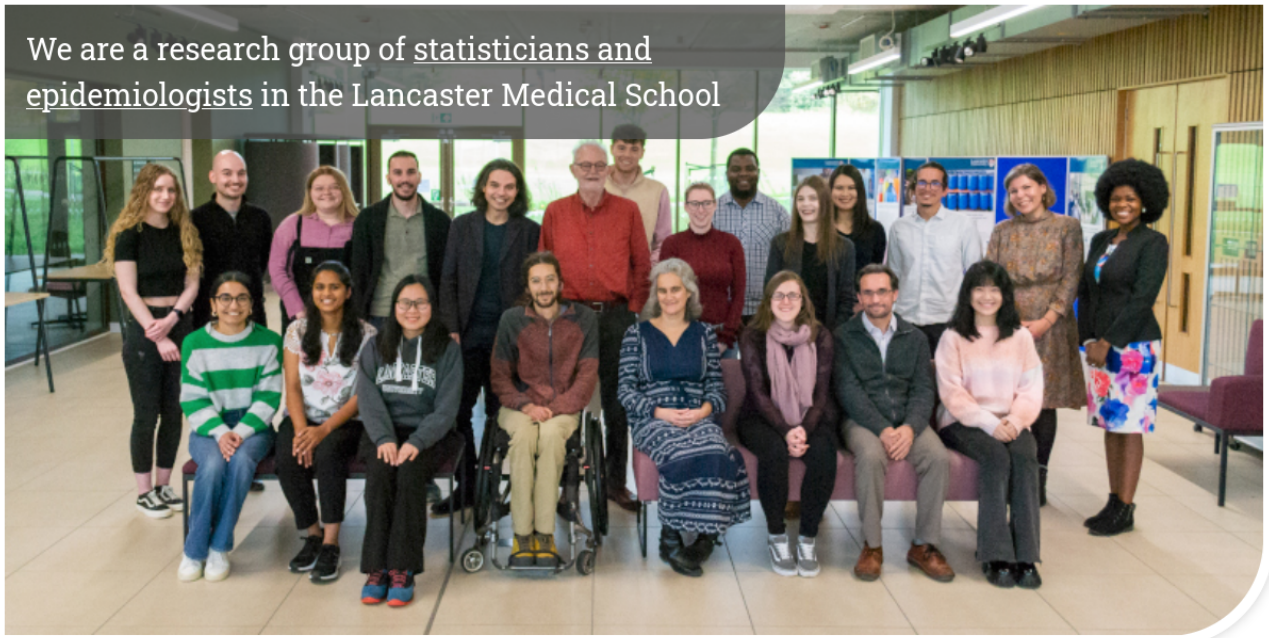
Model-based geostatistical inference with low prevalence data: a case study on lymphatic filariasis

Emanuele Giorgi

Centre for Health Informatics, Computing and Statistics

CHICAS
centre for
health informatics,
computing, and statistics



WHO Collaborating Centre

Current NTD Projects

- Brazilian Leptospirosis Study : Ecoepidemiology of Leptospirosis in the Urban Slums of Brazil
- Geostat NTD Hub : A Geostatistical Web Framework for Prevalence Surveys for Neglected Tropical Diseases
- Loa-loa Mapping : Developing methods to combine prevalence data from multiple diagnostics
- National Snakebite Study in Sri Lanka : Developing a Risk Map for Snakebites
- Neglected Tropical Disease Modelling Consortium : Survey Design and Analysis for Disease Prevalence

Overview of the talk

- Part 1: Introduction to Neglected Tropical Diseases
- Part 2: Model-based geostatistics for disease mapping
- Part 3: Case studies on lymphatic filariasis

Acknowledgements

- Lucinda Hadely (Senior Research Associate)
- Funders: TaskForce for Global Health and USAID

Neglected tropical diseases

Disease	CDC	WHO
Buruli ulcer (*Mycobacterium ulcerans* infection)	+	+
Chikungunya [a]	–	+
Chagas disease	+	+
Cysticercosis	+	+
Dengue fever	+	+
Dracunculiosis (or guinea worm disease) [b]	+	+
Echinococcosis	+	+
Fascioliasis	+	+
Foodborne trematodiasis [a]	–	+
Human African trypanosomiasis (or sleeping sickness)	+	+
Leishmaniasis (or kala-azar)	+	+
Leprosy	+	+
Lymphatic filariasis [b]	+	+
Mycetoma	+	+
Onchocerciasis (or river blindness) [b]	+	+
Rabies	+	+
Schistosomiasis [b]	+	+
Soil-transmitted helminthiasis [b]	+	+
Trachoma [b]	+	+
Yaws	+	+

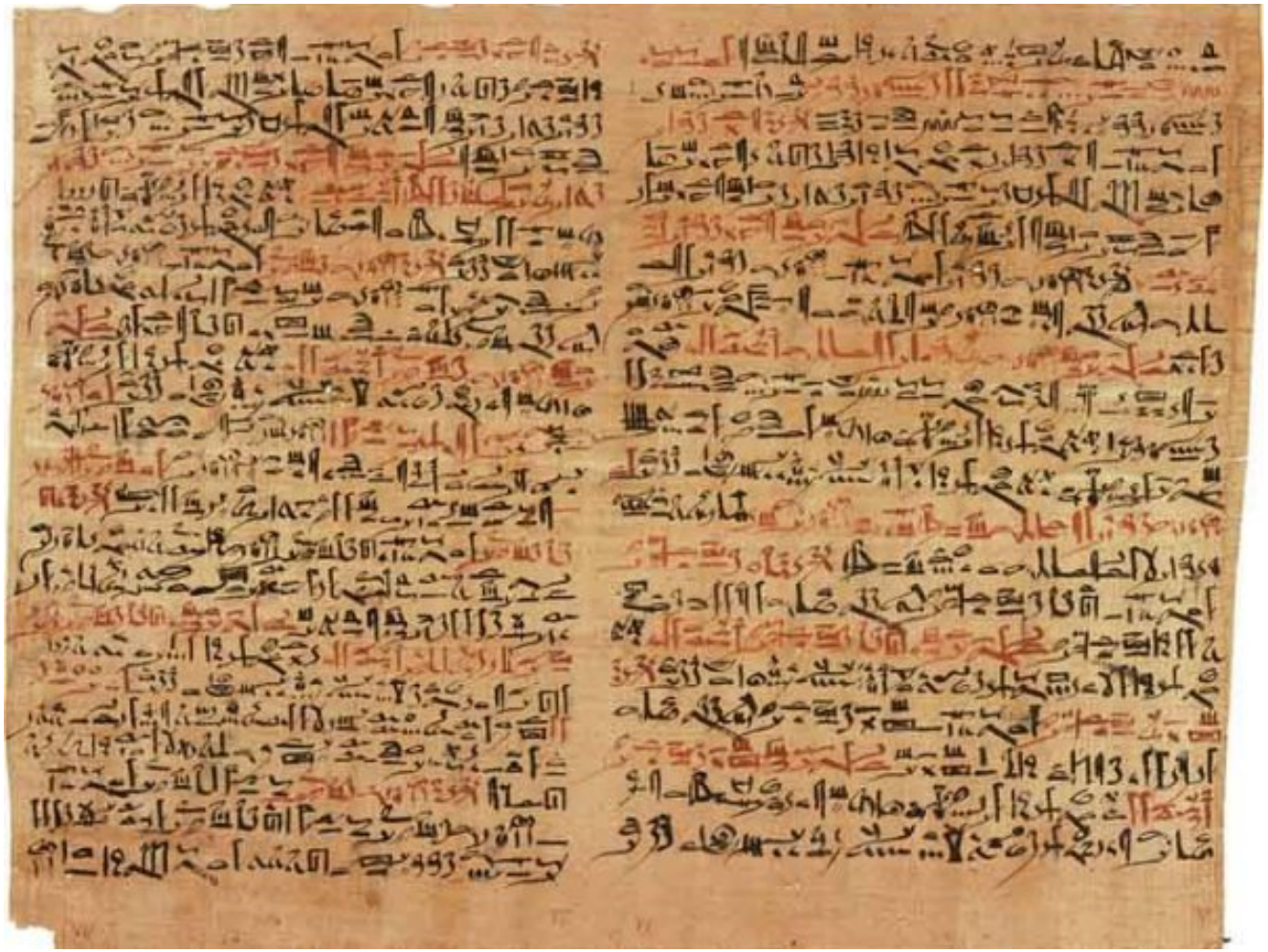
Data source: CDC, 2017 [10]; WHO, 2017 [3]; [a] Not mentioned under CDC list of neglected tropical diseases (NTDs); [b] Diseases that can be controlled or eliminated through mass drug administration (MDA), or other interventions.

Common features of NTDs

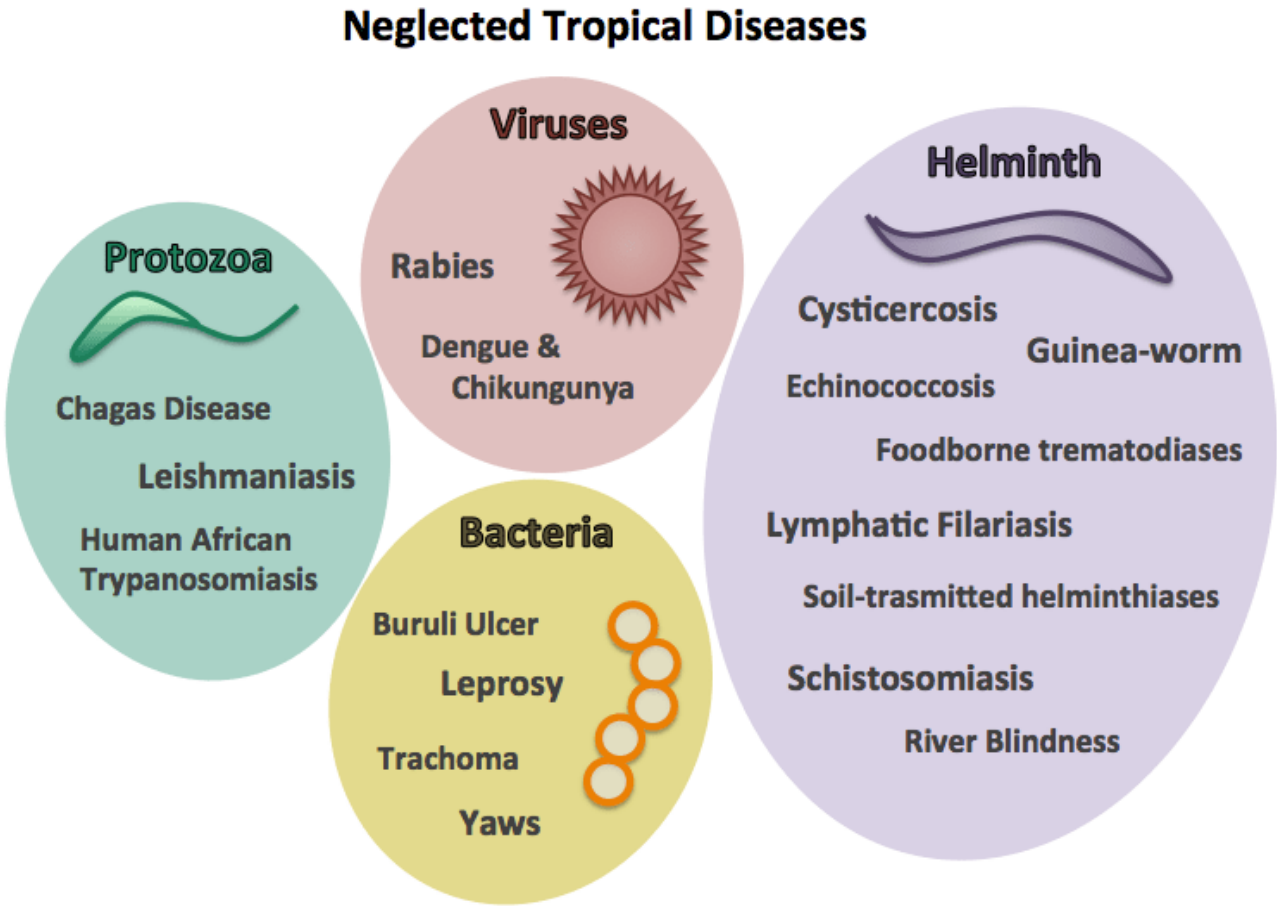
Common features of NTDs

Common features of NTDs

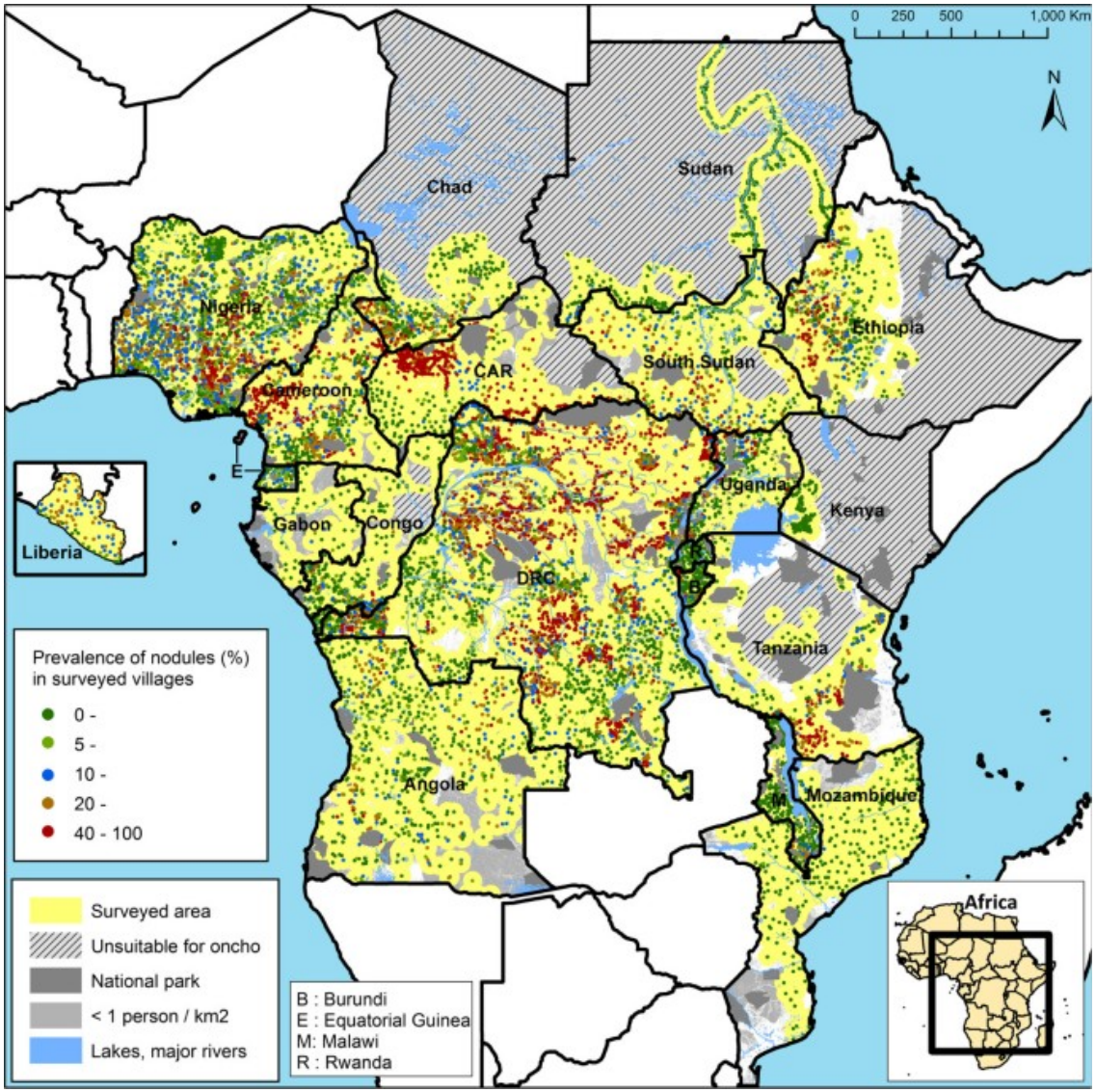
Common features of NTDs



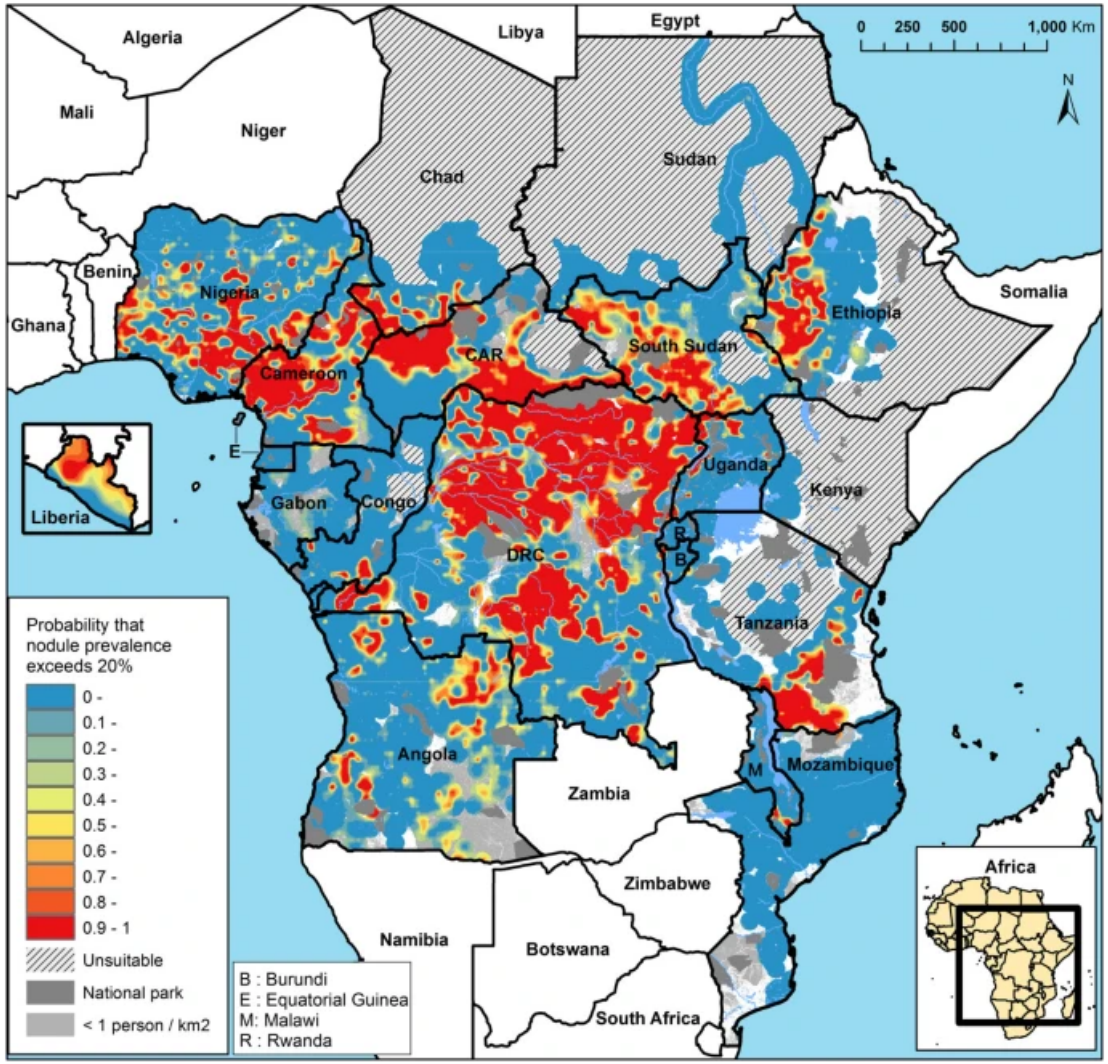
Common features of NTDs



How can we map NTDs?



How can we map NTDs?



How can we map NTDs?

How can we map NTDs?

- x_i: village location

How can we map NTDs?

- x_i: village location
- n_i: number of people tested

How can we map NTDs?

- x_i: village location
- n_i: number of people tested
- y_i: number of positive cases

How can we map NTDs?

- x_i: village location
- n_i: number of people tested
- y_i: number of positive cases
- $p(x)$: prevalence at a location x

How can we map NTDs?

- x_i: village location
- n_i: number of people tested
- y_i: number of positive cases
- $p(x)$: prevalence at a location x
- $d(x)$: spatially referenced covariate

How can we map NTDs?

- x_i: village location
- n_i: number of people tested
- y_i: number of positive cases
- $p(x)$: prevalence at a location x
- $d(x)$: spatially referenced covariate

Problem

1. How can we predict $p(x)$ using (x_i, n_i, y_i)?
2. How can we use $d(x)$ to improve our predictive inferences on $p(x)$?

Defining geostatistical problems

The ingredients:

Defining geostatistical problems

The ingredients:

- S = process of nature (e.g. disease risk)

Defining geostatistical problems

The ingredients:

- S = process of nature (e.g. disease risk)
- Y = data

Defining geostatistical problems

The ingredients:

- S = process of nature (e.g. disease risk)
- Y = data
- A statistical model $[S, Y] = [S] \times [Y|S]$

Standard geostatistical model for prevalence mapping

Standard geostatistical model for prevalence mapping

- Data: x_i = location of the cluster; n_i = number of sampled individuals at x_i; y_i=number of positively tested individuals at x_i

Standard geostatistical model for prevalence mapping

- Data: x_i = location of the cluster; n_i = number of sampled individuals at x_i; y_i=number of positively tested individuals at x_i
- $d(x_i)$ = vector covariates

Standard geostatistical model for prevalence mapping

- Data: x_i = location of the cluster; n_i = number of sampled individuals at x_i; y_i=number of positively tested individuals at x_i
- $d(x_i)$ = vector covariates
- $S(x)$ = spatial stochastic process

Standard geostatistical model for prevalence mapping

- Data: x_i = location of the cluster; n_i = number of sampled individuals at x_i; y_i=number of positively tested individuals at x_i
- $d(x_i)$ = vector covariates
- $S(x)$ = spatial stochastic process
- Z_i = unstructured random effects

Standard geostatistical model for prevalence mapping

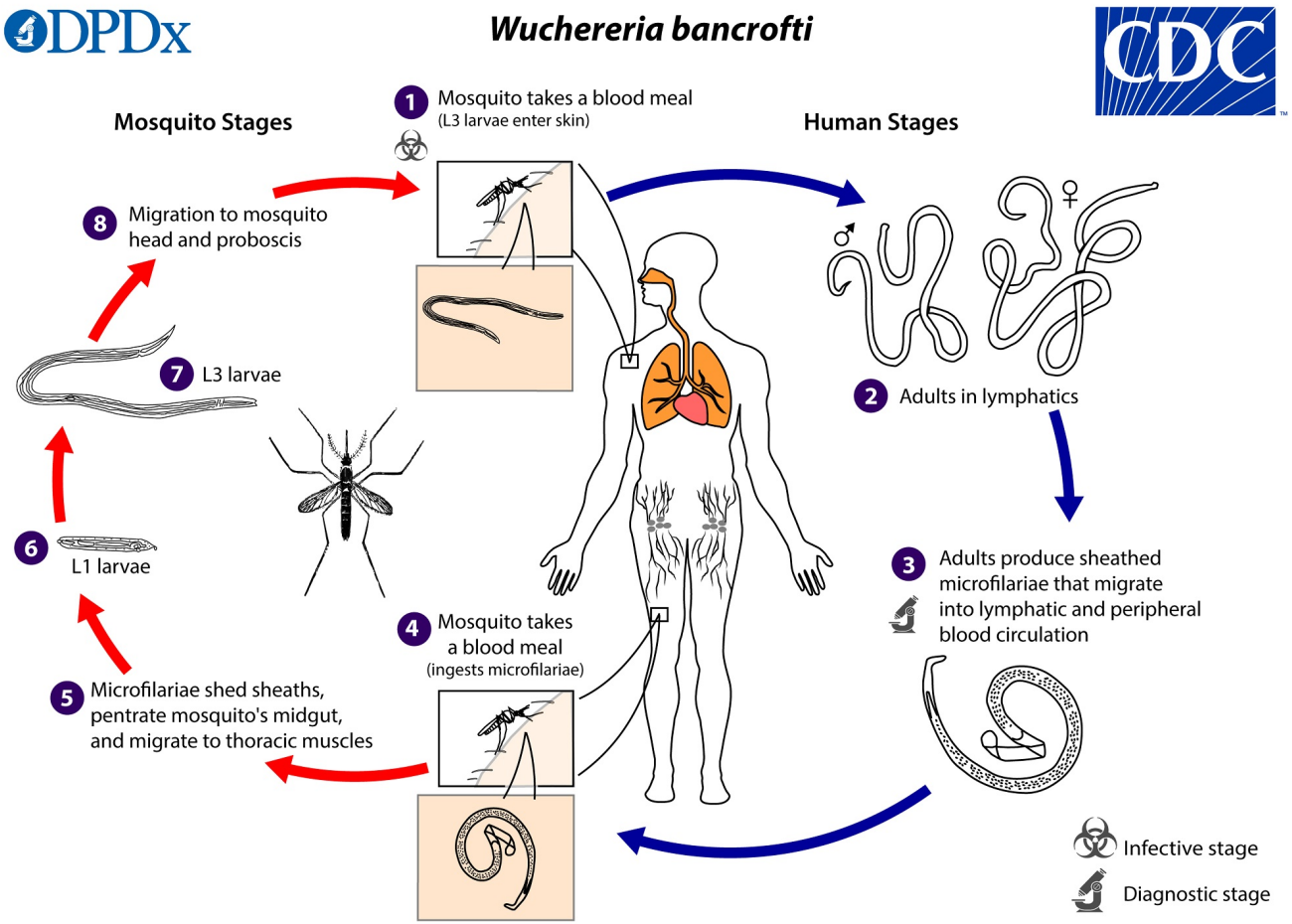
- Data: x_i = location of the cluster; n_i = number of sampled individuals at x_i; y_i=number of positively tested individuals at x_i
- $d(x_i)$ = vector covariates
- $S(x)$ = spatial stochastic process
- Z_i = unstructured random effects
- Assumption: $Y_i|S(x_i), Z_i \sim \text{Bin}(n_i, p(x_i))$

$$\log\left\{\frac{p(x_i)}{(1-p(x_i)}\right\} = d(x_i)^\top\beta + S(x_i) + Z_i$$

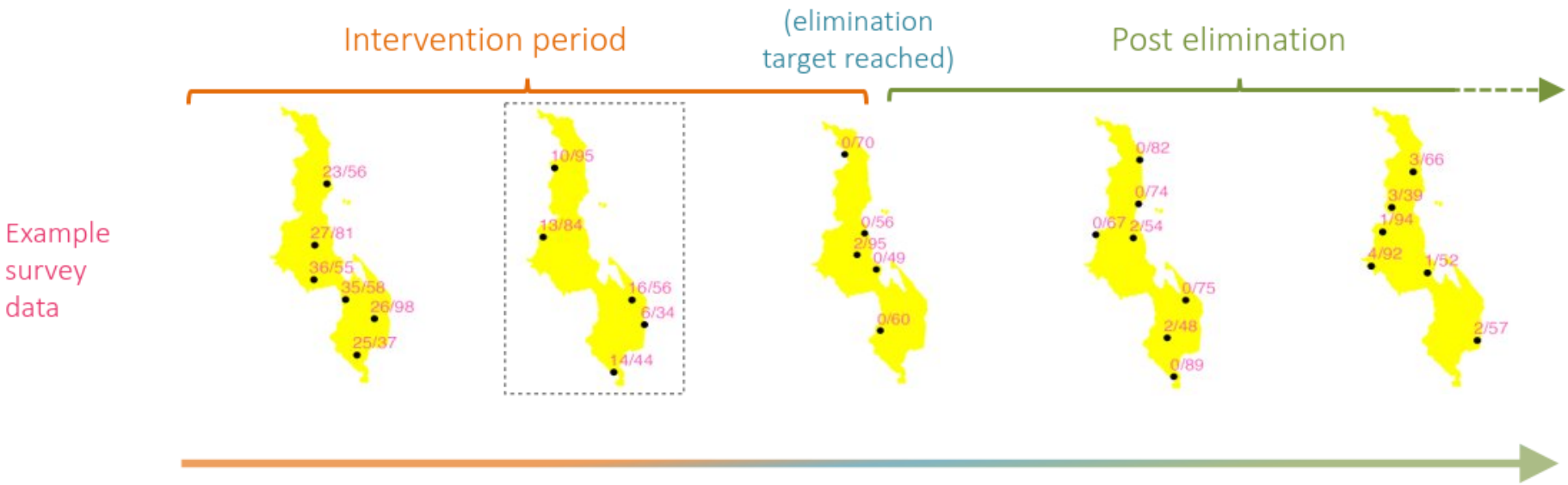
To predict or to explain?

- **Explanatory modelling:** emphasis is placed on understanding the relationships between the health outcome and risk factors
- **Predictive modelling:** maximize the predictive performance of the model

Lymphatic filariasis



LF prevalence surveys



- Elimination is declared when the district-wide is below 1%.
  - Questions: 1) How can we use these data to design a surveillance system in a post-elimination setting? 2) Where should we place the sentinel sites and how many?

Geostatistical modelling of repeated cross-sectional surveys

Geostatistical modelling of repeated cross-sectional surveys

- $\mathcal{S}_t = \{S_t(x) : x \in A\}$ (true spatial surface)

Geostatistical modelling of repeated cross-sectional surveys

- $\mathcal{S}_t = \{S_t(x) : x \in A\}$ (true spatial surface)
- $\mathcal{Y}_t = \{Y_k : k = 0, \ldots, t\}$ (data collected from time 0 to time t)

Geostatistical modelling of repeated cross-sectional surveys

- $\mathcal{S}_t = \{S_t(x) : x \in A\}$ (true spatial surface)
- $\mathcal{Y}_t = \{Y_k : k = 0, \ldots, t\}$ (data collected from time 0 to time t)
- $X_t = (x_{t,1}, \ldots, x_{t,n})$ (locations of the data at time t)

Geostatistical modelling of repeated cross-sectional surveys

- $\mathcal{S}_t = \{S_t(x) : x \in A\}$ (true spatial surface)
- $\mathcal{Y}_t = \{Y_k : k = 0, \dots, t\}$ (data collected from time 0 to time t)
- $X_t = (x_{t,1}, \dots, x_{t,n})$ (locations of the data at time t)
- Assumption 1: $[Y_t | \mathcal{S}_t(X_t)]$ is Binomial with linear predictor

$$\log\left\{\frac{p_t(x_{t,i})}{1 - p_t(x_{t,i})}\right\} = \alpha_t + S_t(x_{t,i})$$

Geostatistical modelling of repeated cross-sectional surveys

- $\mathcal{S}_t = \{S_t(x) : x \in A\}$ (true spatial surface)
- $\mathcal{Y}_t = \{Y_k : k = 0, \ldots, t\}$ (data collected from time 0 to time t)
- $X_t = (x_{t,1}, \ldots, x_{t,n})$ (locations of the data at time t)
- Assumption 1: $[Y_t | \mathcal{S}_t(X_t)]$ is Binomial with linear predictor

$$\log\left\{\frac{p_t(x_{t,i})}{1 - p_t(x_{t,i})}\right\} = \alpha_t + S_t(x_{t,i})$$

- Assumption 2: Autoregressive process of order 1

$$S_t(x) = \gamma S_{t-1}(x) + W_t(x), \gamma > 0$$

with $W_t(x)$ being a Gaussian process with covariance function $\sigma^2 R + \tau^2 I$

Geostatistical modelling of repeated cross-sectional surveys

- $\mathcal{S}_t = \{S_t(x) : x \in A\}$ (true spatial surface)
- $\mathcal{Y}_t = \{Y_k : k = 0, \ldots, t\}$ (data collected from time 0 to time t)
- $X_t = (x_{t,1}, \ldots, x_{t,n})$ (locations of the data at time t)
- Assumption 1: $[Y_t | \mathcal{S}_t(X_t)]$ is Binomial with linear predictor

$$\log\left\{\frac{p_t(x_{t,i})}{1 - p_t(x_{t,i})}\right\} = \alpha_t + S_t(x_{t,i})$$

- Assumption 2: Autoregressive process of order 1

$$S_t(x) = \gamma S_{t-1}(x) + W_t(x), \gamma > 0$$

with $W_t(x)$ being a Gaussian process with covariance function $\sigma^2 R + \tau^2 I$

- Problem: inferring γ and α_t from the data may not empirically feasible.

Parameter estimation of $\theta_{-\gamma}$

Parameter estimation of $\theta_{-\gamma}$

- Let $\theta_{-\gamma}$ denote the model parameters excluding γ.

Parameter estimation of $\theta_{-\gamma}$

- Let $\theta_{-\gamma}$ denote the model parameters excluding γ.
- Note: the marginal variance of $S_t(x)$ is $\omega^2 = \sigma^2/(1-\gamma^2)$.

Parameter estimation of $\theta_{-\gamma}$

- Let $\theta_{-\gamma}$ denote the model parameters excluding γ.
- Note: the marginal variance of $S_t(x)$ is $\omega^2 = \sigma^2/(1-\gamma^2)$.
- We estimate $\theta_{-\gamma}$ using the estimating equation originating from

$$L(Y_0, \dots, Y_k; \theta_{-\gamma}) = \prod_{i=1}^{k} L_1(Y_i; \theta_{-\gamma})$$

Parameter estimation of $\theta_{-\gamma}$

- Let $\theta_{-\gamma}$ denote the model parameters excluding γ.
- Note: the marginal variance of $S_t(x)$ is $\omega^2 = \sigma^2/(1-\gamma^2)$.
- We estimate $\theta_{-\gamma}$ using the estimating equation originating from

$$L(Y_0, \ldots, Y_k; \theta_{-\gamma}) = \prod_{i=1}^{k} L_1(Y_i; \theta_{-\gamma})$$

- We define k in the above equation as the last time point for which the data allow for the estimation of α_k.

Parameter estimation of $\theta_{-\gamma}$

- Let $\theta_{-\gamma}$ denote the model parameters excluding γ.
- Note: the marginal variance of $S_t(x)$ is $\omega^2 = \sigma^2/(1-\gamma^2)$.
- We estimate $\theta_{-\gamma}$ using the estimating equation originating from

$$L(Y_0, \ldots, Y_k; \theta_{-\gamma}) = \prod_{i=1}^{k} L_1(Y_i; \theta_{-\gamma})$$

- We define k in the above equation as the last time point for which the data allow for the estimation of α_k.
- We maximize $L(Y_0, \ldots, Y_k; \theta_{-\gamma})$ with respect to $\theta_{-\gamma})$ using Monte Carlo maximum likelihood.

Proposed inferential methods

Proposed inferential methods

- Defining are suitable prior for γ (the temporal correlation parameter) and for α_t (the overall average prevalence).
  - Uniform discrete prior for γ over $\{0, 1/10, 2/10, \dots, 1\}$.
  - For α_t, we use a tight prior around $\log\{0.01/(1-0.01)\} \approx -4.6$

Proposed inferential methods

- Defining are suitable prior for γ (the temporal correlation parameter) and for α_t (the overall average prevalence).
  - Uniform discrete prior for γ over $\{0, 1/10, 2/10, \dots, 1\}$.
  - For α_t, we use a tight prior around $\log\{0.01/(1-0.01)\} \approx -4.6$
- We use the maximum likelihood estimator distribution as prior for $\theta_{\_\gamma}$.

Proposed inferential methods

- Defining are suitable prior for γ (the temporal correlation parameter) and for α_t (the overall average prevalence).
  - Uniform discrete prior for γ over $\{0, 1/10, 2/10, \dots, 1\}$.
  - For α_t, we use a tight prior around $\log\{0.01/(1-0.01)\} \approx -4.6$
- We use the maximum likelihood estimator distribution as prior for $\theta_{-\gamma}$.
- $\mathcal{Y}_t = \{Y_k : k = 0, \dots, t\}$ (data collected from time 0 to time t)

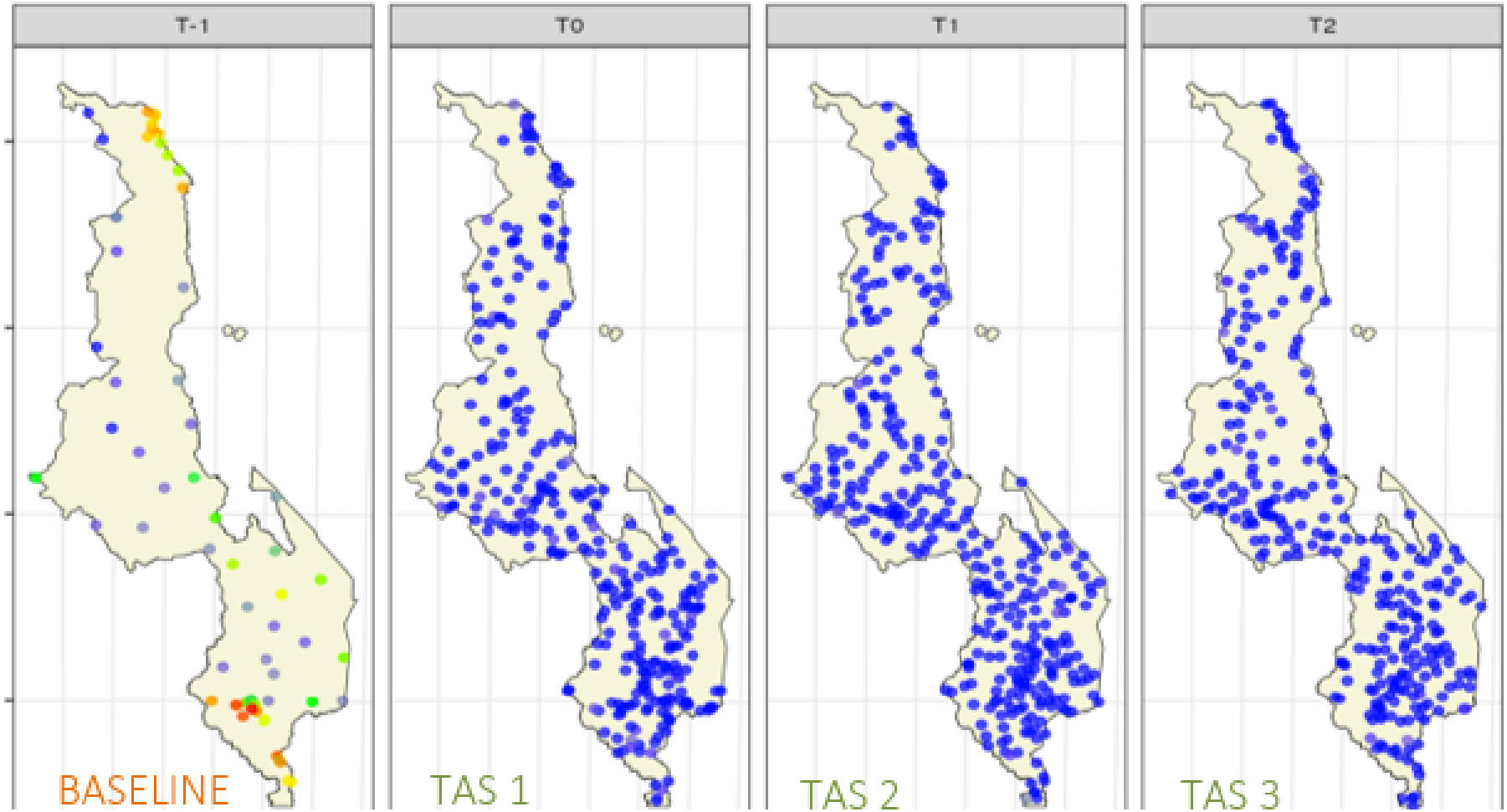
Proposed inferential methods

- Defining are suitable prior for γ (the temporal correlation parameter) and for α_t (the overall average prevalence).
  - Uniform discrete prior for γ over $\{0, 1/10, 2/10, \dots, 1\}$.
  - For α_t, we use a tight prior around $\log\{0.01/(1-0.01)\} \approx -4.6$
- We use the maximum likelihood estimator distribution as prior for $\theta_{\_\gamma}$.
- $\mathcal{Y}_t = \{Y_k : k = 0, \dots, t\}$ (data collected from time 0 to time t)
- At time t, we then sample from $[\mathcal{S}_t|\mathcal{Y}_t]$ to assess the likelihood of resurgence (prevalence above 1%)

Setting up a simulation study

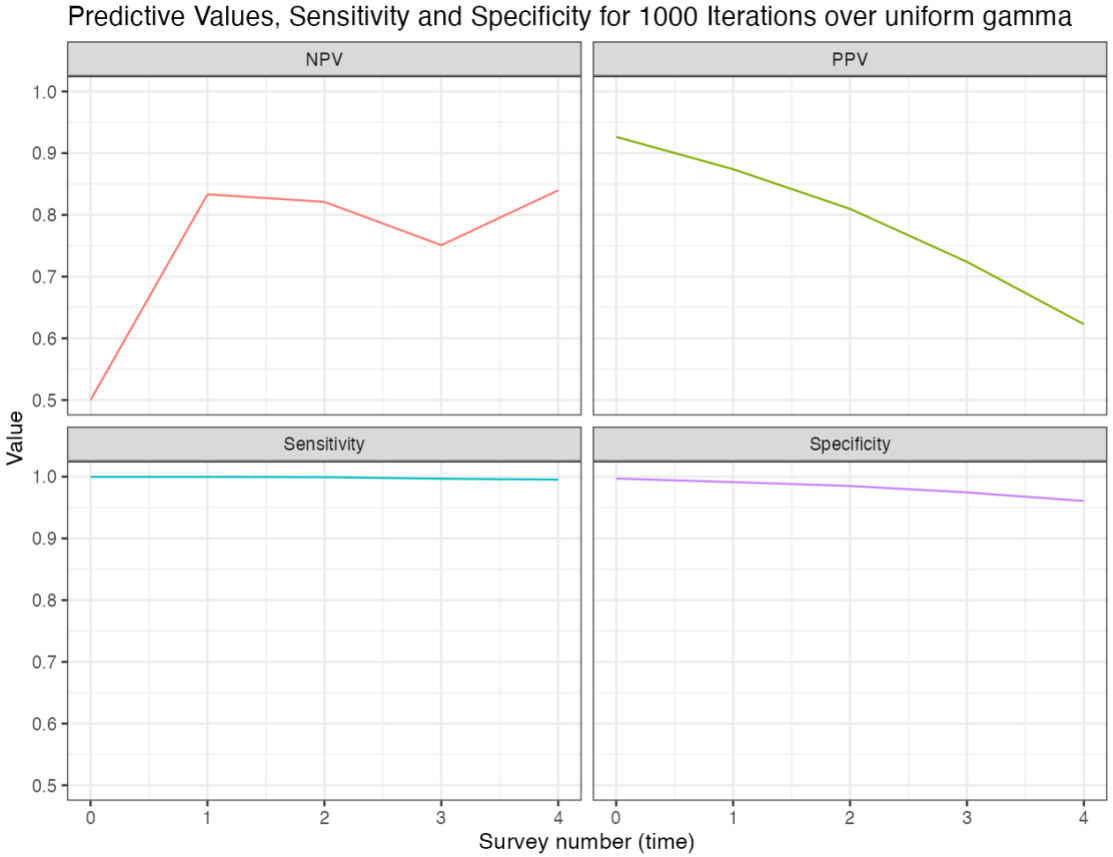
- Objective of the simulation: assess if the modelling framework can detect LF resurgence.
- Parameters of the simulation:
  1) where to place the sentinel sites
  2) how many sentinel sites per district
  3) the resurgence rate
  4) frequency of the sampling at sentinel sites.

LF in Malawi



Inital results from a simulation study

Parameters of the simulation: 1) allocate sentinel sites to locations with highest prevalence; 2) 2 sentinel sites per district; 3) resurgence rate of 5% prevalence increase every year; 4) sampling once per year.



Discussion

- We should also account for heterogeneous intervention coverage (TRANSFIL model)
- Computationally more efficient methods of inference (Kalman filter approximation for Binomial counts?)
- Generalizable to other NTDs.

References

- Galit Shmueli (2010) *To Explain or to Predict?.* Statistical Science 25 (3) 289 - 310 https://doi.org/10.1214/10-STS330
- Giorgi, E., Fronterre, C., Macharia, P., Alegana, V., Snow, R., Diggle, P. (2021) *Model building and assessment of the impact of covariates for disease prevalence mapping in low-resource settings: to explain and to predict.* Journal of the Royal Society Interface. 18:20210104. http://doi.org/10.1098/rsif.2021.0104
- Puranik, A., Diggle, P. J., Odiere, M. R., Gass, K., Kepha, S., Okoyo, C., Mwandawiro, C., Wakesho, F., Omondi, W., Sultani, H. M., Giorgi, E., *Understanding the impact of covariates on the classification of implementation units for soil-transmitted helminths control: A case study from Kenya.* BMC Methodology Research (Under Review). Pre-print available at: https://www.researchsquare.com/article/rs-3334755/v1