The importance of estimation and model choice for small scale epidemics

Kokouvi Gamado

Glenn Marion Thibaud Porphyre

Wednesday, 28th January 2016

Outline

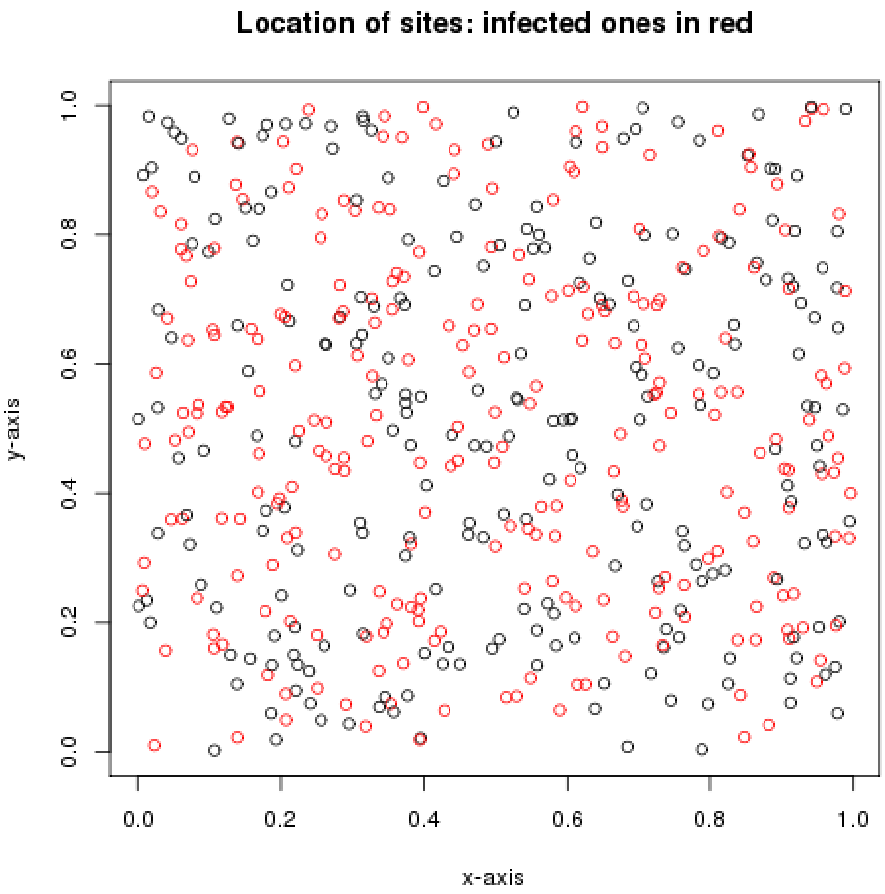
- General framework
- Impact of the different kernels
- Model choice
    - Deviance Information Criteria
    - Latent residuals
- Early phase inference of ongoing epidemics
- Variability between farms
- Within-herd inference for cattle TB
- Conclusions

Between farm model of disease spread

- Susceptible-Infected-Detected model with Spatial infection kernel

- Data on detection of infected premises

- Focus on Bayesian inference using MCMC with efficient mixing



Homogeneous farms model

- Closed population of N individuals
- An individual i makes an infectious contact with a susceptible individual j at rate $\beta_{ij} = \beta_0 h_{ij}$
- h_{ij} is the distance kernel, usually a function of the Euclidian distance. For now,

$$h_{ij} = \exp(-\tau\rho(i, j))$$

- The infectious period follows: $R_i - I_i \approx Ga(\alpha, \gamma)$
- Data augmentation techniques employed for the unknown infection times
- Unknown quantities of interests: (β_0, γ, τ)

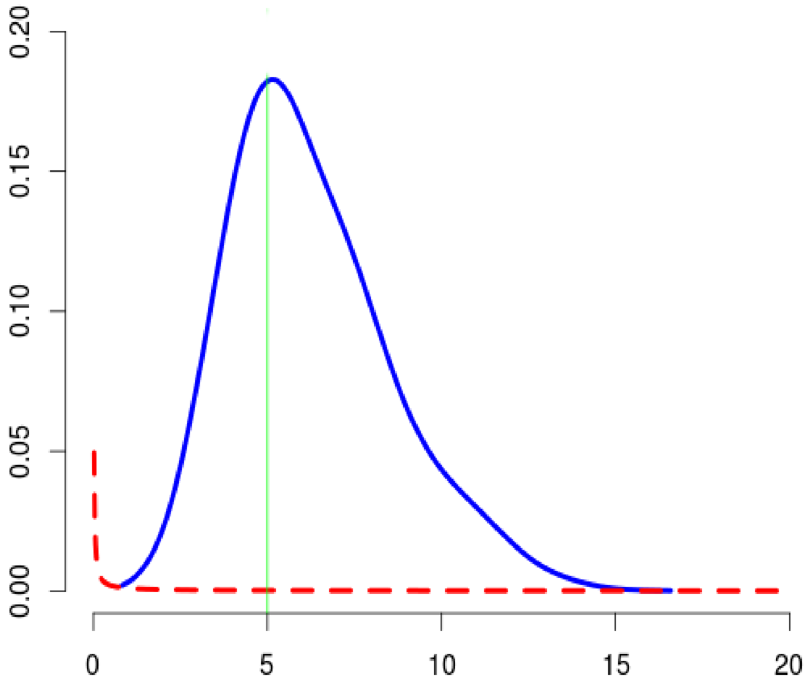
Bayesian Inference with MCMC framework

- Enables statistical inference of
  - parameter values and
  - infection times

- MCMC methods draw successive samples from ‘posterior’ distribution

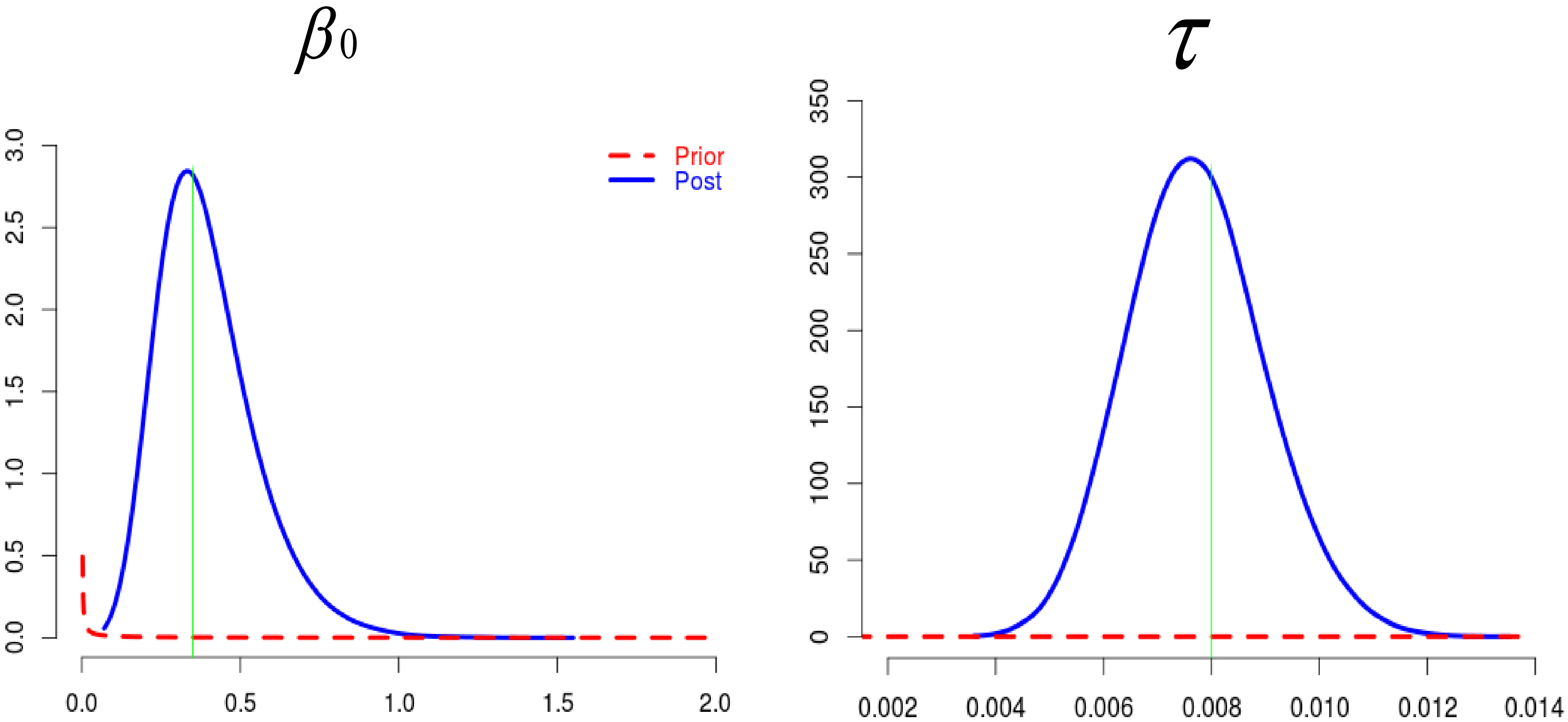
- Want to maximise effective sample size by reducing auto-correlation in MCMC samples (speed up inference).

Inferred (posterior) distribution of γ



Based on simulated data with true values: $\gamma = 5$, $\alpha = 5$, $\beta_0 = 0.35$, #premises = 201, #infected sites n = 43

Inference: posterior distributions



β_0

τ

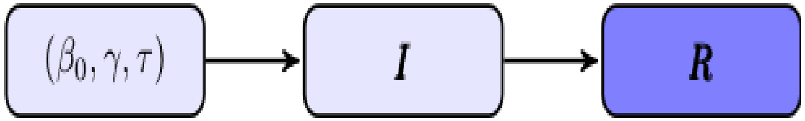
True parameter values in green which are well within the credible intervals of the posterior distributions

Knowledge of β_0 and τ is equivalent to knowledge of the infection process with τ informing about the kernel density

Computationally efficient inference

Centered and non-centered schemes

- Standard algorithms adopt a ‘centred’ approach.

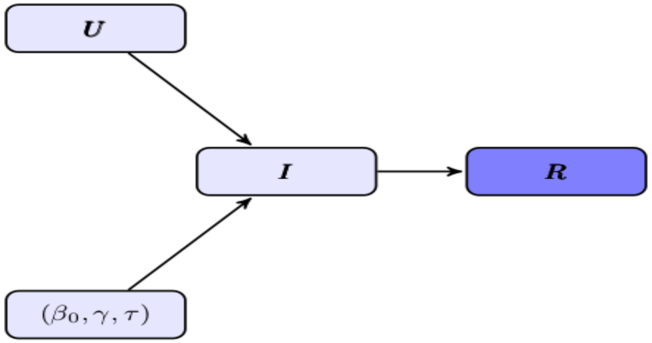


- updating the model parameters
- then updating the latent or unobserved variables

- Non-centering scheme

$$U_i = \gamma(R_i - I_i)$$

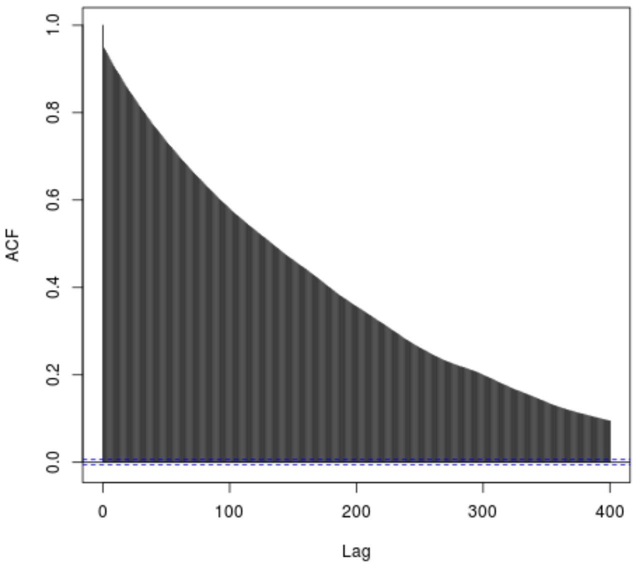
$$(I, \beta_0, \gamma, R) \rightarrow (U, \beta_0, \gamma, R)$$



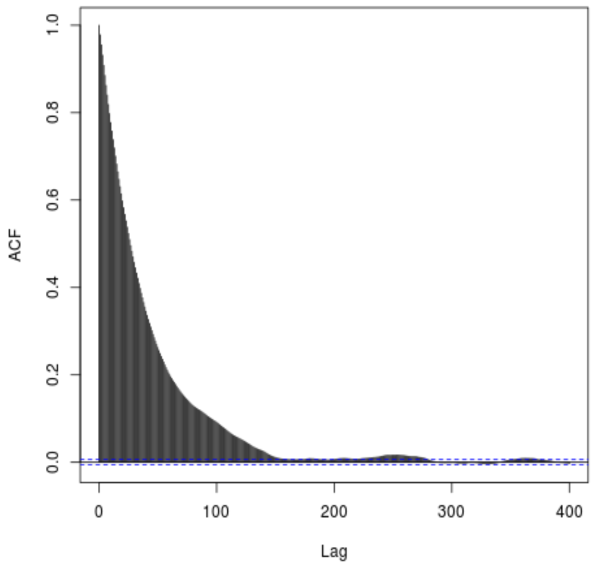
- Update simultaneously parameters and latent variables

Non-centred and hybrid schemes are valuable improvement over standard scheme

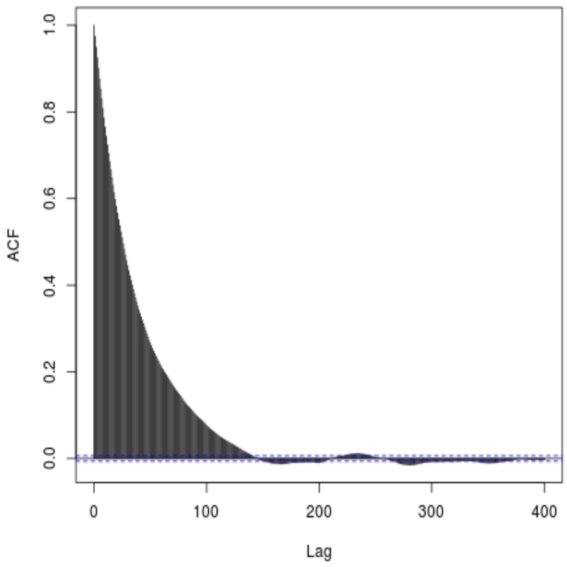
Standard scheme



Non-centred scheme



Hybrid scheme



Non centred parameterisation enables

- correlated moves in Markov chain
- more efficient exploration of posterior
- faster MCMC for complex models
- Non-centered & hybrid schemes reduce autocorrelation

Motivation of using small scale epidemics

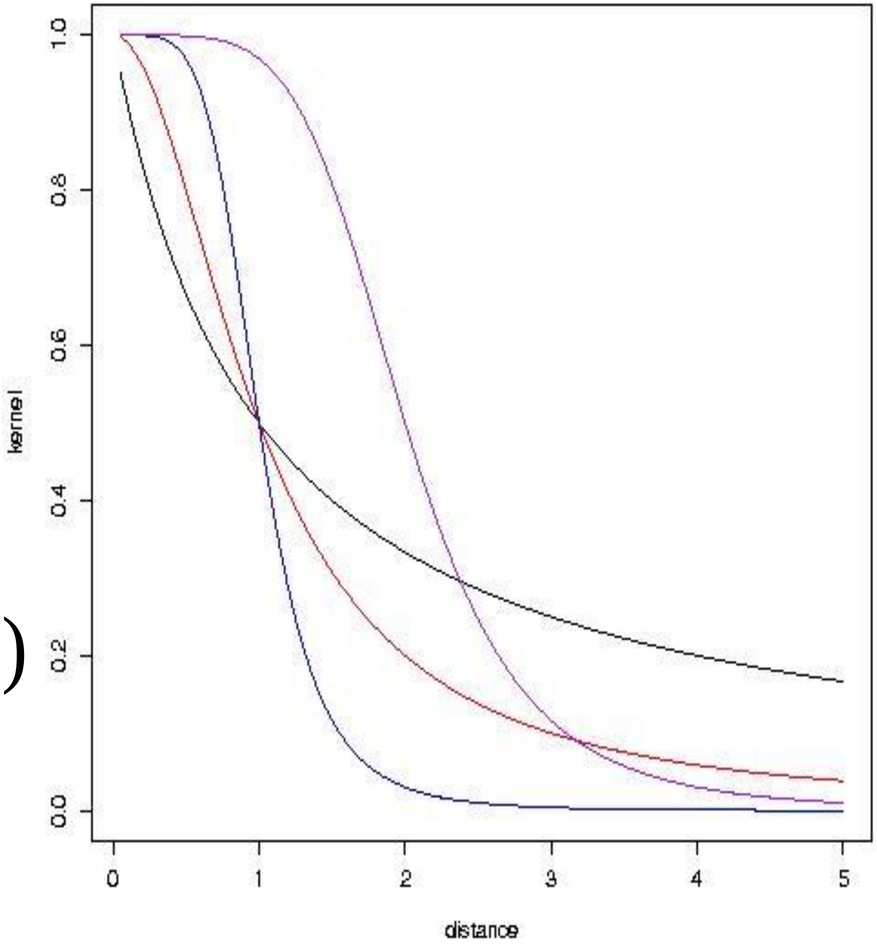
- Epidemics data are usually very limited
- Focus on historical epidemic first: contingency
- Historical epidemics data are useful when preparing for the incursion of emerging or re-emerging diseases
- Useful for risk quantification
- Contribute to design novel, more cost-efficient, control strategies against future epidemics.
- Examples in the literature: FMD, CSF

CSF in East Anglia Northfolk

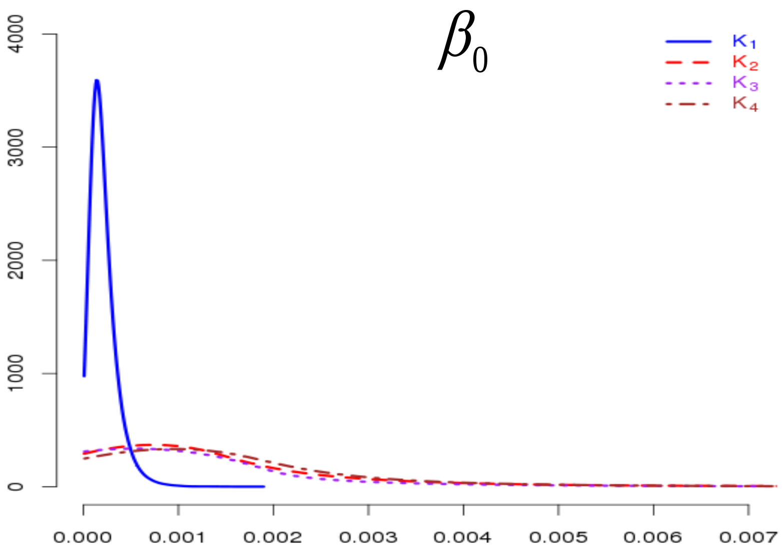
- Description of CSF data in East Anglia
  - Data of CSF epidemic in 2000 (AHVLA)
  - N = 1703 farms with exact location or coordinates
  - 16 reported cases and farms’ CPHs available
- Fit model in the framework described before
- One question we would like to answer is which kernel density “best” fit this data and consequences of choosing wrong kernels

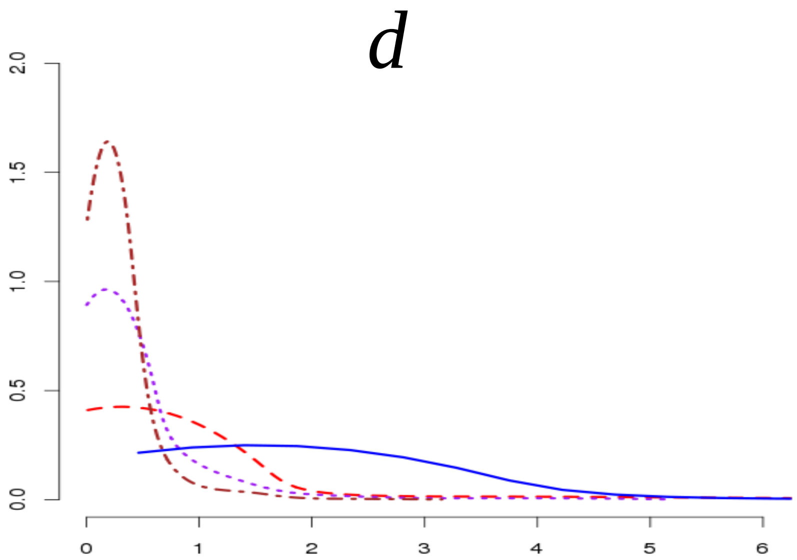
Spatial kernel density functions

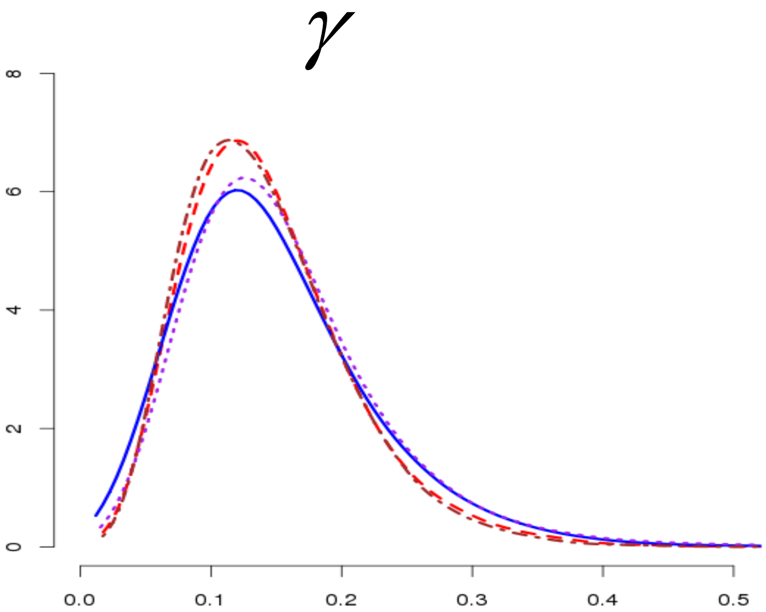
- K1: $\exp(-\tau\rho(i,j))$
- K2: $\dfrac{1}{1+(\rho(i,j)/d)^a}$
- K3: $\dfrac{1}{1+(\rho(i,j)/d)}$
- K4: $1-\exp(-(\rho(i,j)/d)^{-a})$



Comparing parameters with different kernels







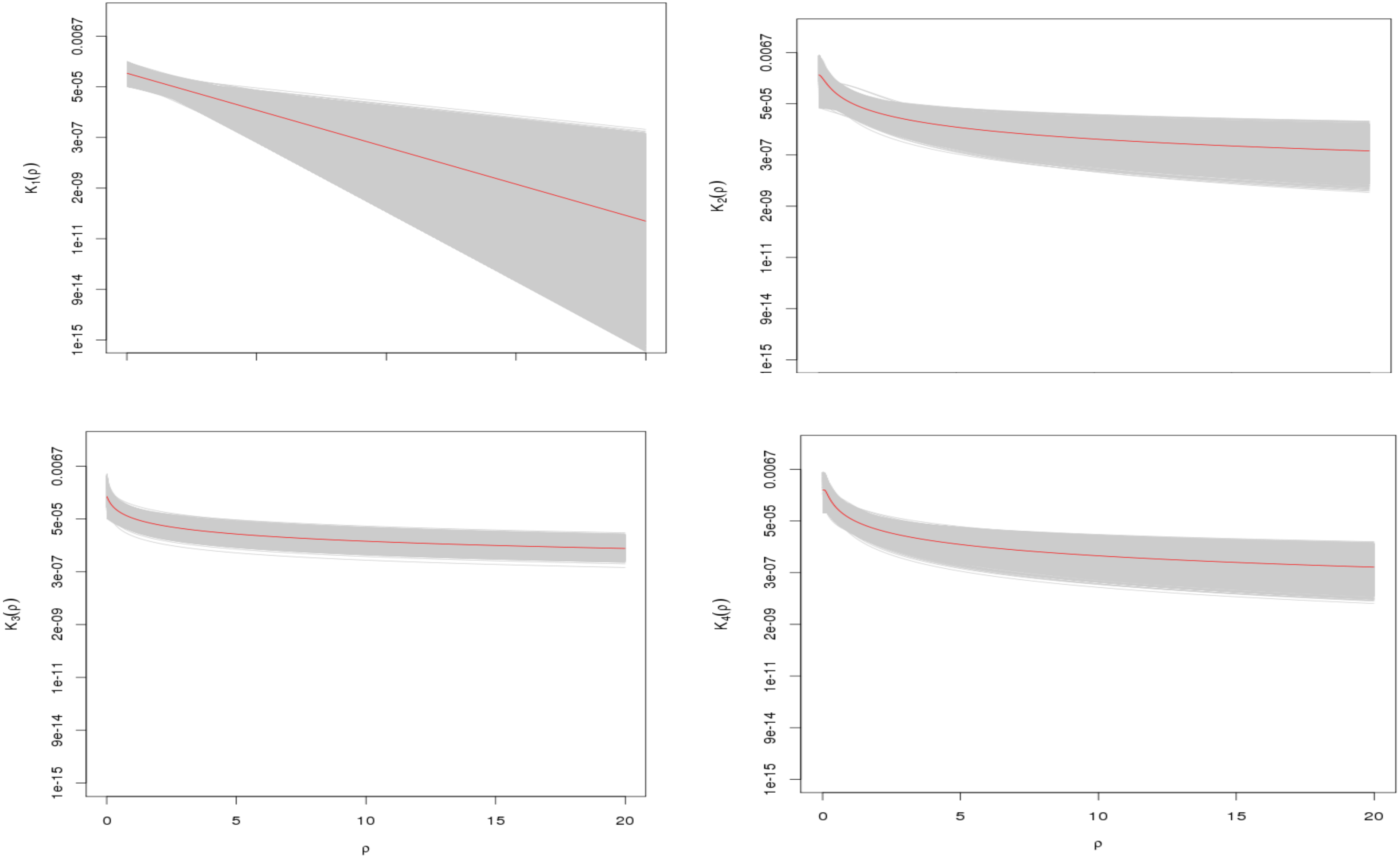
❑ The choice of the kernel does impact the infection process.

❑ β_0 and d seem to produce relative opposite effects

❑ The impact is limited for the recovery process

❑ This brings in the more general question of model choice

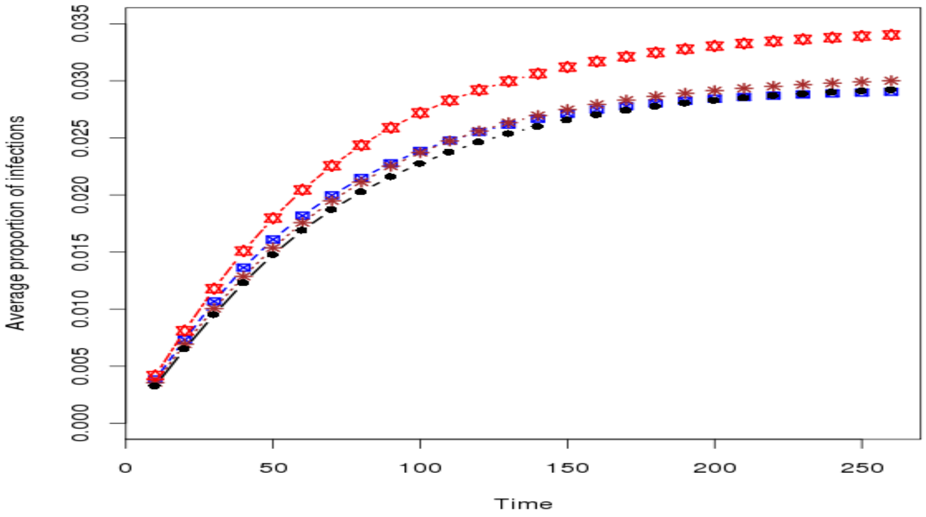
Posterior kernel densities (CSF)



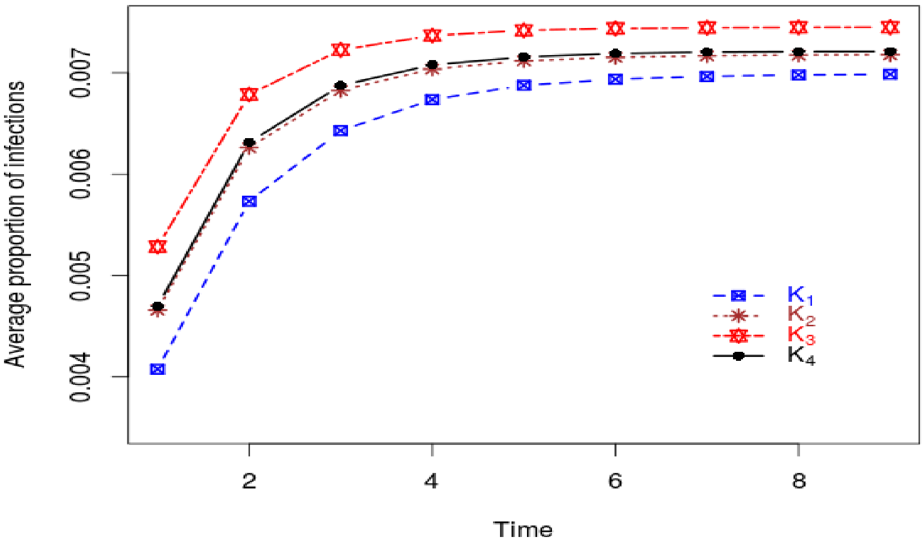
Risk quantification

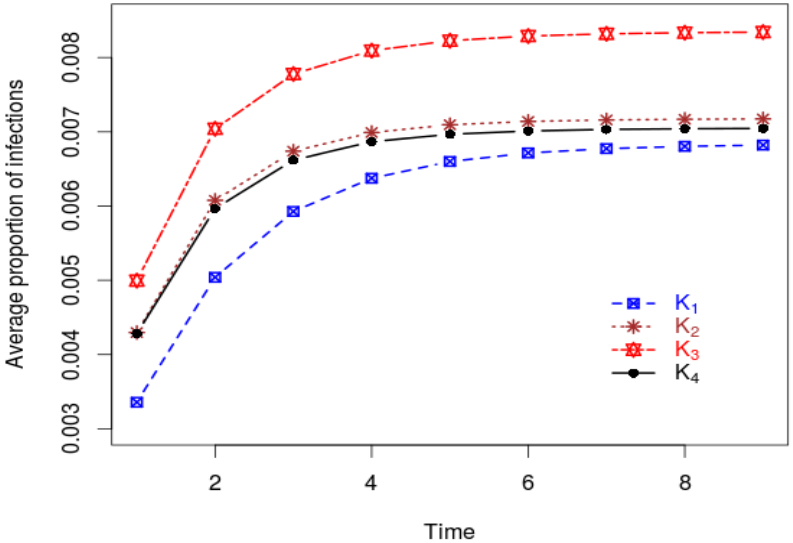
- Draw independently samples from the posterior distribution
- Simulate the epidemic using Gillespie algorithm, always starting with the same index case
- Repeat the procedure a significant number of times for each kernel recording the infected sites and times of events
- Proportions of times each farm is infected are calculated

Average infections with time

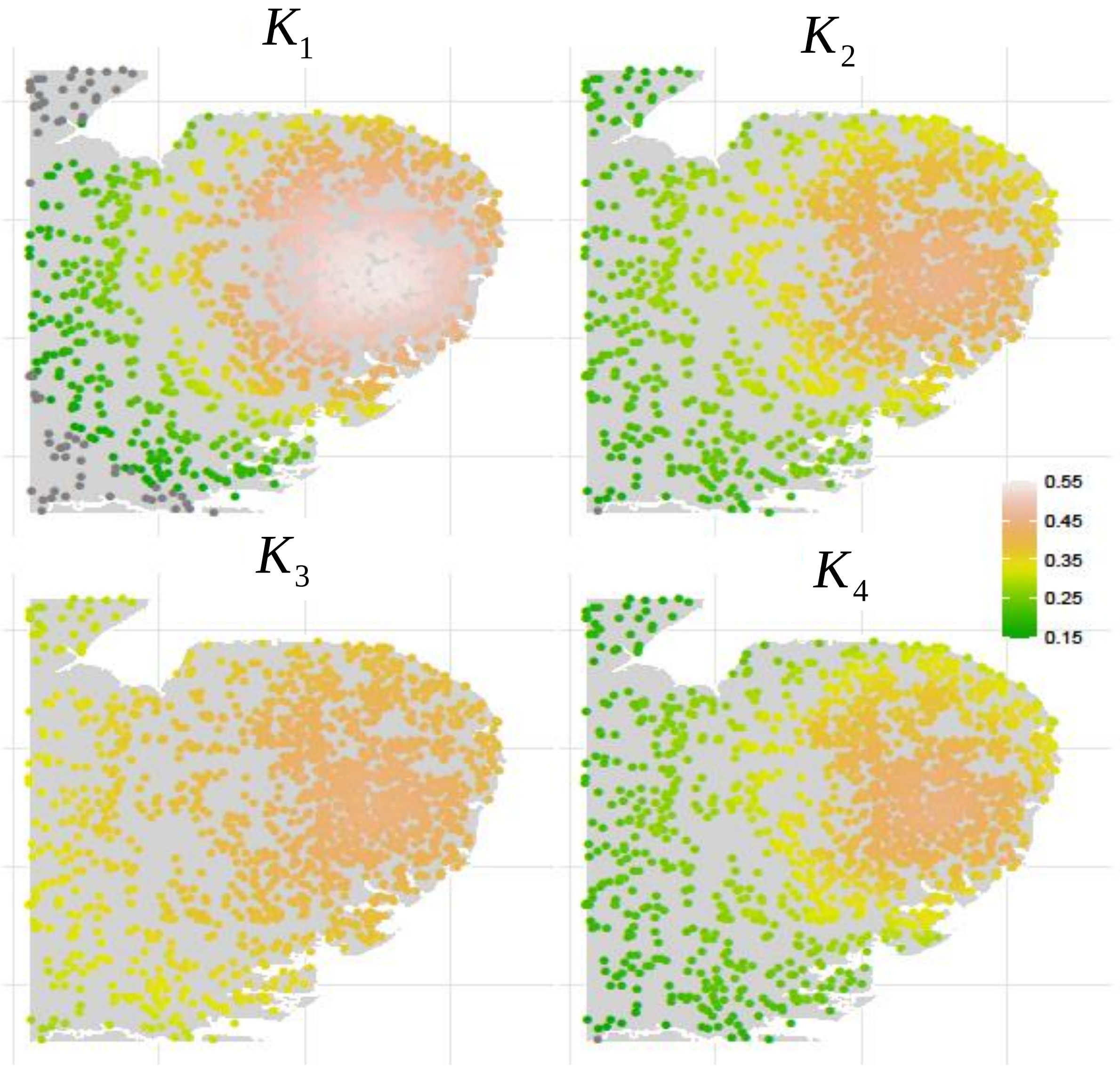


- ➢ Average number of infections differ with kernels over time
- ➢ K_3 presents higher average infections in all cases of real data and simulations
- ➢ Lowest average infections for K_1





Risk maps under various kernels



- n(risk > 0.45)
  - K_1: 866
  - K_2: 105
  - K_3: 1
  - K_4: 15

- The choice of the kernel is important when designing control measures.

Deviance Information Criterion

- The DIC is a model comparison criterion based on trade-off between the fit of the data to the model and the corresponding complexity of the model

  DIC = "goodness of fit" + "complexity"
- Problems: non invariance to reparameterisation, lack of consistency, weak theoretical justification
- Widely used in the literature though
- DIC in the presence of latent variables
- No unique definition: 8 according to Celeux et al. (2006)
- Two of them are adopted: DIC_1 and DIC_2

Bayesian latent residuals

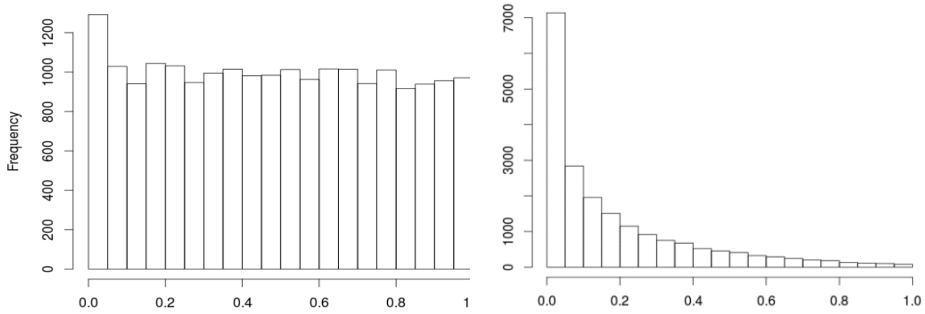
It allows to test different components of the process separately.

- Bayesian latent residuals are the unobserved, independent, uniform random variables that conform with the assumed data generation process
- The uniform deviates are equivalent to the deviates from the non-centred parametrisations
- It is possible to construct so-called infection link residuals focussed on spatial transmission kernel that require the re-construction of the links “who infects who”.
- Best kernel presents uniformly distributed residuals assessed through Kolmogov-Smirnov or Anderson-Darling test

Infer residuals for different processes:

- ➢ embed in data augmentation MCMC scheme
- ➢ Deviations from U(0,1) indicate misspecification

Residuals for dispersal process

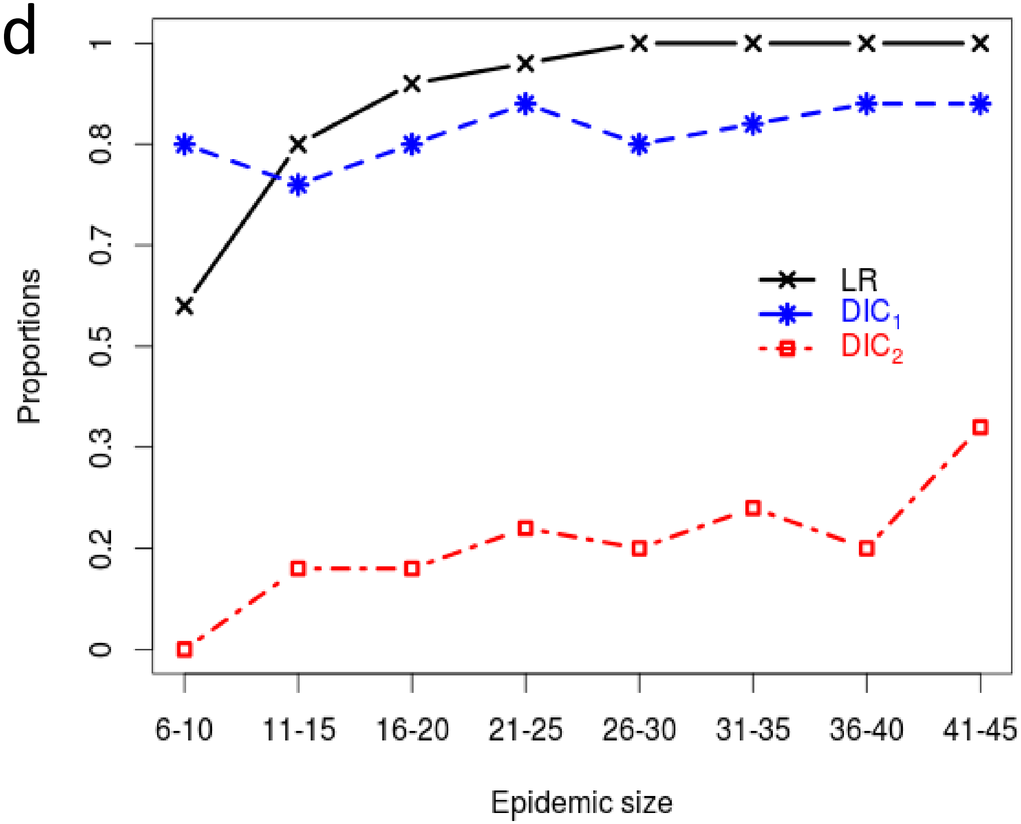


True kernel

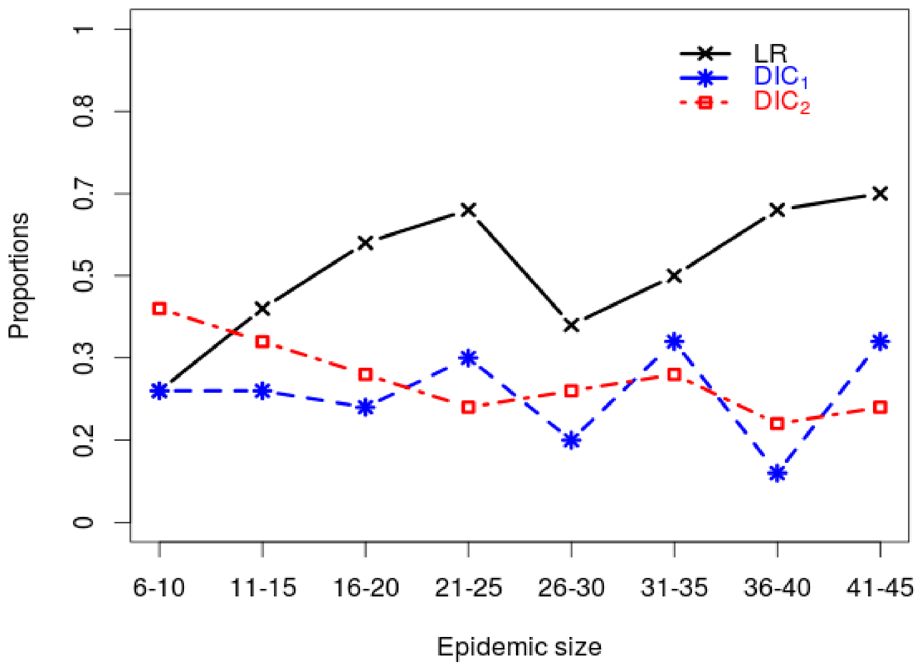
Wrong kernel

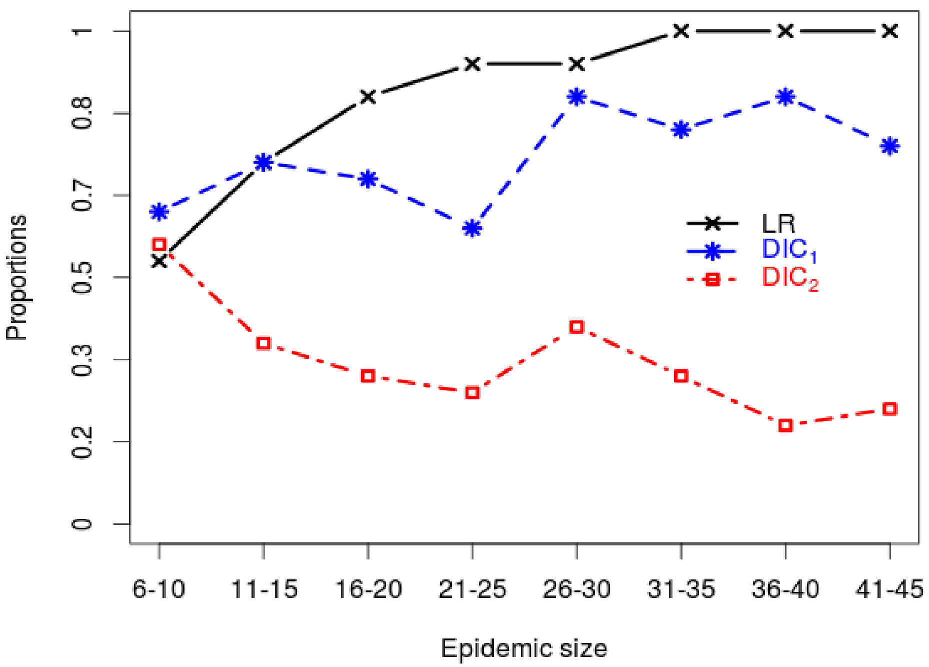
Simulation studies: scenarios 1

- Small and completed epidemics are considered
- Spread of disease first simulated using kernel K_1
- Inferred all parameters with infection times
- Compute all three goodness of fit measures

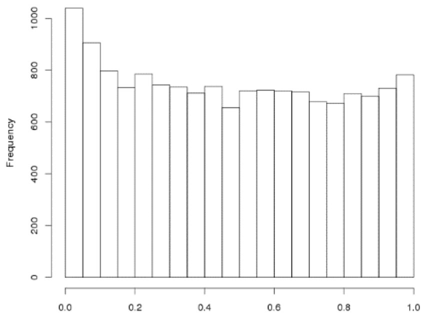


Simulation studies: scenarios 2

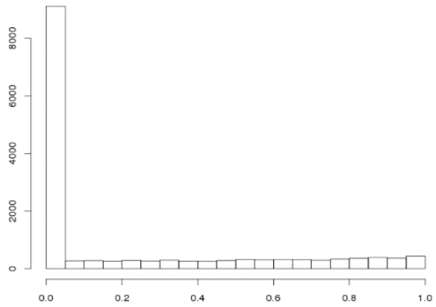




- Using K_2 as the baseline and no clear identification with K_4
- Population density is an important factor for kernels favouring local spread



K_2

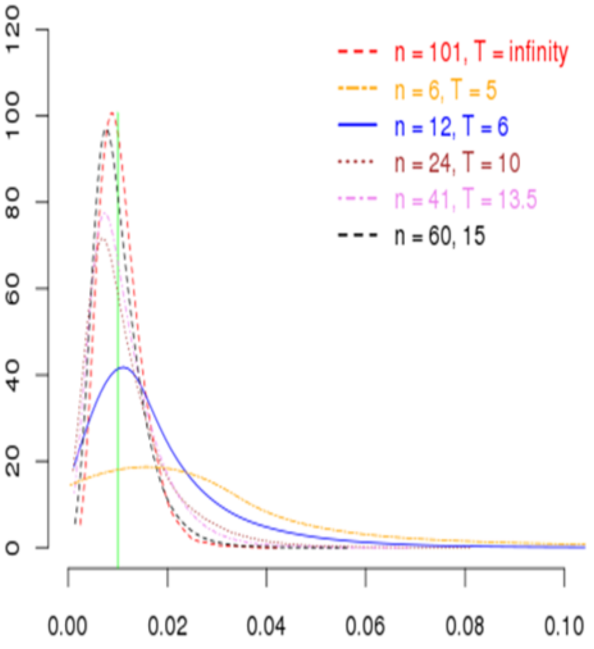


K_4

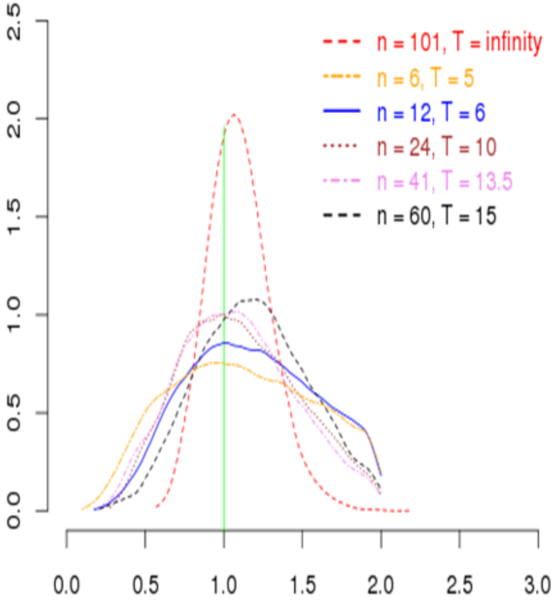
Early phase inference

Need for real-time inference in order to control epidemics through appropriate measures

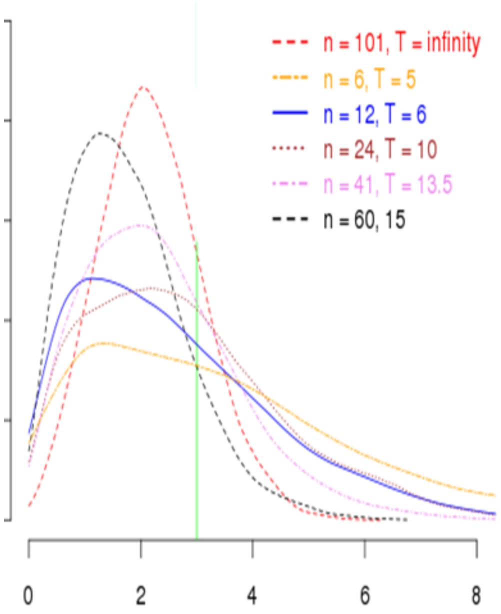
β_0



γ



τ



More data available, more robust inference

Variability between farms

- Most models assume homogeneity of farms
- In reality, each farm is likely to behave differently depending on farm type
- Bio-security practices, areas of location, reaction to epidemic...

Different detection rates per group

As before:

- An individual i makes an infectious contact with a susceptible individual j at rate $\beta_{ij} = \beta_0 \times h_{ij}$
- $h_{ij} = \exp(-\tau \times \rho(i, j))$ is the distance kernel

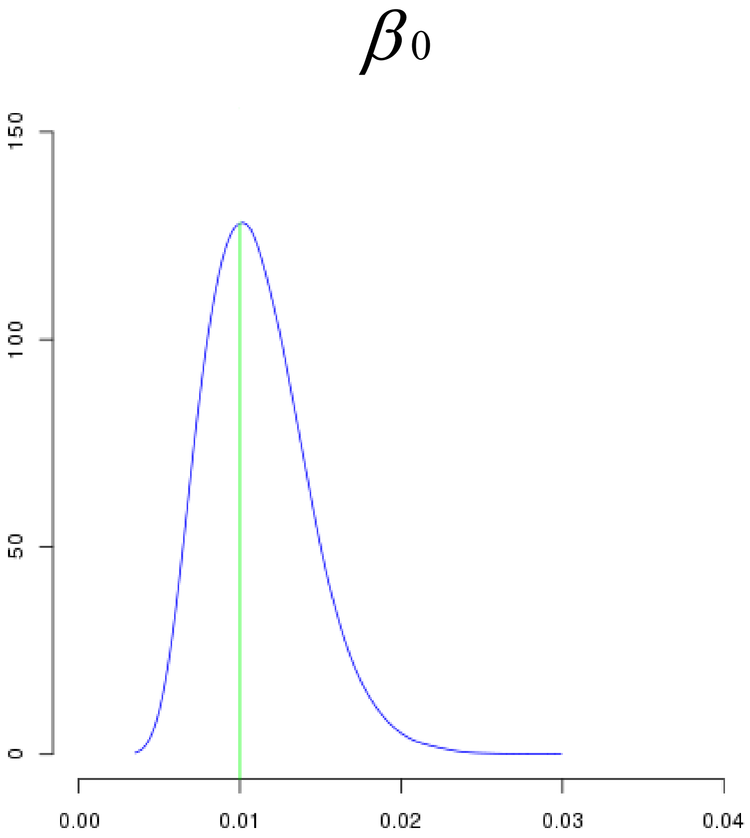
Now also:

- Assume there are 2 groups of farms: K and M
- The infectious period follows: $R_i - I_i \approx Ga(\alpha, \gamma_s)$
  for $s \in \mathrm{K}$ or $s \in \mathrm{M}$

The aim is to learn from data using Bayesian MCMC methods that there are different rates of detection

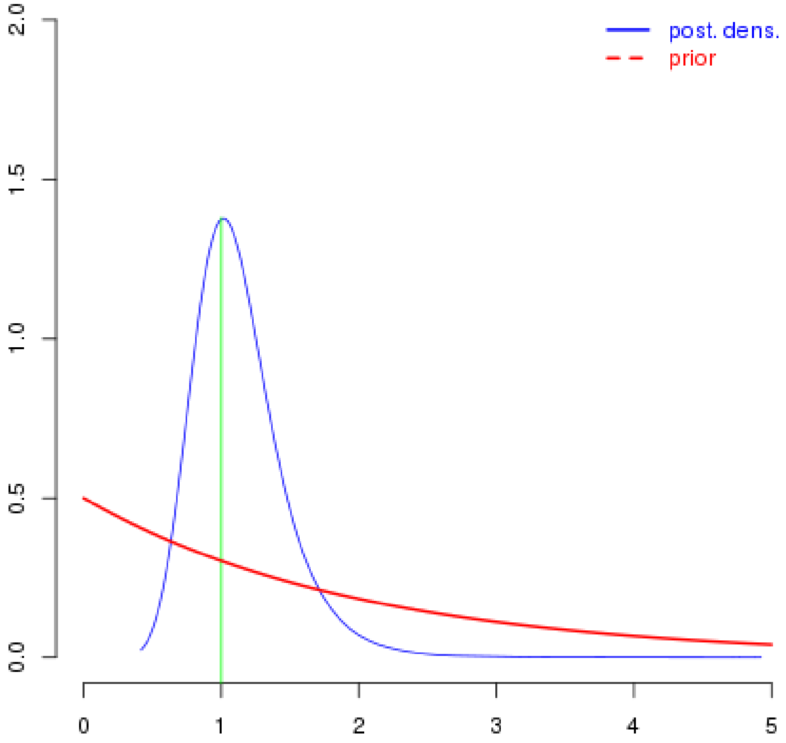
Estimation of the model parameters

- The data consist of the removal times of the infected premises
- There is also knowledge of the groups each site belongs to
- Data augmentation technique with MCMC algorithm using basic non-centering scheme

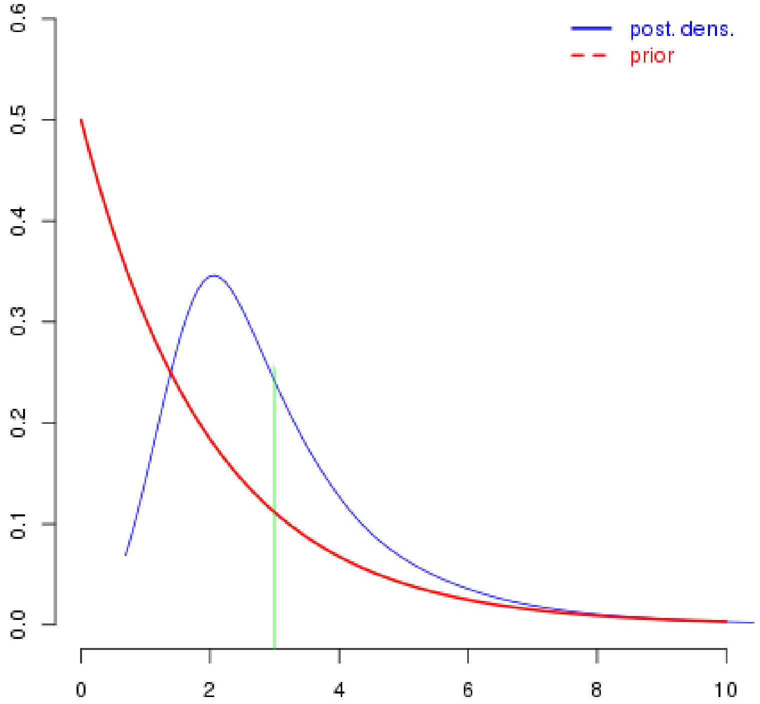


Estimation of the removal rates

γ_K



γ_M



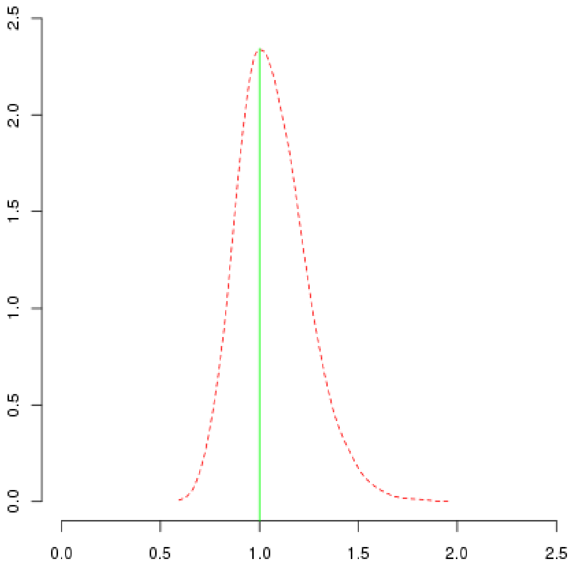
A more general approach would be to assume hierarchical modelling of variability within groups of detection

Varying susceptibility per group of farms

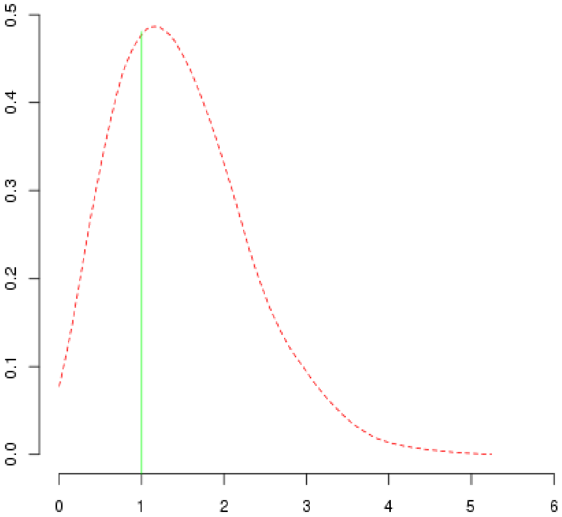
- Assume a closed population of size N divided into L known groups $N = \sum_{G=1}^{L} N_G$
- Model framework same as before except the contact rate depends on the group of the susceptible individual

$$\Pr(j \text{ in } G \in I \text{ in } (t, t+dt) \mid j \in S \text{ at } t) = \beta_0^G h_{ij} dt + o(dt)$$

γ



τ



□ Assume 3 groups with

$N_E = 66, N_G = 98, N_H = 137$

□ #Infections:

$n_I^E = 9, n_I^G = 48, n_I^H = 94$

□ Parameter values:

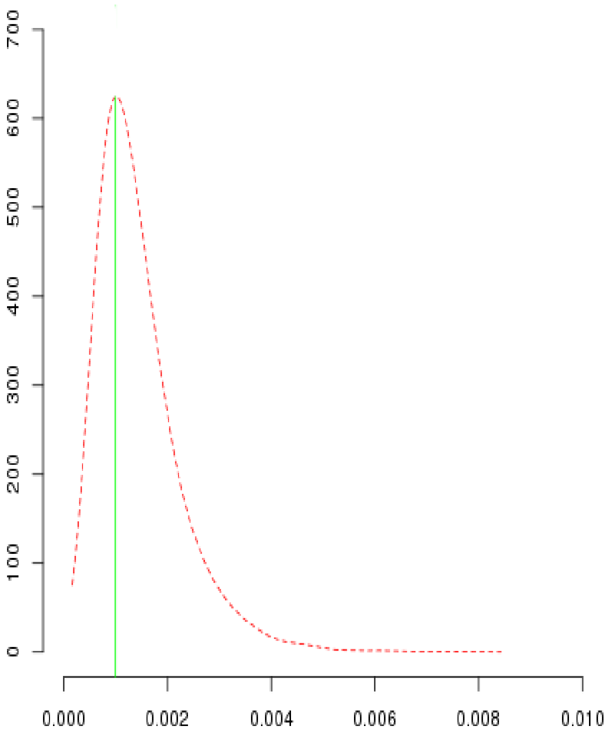
$\beta_0^E = 0.001$ $\beta_0^H = 0.004$

$\beta_0^H = 0.008$ $\gamma = 1$

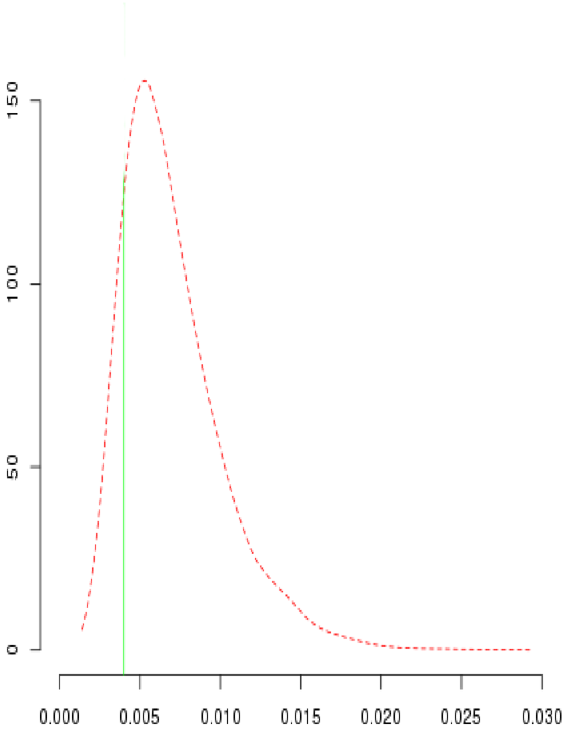
$\tau = 1$

Model picks up difference in parameters

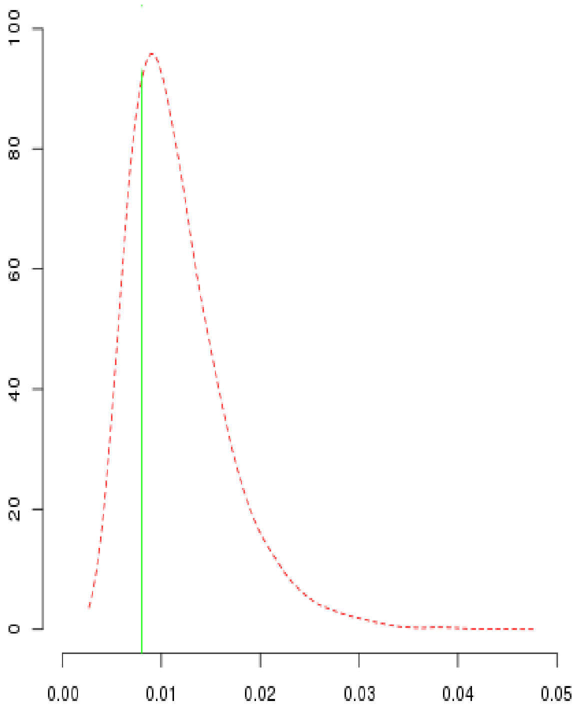
β_0^E



β_0^G



β_0^H

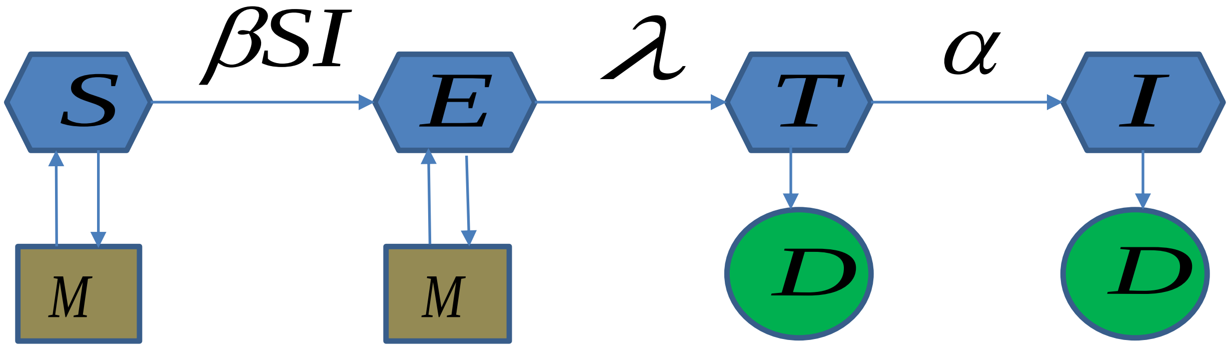


An extension to this model would be to:

❑ assume a random effect model on the susceptibility

❑ Include a varying infectivity in the model

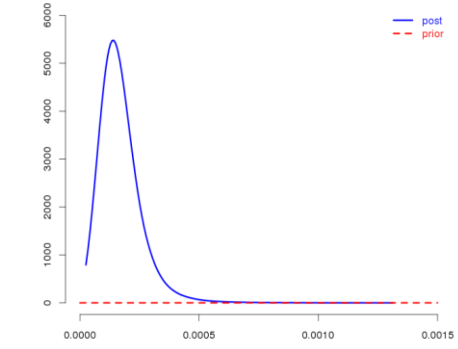
Inferring within herd infection rate for cattle TB



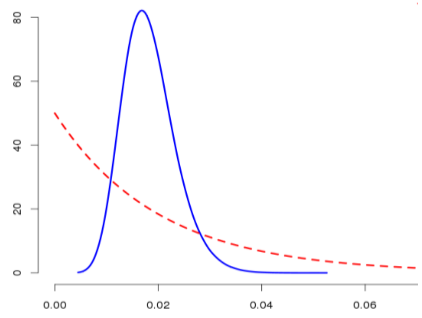
- Disease free herd until earliest movement of the detected individuals (introduction unknown)
- Uniform mixing of animals and infections caused with contact of susceptibles and infectious
- Infection is brought in via movement as no background and hidden infections (perfect sensitivity and specific)
- Data consist of detection times and movements (source VetNet)
- Some animals are considered home bred and disease free status
- Inferring the index case, number of infectious and test sensitives, rates and event times

SETI model results

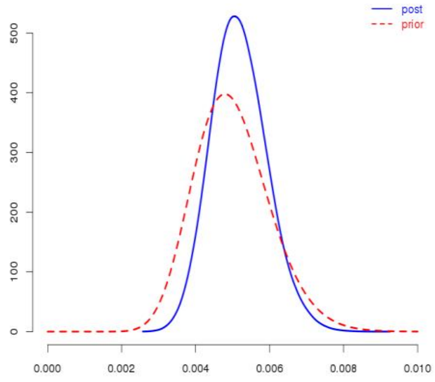
β



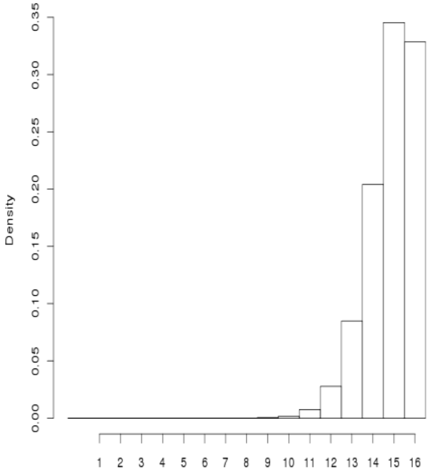
λ



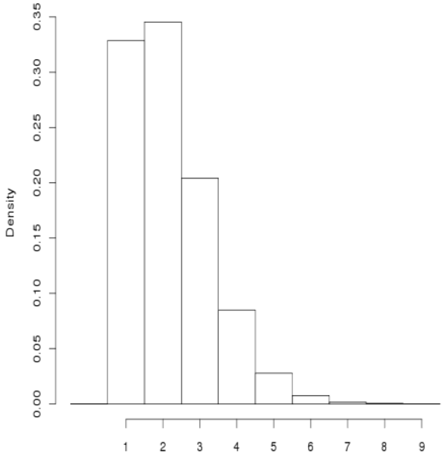
α



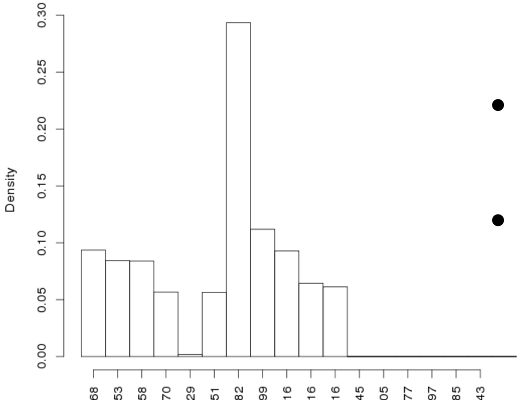
n_{TS}



n_I



index



- Select first introduction among the detected and non-home bred
- Detected can only get exposed after their movement date
- Metropolis within Gibbs
- Reversible jump steps required
- Need of strong prior on α

Initial herd size: 520 total number of reactors: 1 7 max period of disease: 278 days

Conclusions

Directions for future research

- Assessing the amount of data needed for correct kernel identification in real-time epidemic outbreaks

- When do we need to account for variability between farms

- Expand within herd infection model to account for test sensitivity and specificity

- Account for socio-economic effects in modelling epidemics

Hierarchical modelling on detection

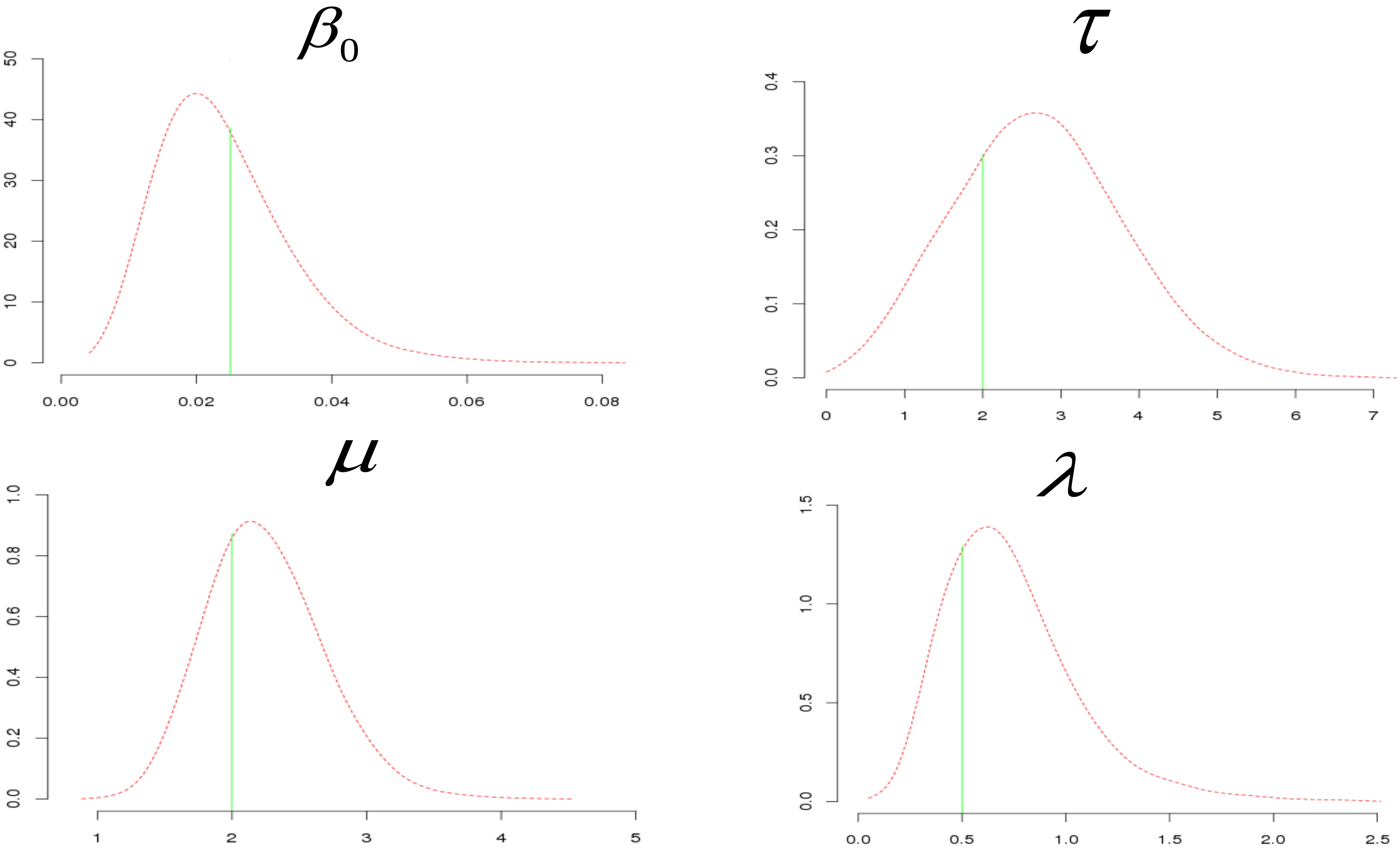
- The infection process is the same as before
- Now, we assume that $R_i - I_i \approx Ga(\alpha, \gamma_i)$

  where $\gamma_i \approx Ga(\mu, \lambda)$
- Interestingly, multiplying the pdfs above, we can integrate out the auxiliary variable γ_i

  using $$\int_0^\infty x^n e^{-ax} dx = \frac{\Gamma(n+1)}{a^{n+1}} (n > -1, a > 0)$$
- We are left to estimate the parameters $(\beta_0, \tau, \mu, \lambda)$

  together with the auxiliary infection times using MCMC

Posterior distributions



β_0

τ

μ

λ

Two DICs

$$DIC_1 = -4E_{\theta,X}\left[\log\left(f(y, X \mid \theta)\right) \mid y\right] + \\ 2E_X\left[\log\left(f(y, X \mid E_\theta[\theta \mid y, X])\right) \mid y\right]$$

$$DIC_2 = -4E_{\theta,X}\left[\log\left(f(y \mid X, \theta)\right) \mid y\right] + \\ 2E_X\left[\log\left(f(y \mid X, \hat{\theta}(y, X))\right) \mid y\right]$$