Model-based Geostatistics for Better Global Health

Peter J Diggle

Lancaster University and Health Data Research UK

IBC, July 2020

Funding: NTD Modelling Consortium, Bill and Melinda Gates Foundation

References

Krige, D.G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, **52**, 119–39.

Matérn, B. (1960). *Spatial variation*. Meddelanden fran Statens Skogsforsknings institut, Stockholm. Band 49, number 5 (Reprinted as Springer Lecture Note, 1978)

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, **58**, 1246–66.

Watson, G.S. (1972). Trend surface analysis and spatial correlation. *Geological Society of America Special Paper*, **146**, 39–46.

Ripley, B.D. (1981). *Spatial Statistics*. New York : Wiley.

Cressie, N.A.C. (1991). *Statistics for Spatial Data*. New York: Wiley.

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

Diggle, P.J. and Ribeiro, P.J. (2007). *Model-based Geostatistics*. New York: Springer.

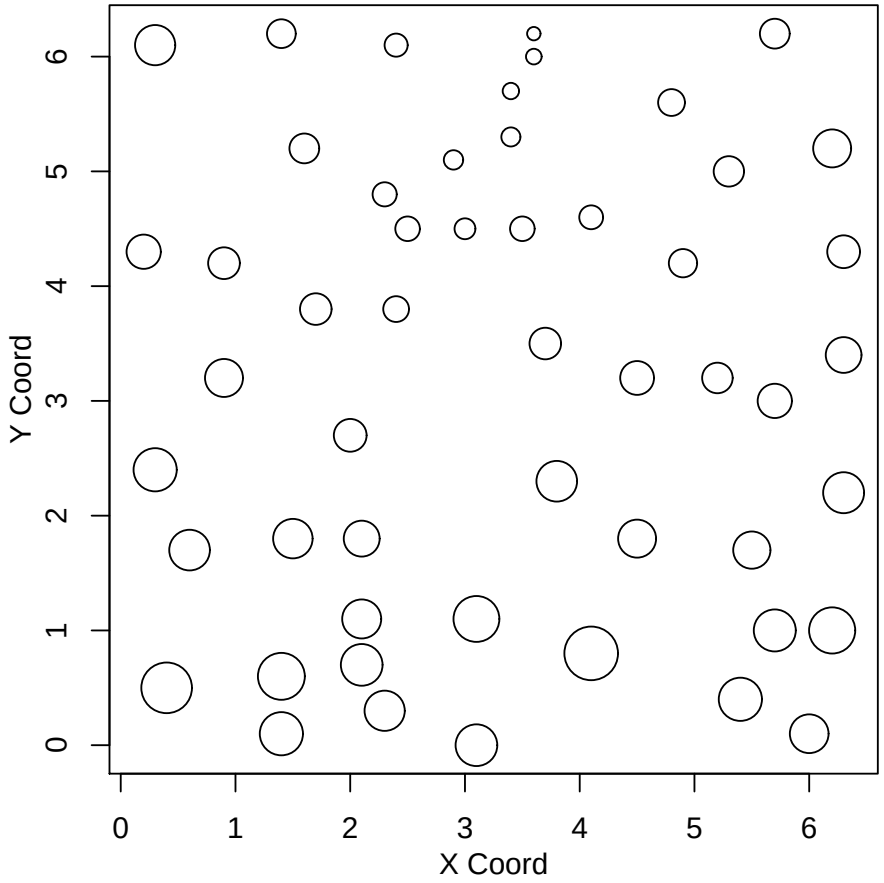
Diggle, P.J., Menezes, R. and Su, T.-L. (2010). Geostatistical analysis under preferential sampling (with Discussion). *Applied Statistics*, **59**, 191–232.

Diggle, P.J. and Giorgi, E. (2019). *Model-based Geostatistics: Methods and Applications in Global Public Health*. Boca Raton: CRC Press

Fronterre, C., Amoah, B., Giorgi, E., Stanton, M.C. and Diggle, P.J. (2020). Design and analysis of elimination surveys for neglected tropical diseases. *Journal of Infectious Diseases*, DOI: 10.1093/infdis/jiz554

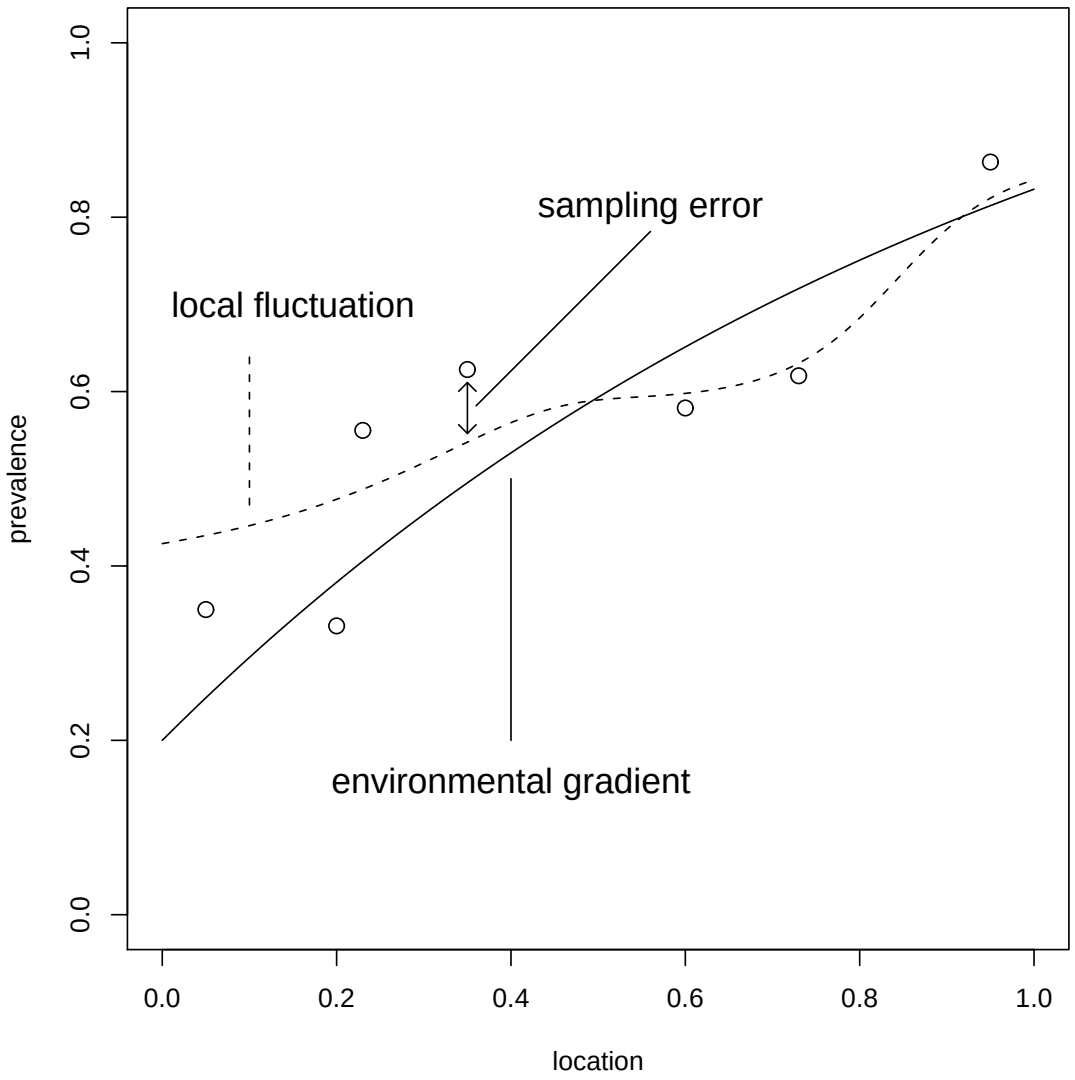
`www.lancaster.ac.uk/staff/diggle/IBC2020slides.pdf`

A simple geostatistical problem



- **Measurements** Y_i **at locations** x_i **in a spatial region** A
- **Unobserved spatially continuous phenomenon** $S(x)$
- **What can we say about the realisation of** $S(x)$ **throughout** A**?**

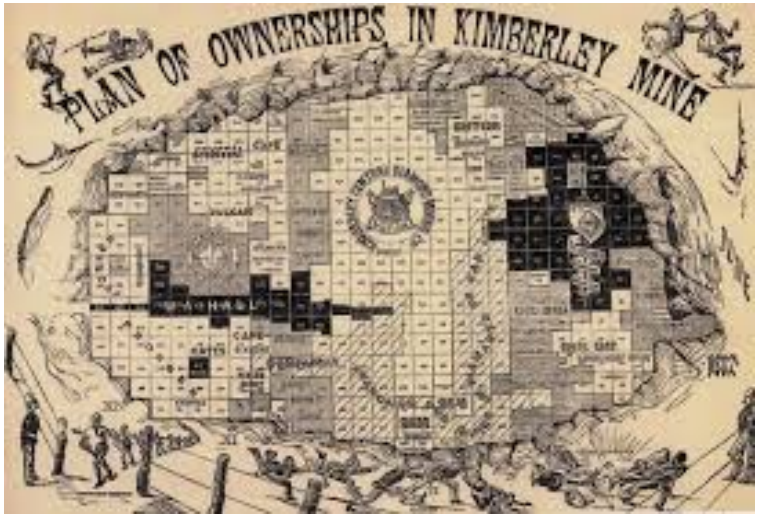
Schematic


sampling error
local fluctuation
environmental gradient
prevalence
location
0.0
0.2
0.4
0.6
0.8
1.0

Geostatistical problems more generally

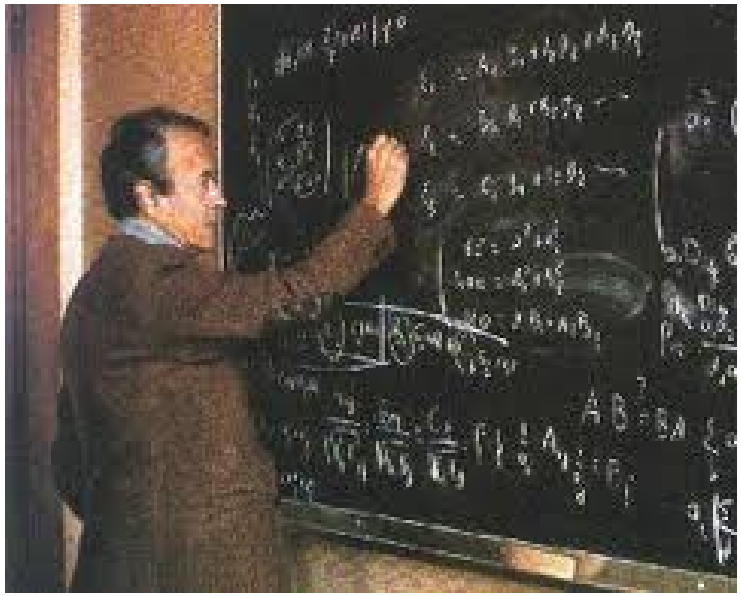
- **Design:** how to choose locations x_i at which to collect measurements
- **Estimation:** how to investigate relationships with covariates when measurements are spatially correlated
- **Prediction:** how to map (expected value of) outcome throughout the study-region

Origins of geostatistics ... mining in South Africa



Krige (1951)

From South Africa to Fontainebleau... classical geostatistics

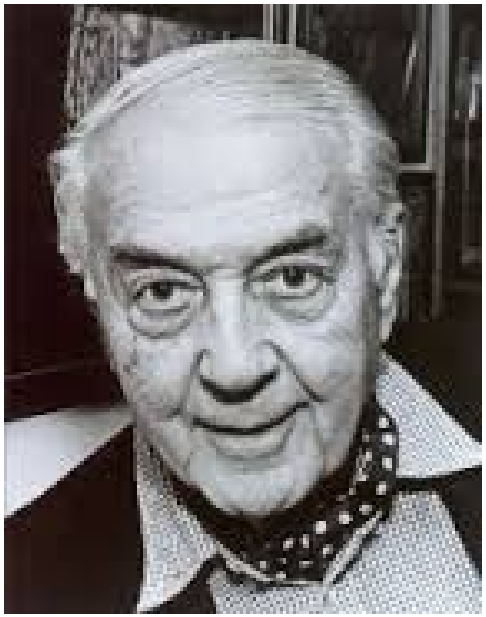


Matheron (1963)

Meanwhile in Sweden...the Royal College of Forestry, Stockholm

Matérn (1960)

Into the statistical mainstream

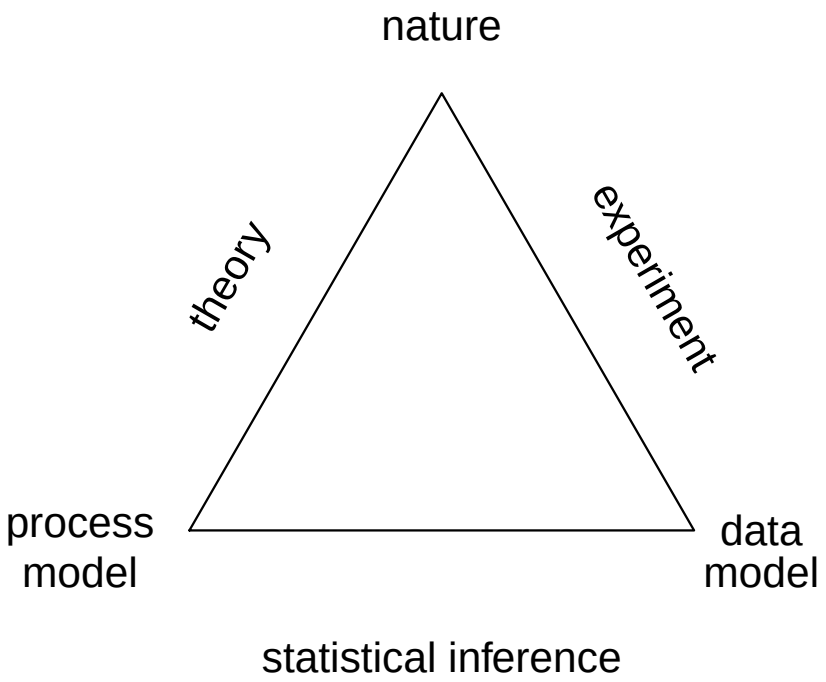


Watson (1972)

Ripley (1981)

Cressie (1991)

Model-based geostatistics: the application of general principles of statistical modelling and inference to geostatistical problems (Diggle, Moyeed and Tawn, 1998)



S: **an unobserved process**
Y : **data relevant to S**

Hierarchical formulation

$$[S, Y] = [S][Y|S] \Rightarrow [T|Y]$$

Parameter estimation: **Monte Carlo maximum likelihood**

Prediction

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

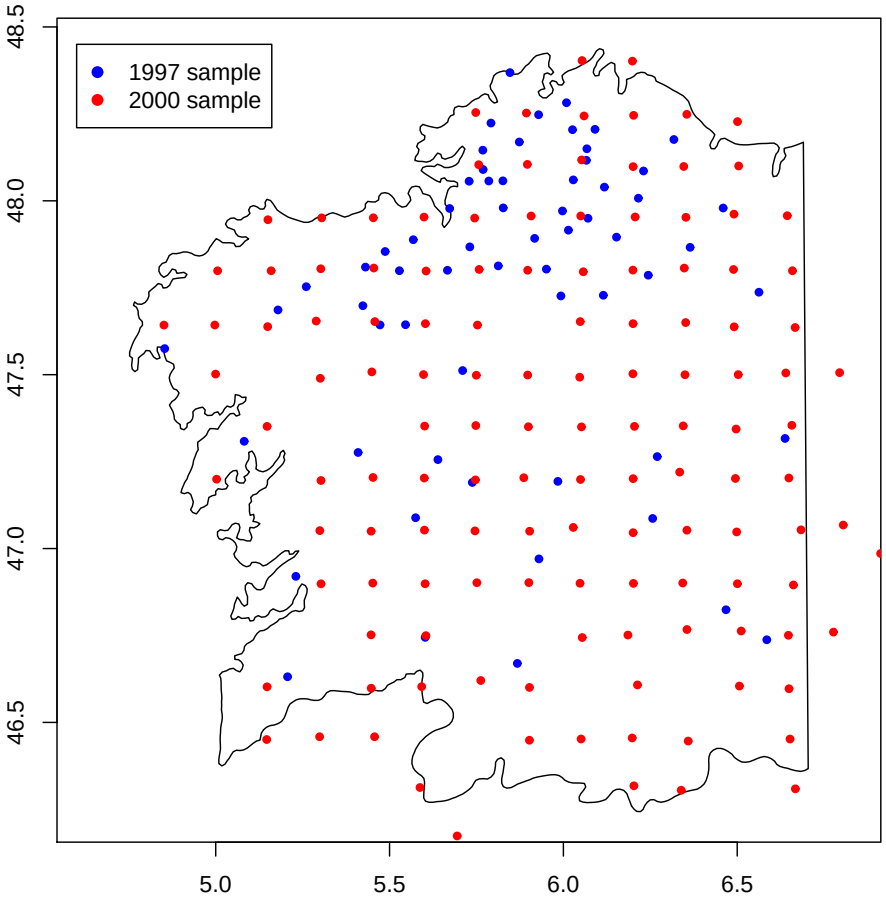
Peter McCullagh

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

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

Parameter uncertainty? $[T|Y] = \int [T|Y;\hat{\theta}][\hat{\theta}|Y]d\hat{\theta}$

Environmental monitoring in Galicia, northern Spain



- 1997 sampling concentrated towards northern Galicia
- potential for selection bias?

Preferential sampling

locations X **signal S** **measurements Y**

- **Conventional model:**

$$[X, S, Y] = [S][X][Y|S] \quad (1)$$

- **Preferential sampling model:**

$$[X, S, Y] = [S][X|S][Y|S, X] \quad (2)$$

Key point for inference: **even if $[Y|S, X]$ in (2) and $[Y|S]$ in (1) are algebraically the same, the term $[X|S]$ in (2) cannot be ignored for inference about $[S, Y]$, because of the shared dependence on the unobserved process S**

Preferential sampling model

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

- $[S]$ = stationary Gaussian process
- $[X|S]$ = inhomogenous Poisson process with intensity

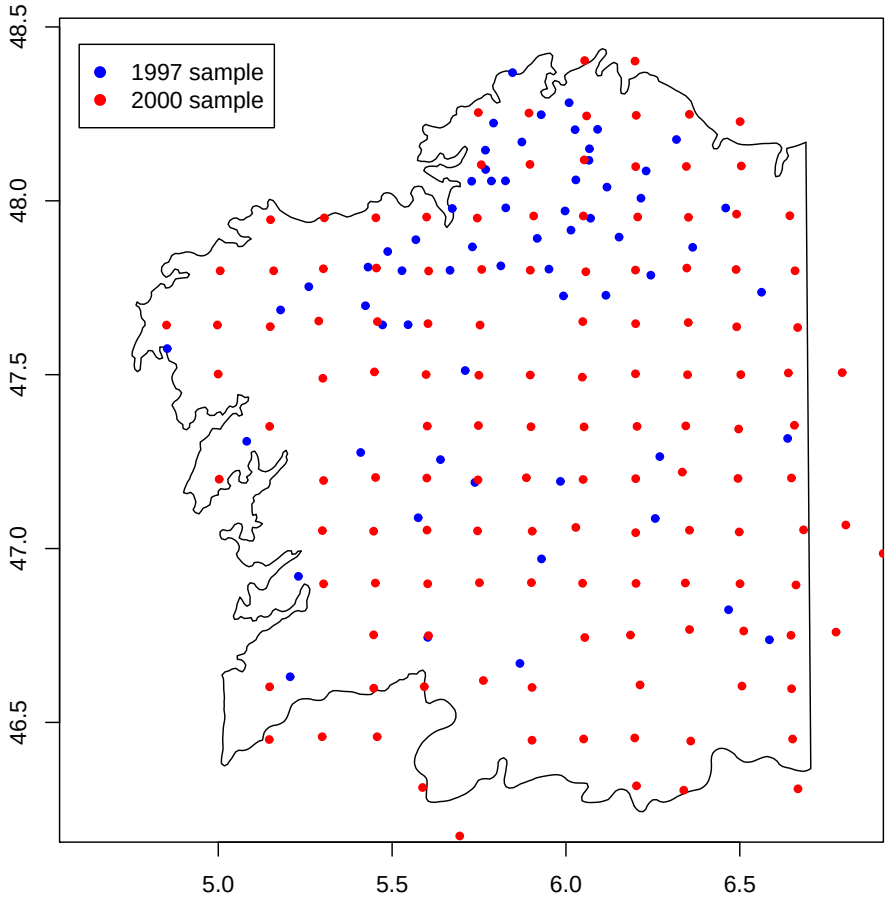
$$\lambda(x) = \exp\{\alpha + \beta S(x)\}$$

- $[Y|S, X] = \mathrm{N}\{\mu + S(x), \tau^2\}$ (independent Gaussian)

Likelihood inference: Importance sampler for direct Monte Carlo evaluation of likelihood

Diggle, Menezes and Su (2010)

Environmental monitoring in Galicia



Modelling strategy

- 2000 sampling is non-preferential
- 1997 sampling may be preferential
- some parameters in common between 1997 and 2000?

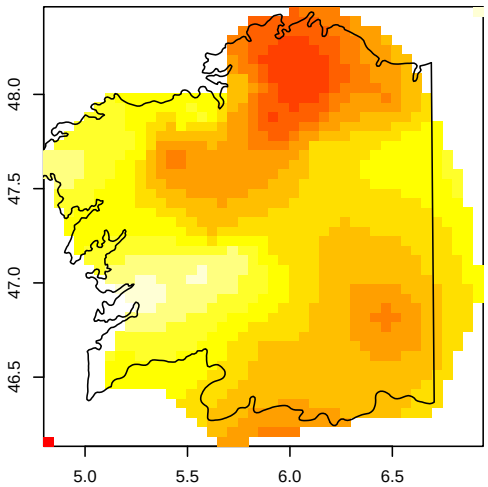
Likelihood ratio testing

Model-fits for log-transformed lead concentrations

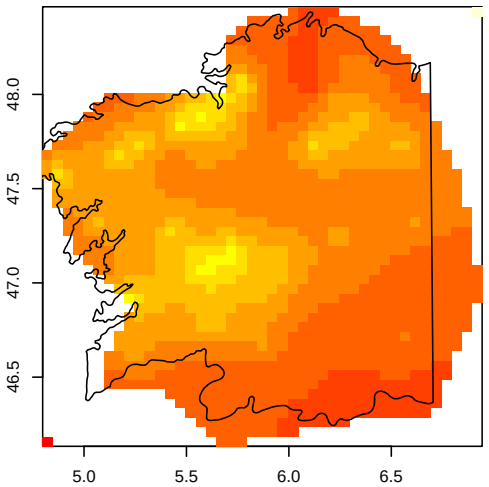
1997 preferential?	parameterisation	$\log L(\hat{\theta})$
no	free parameters	−96.79
	common covariance structure	100.62
yes	free parameters	−82.95
	common covariance structure	−86.04

Spatial prediction of lead concentrations

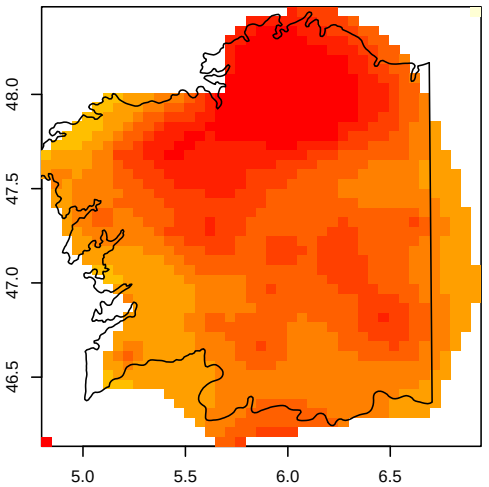
Common scale runs from −0.756 (red) to 8.358 (white)



preferential



non-preferential



difference

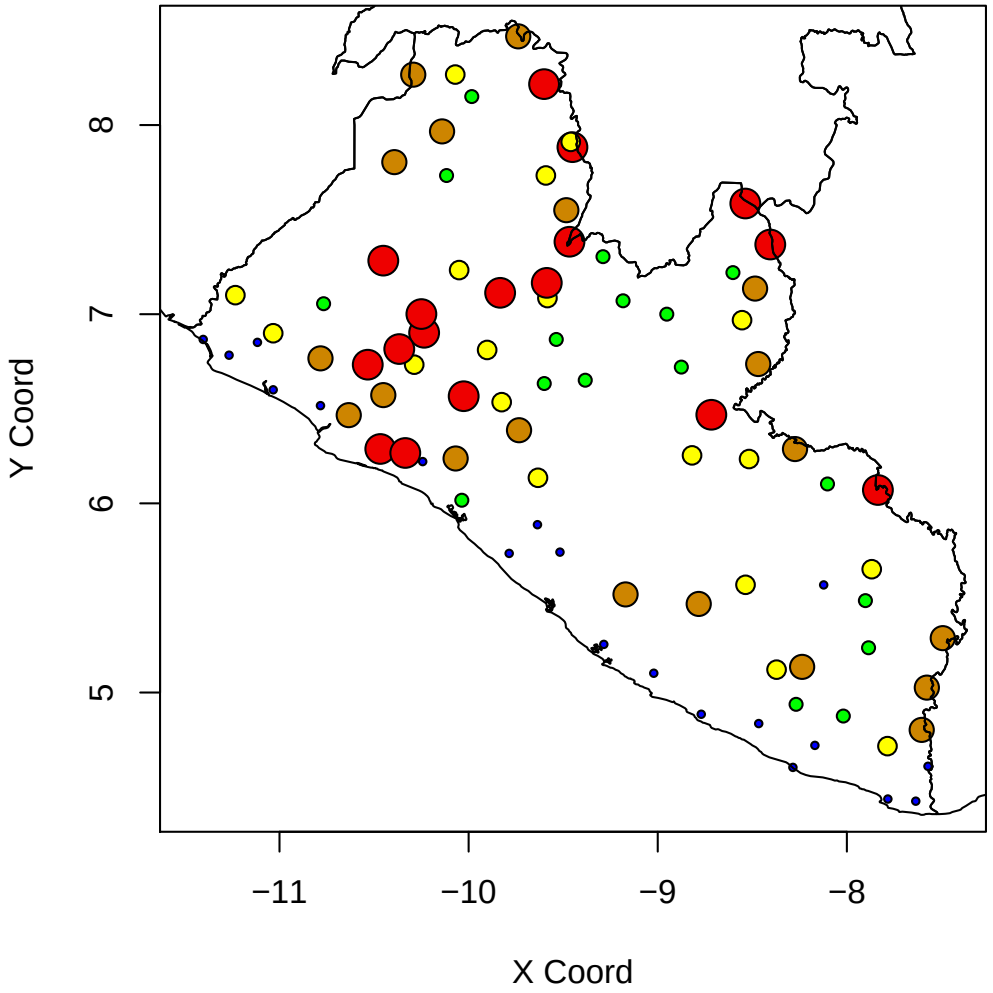
For better global health

Why use model-based geostatistics?

- absence of registry data $\Rightarrow$ modelling assumptions can compensate for sparseness of data
- limited resources for field-work $\Rightarrow$ statistical efficiency is paramount
- classical survey sampling methods are generally inefficient when data are spatially correlated (Matérn, 1960)
- borrowing strength : prevalence data at any one location predicts prevalence at nearby locations

  (aka "the first law of geography")
- $\Rightarrow$ operational performance of classical methods can be matched (or bettered) using smaller sample size

Prevalence surveys



Aim: map variation in prevalence throughout designated region to inform treatment strategy

- how many samples?
- where to take them?
- how to analyse the data?

Geostatistical model for prevalence data

- **Latent spatially correlated process**

  $S(x) \sim \mathrm{SGP}\{0, \sigma^2, \rho(u, v))\}$

- **Latent spatially independent random effects**

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

- **Linear predictor and link function**

  $d(x) =$ environmental variables

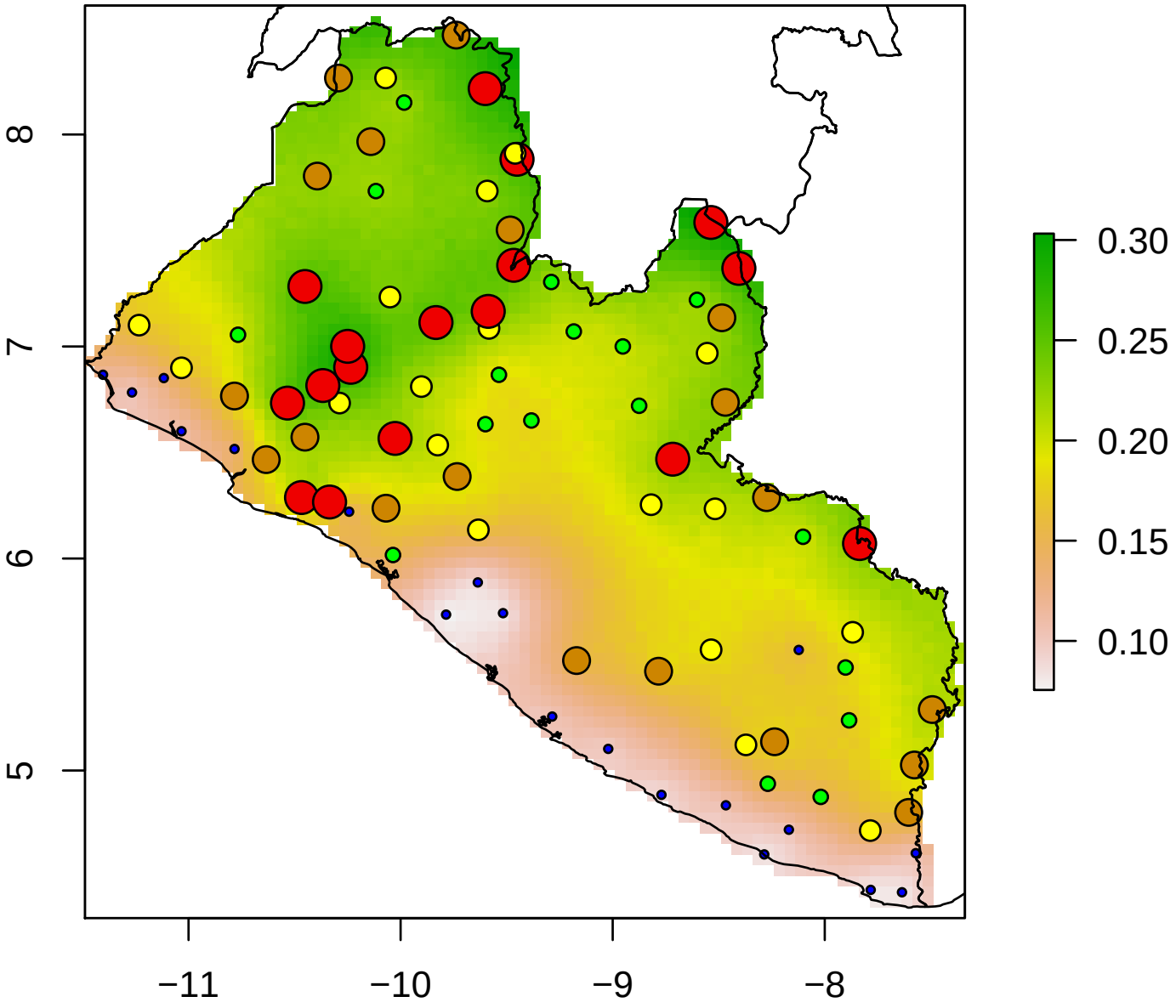
  $\eta(x) = d(x)'\beta + S(x) + U(x)$

  $p(x) = \log[\eta(x)/\{1 - \eta(x)\}]$

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

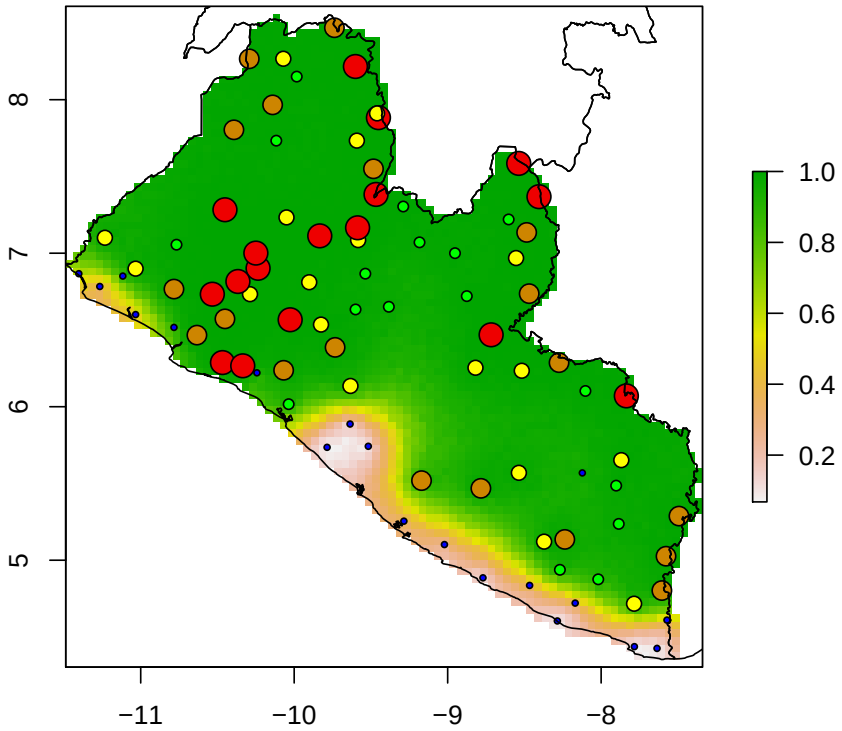
  $Y_i|S(\cdot) \sim \mathrm{Bin}\{n_i, p(x_i)\}$ (binomial sampling)

Liberia: onchocerciasis prevalence map

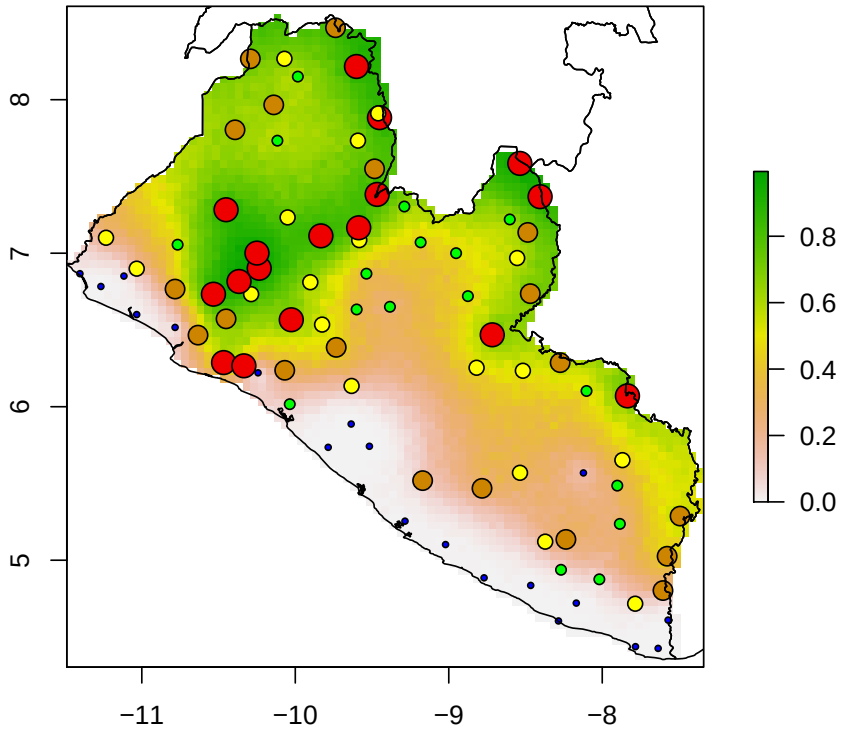


Mapping uncertainty: probability-of-exceedance

P(prevalence>10%)

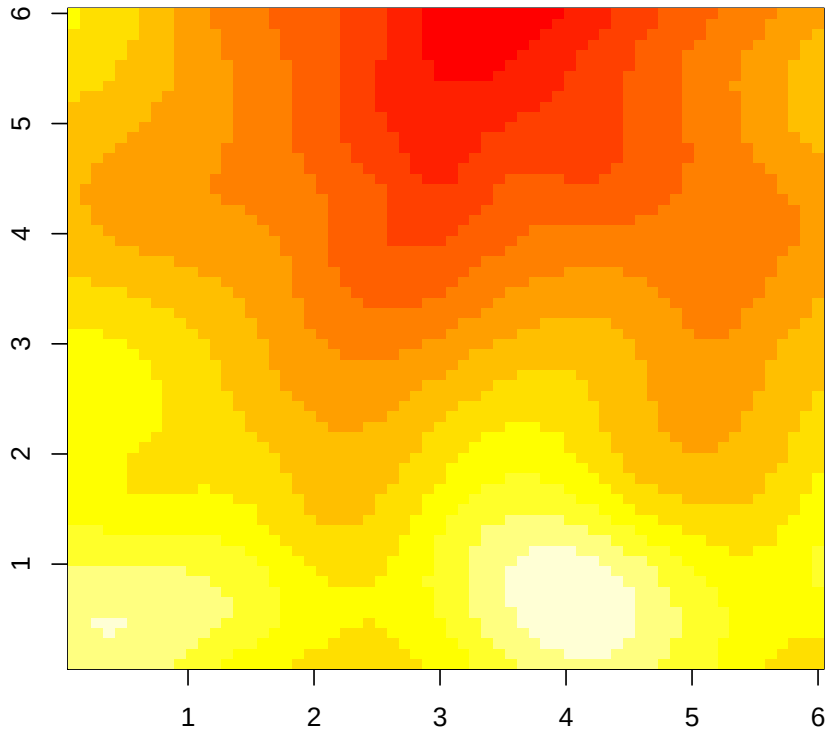


P(prevalence>20%)



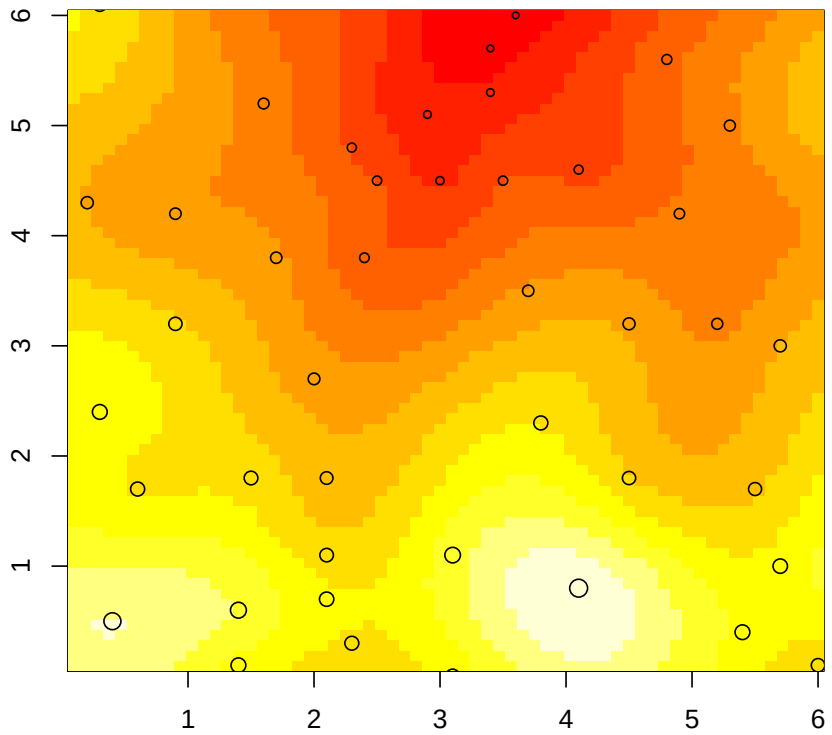
A simulated example

Area-wide average prevalence = 41.7%



A simulated example

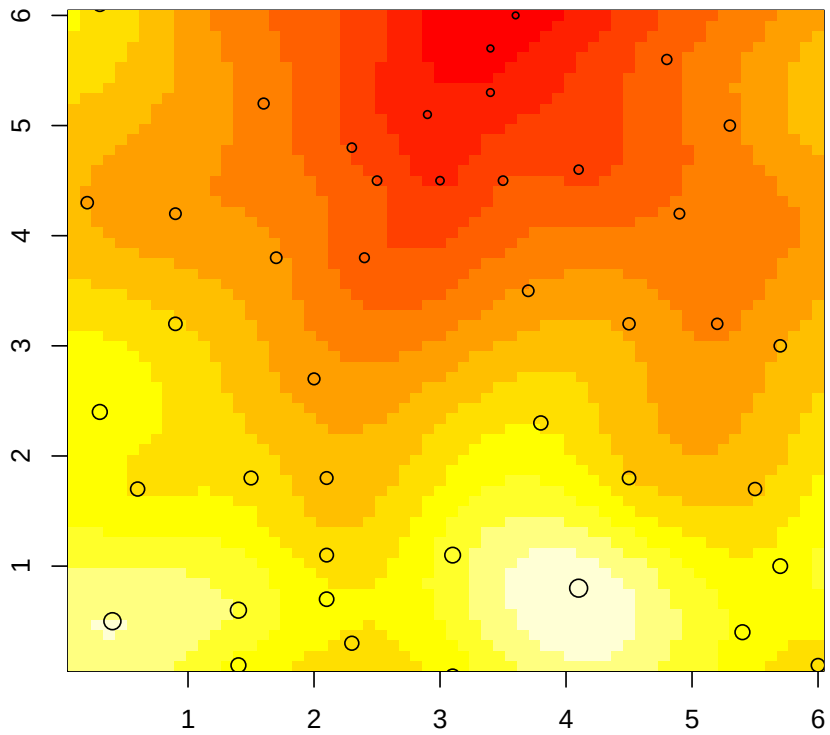
Area-wide average prevalence = 41.7%



Measure prevalence at sample locations

A simulated example

Area-wide average prevalence = 41.7%



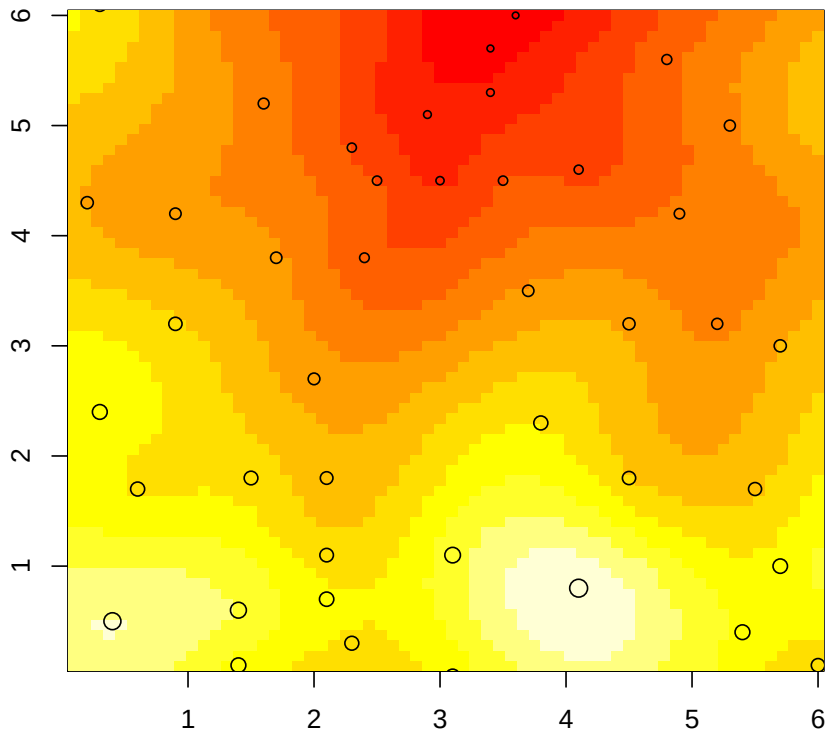
Measure prevalence at sample locations

Survey sampling analysis

$$41.3 \pm 8.6$$

A simulated example

Area-wide average prevalence = 41.7%



Measure prevalence at sample locations

Survey sampling analysis

$$41.3 \pm 8.6$$

Geostatistical analysis

$$41.6 \pm 1.4$$

Where does the extra efficiency come from?

- **Exploiting spatial correlation**

  – **Data** at location x are informative of prevalence at "neighbouring" locations x'

  – **Model-fitting** process allows data to choose optimal definition of "neighbouring"

- **Asking the right question**

  – **Classical approach:** how precisely can we **estimate** prevalence under repeated realisations?

  – **Geostatistical approach:** how precisely can we **predict** prevalence in this particular realisation?

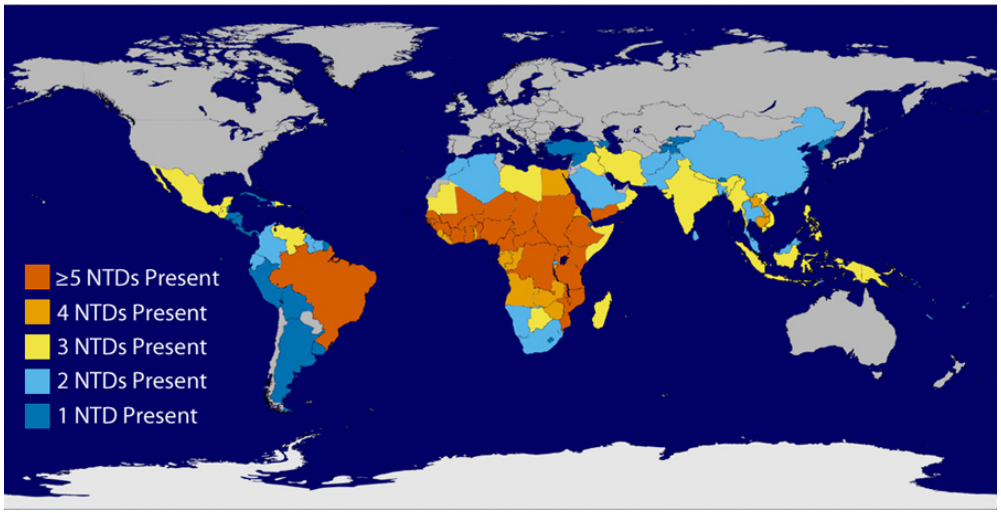
Neglected Tropical Diseases

"A diverse group of communicable diseases that prevail in tropical and subtropical conditions..."

`https://www.who.int/neglected_diseases/diseases/en/`

"One-sixth of the world's population, mostly in developing countries, are infected with one or more of the NTDs."

Mitra and Mawson, 2017



2030 Sustainable Development Goals

Eradication: permanent reduction of worldwide incidence to zero

Elimination: reduction of incidence in a specified geographic area to zero

Elimination as a public health problem: reduction of incidence in a specified geographic area to an agreed level

Lymphatic filariasis in Ghana

- Analyse Ghana-wide pre-intervention data on LF prevalence
- Simulate geographical distribution of LF at or near elimination status
- Compare two design strategies:
  - Current WHO guidelines
  - Model-based geostatistics

Fronterre et al, 2020

Sampling scheme: current WHO guidelines

- 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 observed total number of positives

Elimination target

For each EU

Communities $j = 1, \ldots, N$ of size n_j at locations x_j

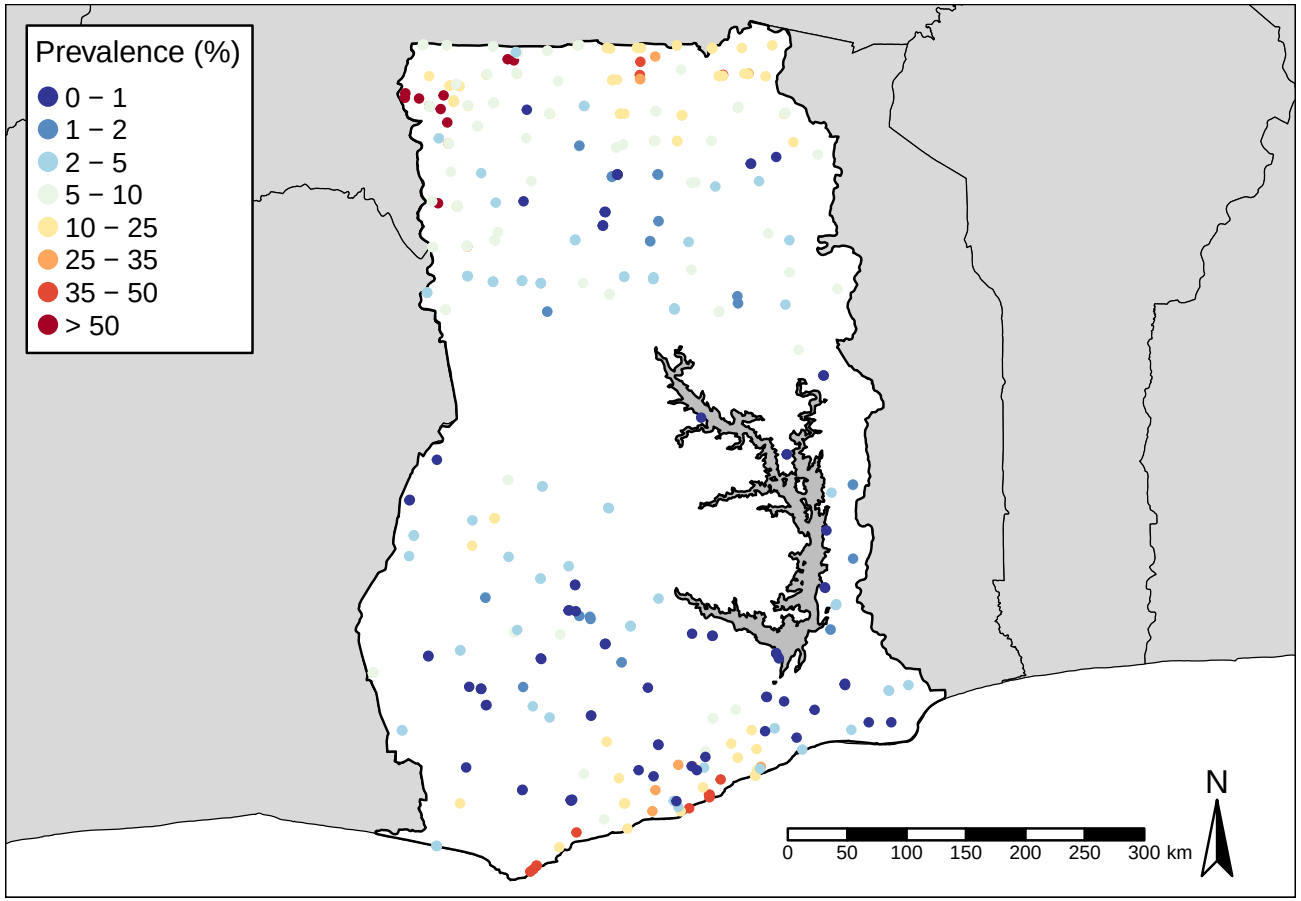
$$T = \frac{\sum n_j P(x_j)}{\sum n_j}$$

Elimination $\Leftrightarrow T < 0.02$

Probabilistic prediction

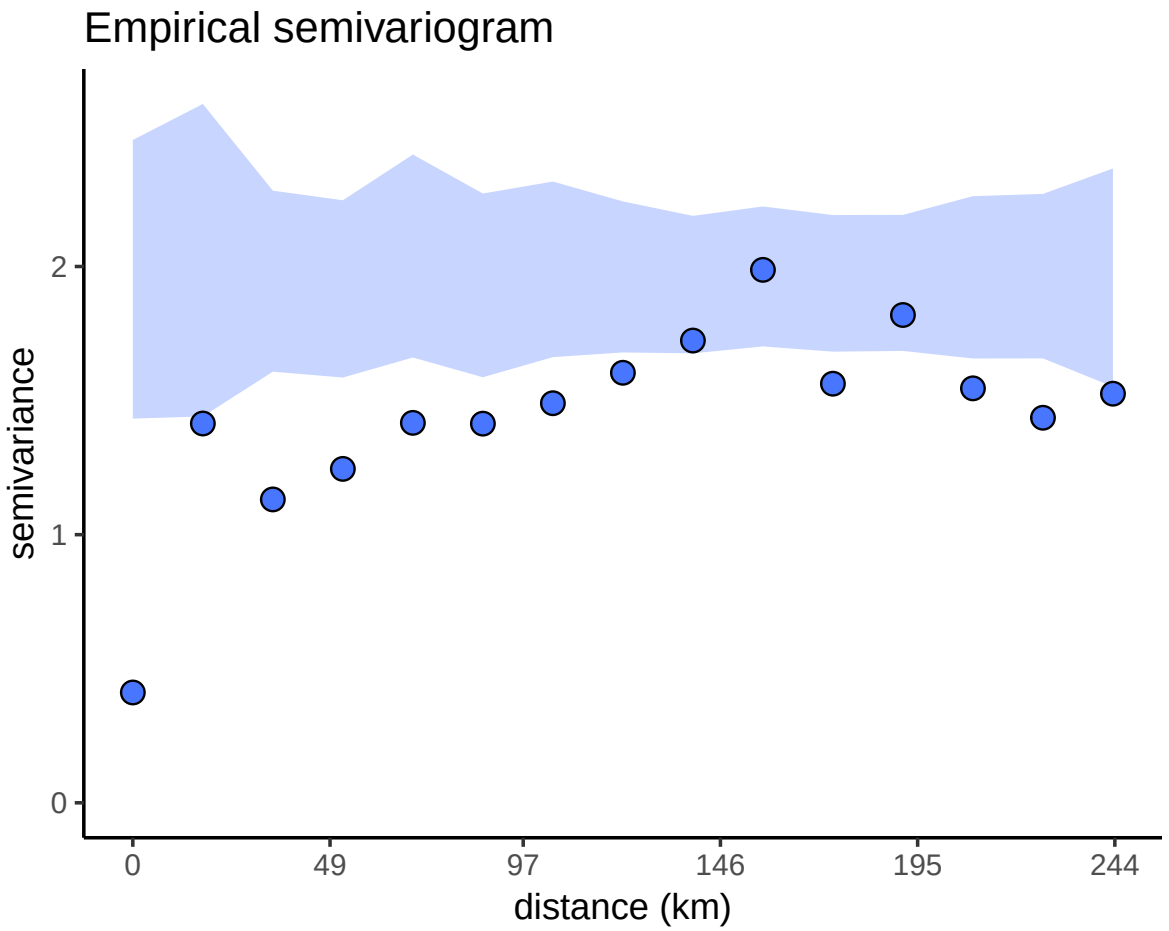
- Draw samples from predictive distribution of T
- Choose probability threshold q
- **Elimination indicated** $\Leftrightarrow \mathrm{Prob}(T < 0.02 | \text{data}) > q$

Ghana: baseline LF prevalence at 403 locations

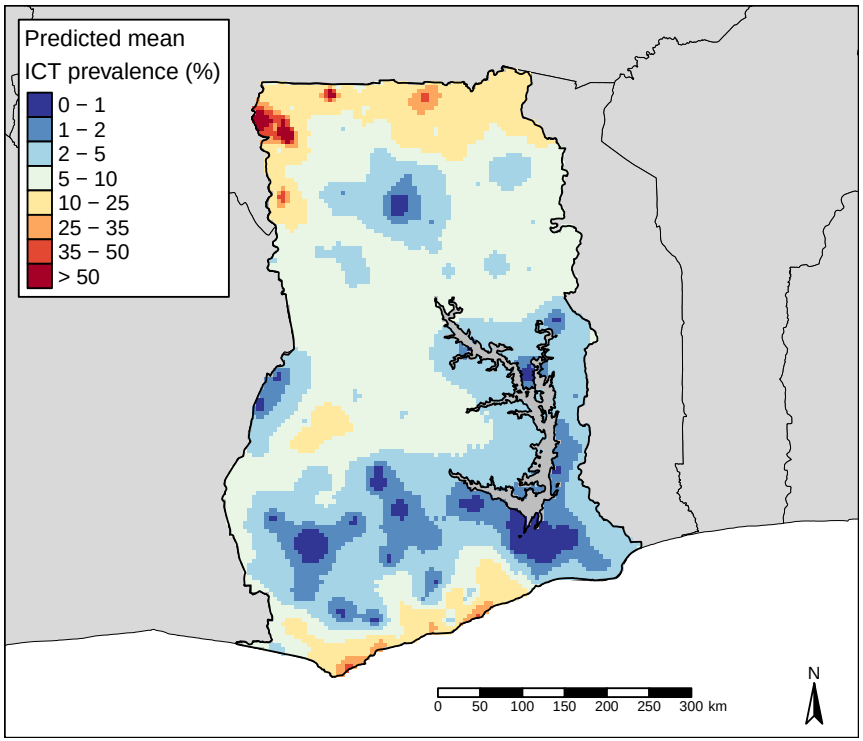


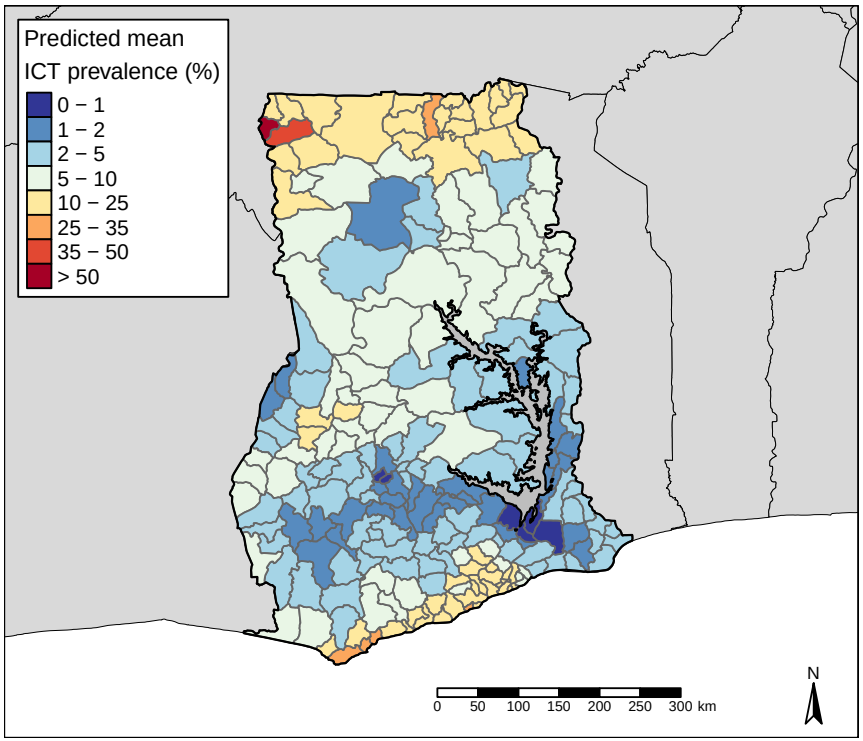
Ghana LF: spatial correlation structure

$$V(u) = \frac{1}{2}\mathrm{Var}\{S(x) - S(x-u)\}$$

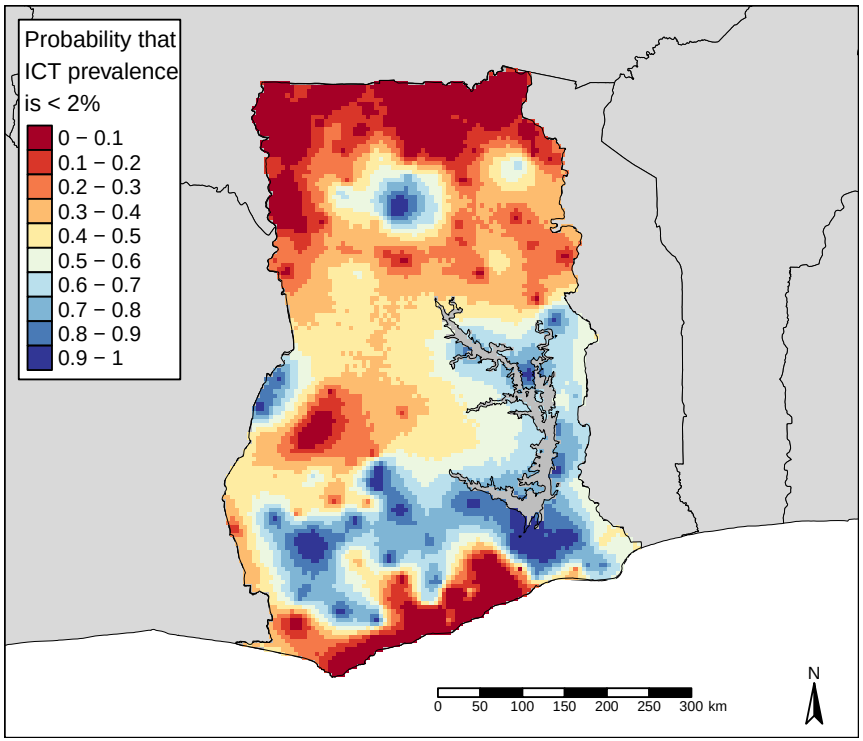


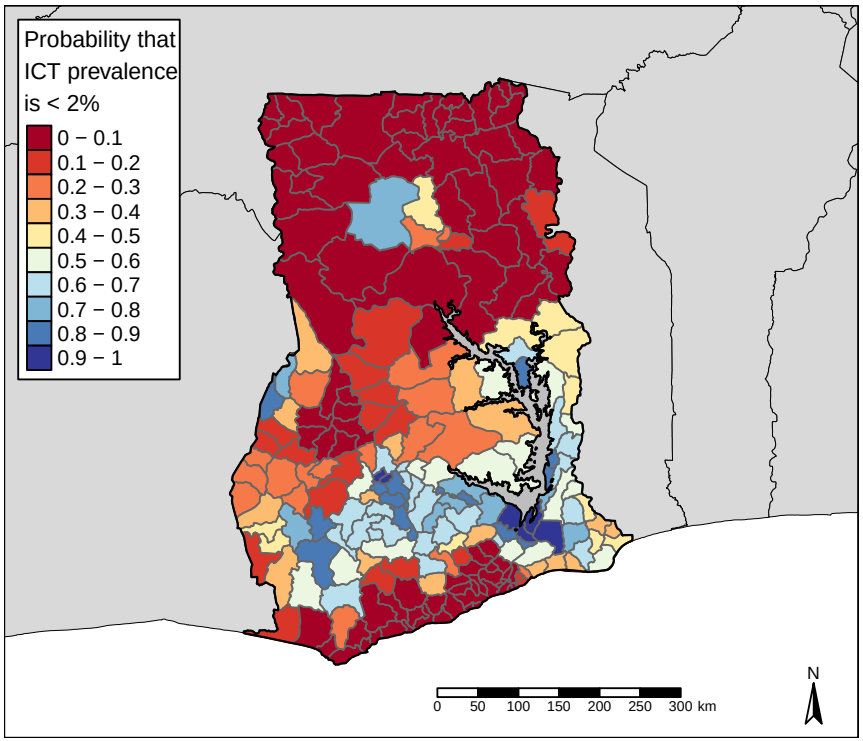
Baseline prevalence maps





Baseline elimination status maps





Simulating progression towards elimination

Geostatistical model fitted to pre-intervention data

$$\log[P(x)/\{1 - P(x)\}] = \alpha + S(x)$$

- Pre-intervention **log-odds surface**

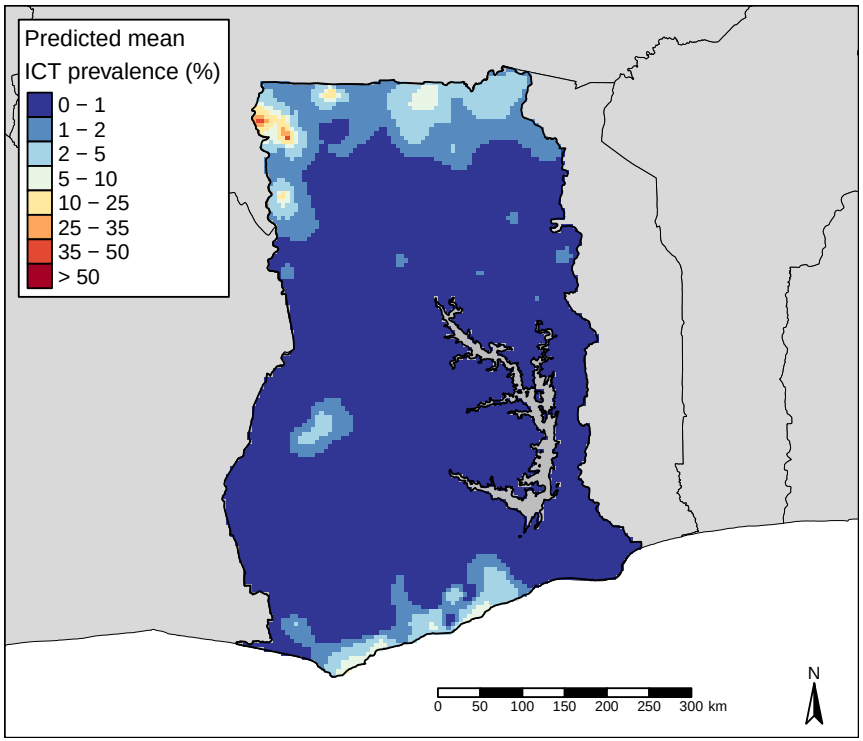
$$L(x) = \log[P(x)/\{1 - P(x)\}]$$

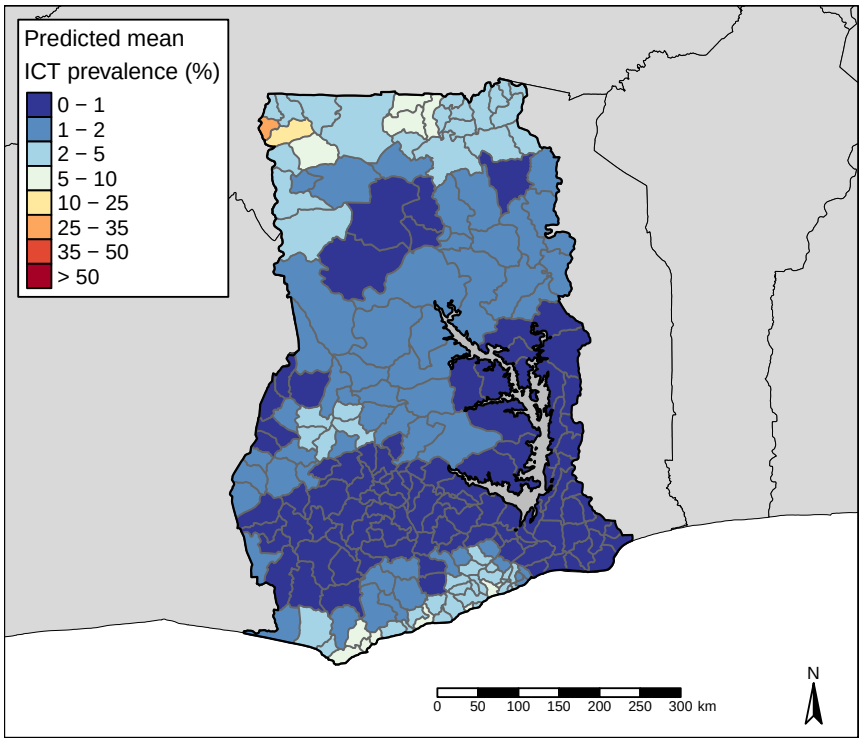
- Calculate

$$L_0(x) = L(x) - c, \quad P_0(x) = \exp\{L_0(x)\}/[1 + \exp\{L_0(x)\}],$$

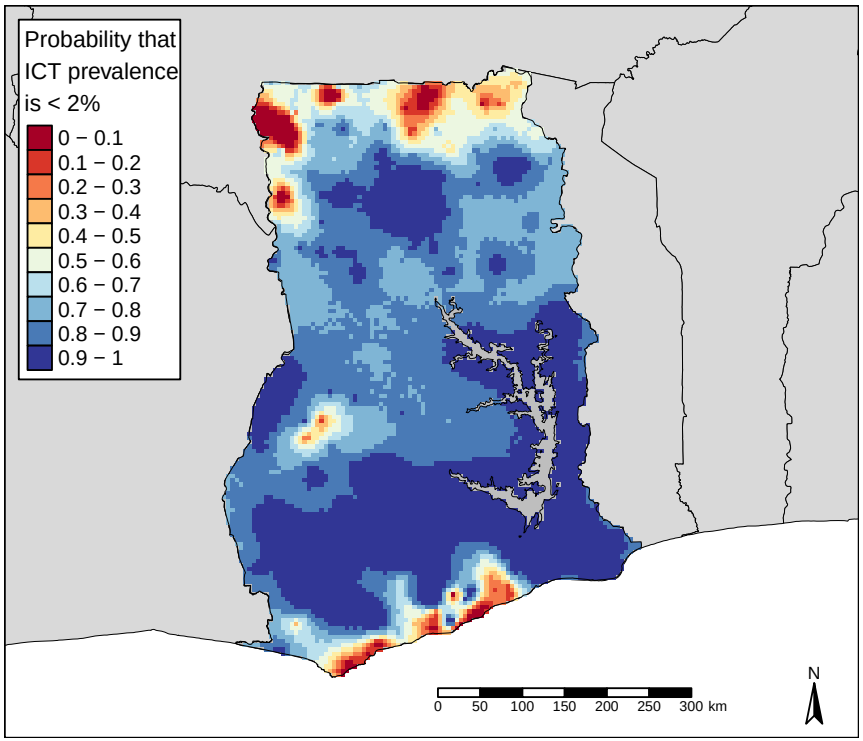
such that population-weighted average of $P_0(x)$ is 0.02.

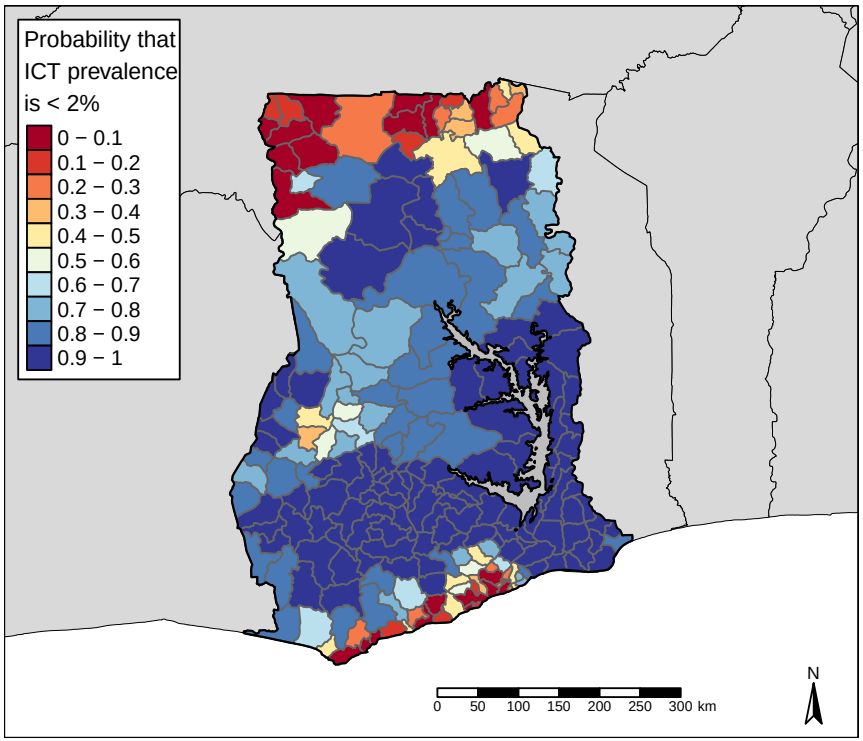
Projected prevalence maps





Projected elimination status maps





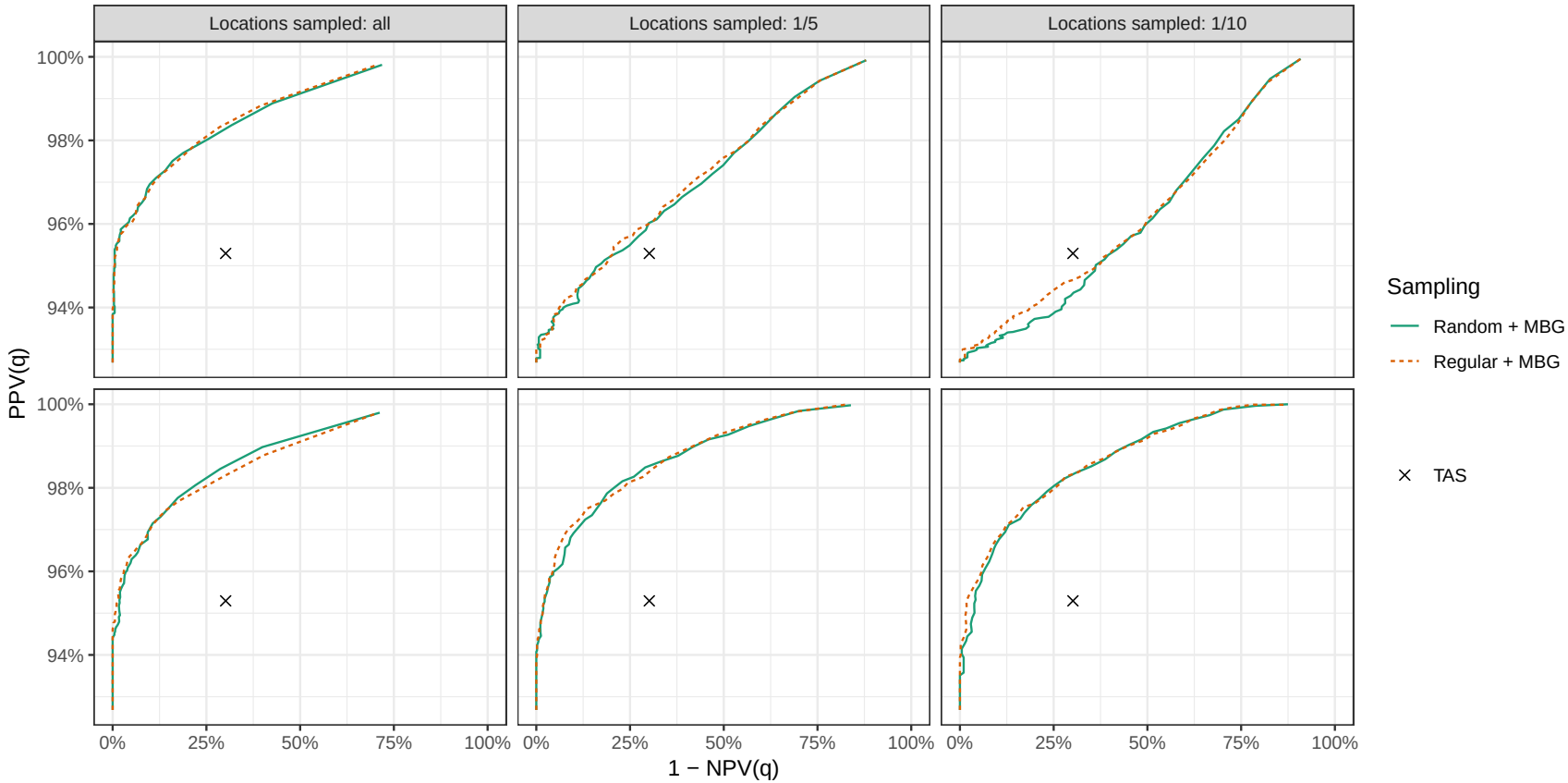
Evaluation of predictive performance

- simulate test results over region of interest
- for each EU
  - apply current WHO guidelines 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 NPV and PPV

ROC curves: positive and negative predictive values

PPV = percent correct positive indications

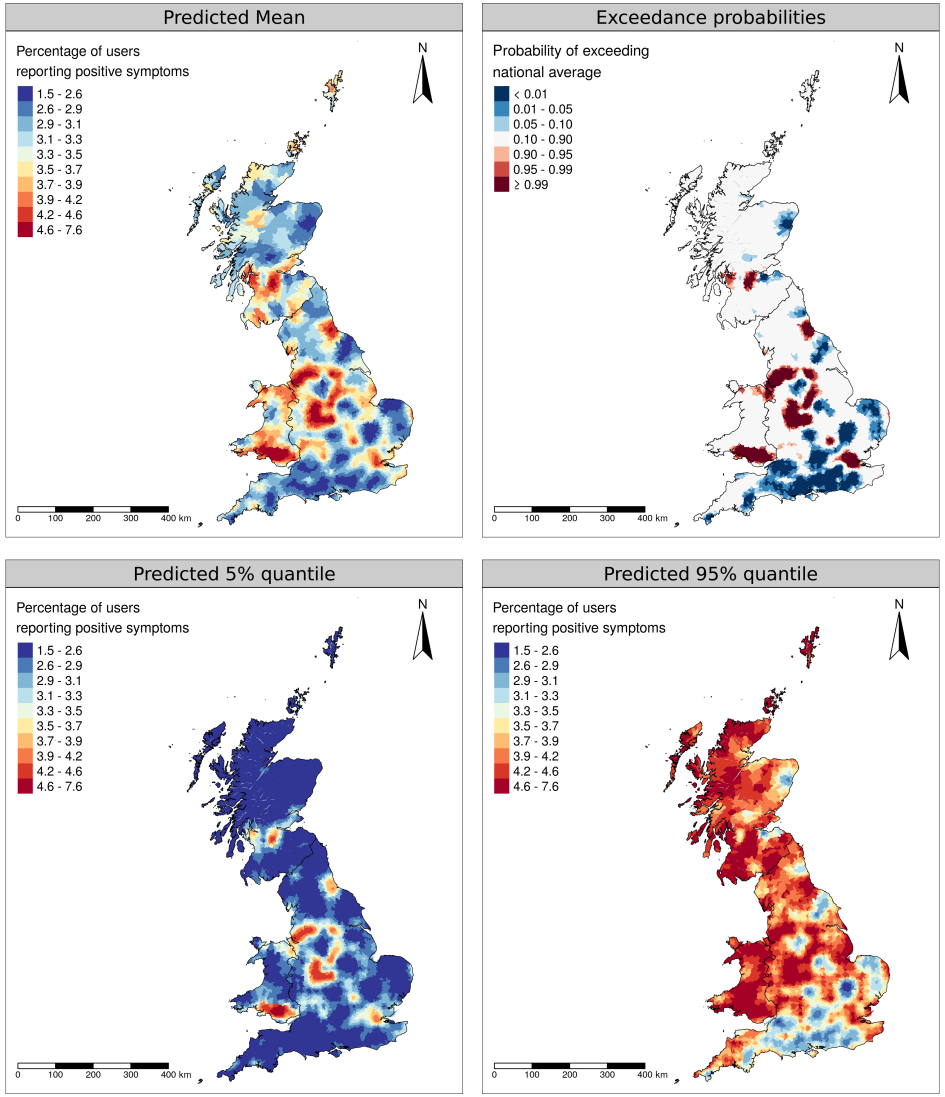
NPV = percent correct negative indications



Top row: number of schools sampled per district: $30 \rightarrow 6 \rightarrow 3$

Bottom row: constant total number of children sampled

Meanwhile back in the UK...the arrival of COVID-19



Concluding remarks

Methodology

1. Developed over 20 years, growing number of case-studies
2. Extensions include:
   - multiple diagnostics;
   - randomised and convenience survey data;
   - co-morbidity surveys;
   - adaptive design;
   - spatio-temporal models
   - real-time surveillance for any georeferenced numerator/denominator data

Software

1. Open-source R package `PrevMap`
2. Plans to develop interactive user-interface for in-country use

Diggle and Giorgi, 2019