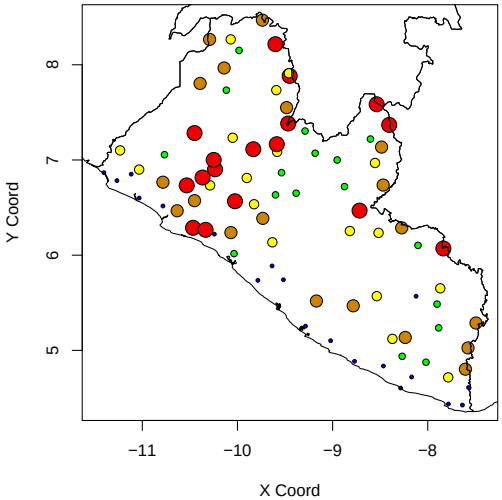
Estimation: what can you say about the parameters that define $[S]$ and $[Y|S]$?

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

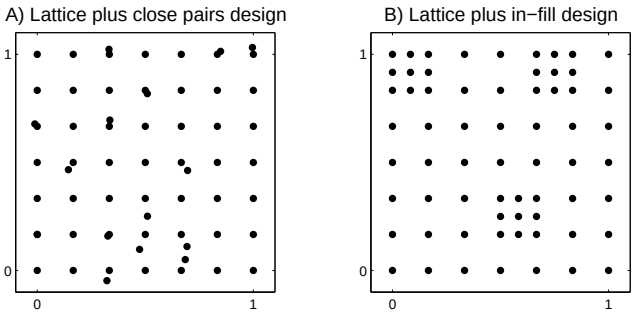
Performance requirement for a survey

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

Kenya EUs

Where to sample?

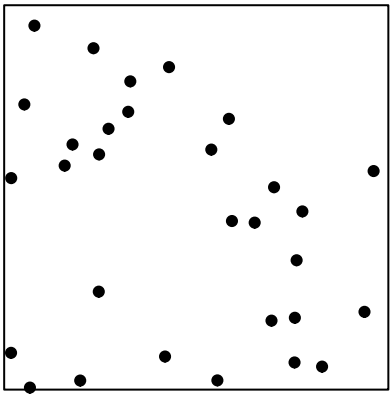
Examples of lattice-based designs



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

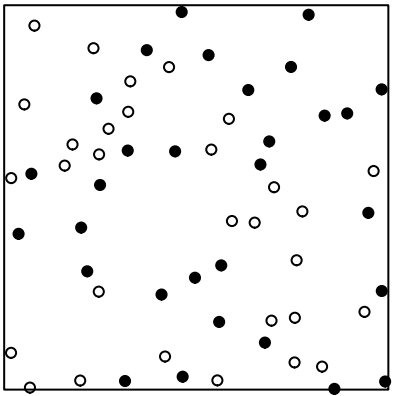
Lattice-free spatially regulated sampling designs

Sample at random subject to a minimum distance constraint



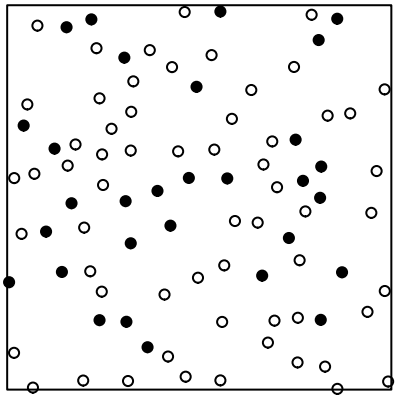
Lattice-free spatially regulated sampling designs

Sample at random subject to a minimum distance constraint



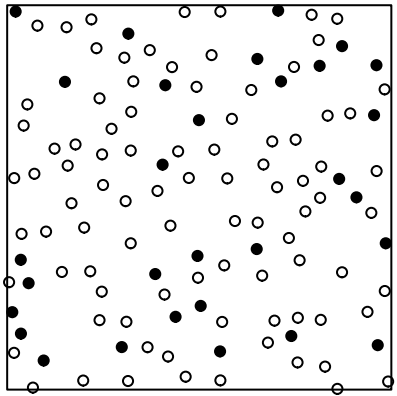
Lattice-free spatially regulated sampling designs

Sample at random subject to a minimum distance constraint



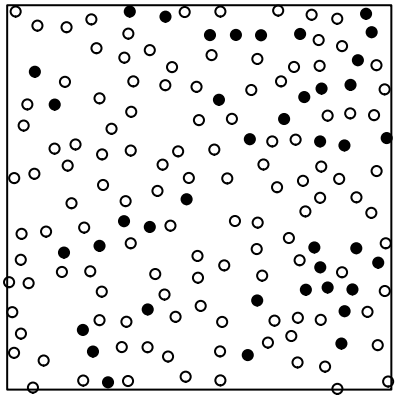
Lattice-free spatially regulated sampling designs

Sample at random subject to a minimum distance constraint



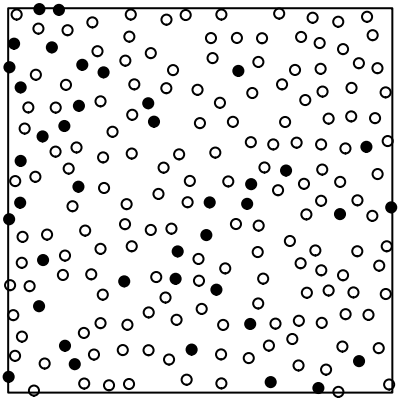
Lattice-free spatially regulated sampling designs

Sample at random subject to a minimum distance constraint

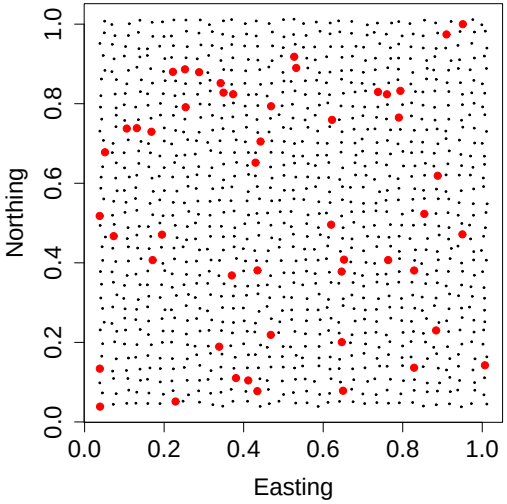


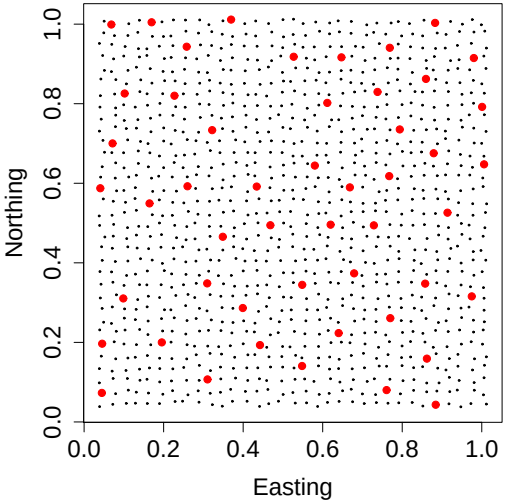
Lattice-free spatially regulated sampling designs

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: $(x_i, d(x_i), n_i, Y_i) : i = 1, ..., n$

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: $(x_i, d(x_i), n_i, Y_i) : i = 1, ..., n$

Model:

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

Analysing prevalence data 3: geostatistics

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

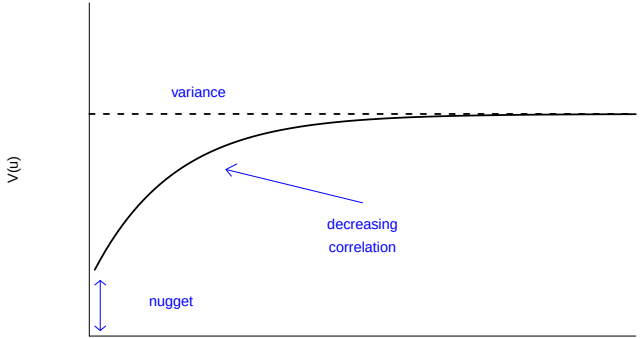
Model:

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

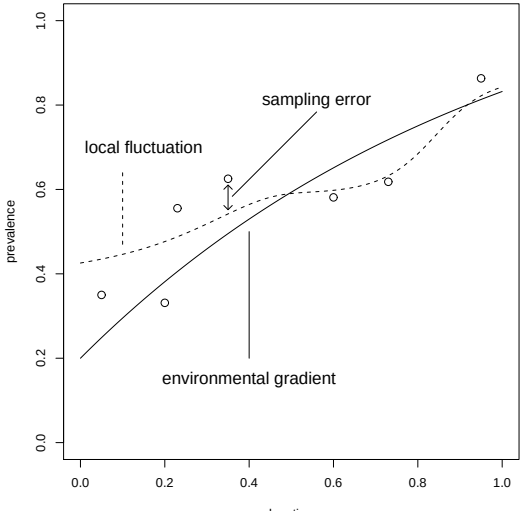
Diggle and Giorgi, 2019

Spatial covariance structure: the variogram

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



Schematic representation of a geostatistical model



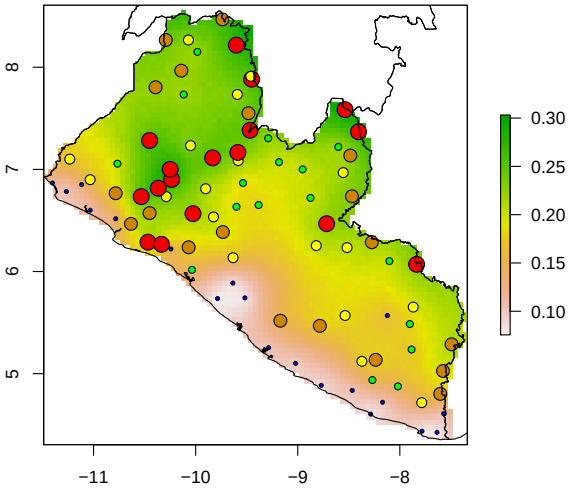
Model-based Geostatistics

- 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"

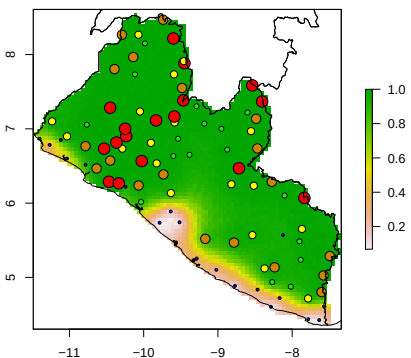
Peter McCullagh

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

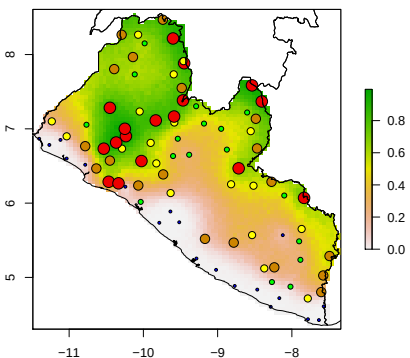
Model: $[S, Y] = [S][Y|S]$
Prediction: $[S, Y] \Rightarrow [S|Y] \Rightarrow [T|Y]$

Liberia: onchocerciasis exceedance maps

P(prevalence>10%)



P(prevalence>20%)



Sample size?

Answer often has two dimensions:

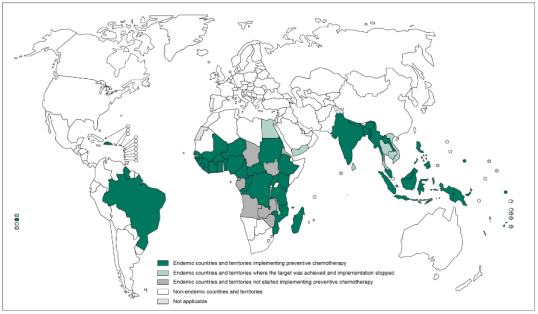
- $n =$ how many locations;
- $m =$ how many samples per location
- Typically, sampling cost is proportional to

$$n + c * n * m$$

- 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



The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps represent approximate border lines for which there may not yet be full agreement. © WHO 2015. All rights reserved

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

World Health Organization

Elimination as a public health problem if antigene 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

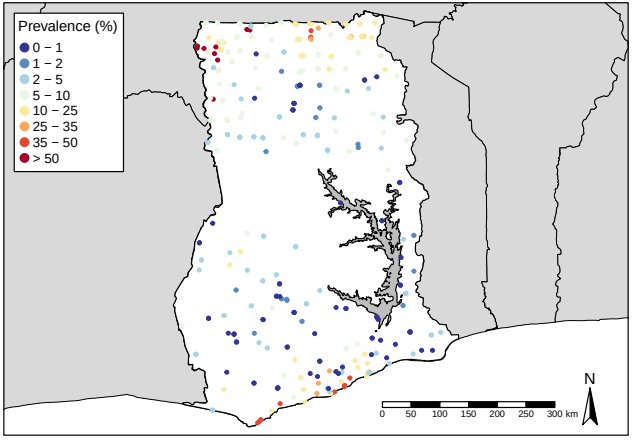
Communities $j = 1, ..., N$ of size n_j at locations x_j, prevalence surface $P(x)$

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

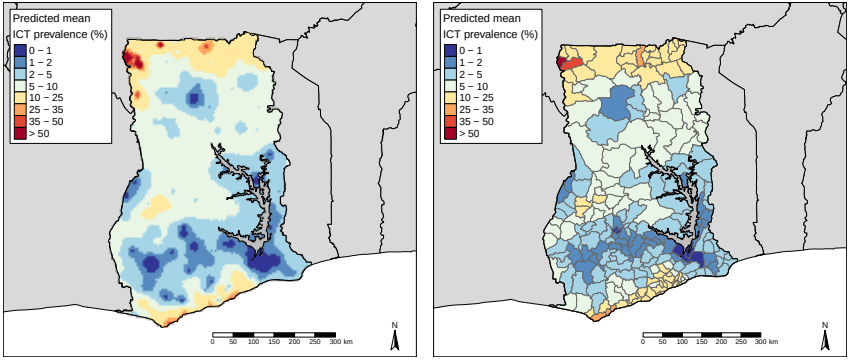
Probabilistic prediction

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

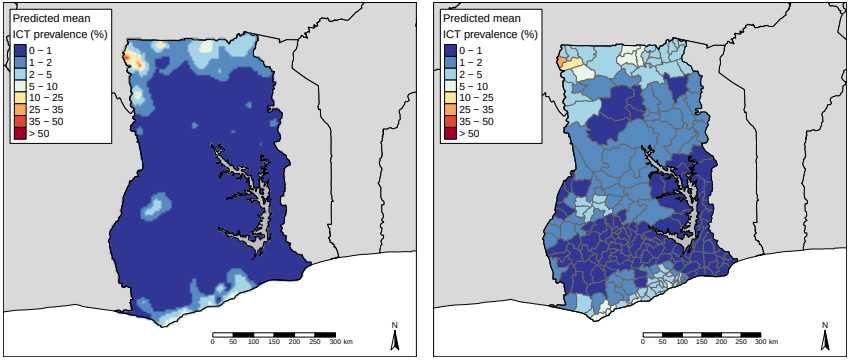
Ghana: baseline LF prevalence at sampled locations



Baseline prevalence maps



Projected prevalence maps

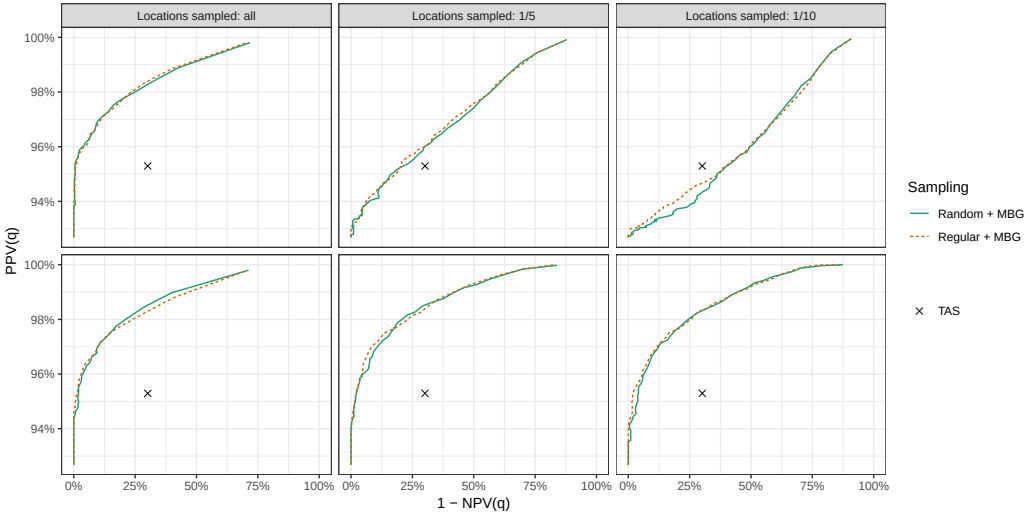


Evaluation of predictive performance

- 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



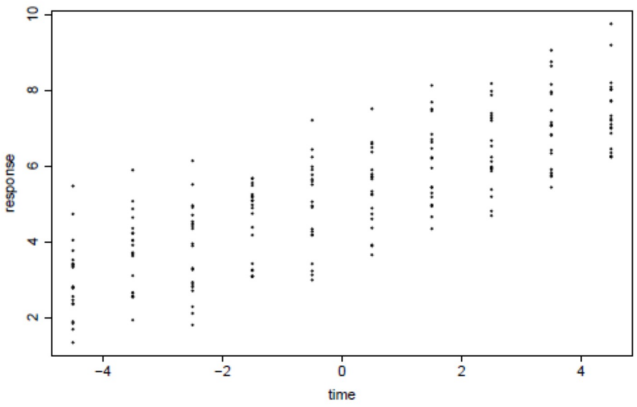
From spatial to spatio-temporal

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

- longitudinal sampling
  - $t = 1$: choose a set of sampling locations
  - $t = 2, 3, ..., m$: 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 $t = 1, 2, ..., m$
  - 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
  - $t = 2, 3, ..., m$: 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

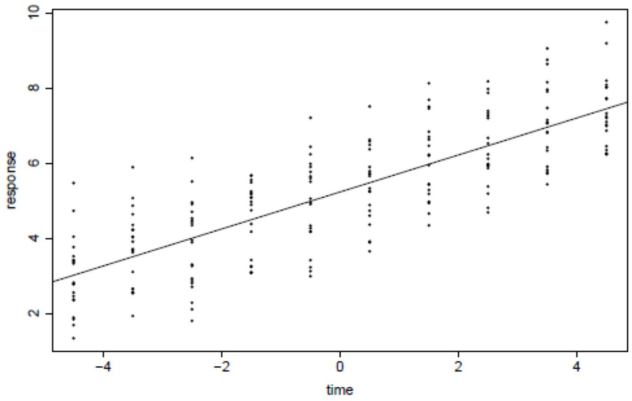
Longitudinal vs repeated cross-sectional: a non-spatial example

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



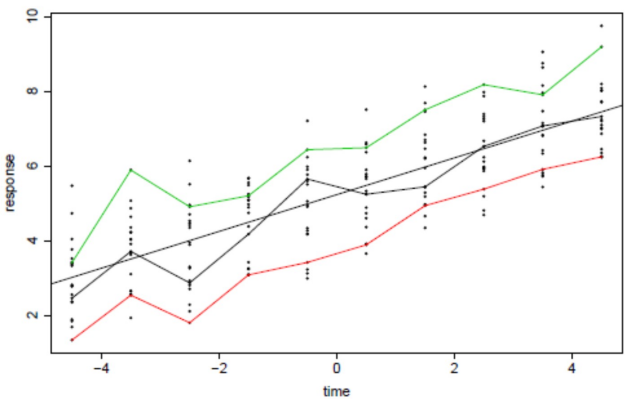
Longitudinal vs repeated cross-sectional: a non-spatial example

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



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

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



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

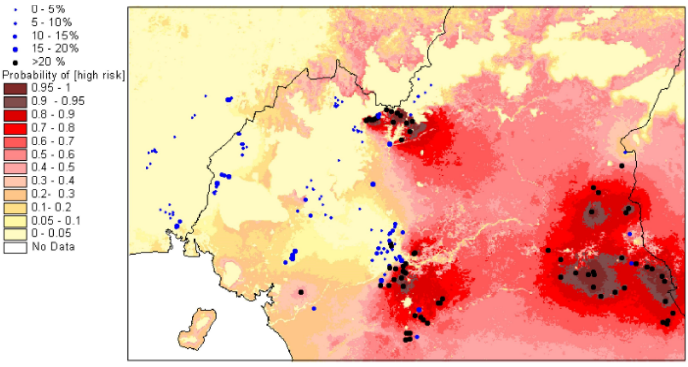
$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \qquad i = 1, ..., m \quad t = 1, ..., n$$

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?

Reading list

There are many books devoted to spatial and/or spatio-temporal statistical methods. All of the following are, in my opinion, very good books on their own terms, but differ in what they cover, and in what style. I have added a short comment on each.

Baddeley, A., Rubak, E. and Turner, R. (2016). *Spatial Point Patterns: methodology and Applications with R* . Boca Raton: CRC Press
Very detailed description of methods for analysing point process data, closely linked to the authors' `spatstat` package

Banerjee, S., Carlin, B.P and Gelfand, A.E. (2015). *Hierarchical Modeling and Analysis for Spatial Data (second edition)*. Boca Raton: CRC Press.
Particularly good on analysis of spatially discrete (registry) data, all from a Bayesian inferential viewpoint.

Cressie, N.A.C. (1991). *Statistics for Spatial Data*. New York: Wiley.
Very wide-ranging and detailed coverage of the while field - one for dipping into as a reference book rather than reading cover-to-cover.

Cressie, N. and Wikle, C.K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken: Wiley.
A follow-on to Cressie (1991)

Diggle, P.J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns (3rd edition)* Boca Raton: CRC Press
Less technical detail than Baddeley, Rubak and Turner (2016), but more discusson of different mehods' strengths and weaknesses, very loosely linked to the *splancs* package

Diggle, P.J. and Giorgi, E. (2019). *Model-based Geostatistics: Methods and Applications in Global Public Health.* Boca Raton: CRC Press
Aimed at applied statisticians or population health scientists with an understanding of statistical methods, linked to the `PrevMap` package.

Diggle, P.J. and Ribeiro, P.J. (2007). *Model-based Geostatistics*. New York: Springer.
Aimed at a statistically trained audience, with closely linked to the `geoR` package.

Moraga, P. (2019). *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. Boca Raton: Chapman Hall/CRC
R-InLA is a popular computing environment for anysing data using (not necessarily spatial) hierarchically structured models

Sahu, S.K. (2022). *Bayesian Modeling of Spatio-Temporal Data with R.* Boca Raton: CRC Press LLC.

Particularly strong on spatially discrete (Registry) data, linked to the author's `spTimer` package

Shaddick, G. and Zidek, J.V. (2016). *Spatio-Temporal Methods in Environmental Epidemiology*. Boca Raton: CRC Press.
An alternative to Cressie and Wikle (2011), with an explicit focus on data from the physical environmental sciences

Waller, L. and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*. New York: Wiley.
A very good, wide-ranging introduction aimed at a population health sciences audience.