Emulation-based Inference for Spatial Infectious Disease Models

Dr. Rob Deardon

Department of Production Animal Health, Faculty of Veterinary Medicine
Department of Mathematics & Statistics, Faculty of Science

(Joint work with Gyanendra Pokharel, University of Guelph)

Infectious Disease Transmission Models

Goal:

Use data to build a mathematical model for how disease spreads through some population

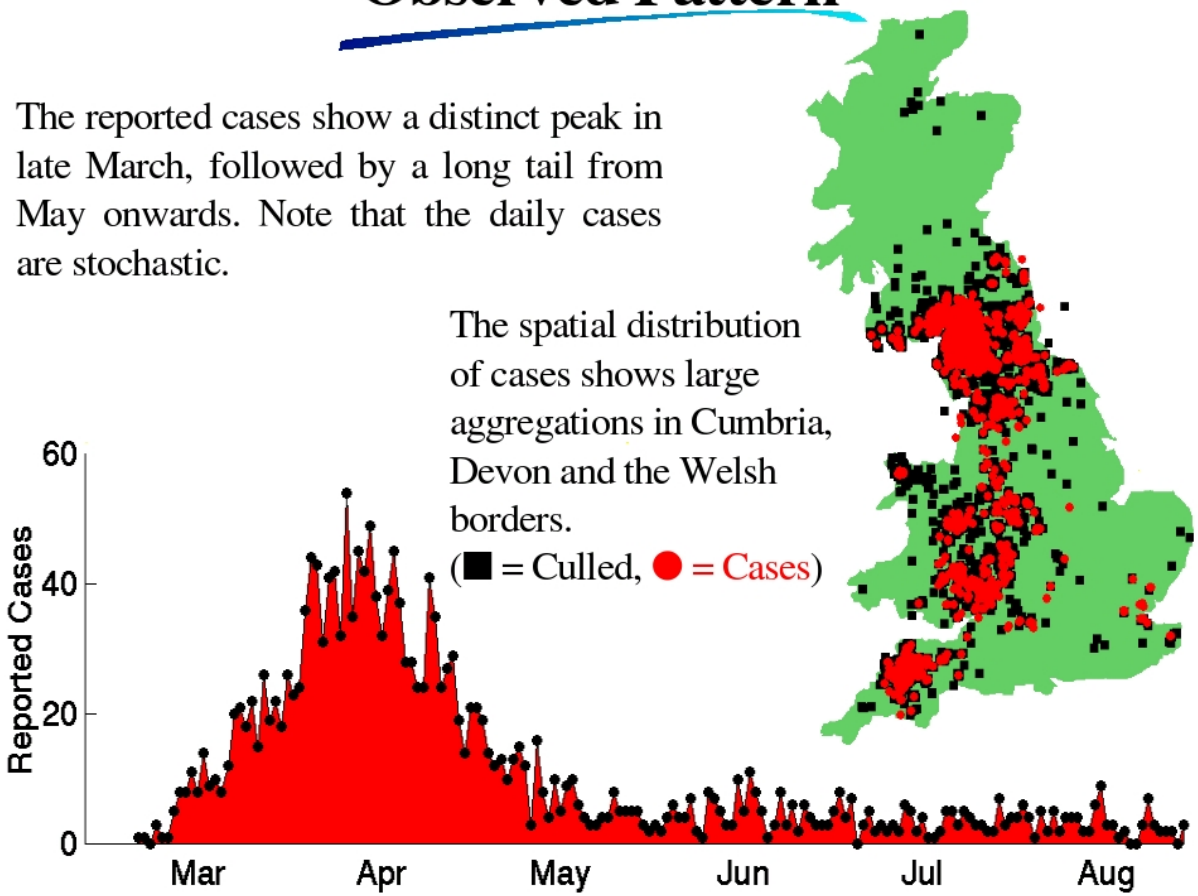
Why?

To help us understand how disease spreads
To help predict what may occur
To help understand how to control disease
To help design optimal vaccination/culling/surveillance policies
To quantify risk/uncertainty associated with any of the above

Observed Pattern

The reported cases show a distinct peak in late March, followed by a long tail from May onwards. Note that the daily cases are stochastic.

The spatial distribution of cases shows large aggregations in Cumbria, Devon and the Welsh borders.
(■ = Culled, ● = Cases)



Discrete Time Individual-based Modelling Framework

- We assume a population of n individuals : $i = 1, \ldots, n$:
  (e.g. herds; animals in herd; plants; fields of plants;
  humans; schools; census divisions)
- We assume that at discrete time point t, $t = 1, \ldots t_{max}$, each individual can be in one of three states:

S	Susceptible	doesn't have disease; can contract it
I	Infectious	has contracted the disease; can pass it on
R	Removed	been removed from the susceptible population e.g. died from the disease; e.g. isolated from the susceptible population; e.g. recovered and developed immunity

- Individuals moves through states in the following way:
  $S \rightarrow I \rightarrow R$

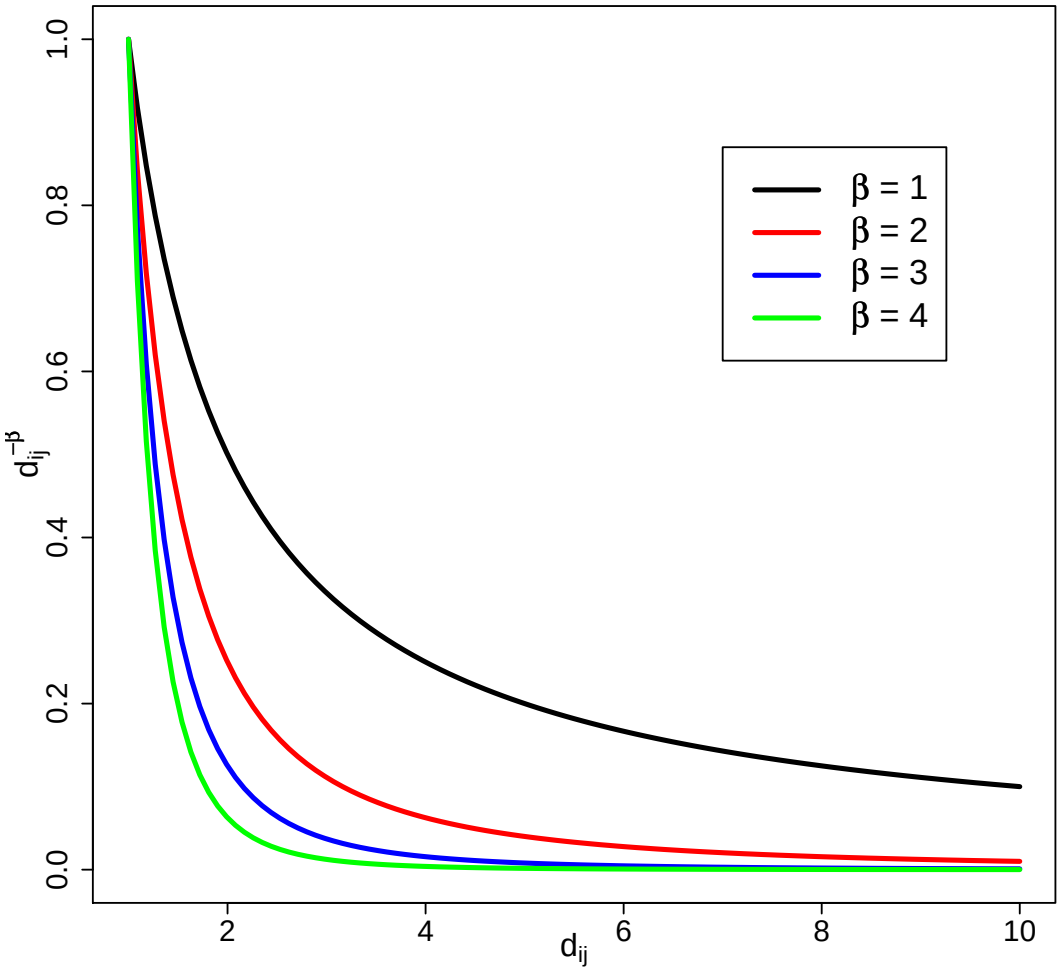
Model 1: Power-law Spatial Model

- The probability of susceptible individual i being infected at time t is given by:

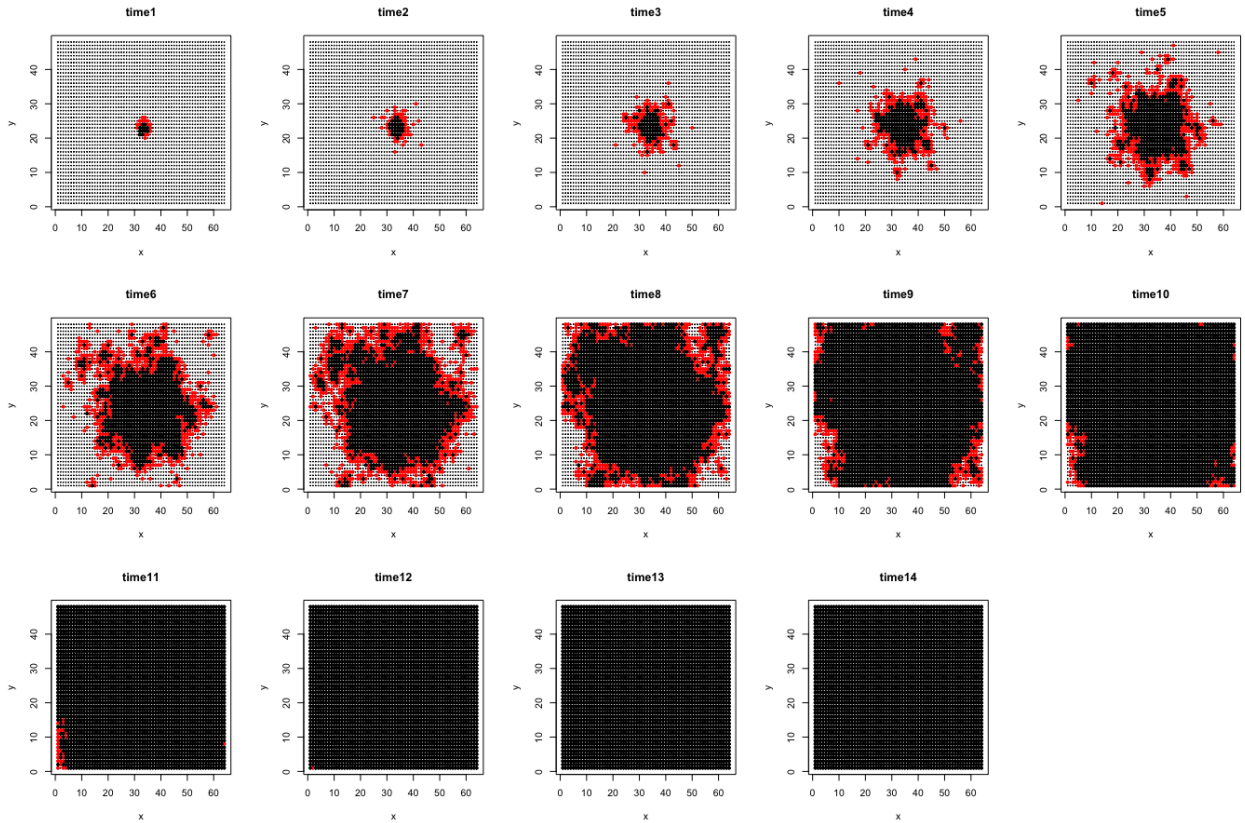
$$P(i,t) = 1 - \exp\left[-\alpha \sum_{j \in I(t)} d_{ij}^{-\beta}\right]$$

- where
  - $I(t)$ is the set of infectious individuals at time t
  - $\mathcal{K}_{ij} = d_{ij}^{-\beta}$ is a power-law **infection/distance kernel**
  - d_{ij} is the distance between individuals i and j
  - $\alpha > 0$ is an 'infectivity' parameter
  - $\beta > 0$ is a 'spatial' parameter

Infection Kernel versus Distance ($\mathcal{K}_{ij} = d_{ij}^{-\beta}$)



Power-law spatial-ILM simulation across grid



From Vrbik et al (2012) in Bayesian Analysis, 7(3), 615 638..

Model 2: Network Model

- The probability of susceptible individual i being infected at time t is given by:

$$P(i,t) = 1 - \exp\left[-\alpha \sum_{j \in I(t)} c_{ij}\right]$$

- where
  - $I(t)$ is the set of infectious individuals at time t
  - $\mathcal{K}_{ij} = c_{ij}$ is the $(i,j)^{\text{th}}$ element of a contact matrix
  - $c_{ij} = \begin{cases} 1 & \text{if i and j have contact} \\ 0 & \text{otherwise} \end{cases}$
  - $\alpha > 0$ is an 'infectivity' parameter

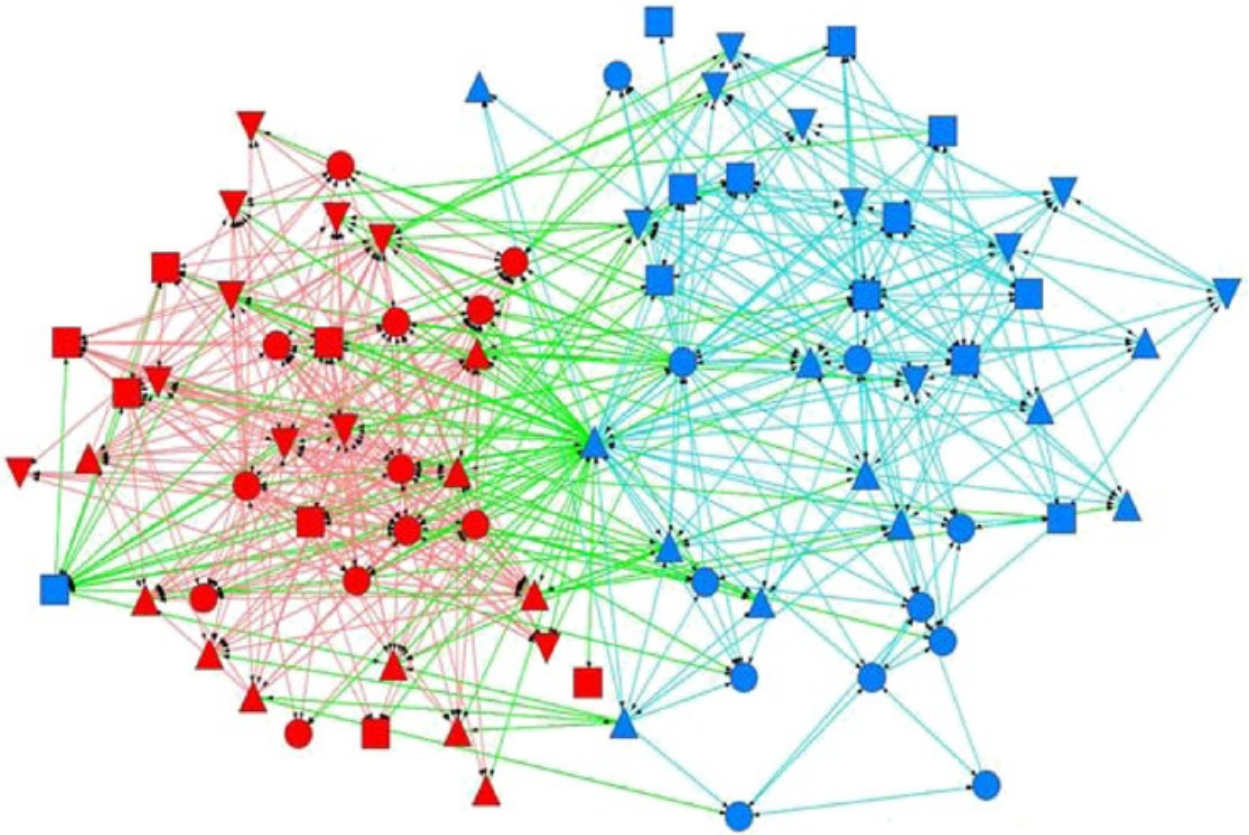
Model 2b: Network Model

- The probability of susceptible individual i being infected at time t is given by:

$$P(i,t) = 1 - \exp\left[-\sum_{j \in I(t)} \left(\alpha_1 c_{ij}^{(1)} + \alpha_2 c_{ij}^{(2)} + \alpha_3 c_{ij}^{(3)} + \ldots\right)\right]$$

- where
  - $I(t)$ is the set of infectious individuals at time t
  - $\mathcal{K}_{ij} = c_{ij}^{(k)}$ is the $(i,j)^{\text{th}}$ element of a contact matrix k
  - $c_{ij}^{(k)} = \begin{cases} 1 & \text{if i and j have contact within network } k \\ 0 & \text{if i and j otherwise} \end{cases}$
  - $\alpha_1, \alpha_2, \ldots > 0$ are 'infectivity' parameter

†Friendship network – Pennsylvanian Elementary School



†Cauchemez et al., 2011, *Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza*, PNAS, 108(7): 2825-30.

Model 4: Coalescent (Genetic) Network-Spatial Model

- Imagine the **'true'** probability of infection is given by:

$$P(i,t) = 1 - \exp\left[-\sum_{j \in I(t)} \left(\alpha_0 d_{ij}^{-\beta} + \alpha_1 c_{ij} + \alpha_2 X_{1j}\right)\right]$$

but we don't observe c_{ij} and X_{1j}

Model 4: Coalescent (Genetic) Network-Spatial Model

- Imagine the **'true'** probability of infection is given by:

$$P(i,t) = 1 - \exp\left[-\sum_{j \in I(t)} \left(\alpha_0 d_{ij}^{-\beta} + \alpha_1 c_{ij} + \alpha_2 X_{1j}\right)\right]$$

but we don't observe c_{ij} and X_{1j}

So we fit a model with

$$P(i,t) = 1 - \exp\left[-\sum_{j \in I(t)} \left(\alpha_0 d_{ij}^{-\beta}\right)\right]$$

- Spatial effect will be estimated with less precision (and be biased).

Model 4: Coalescent (Genetic) Network-Spatial Model

- Now imagine we have collected sequence information on the pathogen in the blood of infected invdividuals
- Thus we can fit a model:

$$P(i,t) = 1 - \exp\left[-\sum_{j \in I(t)} \left(\left[\alpha_0 d_{ij}^{-\beta}\right] g_{ij}\right)\right]$$

where

$g_{ij} \in \{0,1\}$ is a measure of genetic similarity between pathogen sequences i and j

- Therefore, should get improved estimation of spatial effect...

Model 5: From Deardon *et al.* (2010)

The probability of susceptible individual i being infected at time t is given by:

$$P(i,t) = 1 - \exp\left(-\mathbf{SN}_i^{\boldsymbol{\psi}_S}\left[\left\{\sum_{j\in I(t)} \mathbf{TN}_j^{\boldsymbol{\psi}_T} K_A(d_{ij})\right\} + \epsilon|I(t)|\right]\right). \qquad (1)$$

where

$$\kappa(i,j) = K_A(d_{ij}) = \begin{cases} k_0 & 0 < d_{ij} \le \delta_0 \\ d_{ij}^b & \delta_0 < d_{ij} \le \delta_{max} \\ 0 & \text{otherwise} \end{cases};$$

$$\mathbf{SN}_i^{\boldsymbol{\psi}_S} = (S_s \;\; S_c)\begin{pmatrix} {N_{i,s}}^{\psi_{S,s}} \\ {N_{i,c}}^{\psi_{S,c}} \end{pmatrix}; \; \mathbf{TN}_j^{\boldsymbol{\psi}_T} = (T_s \;\; T_c)\begin{pmatrix} {N_{j,s}}^{\psi_{S,s}} \\ {N_{j,c}}^{\psi_{S,c}} \end{pmatrix};$$

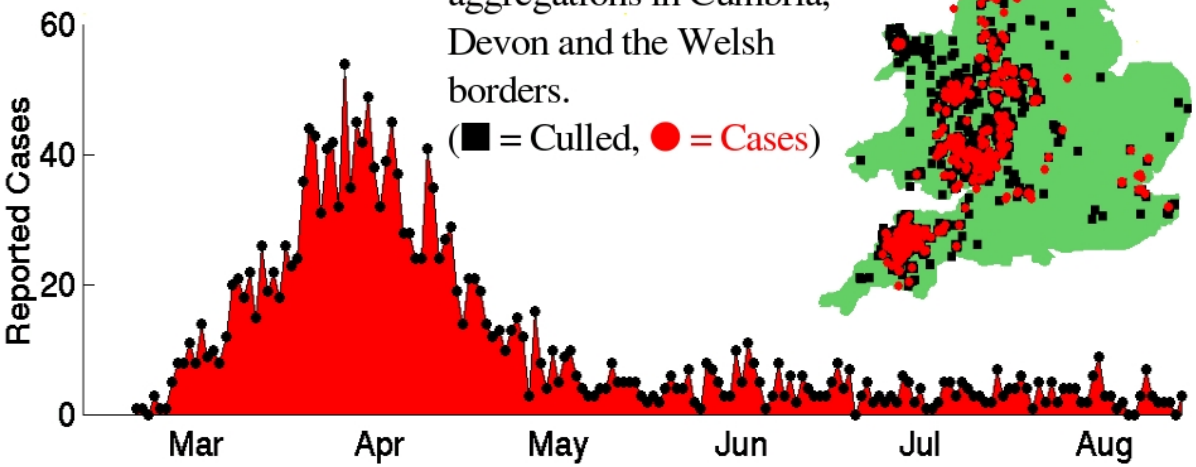
and $|I(t)|$ is the number of elements of the set, $I(t)$.

UK 2001 foot-and-mouth disease epidemic

Observed Pattern

The reported cases show a distinct peak in late March, followed by a long tail from May onwards. Note that the daily cases are stochastic.

The spatial distribution of cases shows large aggregations in Cumbria, Devon and the Welsh borders.
(■ = Culled, ● = Cases)



Infectious Period ($I \rightarrow R$)

- Problem: infection times and removal times (and thus infectious periods) are usually unknown.
- Can be modelled in various ways
  - Typically: assume infectious periods are 'random effects' that follow some distribution (typically exponential) then use data augmentation to infer infection times, removal times and infectious periods
- Here (for simplicity) we will assume:
  - we know when individuals become infected
    and that they all have the same known infectious period.

Model of focus: Power-law Spatial Model

- The probability of susceptible individual i being infected at time t is given by:

$$P_{it} = 1 - \exp\left[-\alpha \sum_{j \in I(t)} d_{ij}^{-\beta}\right]$$

- where
  - $I(t)$ is the set of infectious individuals at time t
  - $\kappa(i,j) = d_{ij}^{-\beta}$ is a power-law **infection/distance kernel**
  - d_{ij} is the distance between individuals i and j
  - $\alpha > 0$ is an 'infectivity' parameter
  - $\beta > 0$ is a 'spatial' parameter

Likelihood

The likelihood is given by:

$$\pi(\mathbf{Y}|\theta) = \prod_t \left[\prod_{i \in S(t+1)} 1 - P_{it} \right] \left[\prod_{i \in I(t+1) \setminus I(t)} P_{it} \right]$$

where:

$S(t+1)$ is the set of susceptible individuals at time, $t+1$

$I(t+1) \setminus I(t)$ is the set of newly infected individuals at time, $t+1$

N.B. Assuming we know when individuals are infected/infectious.

Bayesian Framework, Data and Parameters

Set in a Bayesian framework:

$$\pi(\boldsymbol{\theta}|Y) = \frac{\pi(Y|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta})}{\pi(Y)}$$

(posterior $\propto$ likelihood $\times$ prior)

- Data: Y parameter vector: $\boldsymbol{\theta} = (\alpha, \beta)$
- $\pi(Y) = \int \pi(Y|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is a normalization constant

Bayesian Framework, Data and Parameters

Set in a Bayesian framework:

$$\pi(\boldsymbol{\theta}|Y) = \frac{\pi(Y|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta})}{\pi(Y)}$$

(posterior $\propto$ likelihood $\times$ prior)

- Data: Y parameter vector: $\boldsymbol{\theta} = (\alpha, \beta)$
- $\pi(Y) = \int \pi(Y|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is a normalization constant

In the rest of this talk:

- Want to use Metropolis-Hastings MCMC to sample from $\pi(\boldsymbol{\theta}|Y)$
- We put vague priors $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$

Issues with Statistical Modeling of Infectious Diseases

- Recall that we are assuming we know infection times and infectious periods (i.e. no data augmentation).
- However, even for moderately sized populations and epidemic lengths, likelihood calculation can be computationally prohibitive.
- Bad since here the likelihood function is calculated numerous times in an MCMC chain.

Issues with Statistical Modeling of Infectious Diseases

- Possible solutions:
  - Simplify model – e.g. homogeneous mixing
  - Data aggregation
  - Approximate Bayesian computation
    (e.g. McKinley *et al.*, 2009; Numminen *et al.*, 2013)
  - Linearization of kernel
    (e.g. Deardon *et al.*, 2010; Kwong & Deardon, 2012)
  - Sampling-based likelihood approximation

Issues with Statistical Modeling of Infectious Diseases

- Possible solutions:
  - Simplify model – e.g. homogeneous mixing
  - Data aggregation
  - Approximate Bayesian computation
    (e.g. McKinley *et al.*, 2009; Numminen *et al.*, 2013)
  - Linearization of kernel
    (e.g. Deardon *et al.*, 2010; Kwong & Deardon, 2012)
  - Sampling-based likelihood approximation
  - **Emulation (build fast model of likelihood)**

Emulation-based Inference

- Here, we propose to use inference based on so-called emulation techniques.
- The method involves:
  - replacing the likelihood with a Gaussian Process (GP) approximation **(EMULATOR)** of the likelihood function
  - within an otherwise Bayesian MCMC framework

GP Emulator

- Design Matrix:

$$X = \begin{pmatrix} 1 & \theta_1^{(1)} & \theta_2^{(1)} & \dots & \theta_n^{(1)} \\ 1 & \theta_1^{(2)} & \theta_2^{(2)} & \dots & \theta_n^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \theta_1^{(p)} & \theta_2^{(p)} & \dots & \theta_n^{(p)} \end{pmatrix}$$

GP Emulator

- Design Matrix:

$$X = \begin{pmatrix} 1 & \theta_1^{(1)} & \theta_2^{(1)} & \dots & \theta_n^{(1)} \\ 1 & \theta_1^{(2)} & \theta_2^{(2)} & \dots & \theta_n^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \theta_1^{(p)} & \theta_2^{(p)} & \dots & \theta_n^{(p)} \end{pmatrix}$$

- For each $\mathbf{\Theta}^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, ..., \theta_n^{(i)})$ we simulate an epidemic to get data (or set of summary statistics of data):

$$Y_{\text{sim}}(\mathbf{\Theta}^{(i)}) = (\delta_1^{(i)}, \delta_2^{(i)}, ..., \delta_{t_{max}}^{(i)})$$

GP Emulator

- Design Matrix:

$$X = \begin{pmatrix} 1 & \theta_1^{(1)} & \theta_2^{(1)} & \dots & \theta_n^{(1)} \\ 1 & \theta_1^{(2)} & \theta_2^{(2)} & \dots & \theta_n^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \theta_1^{(p)} & \theta_2^{(p)} & \dots & \theta_n^{(p)} \end{pmatrix}$$

- For each $\mathbf{\Theta}^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, ..., \theta_n^{(i)})$ we simulate an epidemic to get data (or set of summary statistics of data):

$$\mathrm{Y}_{\mathrm{sim}}(\mathbf{\Theta}^{(i)}) = (\delta_1^{(i)}, \delta_2^{(i)}, ..., \delta_{t_{max}}^{(i)})$$

- Then calculate a distance metric between simulated and observed data:

$$D(\mathbf{\Theta}^{(i)}) = ||\mathrm{Y}_{sim}(\mathbf{\Theta}^{(i)}) - \mathrm{Y}_{obs}||^2,$$

GP Emulator

- Design Matrix:

$$X = \begin{pmatrix} 1 & \theta_1^{(1)} & \theta_2^{(1)} & \dots & \theta_n^{(1)} \\ 1 & \theta_1^{(2)} & \theta_2^{(2)} & \dots & \theta_n^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \theta_1^{(p)} & \theta_2^{(p)} & \dots & \theta_n^{(p)} \end{pmatrix}$$

- For each $\mathbf{\Theta}^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, ..., \theta_n^{(i)})$ we simulate an epidemic to get data (or set of summary statistics of data):

$$\mathrm{Y}_{\mathrm{sim}}(\mathbf{\Theta}^{(i)}) = (\delta_1^{(i)}, \delta_2^{(i)}, ..., \delta_{t_{max}}^{(i)})$$

- Then calculate a distance metric between simulated and observed data:

$$D(\mathbf{\Theta}^{(i)}) = ||\mathrm{Y}_{sim}(\mathbf{\Theta}^{(i)}) - \mathrm{Y}_{obs}||^2,$$

- This gives us our **training data**:

$$\mathbf{D}(\mathbf{\Theta}) = [D(\mathbf{\Theta}^{(1)}), D(\mathbf{\Theta}^{(2)}), ..., D(\mathbf{\Theta}^{(p)})]^T$$

where $\mathbf{\Theta} = (\mathbf{\Theta}^{(1)}, \mathbf{\Theta}^{(2)}, ..., \mathbf{\Theta}^{(p)})^T$

GP Emulator

- Fit GP model: $\mathbf{D}|\Theta, \boldsymbol{\beta}_G, \boldsymbol{\psi}_G \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}_G, \Sigma(\boldsymbol{\psi}_G))$,

  where $\boldsymbol{\beta}_G = (\beta_0, \beta_1, ..., \beta_n)$,

$$(\Sigma(\boldsymbol{\psi}_G))_{ij} = \begin{cases} \tau^2_{GP} \exp\left(-\sum_{k=1}^{n} \eta_k \left(\theta_k^{(i)} - \theta_k^{(j)} \right)^2 \right), & \text{if } i \neq j, \\ \tau^2_{GP} + \tau^2_{\epsilon}, & \text{otherwise} \end{cases}$$

- $\tau^2_{GP} = \text{Var}[\mathbf{D}(\Theta)]$, unconditional variance of GP,
- η_k are smoothing parameters,
- τ^2_{ϵ} is a nugget parameter, representing variance due to the stochasticity of the response.

Predictive distribution

- The predictive distribution for a new data set Y^* producing distance D^* for parameters Θ^*, $f_E(D^*; \Theta^*)$ is Gaussian and so can be readily computed
- We can therefore approximate the computationally costly likelihood function using this predictive distribution
- Naively, may use $f_E(0; \Theta^*)$
- However, since our GP emulator is an approximation to the underlying likelihood function works better to us $f_E(\delta; \Theta^*)$

  where δ is a discrepancy parameter to be estimated
- δ is usually *a priori* constrained to be 'small'

Predictive distribution

- MLE of $\boldsymbol{\beta}_G$ and $\boldsymbol{\psi}_G$ are $\hat{\boldsymbol{\beta}}_G = (\hat{\beta}_{0G}, \hat{\beta}_{1G}, ..., \hat{\beta}_{nG})$ and $\hat{\boldsymbol{\psi}}_G = (\hat{\tau}^2_{GP}, \hat{\tau}^2_{\epsilon}, \hat{\boldsymbol{\eta}})$.

- Any new data producing distance D^* at unknown parameter Θ^*, has the normal predictive distribution $D^* | \mathbf{D}, \boldsymbol{\Theta}, \boldsymbol{\Theta}^* \sim N(\mu^*, \Sigma^*)$, where

$$\mu^* = \hat{\beta}_{0G} + \theta^*_1 \hat{\beta}_{1G} + \theta^*_2 \hat{\beta}_{2G} + ... + \theta^*_n \hat{\beta}_{nG} + \hat{\tau}^2_{GP} \mathbf{r}^T (\Sigma(\hat{\psi}_G))^{-1} (\mathbf{D} - \mathbf{X}\hat{\boldsymbol{\beta}}_G),$$
$$\Sigma^* = \hat{\tau}^2_{GP} + \hat{\tau}^2_{\epsilon} - \hat{\tau}^4_{GP} \mathbf{r}^T (\Sigma(\hat{\psi}_G))^{-1} \mathbf{r}$$
$$\mathbf{r} = (r_1, r_2, ..., r_i, ..., r_p), \text{ and } r_i = \text{cor}(D(\boldsymbol{\Theta}^*), D(\boldsymbol{\Theta}^{(i)})).$$

- We can use the predictive distribution: $f_E(D^*; \Theta^*)$ as an emulator (approximation to the likelihood).
- Discrepancy: $D^* := \lambda \quad \rightarrow$ use $f_E(\delta; \Theta^*)$ where δ is a parameter to be estimated.

Bayesian MCMC framework

So we replace our previous Bayesian framework:

$$\pi(\alpha, \beta | Y) \propto \pi(Y | \alpha, \beta)\ \pi(\alpha)\pi(\beta)$$

(posterior proportional to likelihood $\times$ prior)

Bayesian MCMC framework

So we replace our previous Bayesian framework:

$$\pi(\alpha, \beta|Y) \propto \pi(Y|\alpha, \beta)\ \pi(\alpha)\pi(\beta)$$

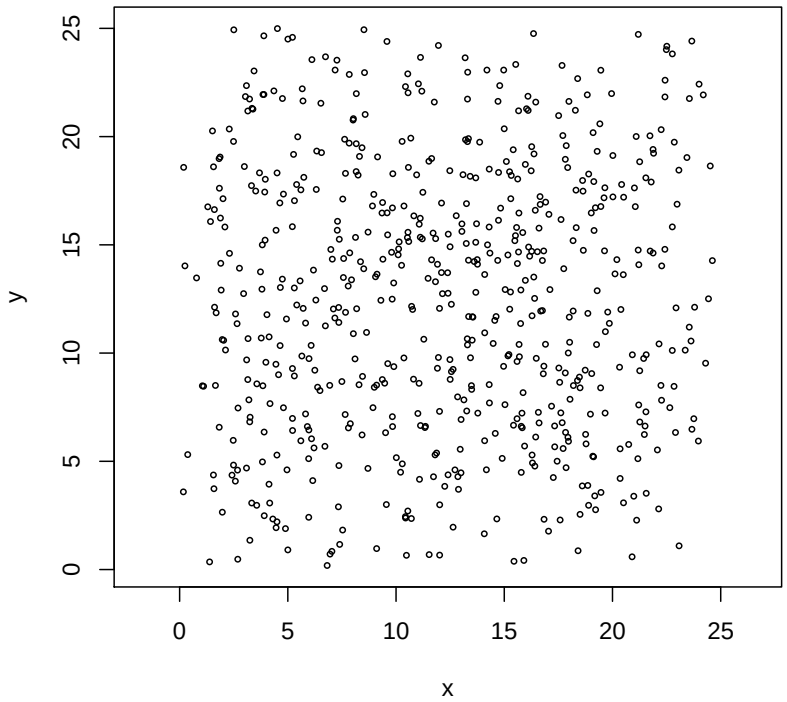
(posterior proportional to likelihood $\times$ prior)

....with an approximate Bayesian framework:

$$\pi(\alpha, \beta|Y) \widetilde{\propto} f_E(\delta; \alpha, \beta)\ \pi(\alpha)\pi(\beta)\pi(\delta)$$

(posterior approximately proportional to emulator $\times$ prior)

Epidemic simulation I



Epidemic simulated with a fixed infectious period of $\gamma_I = 2$ from power-law model:

$$P_{it} = 1 - \exp\left[-\alpha \sum_{j \in I(t)} d_{ij}^{-\beta}\right]$$

Epidemic simulation II

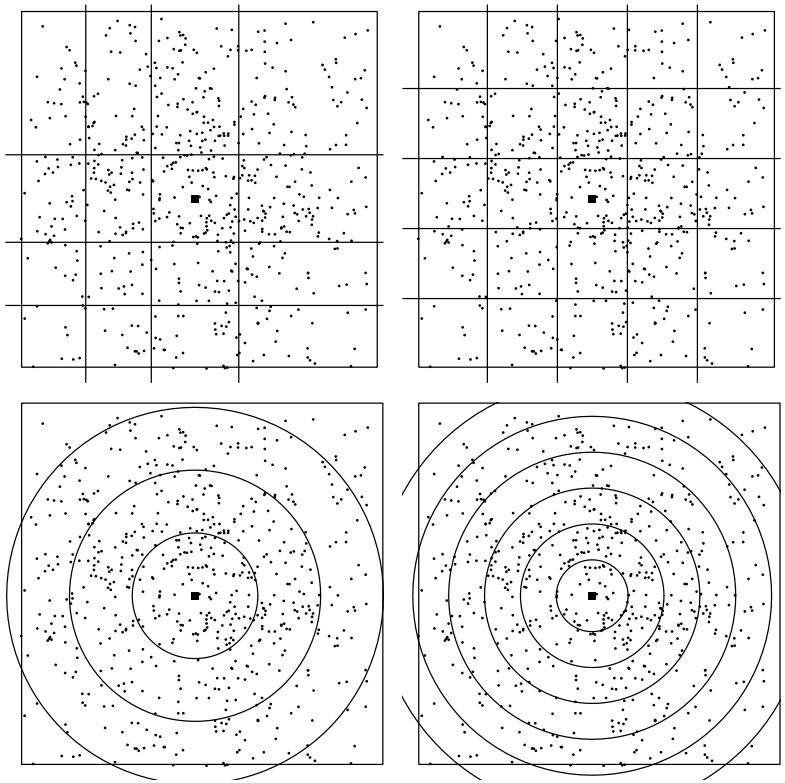
- For observed data:
  infectivity, spatial parameters were $\alpha = 0.2$ and $\beta = 2.5$.

- For simulated data and design matrix:
  - $\alpha \in [0.1, 1.0]$ and $\beta \in [2.1, 3.0]$
  - on a regular grid

  - To evaluate the robustness, 10, 15, 20, and 25 points in the parameter intervals were used.

- One individual approximately in the centre was set as the initial seed for each simulation.

Spatial Stratification for building GP Emulator



- Rectangular:
  - Regular: each stratum has equal area (size).
  - Irregular: First infection is approximately at the centre of one of the strata.
- Circular: Concentric circles with rings of equal width, centre is the location of the first infection.

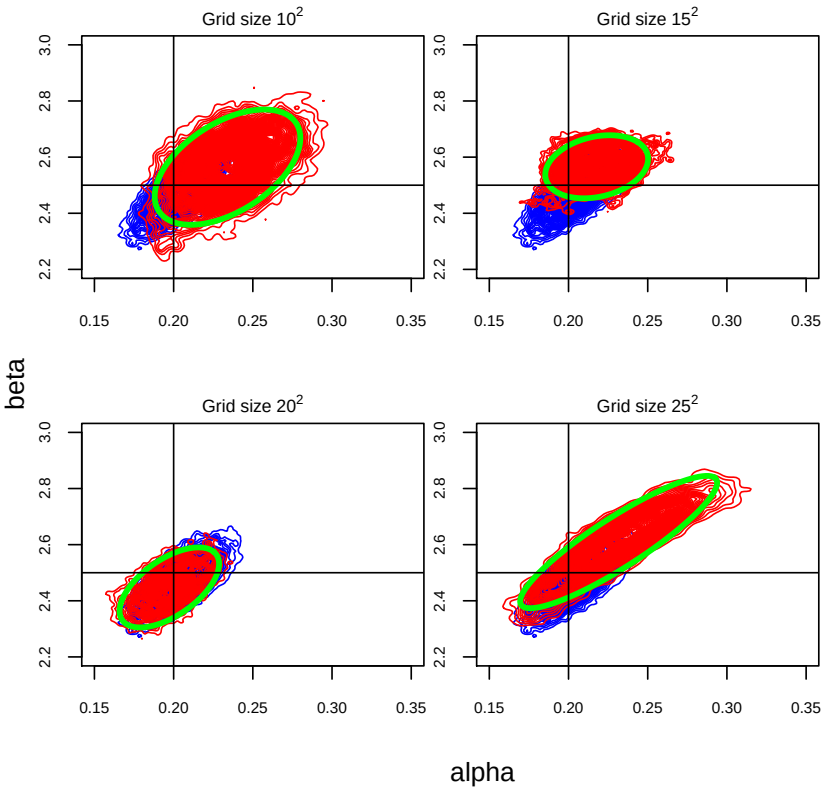
Global and Stratified Epidemic Curves

- Global epidemic curve:
  - Simulated data: $\Delta(\mathbf{\Theta}^{(i)}) = (\delta_1^{(i)}, \delta_2^{(i)}, ..., \delta_{t_{max}}^{(i)})$.
  - Observed data: $Z(\mathbf{\Theta}^*) = (z_1, z_2, ..., z_{t_{max}})$.
  - $D^{(i)} = ||\Delta(\mathbf{\Theta}^{(i)}) - Z(\mathbf{\Theta}^*)||^2$.
- Stratified epidemic curve:
  - Simulated data in k^{th} stratum: $\Delta_k^{(i)} = (\delta_{k1}^{(i)}, \delta_{k2}^{(i)}, ..., \delta_{k\nu}^{(i)})$.
  - Observed data in k^{th} stratum: $\mathcal{Z}_k(\mathbf{\Theta}^*) = (z_{k1}, z_{k2}, ..., z_{k\nu})$.
  - Full simulated epidemic data: $\bar{\Delta}^{(i)} = (\Delta_1^{(i)}, \Delta_2^{(i)}, ..., \Delta_s^{(i)})$.
  - Full observed epidemic data: $\bar{Z}(\mathbf{\Theta}^*) = (\mathcal{Z}_1, \mathcal{Z}_2, ..., \mathcal{Z}_s)$.
  - $D^{(i)} = ||\bar{\Delta}^{(i)} - \bar{Z}(\mathbf{\Theta}^*)||^2$.
- The response variable for GP model: $\mathbf{D} = [D^{(1)}, D^{(2)}, ..., D^{(p)}]^T$.

Model Fitting: Simulated Data

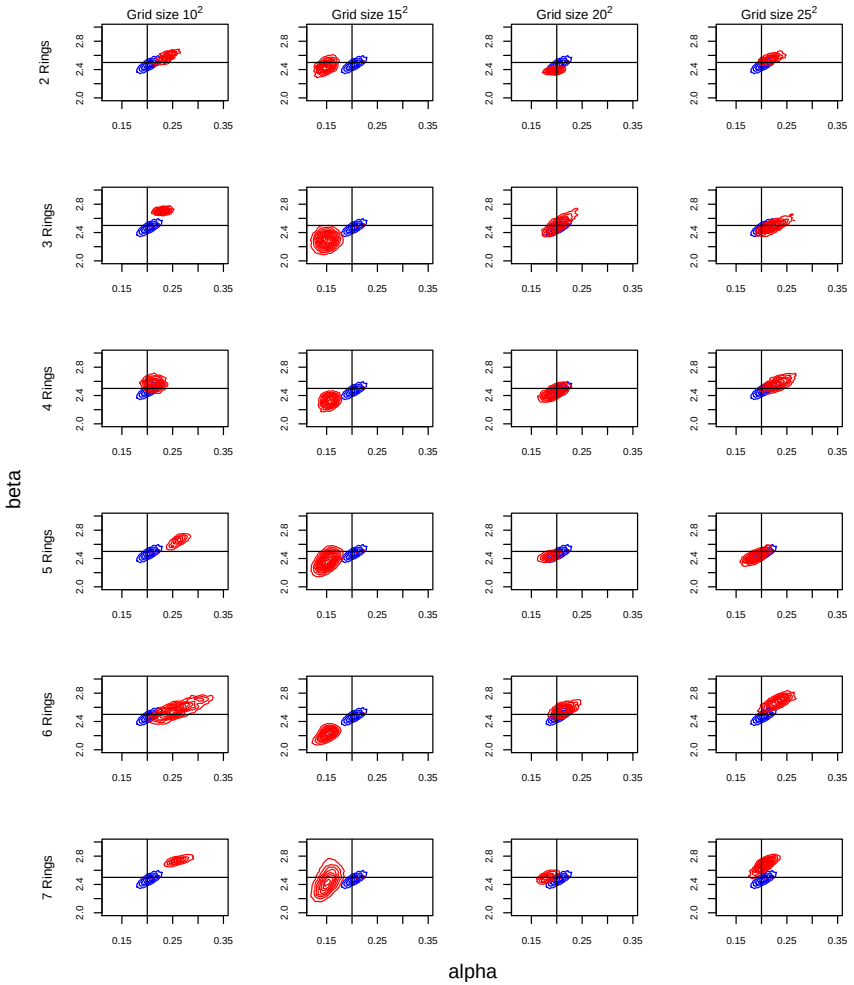
- The full Bayesian model and emulation-based model were fitted to this data set via RW-MH-MCMC.
- Vague exponential priors with mean 10^5 were placed on α and β.
- An exponential prior, Exp(10) was used for the discrepancy λ.
- MCMC run for 200,000 iterations and convergence visually ascertained

Results: Global population



- Blue: true posterior surface.
- Red: Emulation-based posterior surface.
- Green: 95% confidence ellipse.
- Black lines: True parameter values.
- Parameter grid size matters:
  - Lower resolutions $\Rightarrow$ bias.
  - Higher resolution $\Rightarrow$ biased and time consuming.
- 20^2 grid size works well.

Results: Circular stratification



- Blue: true posterior surface.
- Red: Emulation-based posterior surface.
- Black lines: True parameter values.
- Performance was reasonably good for some stratification settings (4 rings and 20^2 grid).
- Overall, little improvement obvious.

Results: Computation time

Table: CPU time to run 200,000 MCMC iterations for both the full Baysian and emulation-based models with different grid sizes in the parameter space.

Model	Grid size	Time in seconds
Bayesian Model	-	3533.69
Emulation-based Model	10	13.99
	15	57.47
	20	168.15
	25	392.61

Introduction: Tomato Spotted Wilt Virus (TSWV) I

- TSWV is one of the most widespread and significantly economically damaging plant virus infecting over 1000 plant species.

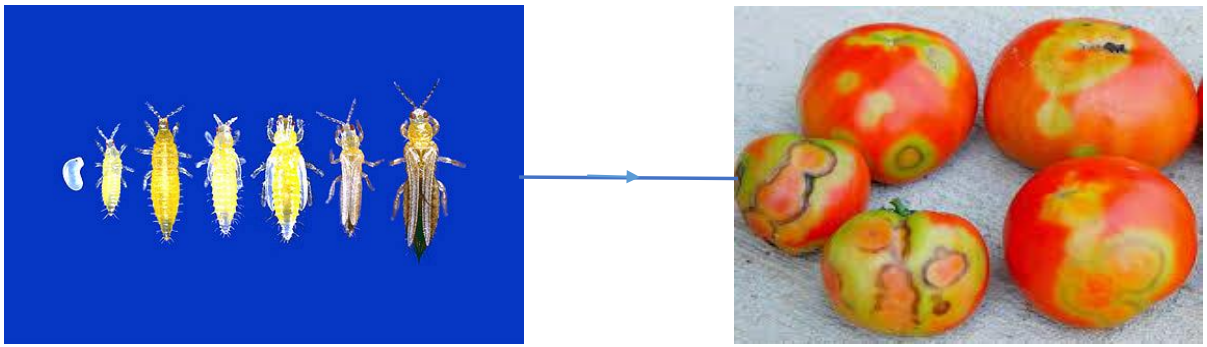
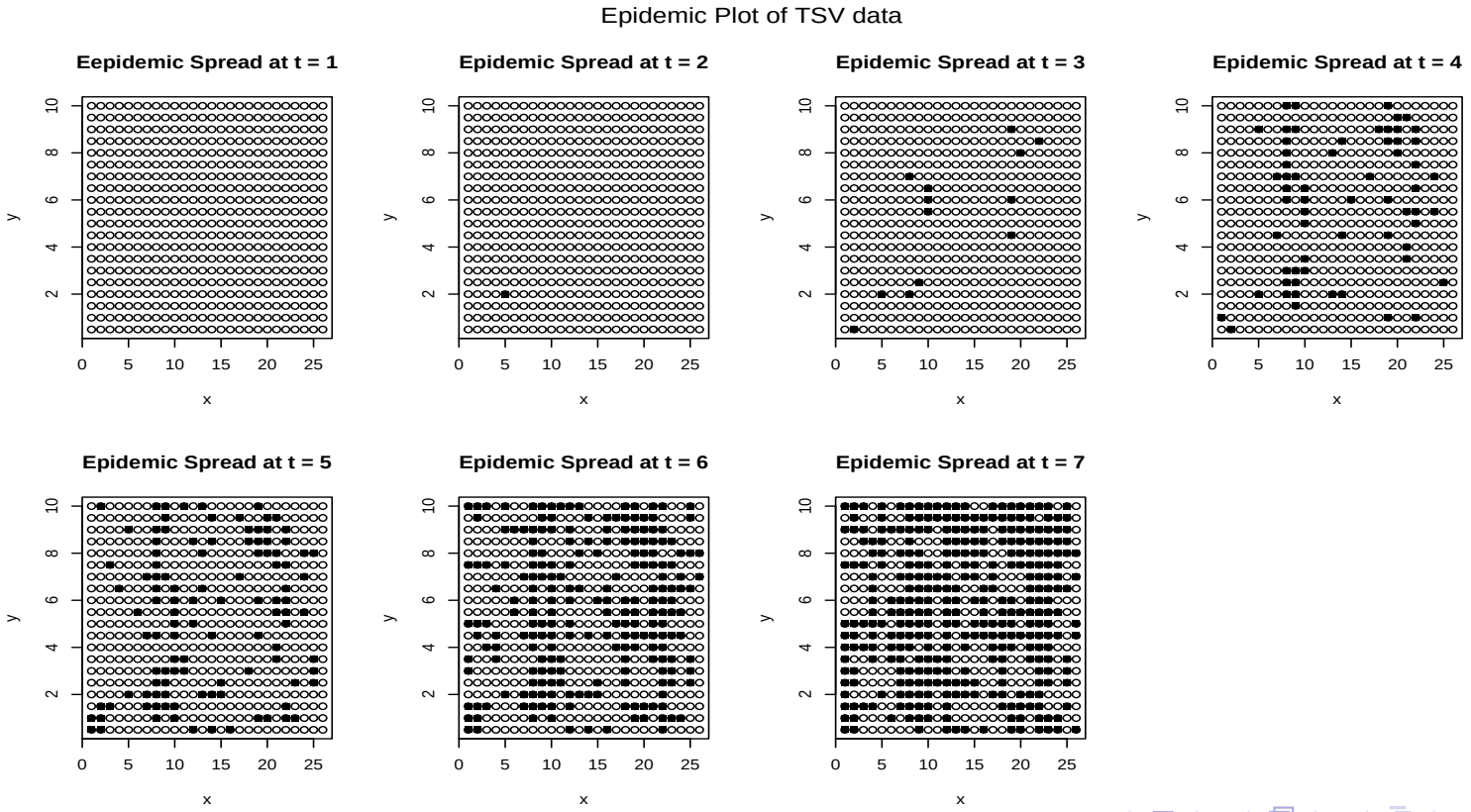


Figure: Pictures are taken from google web.

Introduction: Tomato Spotted Wilt Virus (TSWV) II

- Data from a 1993 study, described in Hughes et al. (1997), of TSWV in pepper plants consist 520 individuals in a uniform grid of 26×10.
- Epidemic ran for $t = 1, 2, ..., 7$ in increments of 14 days.



Model Fitting: TSWV

- The full Bayesian model and emulation-based model were fitted to this data set via RW-MH-MCMC.
- A fixed infectious period, $\gamma_I = 3$ and distance-based power-law kernel $\kappa(i,j) = d_{ij}^{-\beta}$ were used.
- Design matrix: $\alpha \in [0.005, 0.5]$ and $\beta \in [1.0, 2.0]$ with 20^2 grid size.
- Vague exponential priors with mean 10^5 were placed on α and β; and an exponential prior Exp(100) was used for the discrepancy λ.

Results:TSWV

Model	Stratification	Parameter Estimate, ... (..., ...) = mean (95% PI)		
		α	β	λ
Bayesian Model	-	0.0194 (0.0092, 0.0296)	1.3597 (0.8567, 1.7553)	-
Emulation-based Model	Global	0.0227 (0.0165, 0.0287)	1.1963 (1.1123, 1.2886)	4258.1 (3273.9, 5334.6)
	2 Rings	0.0254 (0.0173, 0.0350)	1.2055 (1.0851, 1.3242)	3540.6 (2611.5, 4557.3)
	3 Rings	0.0209 (0.0137, 0.0287)	1.1803 (1.0533, 1.2898)	3136.9 (2377.9, 3951.6)
	4 Rings	0.0241 (0.0173, 0.0311)	1.2335 (1.1458, 1.3253)	2992.7 (2412.3, 3619.3)
	5 Rings	0.0213 (0.0124, 0.0305)	1.1066 (0.9196, 1.2606)	2723.4 (1985.5, 3548.0)
	6 Rings	0.0254 (0.0159, 0.0347)	1.2274 (1.1213, 1.3337)	2187.5 (1612.5, 2848.3)
	7 Rings	0.0196 (0.0083, 0.0301)	1.2059 (1.0846, 1.3339)	2235.1 (1671.3, 2857.9)
	2×2	0.0240 (0.0158, 0.0325)	1.2268 (1.1315, 1.3175)	2671.5 (1972.9, 3396.6)
	3×3	0.0245 (0.0140, 0.0363)	1.2204 (1.0648, 1.3713)	2000.6 (1537.8, 2526.3)
	4×4	0.0195 (0.0032, 0.0368)	1.1646 (0.8675, 1.4279)	1589.2 (1190.2, 2044.9)
	5×5	0.0172 (0.0019, 0.0335)	1.1731 (0.9737, 1.3854)	1284.6 (1013.1, 1606.1)

- The full Bayesian analysis took about 41 times longer (6834 seconds) than the emulation-based methods (166 seconds).

Conclusions

1. Emulation-based method offer a much quicker mode of analysis than the full Bayesian MCMC analysis.
2. The emulation-based methods can successfully infer the biological characteristics of simple spatial infectious disease systems.
3. Spatial stratification did not noticeably improve the model fit.
4. Care in defining the design matrix is needed to achieve accurate and computationally efficient emulation-based inference.

Future work

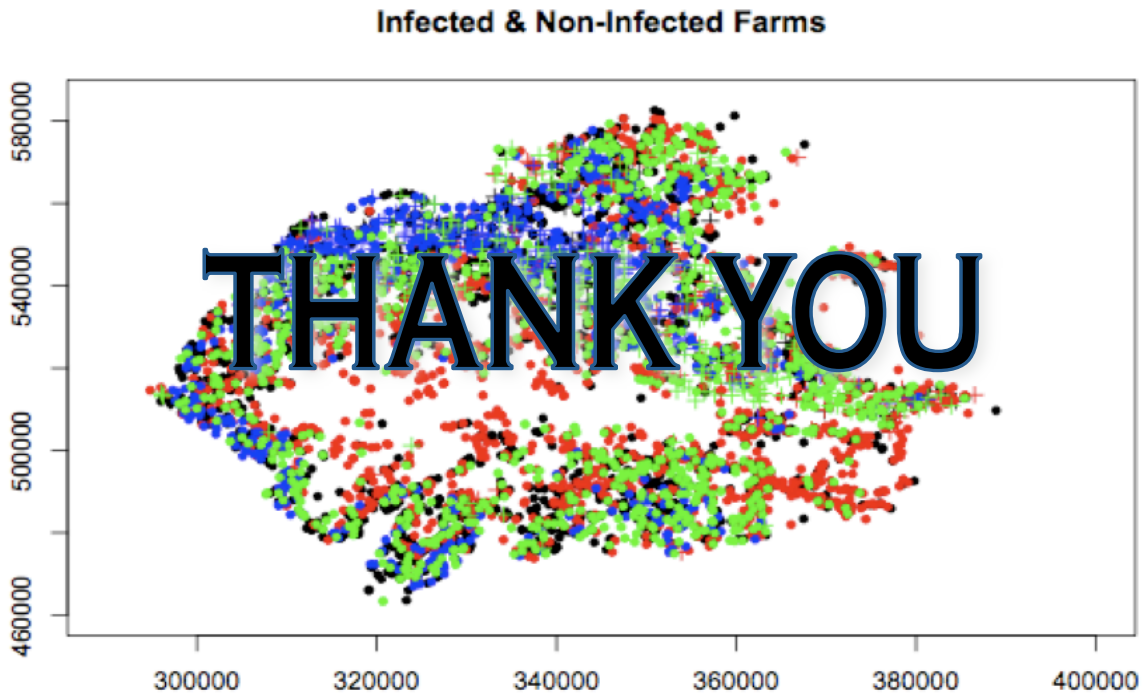
1. Compare model fit for different models
2. Much bigger, more complex systems
   - Observation models to account for unknown infection times, infectious periods, under-reporting, etc.
   - Continuous time disease models.
   - Network-based complex disease systems.
3. For complex and large number of parameter system, GP covariance matrix inversion become a computational bottleneck in itself. Consider methods to address...
4. Systematic comparison with other available methods for speeding up computation time.

Selected References

- Deardon et al (2010). Inference for individual level models of infectious diseases in large populations. *Statistica Sinica*, 20(1), 239 - 261.
- Kwong & Deardon (2012). Linearized forms of individual-level models for large-scale spatial infectious disease systems. *Bulletin of Mathematical Biology*, 74(8), 1912 - 37.
- E. Numminen, L. Cheng, M. Gyllenberg, and J. Corander (2013). Estimating the transmission dynamics of *Streptococcus Pneumoniae* from strain prevalence data. *Biometrics*, 69(3):748-757, 2013.
- Jandarov, R., Haran, M., Bjrnstad, O., and Grenfell, B. (2014). Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease. *Journal of Royal Statitical Society: Series C (Applied Statistics)*, 63(3):423 - 444.
- Bayarri, M., Berger, J., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J., Lin, C. and Tu, J. (2007) A framework for validation of computer models. *Technometrics*, 49, 138-154.
- Kennedy, M. C. and O'Hagan, A. (2001) Bayesian calibration of computer models (with discussion). *J. R. Statist. Soc. B*, 63, 425-464.
- Sacks, J., Welch, W., Mitchell, T. and Wynn, H. (1989) Design and analysis of computer experiments. *Statistical Science*, 4, 409-423.

Acknowledgements

- This work has been funded by:
  - Ontario Ministry of Agriculture, Food & Rural Affairs (OMAFRA)
  - Natural Sciences & Engineering Council of Canada (NSERC)
  - Canadian Foundation for Innovation (CFI)


Infected & Non-Infected Farms
THANK YOU
580000
540000
500000
460000
300000
320000
340000
360000
380000
400000