University of BRISTOL

MACHINE LEARNING & GLOBAL HEALTH NETWORK

Using Hawkes Processes to determine drivers of transmission

H. Juliette T. Unwin

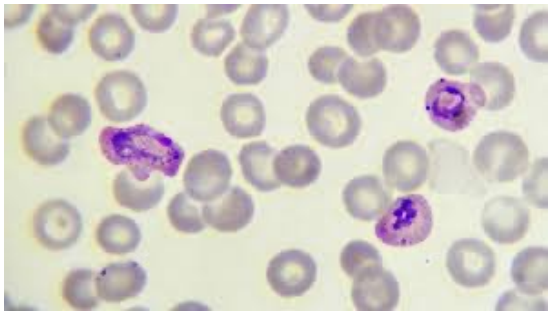
School of Mathematics

juliette.unwin@bristol.ac.uk

bristol.ac.uk

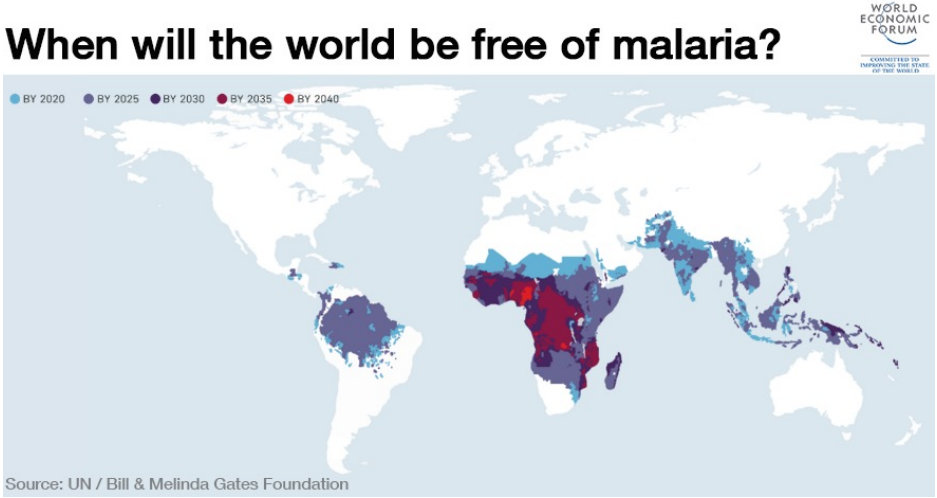
Malaria

- Malaria is a life-threatening disease spread to humans by some types of mosquitoes. It is mostly found in tropical countries.
- The infection is caused by a parasite and does not spread from person to person but requires a vector (the mosquito)
- Symptoms can be mild or life-threatening. Mild symptoms are fever, chills and headache. Severe symptoms include fatigue, confusion, seizures, and difficulty breathing.
- Infants, children under 5 years, pregnant women, travellers and people with HIV or AIDS are at higher risk of severe infection.



Context (near elimination settings)

- According to WHO's latest *World malaria report*, there were an estimated **241 million malaria cases** and **627 000 malaria deaths** worldwide in 2020.
- Malaria elimination is defined as **the interruption of local transmission of a specified malaria parasite species in a defined geographical area as a result of deliberate activities**.



Research question

From a list of all cases of malaria in a country

Want to determine which cases were acquired within the country (person to person transmission) versus those who acquired the disease elsewhere and brought it back with them (imported malaria cases).

Traditionally done using travel history

Hawkes Process

Self exciting point process

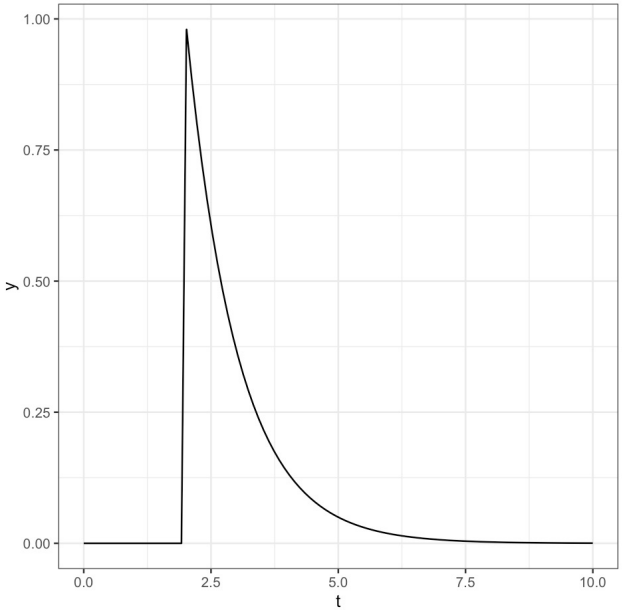
Hawkes Processes

$$\lambda(t) = \mu(t) + \sum_{t>t_i} g(t - t_i)$$

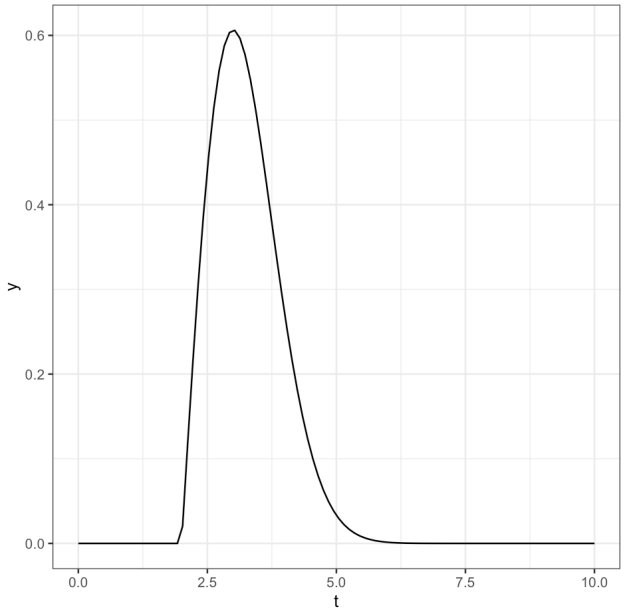
Intensity of infection = Background contribution to intensity. e.g imported malaria cases or spill-over + Contribution to intensity from person to person e.g. direct transmission within country transmission

The kernel

$$\lambda(t) = \mu(t) + \sum_{t>t_i} g(t - t_i)$$



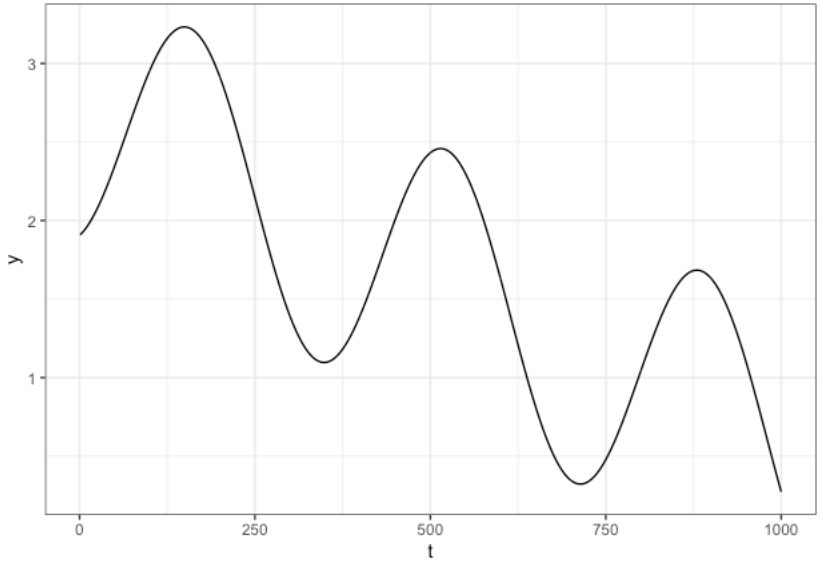
$$g = \alpha e^{-\delta(t-(t_i+\Delta))}$$



$$g = \alpha(t - (t_i + \Delta))e^{-\delta(t-(t_i+\Delta)^2/2}$$

Background term

$$\lambda(t) = \mu(t) + \sum_{t>t_i} g(t - t_i)$$



$$\mu = A + Bt + C\cos\frac{2\pi t}{p} + D\sin\frac{2\pi t}{p}$$

Case reproduction number

Branching factor

$$n^* = \int_0^\infty g(\tau)\, d\tau$$

Analogous to the case reproduction number

$$R = \frac{\alpha}{\delta}$$

The type of data (order 1000 people)

ID	Date of symptoms onset	Date of test	Date of hospitalisation	Sex
01	12/7/23	15/7/23	16/7/23	M
02	16/7/23	17/7/23	17/7/23	F
03	12/7/23	19/7/23		F
04		20/7/23	21/7/23	M
05	19/7/23	20/7/23		M

Inference methods

- Maximum likelihood estimation
- Expectation - Maximisation
- Bayesian Inference

Maximum Likelihood Estimation

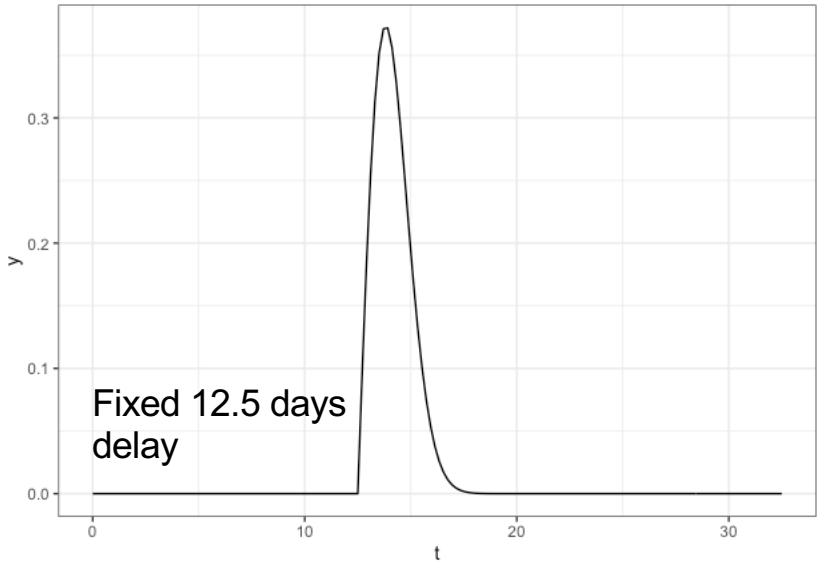
Use an optimizer to minimize the negative log-likelihood

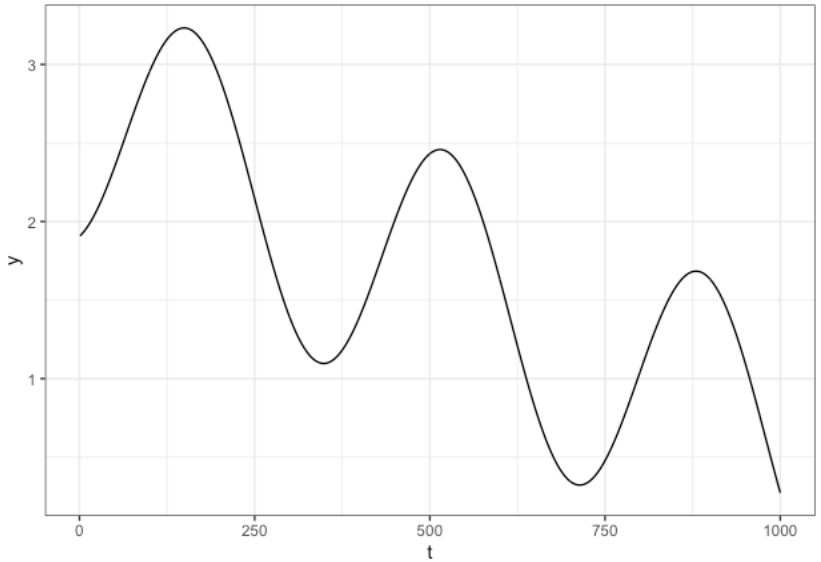
$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \left(-\log L(\theta) \right),$$

$$\log L(\theta) = \sum_{i=1}^{n} \log \lambda(t_i) - \int_0^T \lambda(\tau)\, \mathrm{d}\tau$$

https://link.springer.com/article/10.1007/s11009-011-9272-5

Estimate the background term and kernel





How to you account for uncertainty?

ALGORITHM 6. Using the parameter values $\hat{\Theta}$ from a previously fitted model, and starting with $i = 1$:

1. Using a simulation algorithm from Section 3.3, simulate a new dataset in the same spatio-temporal region.
2. Fit the same model to this new data, obtaining new parameter values $\hat{\Theta}^{(i)}$.
3. Repeat steps 1 and 2 with $i = i + 1$, up to some prespecified number of simulations B (e.g., 1000).

   (Alternately, the algorithm can be adaptive, by checking the confidence intervals after every b steps and stopping when they seem to have converged.)
4. Calculate bootstrap 95% confidence intervals for each parameter by using the 2.5% and 97.5% quantiles of the estimated $\hat{\Theta}^{(i)}$.

	Fitted value [95% confidence interval]
α	0.0308 [0.0126, 0.0676]
δ	0.0789 [0.0248, 0.1681]
A	2.1369 [-7.1932, 2.8548]
B	-0.0018 [-0.1689, -0.0014]
M	-0.5836 [-0.8269, 10.2468]
N	0.3262 [-1.0208, 9.8495]

https://projecteuclid.org/journals/statistical-science/volume-33/issue-3/A-Review-of-Self-Exciting-Spatio-Temporal-Point-Processes-and/10.1214/17-STS629.full

Expectation - Maximisation

Notation switch:

$$\omega = \frac{\alpha}{\delta}$$

$$\implies \lambda(t) = \mu + \sum_{t>t_i} \omega\delta e^{-\delta(t-t_i)}.$$

- Expectation

$$P_{ij}^k = \frac{\omega^k \delta^k e^{-\delta^k (t_i - t_j)}}{\mu^k + \sum_{h=1}^{i-1} \omega^k \delta^k e^{-\delta^k (t_i - t_h)}}$$

$$P_{ii}^k = \frac{\mu^k}{\mu^k + \sum_{h=1}^{i-1} \omega^k \delta^k e^{-\delta^k (t_i - t_h)}}$$

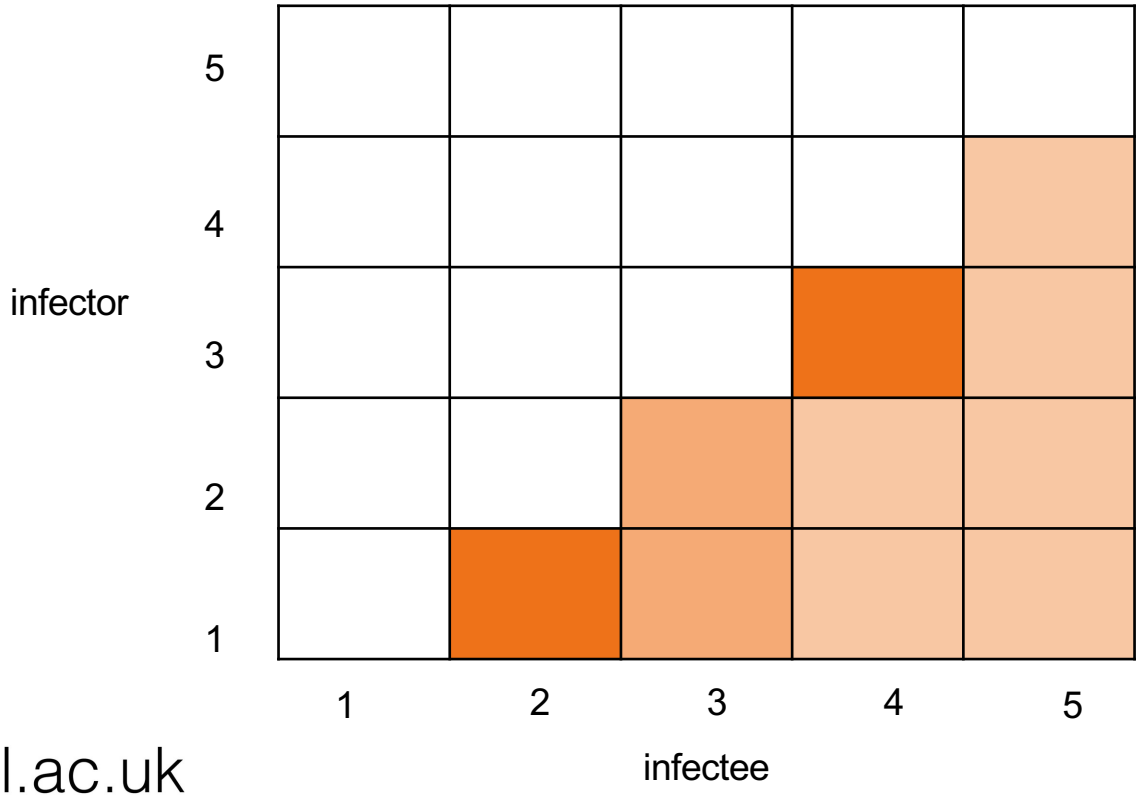
- Maximisation

$$\mu^{k+1} = \frac{1}{T_{max}} \sum_{i=1}^{N} P_{ii}^k$$

$$\omega^{k+1} = \frac{\sum_{i>j} P_{ij}^k}{N - \sum_{i=1}^{N} e^{-\delta^k (T_{max} - t_i)}}$$

$$\delta^{k+1} = \frac{\sum_{i>j} P_{ij}^k}{\sum_{i>j} (t_i - t_j) P_{ij}^k + \omega^k \sum_{i=1}^{N} (T_{max} - t_i) e^{-\delta^k (T_{max} - t_i)}}$$

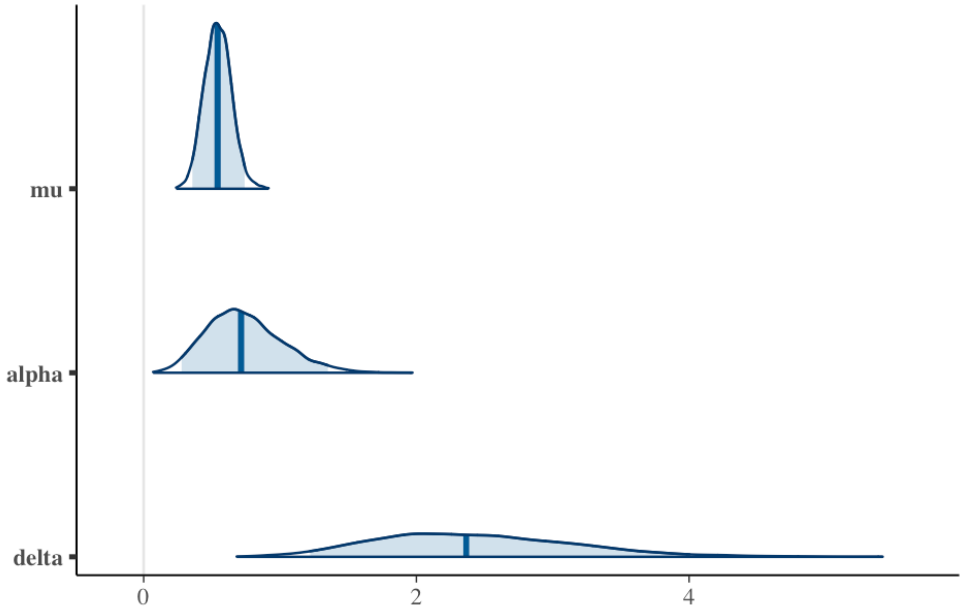
Problem with temporal only data



Bayesian Inference

$$\lambda = \mu + \sum_{t>t_i} \alpha e^{-\delta(t-t_i)}$$
$$= 0.5 + \sum_{t>t_i} 1e^{-2(t-t_i)}$$

```
    real log_likelihood(real mu, real alpha, real delta, vector events_list, int N, real max_T){

        // first term

        vector[N] differences_from_max_T = max_T - events_list;
        vector[N] summands = exp(-delta * differences_from_max_T) - 1;

        vector[N] within_max_T_Mask = non_negative_mask(differences_from_max_T);
        summands = summands .* within_max_T_Mask;

        real first = mu * max_T - (alpha / delta) * sum(summands);

        // second term

        matrix[N, N] differences_mat = self_differences(events_list);
        matrix[N, N] inner_sum_mat = exp(-delta * differences_mat);
        inner_sum_mat = zero_above_diagonal(inner_sum_mat);

        vector[N] term_inside_log = mu + alpha * rowsum(inner_sum_mat);

        vector[N] second_sum_terms = log(term_inside_log);

        real second = sum(second_sum_terms);

        return -first + second;
    }
}

data {
    int<lower=0> N;
    vector[N] events_list;
    real max_T;
}

parameters {
    real mu;
    real <lower=0> alpha;
    real <lower=0> delta;
}

model {
    mu ~ normal( 1 , 1 );
    alpha ~ normal( 1 , 1 );
    delta ~ normal( 2 , 1 );

    target += log_likelihood(mu, alpha, delta, events_list, N, max_T);
}
```



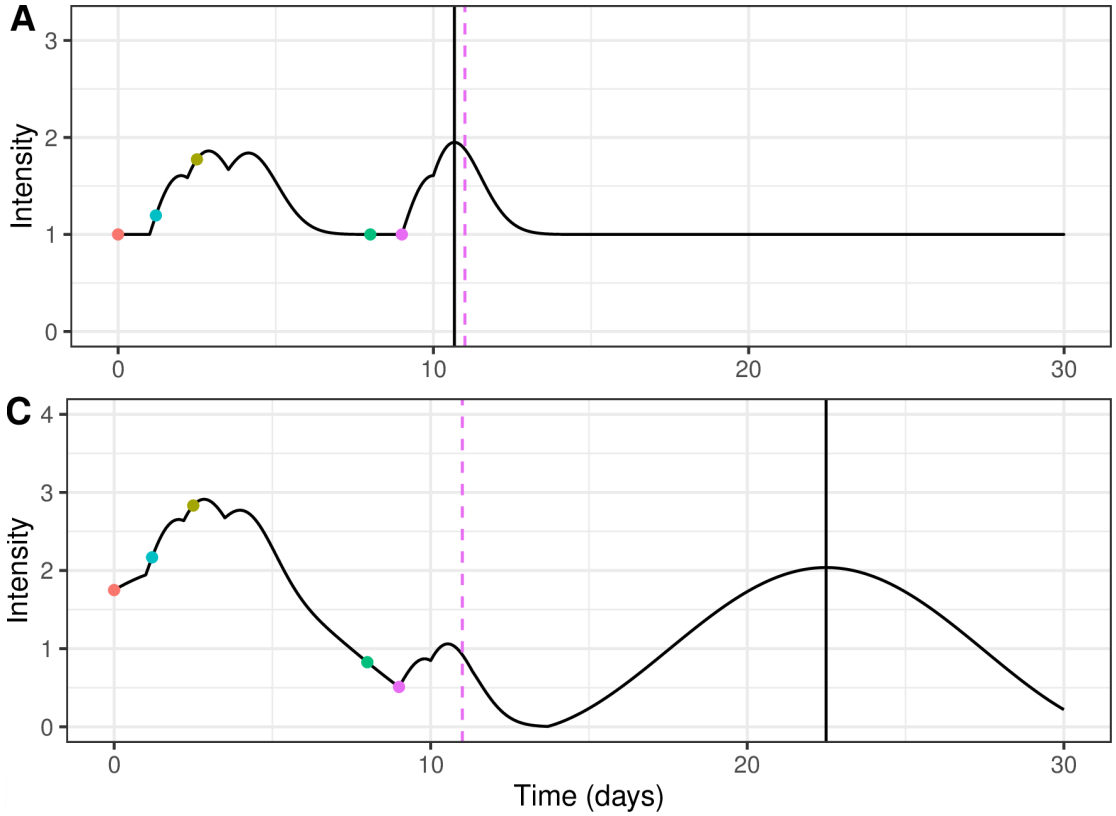
Simulate to see if tease apart contributions from importations or not

One method is thinning.

Supplementary Algorithm 1: Ogata's thinning algorithm adapted for Hawkes Processes

Set current time $t = 0$ and event counter $i = 0$;
while $t \leq T_{max}$ **do**
  (a) Calculate the upper bound of the Hawkes intensity $\lambda^\star = \lambda(t^+)$. If an event occurs at time t it is accounted for;
  (b) Sample inter-arrival time by drawing $u \sim U(0,1)$ and letting $\tau = -\frac{\ln u}{\lambda}$;
  (c) Update current time: $t = t + \tau$;
  (d) Draw $s \sim U(0,1)$.;
  if *If* $s \leq \frac{\lambda(t)}{\lambda^\star}$ **then**
    Accept the current sample and let $t_i = t$ and $i = i + 1$;
  else
    Reject the sample;
  end
end

It’s hard for this malaria set up



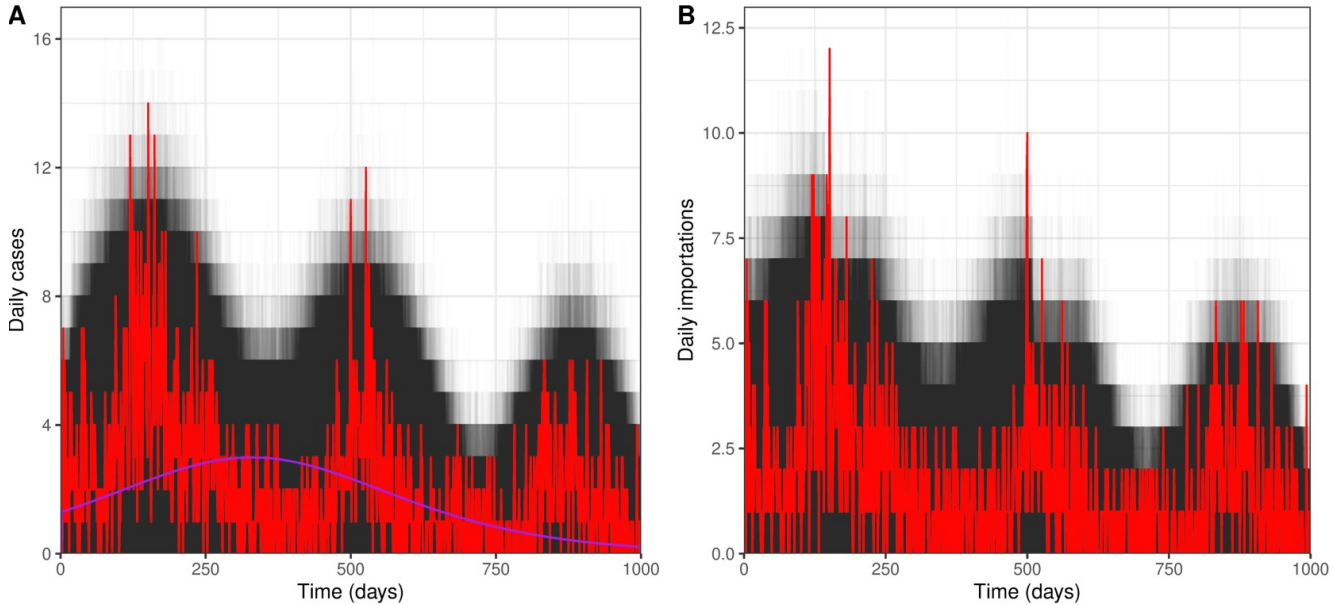
bristol.ac.

Can use cluster-based simulation instead

Algorithm 2 Simulation by Cluster Structure

Inputs: T_{max}, θ
Simulate $t_1, \ldots, t_k$, the times of exogenous events
$G_0 = \{t_1, \ldots, t_k\}$
$N_0 = \mathbf{card}(G_0)$
$\ell = 0$
while $G_\ell \neq \emptyset$ **do**
  for $i = 1$ to N_ℓ **do**
    Simulate C_i, the number of offspring of event i
    Simulate $O_1, \ldots, O_{C_i}$, the inter-arrival times of the offspring events
  end for
  $\ell = \ell + 1$
  $G_\ell = \{G_{\ell-1} + \bigcup_{i=1}^{N_\ell} O_1, \ldots, O_{C_i}\}_{<T_{Max}}$
  $N_\ell = \mathbf{card}(G_\ell)$
end while
return $\bigcup_{\ell=0}^{\ell_{Max}} G_\ell$

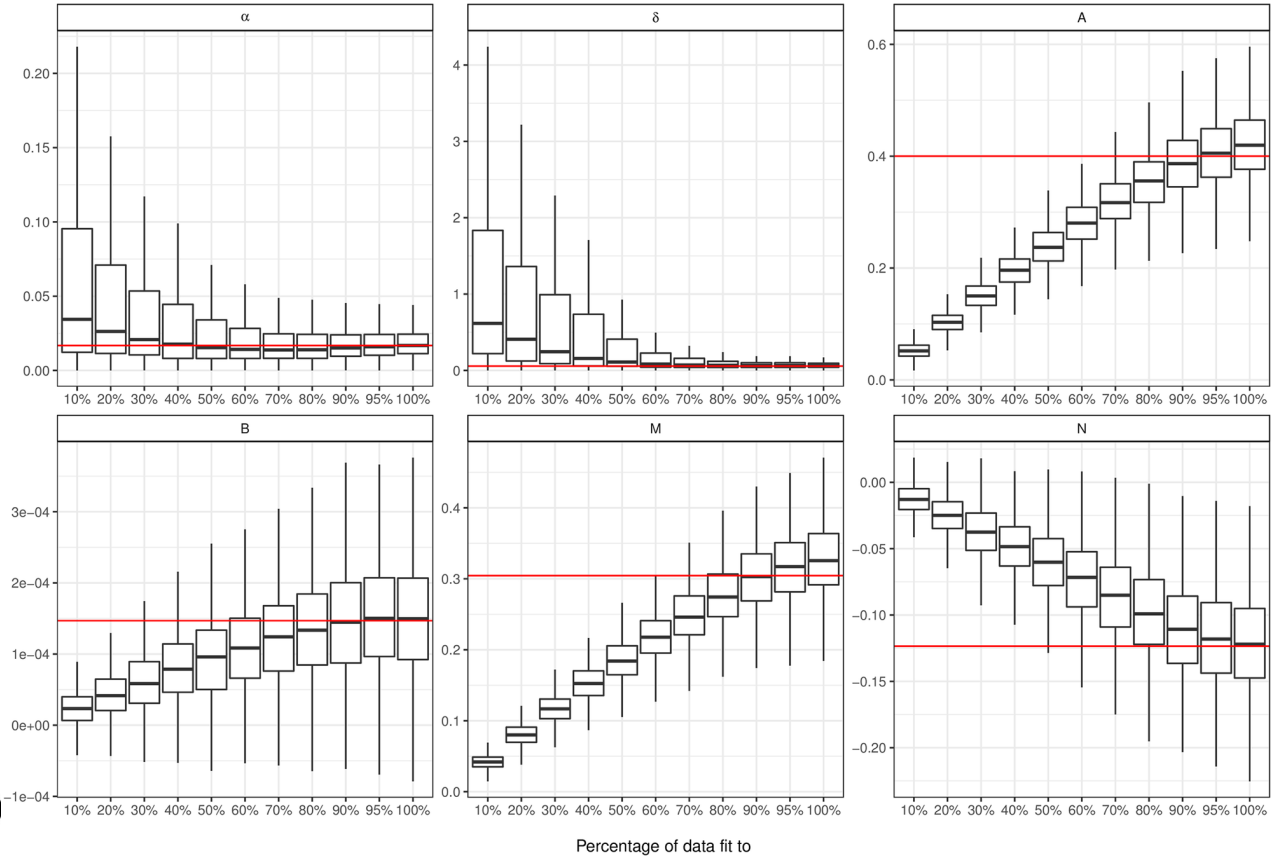
Malaria Cases in Yunnan Province, China



bristol.ac.uk

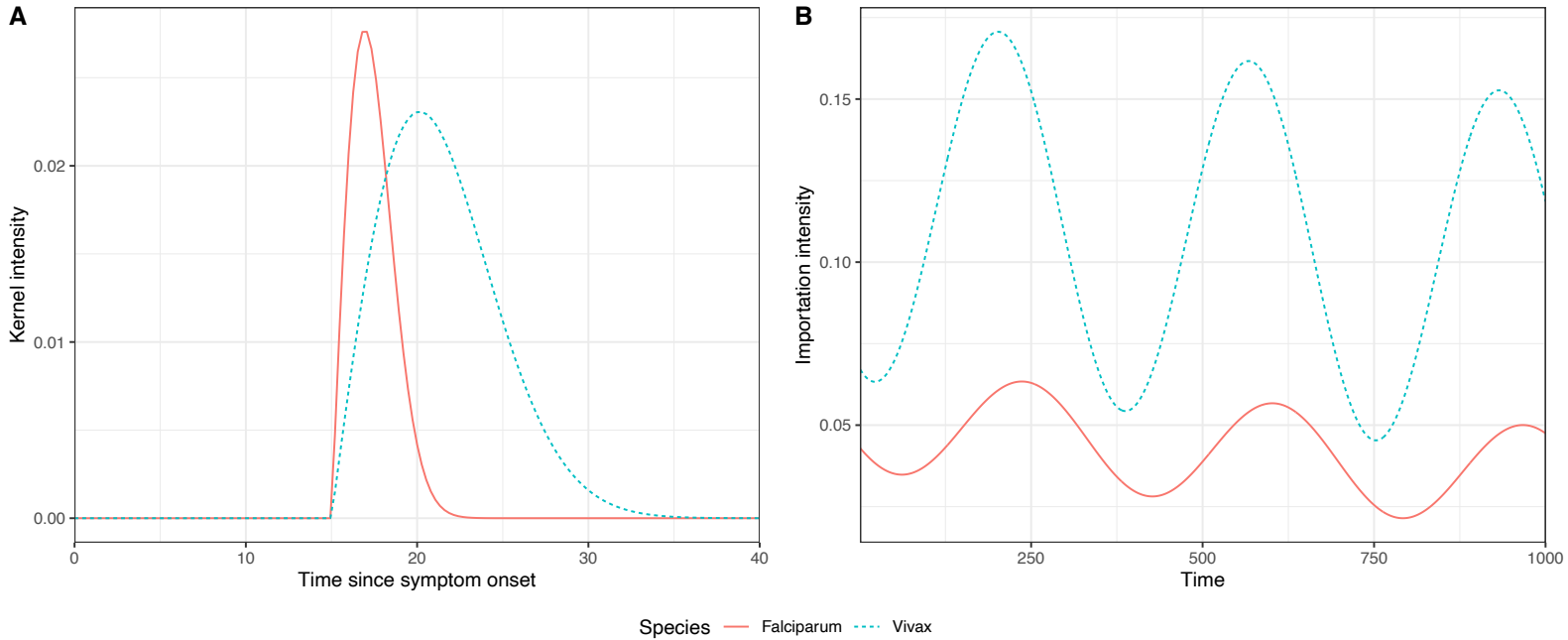
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008830

Missing data

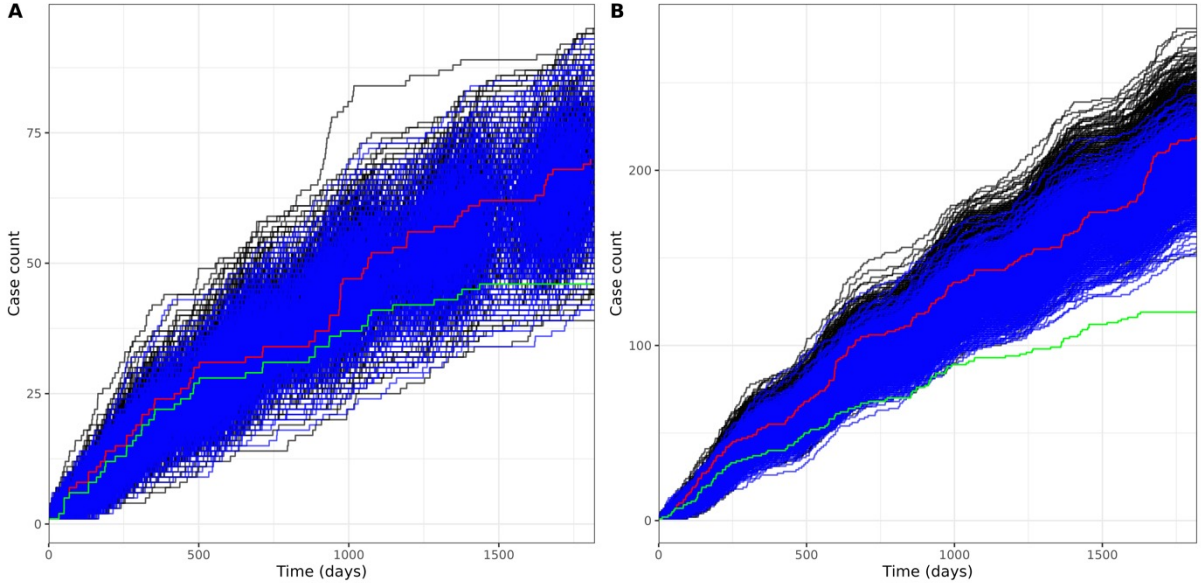


bristol.a

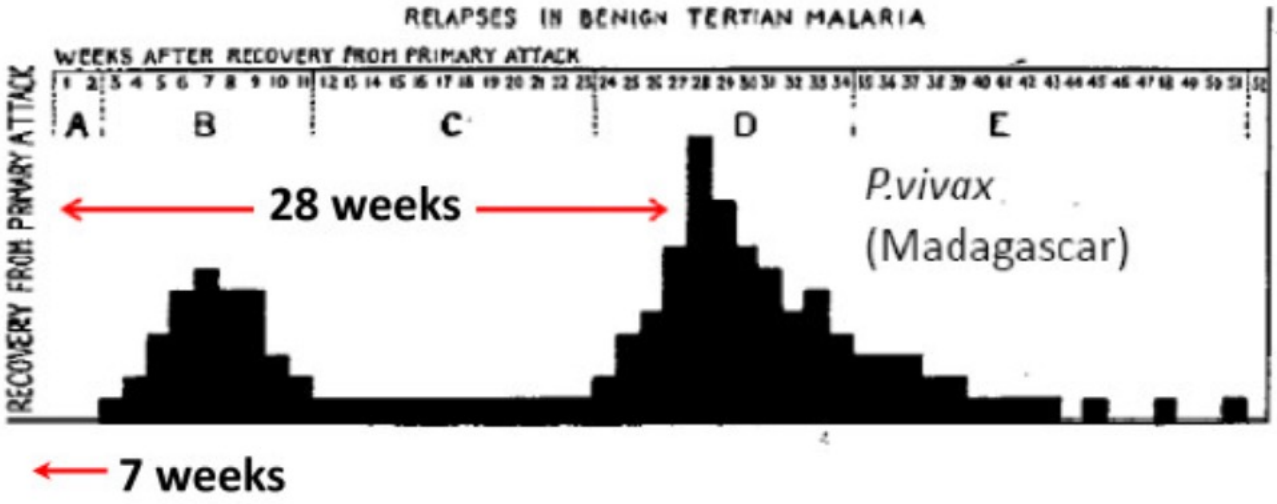
Bhutan


A
Kernel intensity
0.00
0.01
0.02
0
10
20
30
40
Time since symptom onset
B
Importation intensity
0.05
0.10
0.15
250
500
750
1000
Time
Species
Falciparum
Vivax

Bhutan



Vivax malaria relapses



bristol.ac.uk

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3228849/

University of BRISTOL

This led to a COVID-19 renewal model

bristol.ac.uk

March 2020

- Didn’t know too much about this novel pathogen
- Had multiple cases each day
- Were not observing the whole line list

Renewal based models for outbreak response

$$\mathbb{E}[Z(t)] = f(t) = \underbrace{\mu(t)}_{exogenous} + R_0 \underbrace{\int_{\tau=0}^{t} f(t-\tau)g(\tau)d\tau}_{endogenous}$$

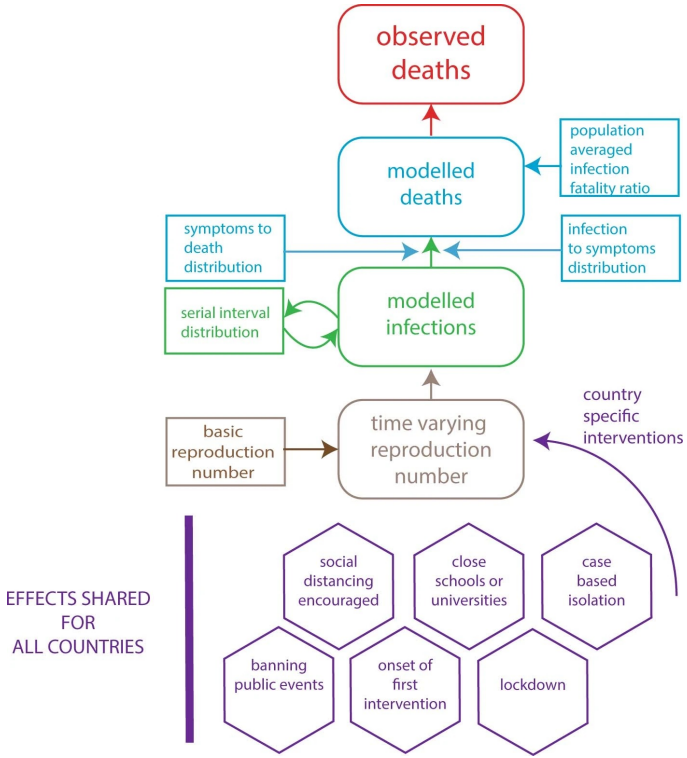
Expectation of number of infections

Imported infections

Basic reproduction number

Serial interval distribution (infection profile)

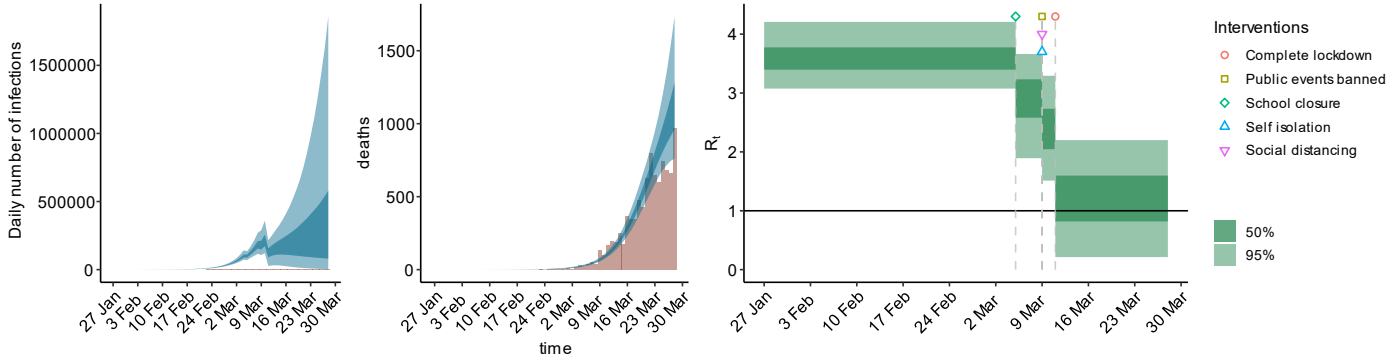
Embedded in a hierarchical framework



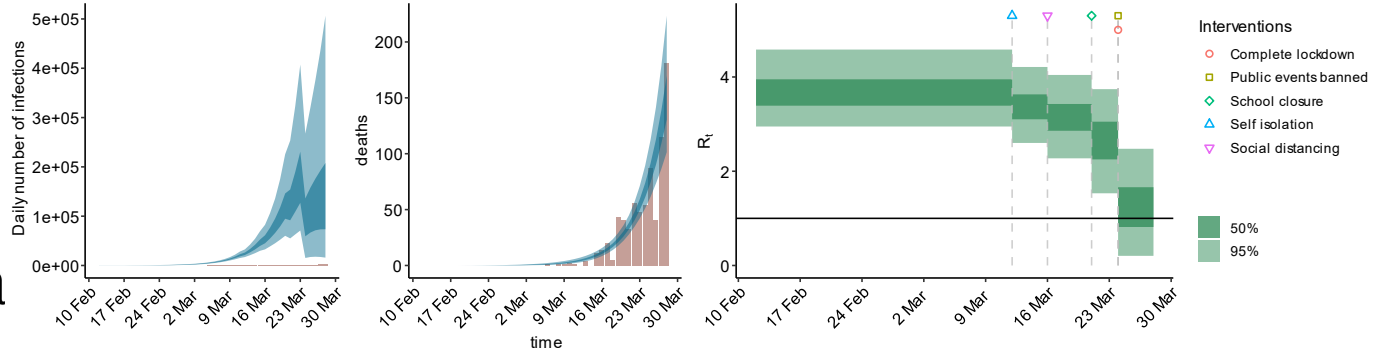
Renewal model

Estimating R_t

(F) Italy



(K) United Kingdom

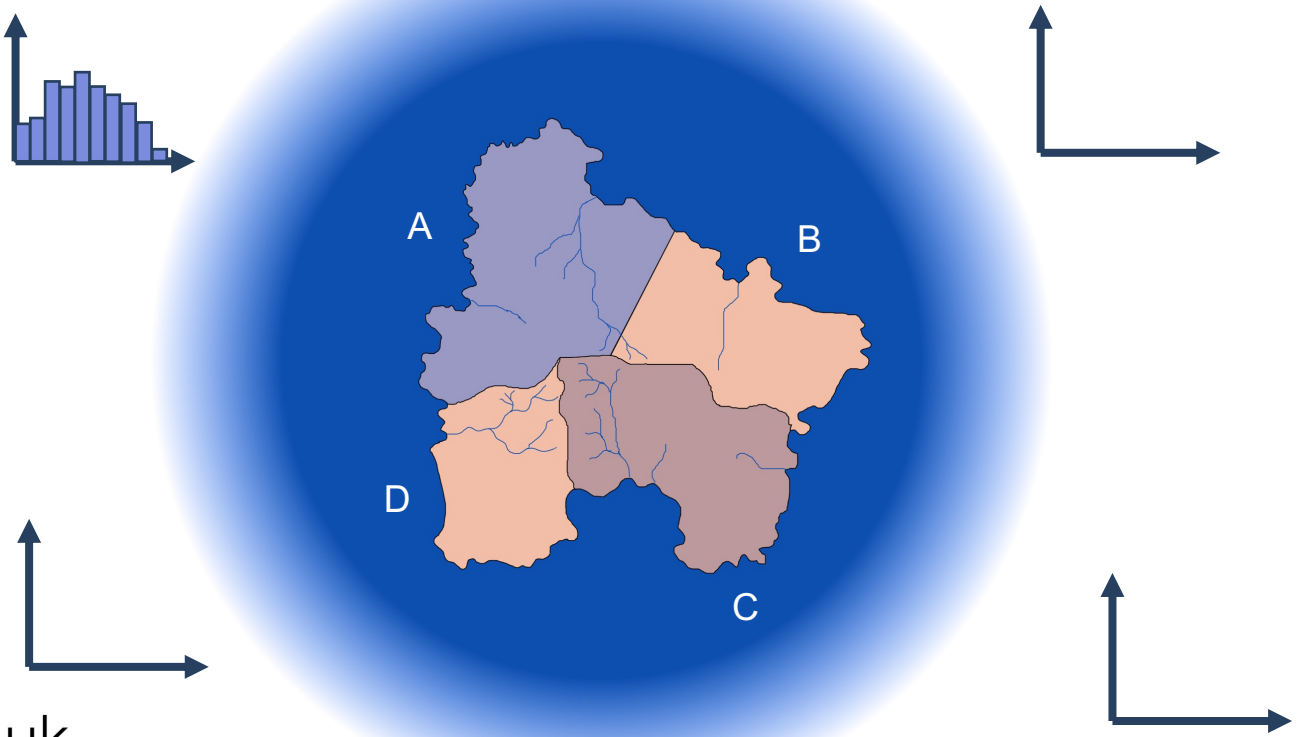


University of BRISTOL

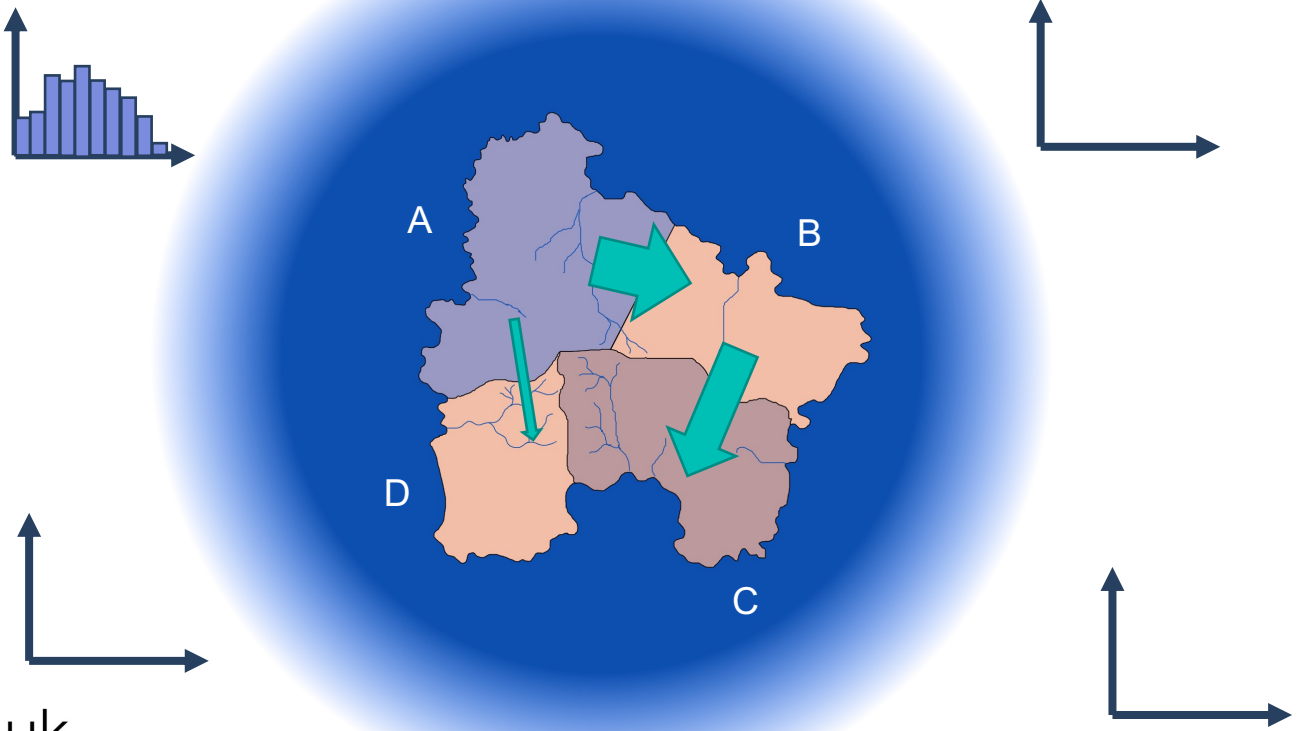
Future directions

bristol.ac.uk

1) A spatial-temporal model



1) A spatial-temporal model



1) A spatio-temporal renewal model

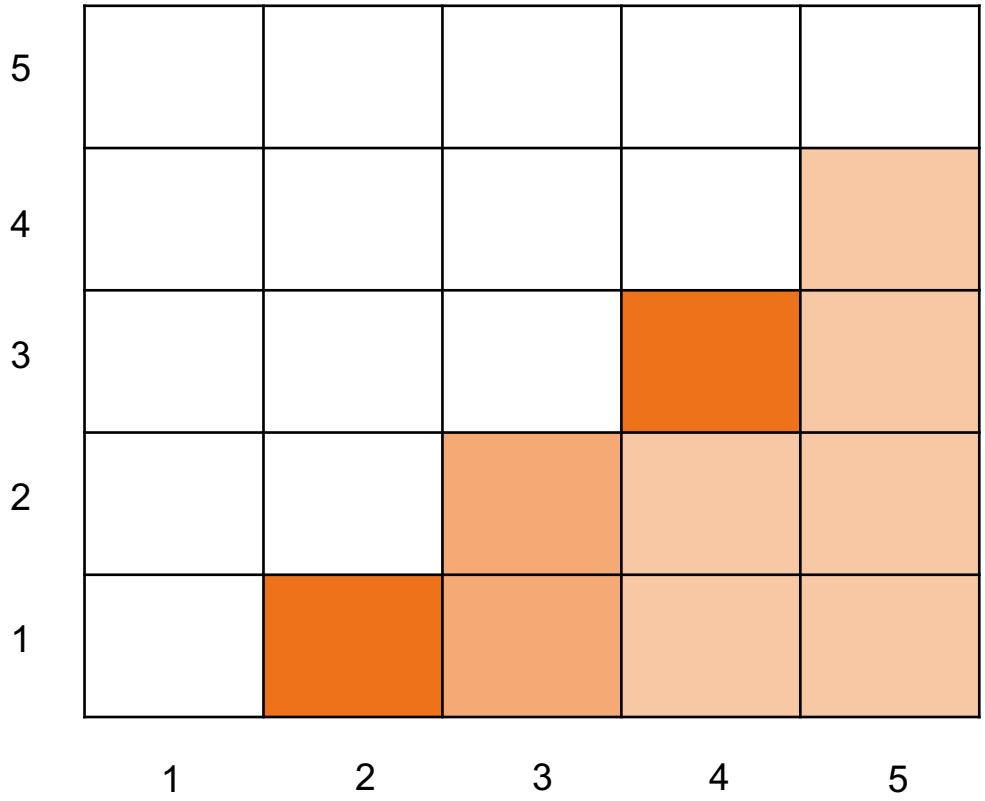
$$i_{t,m} = \underbrace{R_{t,m}\kappa_m^{\text{remain}} \sum_{\tau}^{t-1} i_{\tau,m} g_{t-\tau}}_{\text{within region transmission}} + \underbrace{\sum_{n \neq m} R_{t,m}\kappa_{nm}^{\text{enter}} \sum_{\tau}^{t-1} i_{\tau,n} g_{t-\tau}}_{\substack{\text{transmission from people commuting} \\ \text{in and bringing infections with them}}} + \underbrace{\sum_{n \neq m} R_{t,n}\kappa_{mn}^{\text{leave}} \sum_{\tau}^{t-1} i_{\tau,n} g_{t-\tau}}_{\substack{\text{transmission from people commuting} \\ \text{out and returning with infection}}}$$

1) Problems

- Have discrete spatial regions where health data is recorded so can’t do anything continuous in space
- Need to know Rt for each region to calculate the number of infections (hard to parallelise)
- Gets expensive when have multiple regions because infections from all regions can infect each other region
- Stan is slow

2) Recreating transmission trees

2) Problem with temporal only Hawkes



2) Adding a spatio-temporal component

$$\lambda(t) = \mu(t) + \sum_{t>t_i} g(t-t_i)\, h(x-x_i)$$

2) Is this enough?

- How to encode spatial regions in this continuous framework?
- Would that be enough to recreate the chains?
- Adding extra genetic information?

University of BRISTOL

Any questions / thoughts / ideas?

Thanks to Aisling Stokes and Ethan Honey for some slides / code