ScoVa16 workshop 28th-29th January 2016

Doing statistics with dynamics: Bayesian inference and applications for Markov and semi-Markov processes

Contact: glenn.marion@bioss.ac.uk

ScoVa16 workshop 28th-29th January 2016

Doing statistics with dynamics: Bayesian inference and applications for Markov and semi-Markov processes

Contact: glenn.marion@bioss.ac.uk

Joint work with
Chris Pooley, Stephen Bishop (Roslin)
Max Lau, George Streftaris, Gavin Gibson (Heriot Watt University)
Laura Walton, Kokouvi Gamado, Adam Butler, Stephen Catterall (BioSS)
Leo Zijerveld (UoE, SRUC, BioSS – now Bergen)

Mike Hutchings, Ross Davidison (SRUC)
Piran White (York)
Robbie Macdonald (Exeter)
Dez Delahay (APHA)
Dolores Gavier-Widen (Uppsala)
Lisa Yon & Duncan Hannant (Nottingham)

ScoVa16 workshop 28th-29th January 2016

Doing statistics with dynamics: Bayesian inference and applications for Markov and semi-Markov processes

Contact: glenn.marion@bioss.ac.uk

BioSS

HERIOT WATT UNIVERSITY

SRUC

The University of York

ROSLIN

Animal & Plant Health Agency

UNIVERSITY OF EXETER

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

WildTech

Novel Technologies for Surveillance of
Emerging and Re-emerging Infections of Wildlife

SEVENTH FRAMEWORK PROGRAMME

TB Step
Strategies for the eradication
of bovine tuberculosis

The Scottish Government

defra
Department for Environment
Food and Rural Affairs

Research interests

Process and Systems modelling

Development of methods

- Simplification/formulation, analysis and approximation
- Statistical methods for stochastic process-based models

Application areas

- Ecology:

  alien invasive species and dynamic species distribution models, foraging behaviour, population genetics, species range boundaries, evolution of virulence, ecosystem services e.g. pest regulation, population and community dynamics.

- Epidemiology:

  Epidemics in crops, livestock & wildlife populations, the interaction between epidemiology and ecology e.g. demographic fluctuations in wildlife populations, population response to control, foraging behaviour etc.

Stochastic process-based models

Benefits of stochastic process models

- Dynamics critical to understanding many phenomena!
  - Species distribution modelling: invasive aliens
  - Epidemiology, population ecology, behaviour
  - Phylogenetics …?
- Stochasticity models inherent system variability
  - demographic (intrinsic) and environmental (extrinsic)
- Event-driven
  - More directly related to processes than differential equations
  - Facilitates communication with subject scientists
- Understanding system properties via model analysis
  - Simulation straightforward, analytically intractable but can approx.
  - Deterministic analogues more tractable & provide insights
- Stochastic model: self consistent parameter estimation
  - Model formulation handles spatio-temporal correlations

Example application

The ecology of wildlife disease surveillance: disease dynamics and host demography

BioSS

Motivation

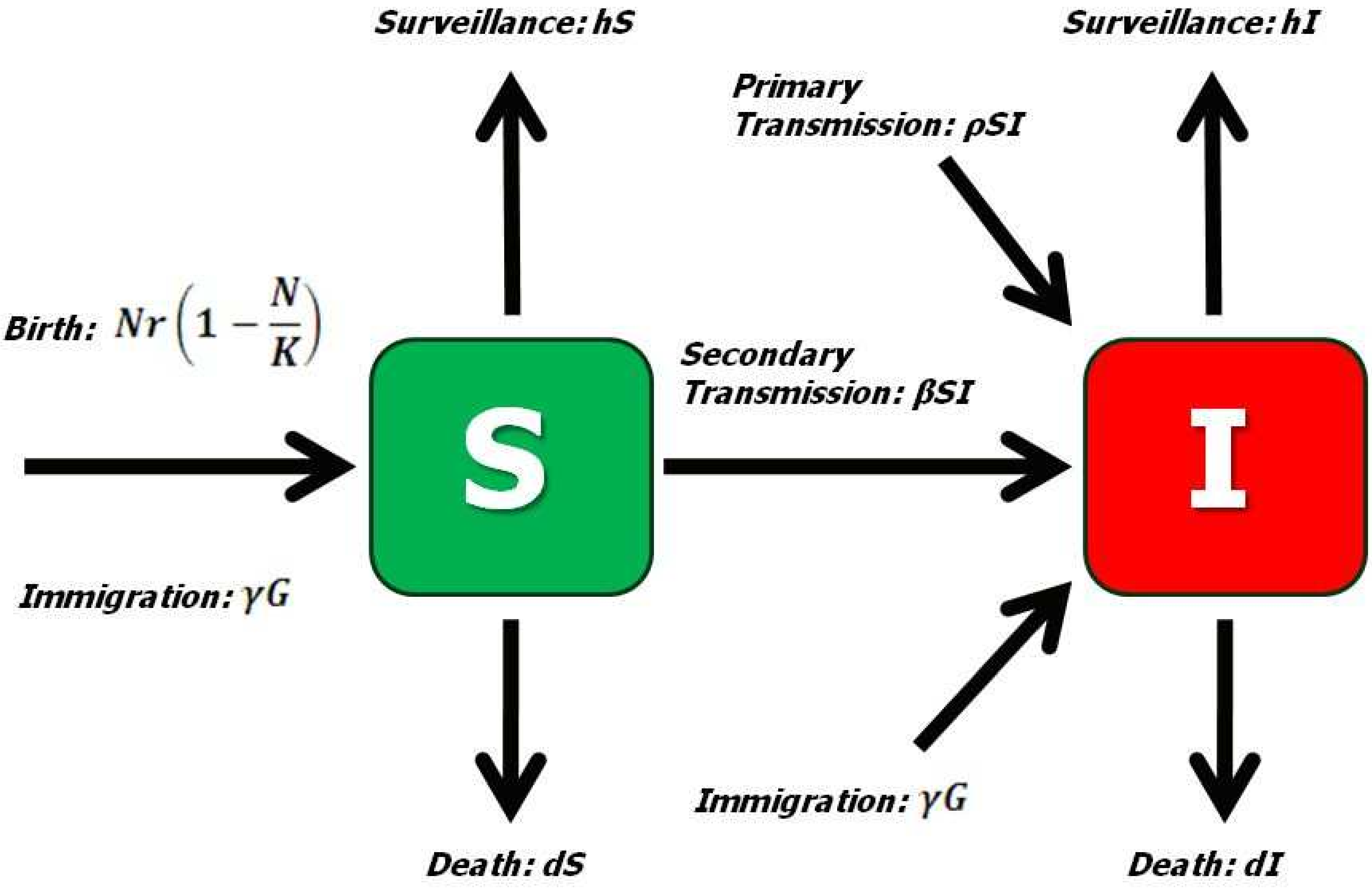
Current theoretical underpinning of wildlife surveillance design tends to:

- be based on relatively simple binomial arguments
- which sometimes account for finite population sizes
- and are typically stratified spatially (by host habitat preferences)
- but **ignore dynamic nature of host populations** (demographic fluctuations & disease dynamics)

Develop models of wildlife disease systems to understand impact of such fluctuations

Population level model

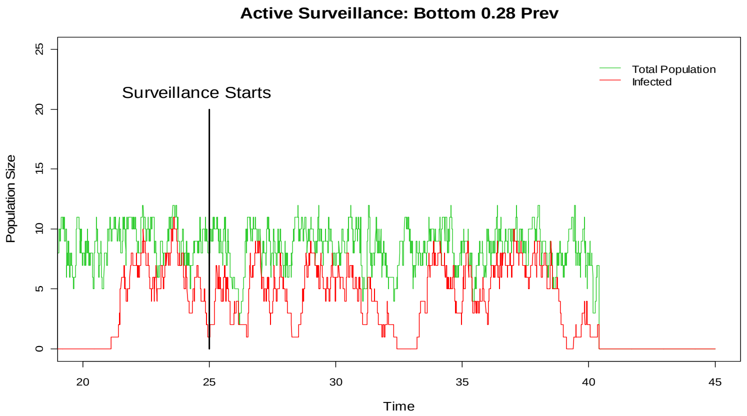
Capturing key demographic impacts on surveillance

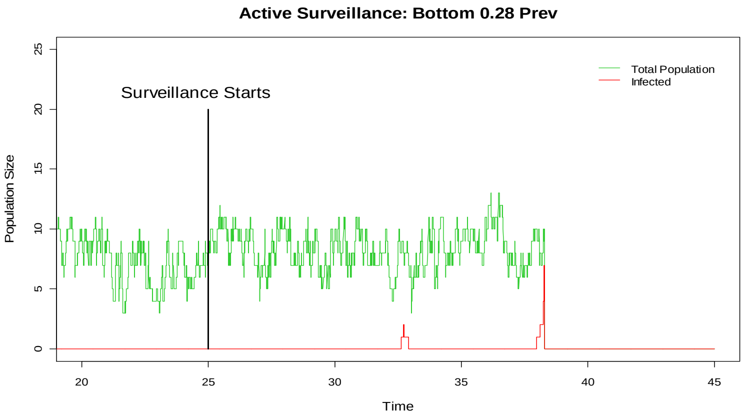


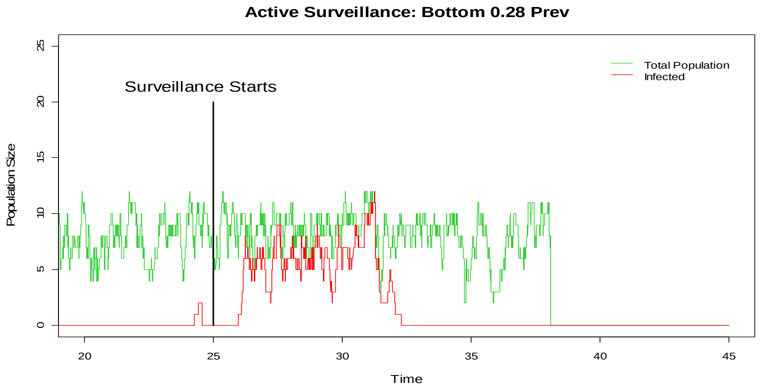
Model properties

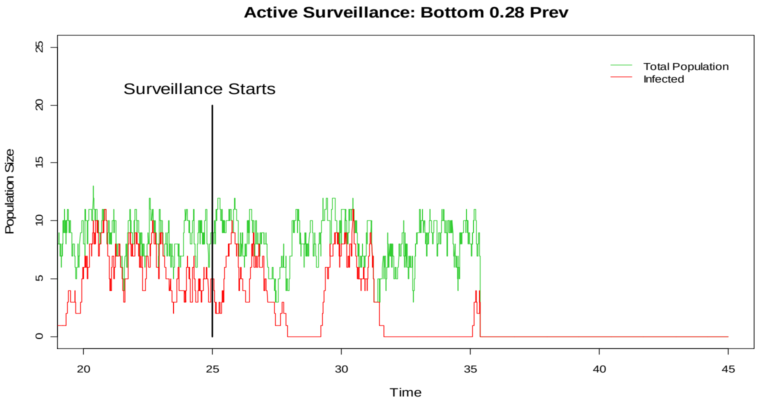
Capturing key aspects of demography & disease dynamics

Variability in population and disease …


Active Surveillance: Bottom 0.28 Prev


Active Surveillance: Bottom 0.28 Prev


Active Surveillance: Bottom 0.28 Prev

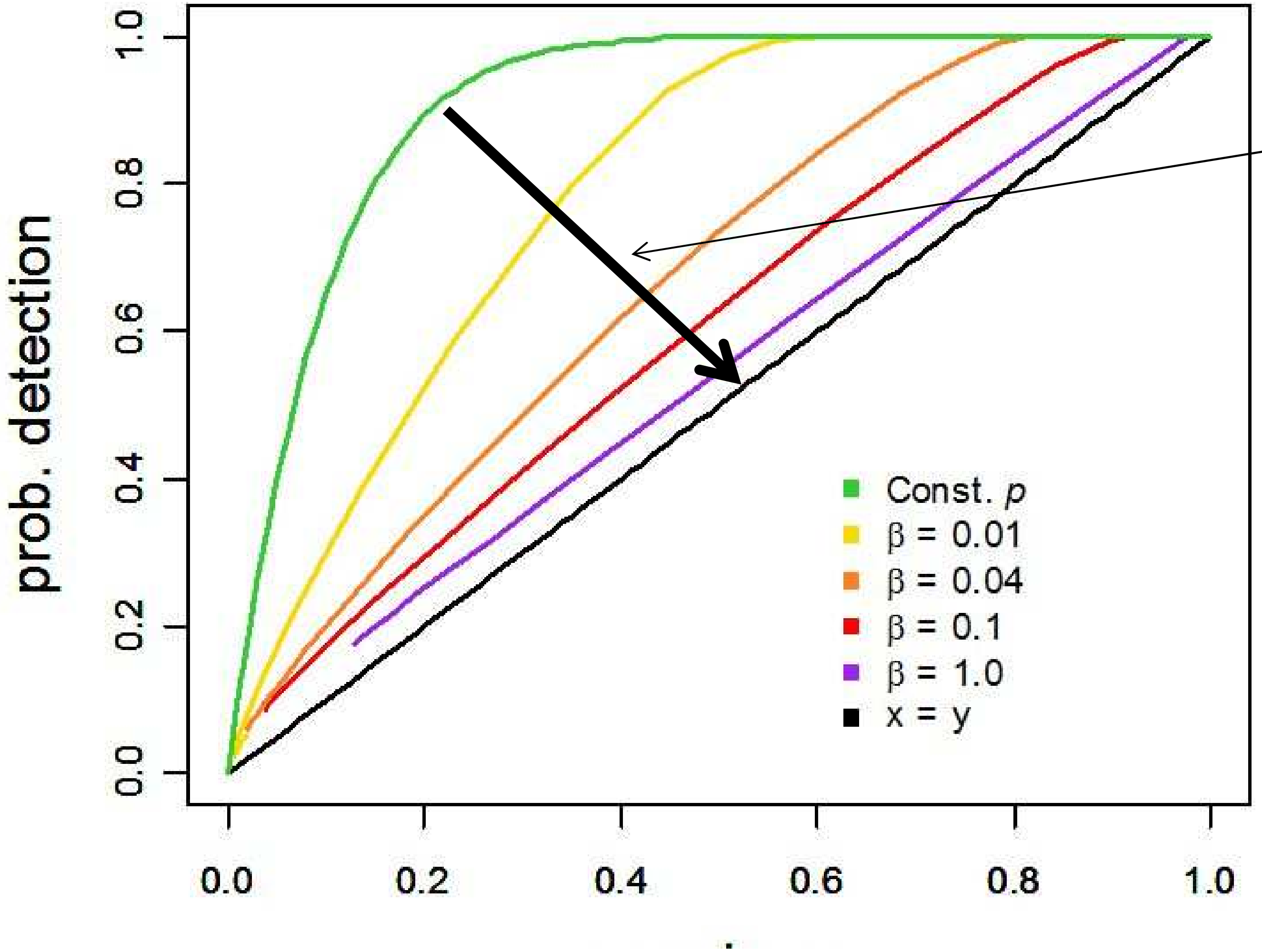

Active Surveillance: Bottom 0.28 Prev

Note: both more and less extreme variations are also possible

Disease detection – suppressed by fluctuations

Probability of detection: no fluctuations in prevalence

$$PD^{Bin} = f(\mathrm{E}[p], m) = 1 - (1 - \mathrm{E}[p])^m$$



Fluctuations increasing

Discrete state-space continuous time Markov processes

- state of system: $s(t)$ at time t and q event types $\{e_i : i = 1, ..., q\}$
- type e_i induces a change δs_{e_i} i.e. $s(t) \rightarrow s(t) + \delta s_{e_i}$
- The rate at which event e_i occurs: $r(e_i, s(t); \mathbf{a})$
- The total event rate at time t is $R(s(t); \mathbf{a}) = \Sigma_{i=1}^{q} r(e_i, s(t); \mathbf{a})$.
- The density associated with occurrence of event e_i at $t + \tau$ is:

$$P(s(t+\tau) = s(t) + \delta s_{e_i} \mid s(t)) \quad = \quad r(e_i, s(t); \mathbf{a}) e^{-\tau R(s(t); \mathbf{a})}$$

Parameter estimation in stochastic pbms

- The complete set of events $\mathcal{E} = \{(E(k), t_k) : k = 1, ..., n\}$.

- A complete realization of the state-space of the stochastic process reconstructed from $\mathcal{E}$ and $s(t_0)$: $S = \{\mathcal{E}, s(t_0)\}$.

- The *complete likelihood* is:

$$P(\mathcal{E} \mid \mathbf{a}, s(t_0)) \;=\; \prod_{k=1}^{n} r(E(k), s(t_{k-1}); \mathbf{a}) e^{-(t_k - t_{k-1}) R(s(t_{k-1}); \mathbf{a})}$$

- Follows from model definition

Inference framework for stochastic process based models

Parameter estimation in stochastic pbms

- The complete set of events $\mathcal{E} = \{(E(k), t_k) : k = 1, ..., n\}$.

- A complete realization of the state-space of the stochastic process reconstructed from $\mathcal{E}$ and $s(t_0)$: $S = \{\mathcal{E}, s(t_0)\}$.

- The *complete likelihood* is:

$$P(\mathcal{E} \mid \mathbf{a}, s(t_0)) \;=\; \prod_{k=1}^{n} r(E(k), s(t_{k-1}); \mathbf{a}) e^{-(t_k - t_{k-1}) R(s(t_{k-1}); \mathrm{a})}$$

$$\mathrm{X}\;\; e^{-(t_f - t_n) R(s(t_n), a)}$$

- If observe system up to time $t_f > t_n$

Inference framework for stochastic process based models

Event and state-space data

- Only observe a subset of events $\mathcal{D}_E$ (the event data)
- hidden events $\mathcal{H}$ unknown part of $\mathcal{E} = (\mathcal{D}_E, \mathcal{H})$
- State space observations: $\mathcal{D}_S = \{\tilde{S}(t_i) : i = 1, ..., N_S\}$
- Need observation model

$$P(\mathcal{D}_S \mid S, \mathbf{a}_N)$$

to relate $\mathcal{D}_S$ to underlying state of system

The combined likelihood

- for parameters $\mathbf{a} = (\mathbf{a}_P, \mathbf{a}_N)$, the complete set of events $\mathcal{E} = (\mathcal{D}_E, \mathcal{H})$ and state-space information $\mathcal{D}_S$

$$\begin{aligned}
L(\mathcal{D}, \mathcal{H}, \mathbf{a}) &\equiv P(\mathcal{D}_S \mid S, \mathbf{a}_N) P(\mathcal{E} \mid \mathbf{a}_P, s(t_0)) \\
&= P(\mathcal{D}_S \mid \mathcal{E}, s(t_0), \mathbf{a}_N) P(\mathcal{E} \mid \mathbf{a}_P, s(t_0)) \\
&= P(\mathcal{D}_S, \mathcal{E} \mid s(t_0), \mathbf{a}_N, \mathbf{a}_P) \\
&= P(\mathcal{D}_S, \mathcal{D}_E, \mathcal{H} \mid s(t_0), \mathbf{a}_N, \mathbf{a}_P) \\
&= P(\mathcal{D}, \mathcal{H} \mid s(t_0), \mathbf{a})
\end{aligned}$$

- where $\mathcal{D} = (\mathcal{D}_E, \mathcal{D}_S)$

A Bayesian approach

- Applying Bayes' rule we obtain the joint *posterior* distribution for the parameters $\mathbf{a}$ and the unobserved events $\mathcal{H}$,

$$P(\mathbf{a}, \mathcal{H} \mid \mathcal{D}) = \frac{P(\mathcal{D}, \mathcal{H} \mid \mathbf{a}) P(\mathbf{a})}{P(\mathcal{D})}$$

- In terms of the complete likelihood, the noise model for state-space observations and the normalisation $P(\mathcal{D})$ and the prior $P(\mathbf{a})$

- Despite not knowing $P(\mathcal{D})$ we can use MCMC to draw samples from posterior

Inference framework for stochastic process based models

A Bayesian approach

- Must specify priors. Typically make independence assumption

$$P(\mathbf{a}) = \prod_{k=1}^{N} P(a_k)$$

Make use of information not contained in data D

- Expert knowledge
- Previous analysis
- Relatively uninformative priors

The parameter posterior

- Obtained by marginalising samples over hidden events $\mathcal{H}$

$$P(\mathbf{a} \mid \mathcal{D}) \;=\; \int_{\mathcal{H}} P(\mathbf{a}, \mathcal{H} \mid \mathcal{D}) d\mathcal{H}$$

- Sampling based approaches allow calculation of essentially *any statistics of interest*
  - e.g. number and nature of missing events
  - possibly of scientific interest but difficult to observe directly

Markov chain Monte Carlo

- To generate samples of missing events $\mathcal{H}_i$ from the posterior for a fixed parameter vector $\mathbf{a}$

- Iterate the following

1. propose $\mathcal{H}_i \rightarrow \mathcal{H}'$ with probability $q(\mathcal{H}_i, \mathcal{H}')$
2. set $\mathcal{H}_{i+1} = \mathcal{H}'$ with probability $\min\{1, \dfrac{P(\mathcal{D}, \mathcal{H}' \mid \mathbf{a}) q(\mathcal{H}', \mathcal{H}_i)}{P(\mathcal{D}, \mathcal{H}_i \mid \mathbf{a}) q(\mathcal{H}_i, \mathcal{H}')}\}$
3. else $\mathcal{H}_{i+1} = \mathcal{H}_i$.

- Note acceptance probability really the ratio of the posteriors $\mathcal{H}'$ and $\mathcal{H}_i$ But normalisation $P(\mathcal{D})$ and prior $P(\mathrm{a})$ cancel from acceptance ratio

Sampling hidden events

- Typically sufficient to specify proposal probabilities so that can add, remove or move events
- Changing numbers of events changes state-space dimension of Markov chain: reversible jump McMC.
- Need to be careful that acceptance ratios account for this but discrete state-space makes this relatively easy.

Gibbs sampling of parameters

- If proposal distribution $q(\mathbf{a}_i, \mathbf{a}') \sim q(\mathbf{a}')$ and is proportional to the posterior

- Then acceptance probability is 1 min$\{1, \dfrac{P(\mathcal{D}, \mathcal{H} \mid \mathbf{a}')P(\mathbf{a}')q(\mathbf{a}', \mathbf{a}_i)}{P(\mathcal{D}, \mathcal{H} \mid \mathbf{a}_i)P(\mathbf{a}_i)q(\mathbf{a}_i, \mathbf{a}')}\}$

- If choose conjugate priors then can identify marginal posterior distributions

- Can show that for linear rate parameters in Markov processes Gamma priors are conjugate

Gamma priors and Gibbs sampling

Event type X governed by linear parameter a_i i.e. $r(X, S, a) = a_i \tilde{r}(X, S, a_{-i})$

If event X occurs n_X times in event history then complete likelihood

$$P(D, H \mid a_i) \propto a_i^{n_X} e^{-a_i \int_{t_0}^{t_f} \tilde{r}(X,S,a_{-i})} dt$$

Gamma prior $P(a_i) = Ga(\alpha, \beta) \propto a_i^{\alpha} e^{-a_i \beta}$

Leads to the posterior

$$\begin{aligned} P(a_i \mid D, H) &\propto a_i^{\alpha + n_X} e^{-a_i(\beta + \int_{t_0}^{t_f} \tilde{r}(X,S,a_{-i}) dt)} \\ &= Ga(\alpha + n_X, \beta + \int_{t_0}^{t_f} \tilde{r}(X, S, a_{-i}) dt) \end{aligned}$$

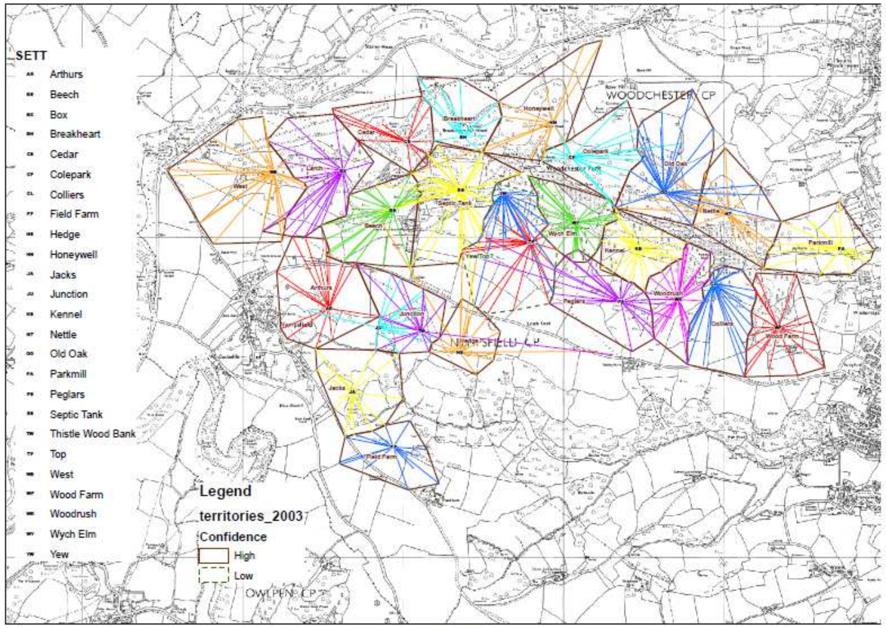
Note: dependence on other events and parameters absorbed into normalisation

Population & disease dynamics in badgers

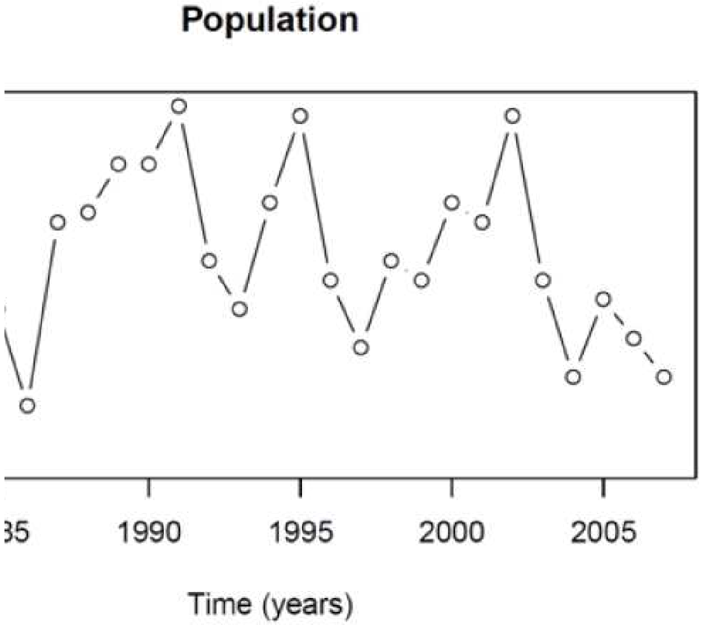
Badgers play a key role in the dynamics of TB in cattle which is a major issue for U.K. agriculture

Woodchester Park data

- Long term study site for badgers and bTB 1982-present
- Very detailed data on **individuals** and **social groups**
- Use population level and individual based models
- Infer key demographic and disease parameters



(a) Woodchester Park baitmarking map



badger
Meles meles
teixó
tejón

How to model?

Model structure and parameter inference

- Dynamic stochastic process-based model
  - Demographic fluctuations
  - Disease dynamics
  - Individual based (capture-mark-recapture)

- Observation model
  - Do not observe all animals
  - Diagnostic tests are imperfect

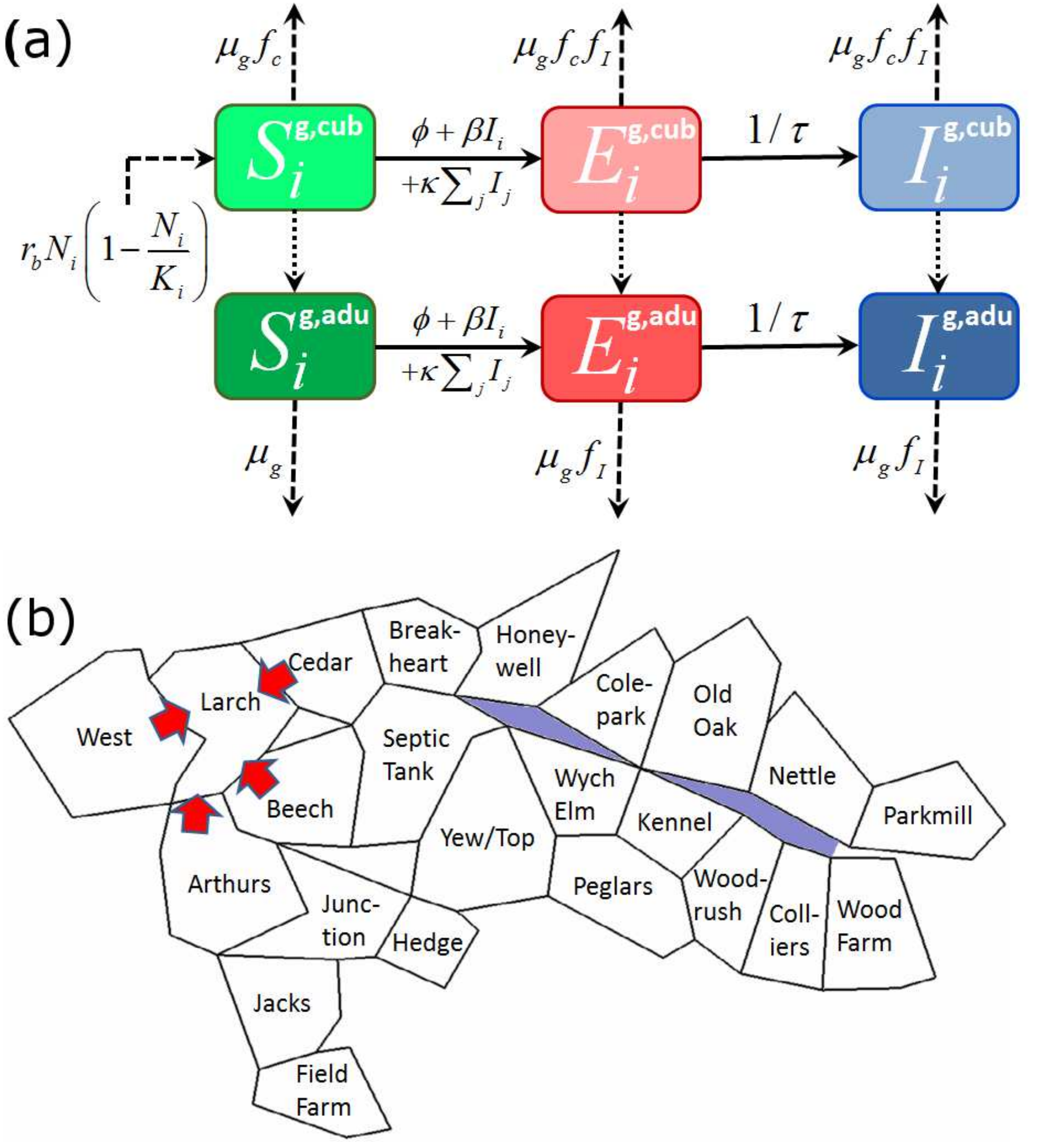
Use dynamic continuous time stochastic models

Can apply Bayesian inference tools

Individual-based model parameterised using Woodchester park data

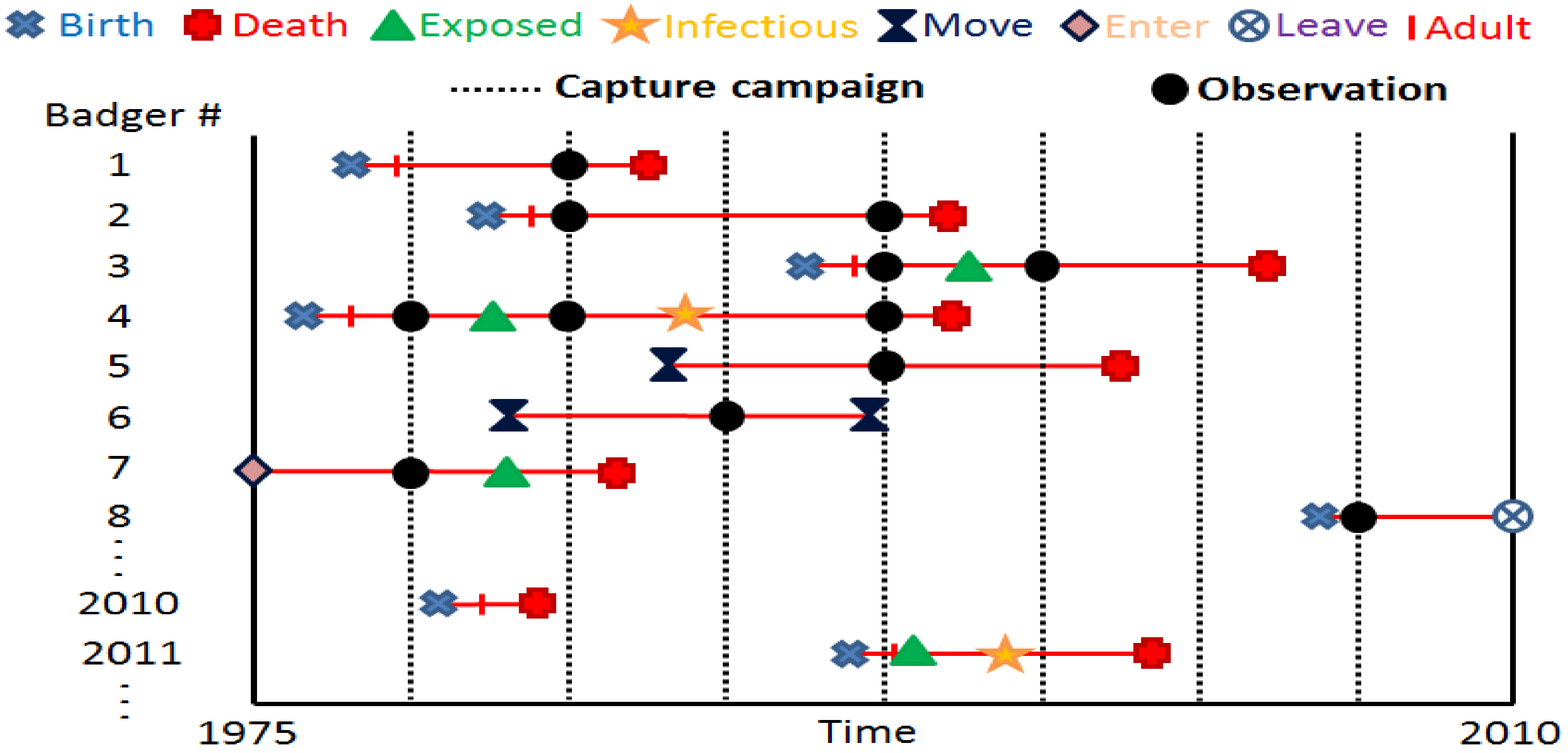
Individual based stochastic model accounts for

- *Demography including births, deaths*
- *Age and sex*
- *Disease induced mortality*
- *Transmission with and between social groups*
- *External transmission*



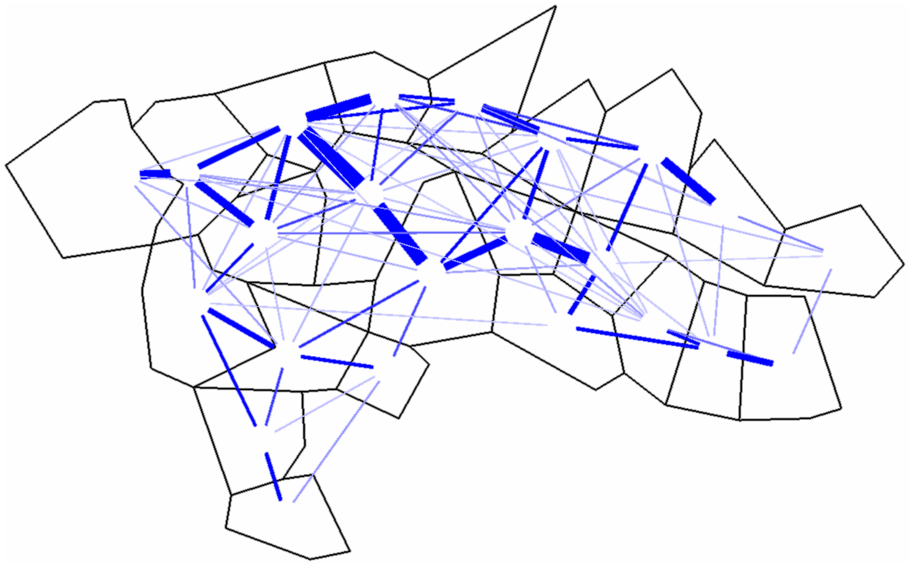
Observation model accounts for

- *Individual trapping*
- *Sex and age differences in trappability*
- *Seasonal variation in capture rates*
- *Se & Sp of diagnostic tests*



Bayesian Statistical treatment used to infer parameter values and missing history/event data e.g. *Births, deaths, infection events etc.*

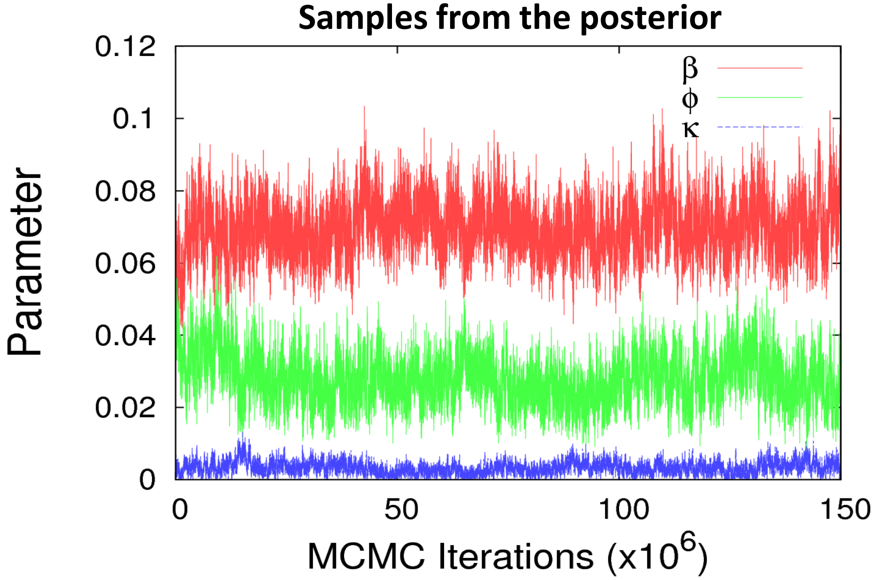
Bayesian Statistical treatment used to infer parameter values and missing history/event data e.g. *Births, deaths, infection events etc.*



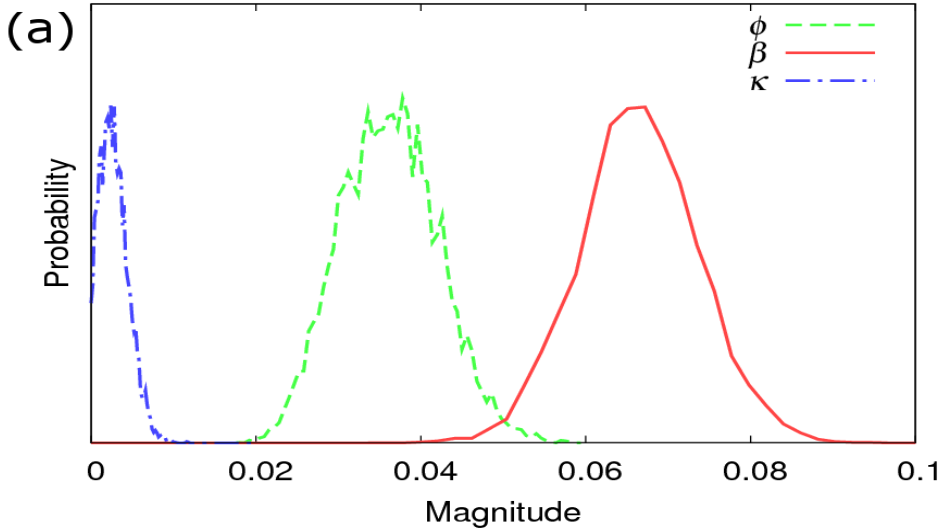
The average inferred movement of badgers between different social groups

Bayesian inference: information encoded in posterior

Sampled using Markov chain Monte Carlo MCMC methods

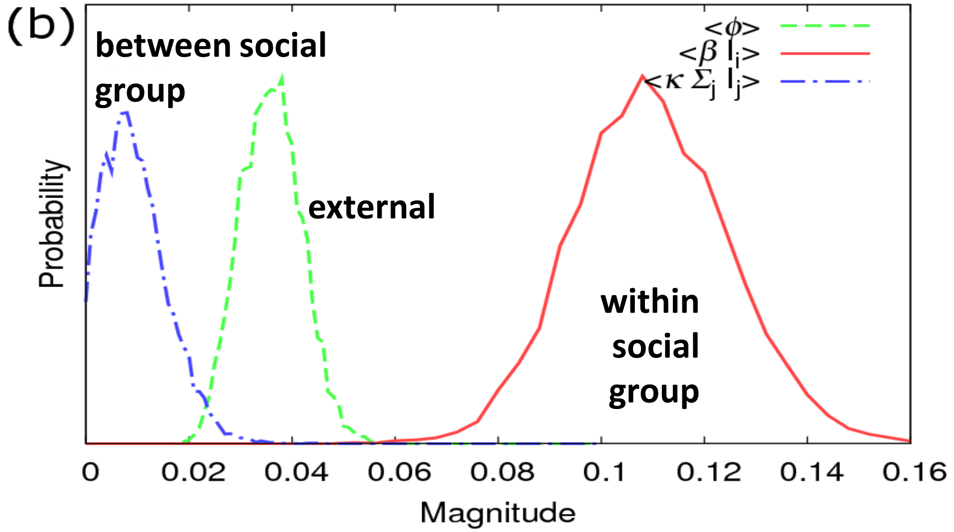


Use samples to estimate statistics of interest



Parameter distributions and related means & credible intervals

Or more complex statistics …



Probability distributions describing the rate of disease transmission associated with different routes of infection

Can use WP data to estimate a wide range of parameter values

Type	Param	Mean	95% CI	Description
Badger Demographic	r_b	0.603	0.542 – 0.671	Birth rate of badgers
	μ_M	0.281	0.263 – 0.311	Mortality rate in male badgers
	μ_F	0.210	0.189 – 0.238	Mortality rate in female badgers
	f_C	1.89	1.68 – 2.18	Factor increase in mortality for cubs (< 1 year old)
	f_I	3.41	2.90 – 4.01	Factor increase in mortality for infected badgers
	$\overline{K}$	28.1	22.9 – 36.1	Mean carrying capacity per unit area of sett
	σ_K	10.8	7.3 – 16.2	Standard deviation in carrying capacity
Disease Transmission	ϕ	0.035	0.024 – 0.049	Rate of disease coming externally
	β	0.064	0.052 – 0.081	Within social group transmission
	κ	0.0035	0 – 0.0072	Between social group transmission
Trapping Probability	d_M	0.38	0.36 – 0.40	Average probability of trapping male badger
	d_F	0.36	0.34 – 0.38	Average probability of trapping female badger
	e_C	1.60	1.49 – 1.72	Factor increase in trapping probability for cubs (< 1 year old)
	e_I	1.39	1.29 – 1.51	Factor increase in trapping probability for infected badgers
	Δp_d	0.34	0.31 – 0.37	Seasonal variation in trapping probability
	θ_d	0.43	0.42 – 0.44	Point in the year with maximum trapping probability
Tests	Se_{ELISA}	0.55	0.51 – 0.60	Sensitivity of Brock-ELISA test
	Sp_{ELISA}	0.946	0.935 – 0.952	Specificity of Brock-ELISA test
	Se_{Cult}	0.37	0.32 – 0.45	Sensitivity of culture test
	Sp_{Cult}	1	1 – 1	Specificity of culture test
	Se_{SP}	0.64	0.50 – 0.79	Sensitivity of Stat-Pak test
	Sp_{SP}	0.959	0.927 – 0.982	Specificity of Stat-Pak test
	$\mathrm{Se}_{\gamma-IFN}$	0.78	0.63 – 0.90	Sensitivity of γ-IFN test
	$\mathrm{Sp}_{\gamma-IFN}$	0.912	0.869 – 0.943	Specificity of γ-IFN test

Summary from WP analysis

Can estimate

- Detection rates and diagnostic test sensitivity and specificity
- Demographic parameters
- Disease parameters including DIM and transmission rates

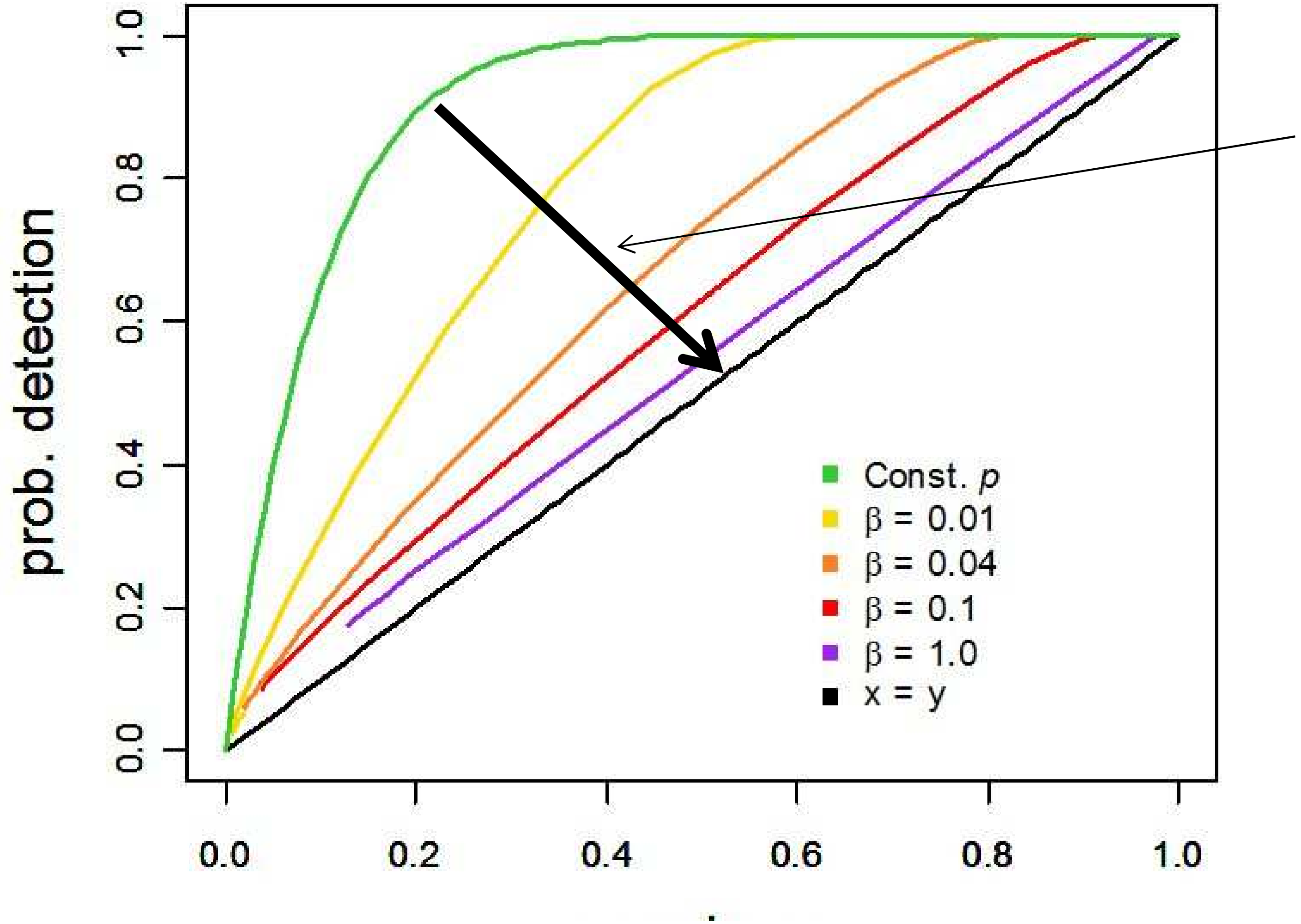
Conclusions and future work

- Transmission within group β dominates between group κ
- External Transmission ϕ i.e. from outside modelled population important
- **Plan to assess correlation of ϕ with infection in surrounding farms**
- **Get better estimates of Se & Sp for new tests using most recent WP data**
- Not able to estimate dispersal or behaviour change associated with culling since no culling events in Woodchester Park data
- Need to turn to other sources of data e.g. RBCT and VetNet

Disease detection – suppressed by fluctuations

Probability of detection: no fluctuations in prevalence

$$PD^{Bin} = f(\mathrm{E}[p], m) = 1 - (1 - \mathrm{E}[p])^m$$



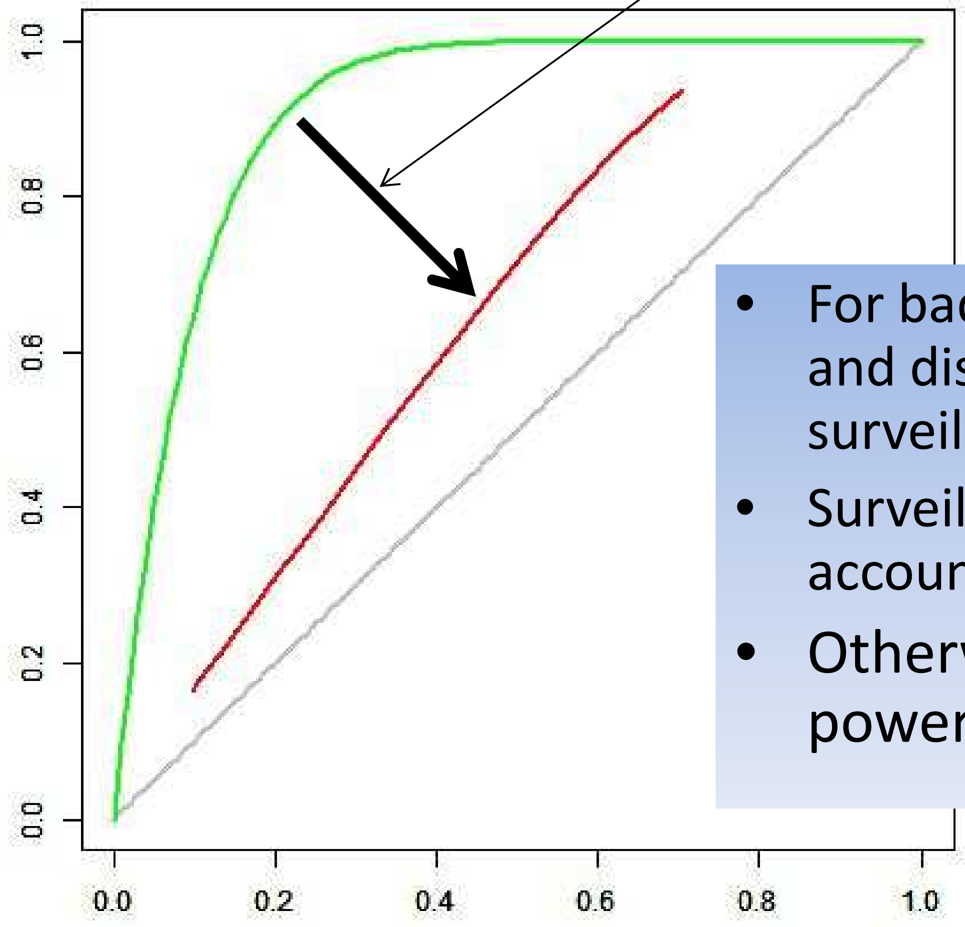
Fluctuations increasing

Disease detection

Probability of detection: no fluctuations in prevalence

Probability of detection: badgers & bTB

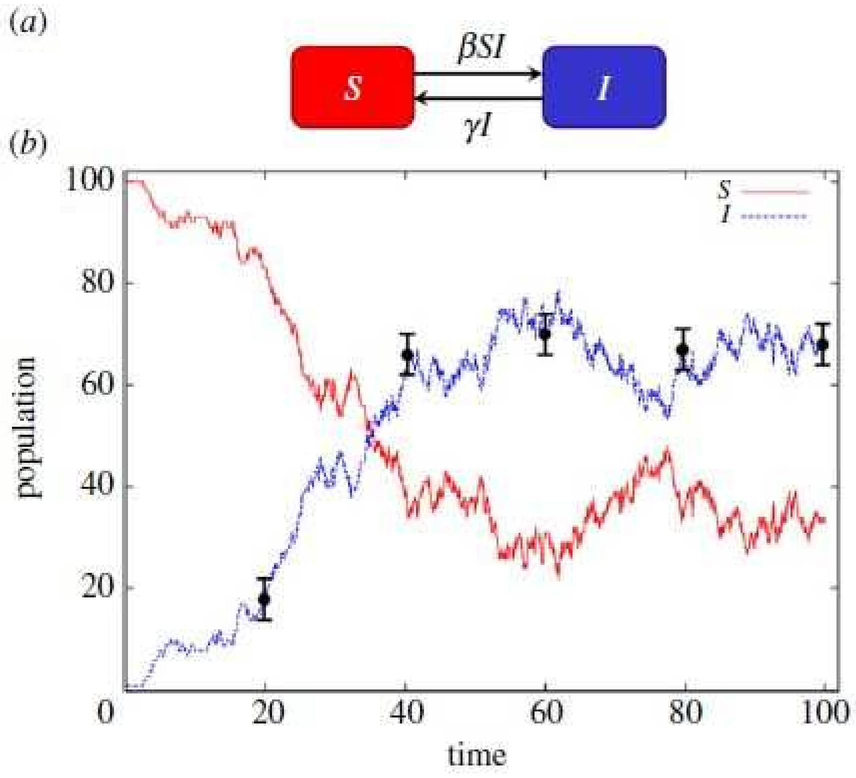
Fluctuations increasing



- For badgers and bTB demographic and disease fluctuations undermine surveillance
- Surveillance design & analysis need to account for such fluctuations
- Otherwise danger of under powered studies

Inference in stochastic process based models

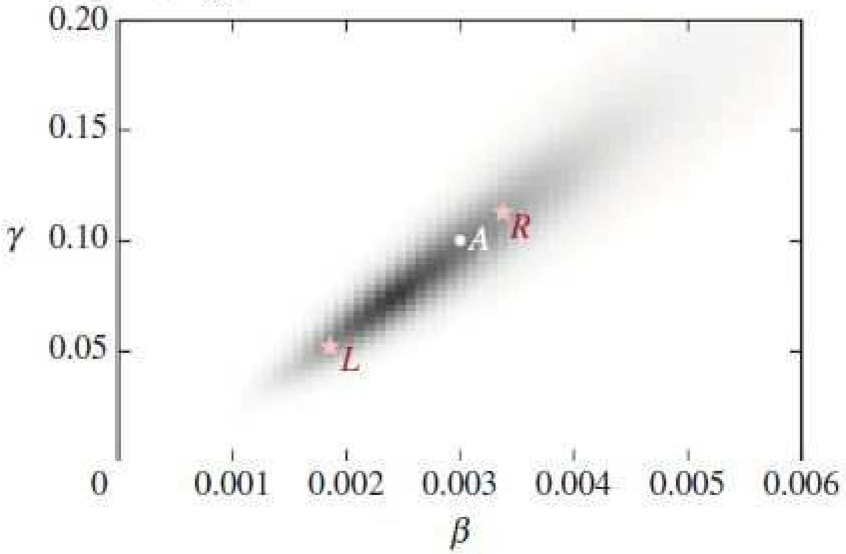
Model-based-proposals



Posterior

$$\pi(\theta, \xi | y) \propto \pi(y | \xi) \pi(\xi | \theta) \pi(\theta)$$

$\hat{\pi}(\theta | y)$ marginal parameter Posterior

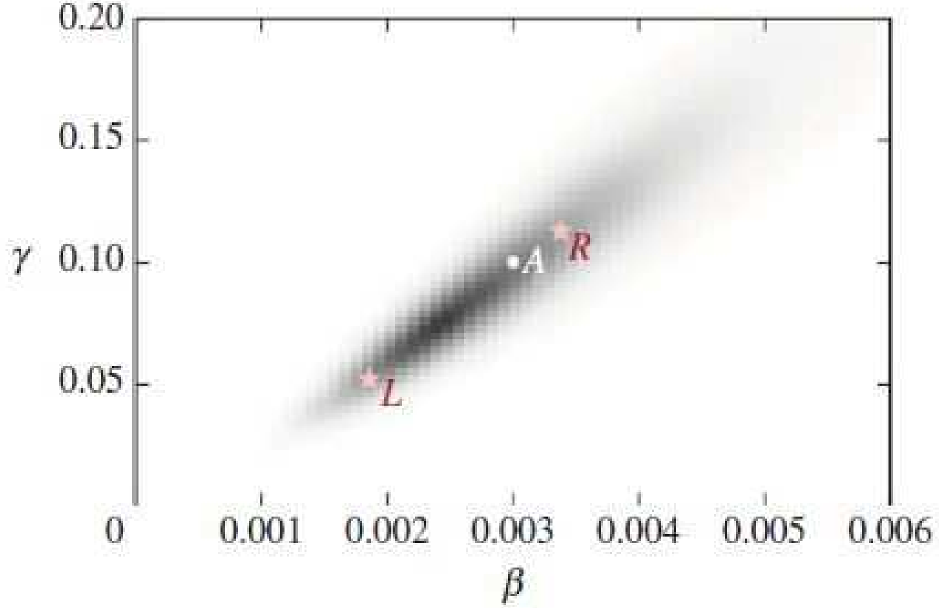


Complete likelihood
$$\pi(\xi | \theta) \propto \mathrm{e}^{-R(\theta, n_T) \times (T - t_{N_\xi})} \times \prod_{k=1}^{N_\xi} r_{\varepsilon_k}(\theta, n_k) \mathrm{e}^{-R(\theta, n_k) \times (t_k - t_{k-1})}$$

Observation model
$$\pi(y | \xi) = \prod_{m=1}^{M} N(v_m; n_{c_m}(\tau_m), \sigma_m^2)$$

Inference in stochastic process based models
Model-based-proposals

$\hat{\pi}(\theta|y)$ marginal parameter posterior



Posterior involves parameters and latent history

$$\pi(\theta, \xi|y) \propto \pi(y|\xi)\pi(\xi|\theta)\pi(\theta)$$

Key problem with M-H

Would like to make large changes in parameters

But this requires large changes in latent history

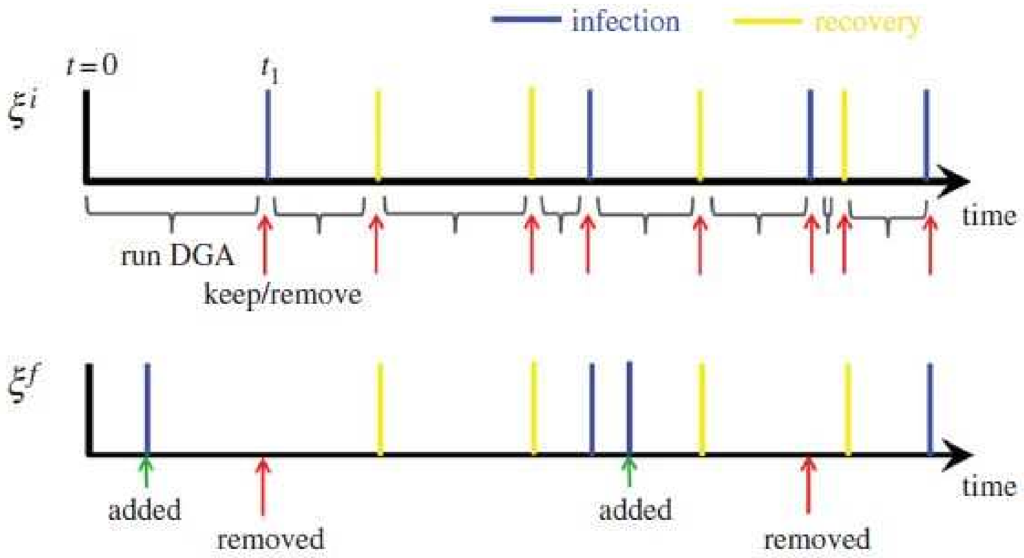
M-H limited to small changes in parameter space

Diffusive exploration of parameter space – slow!

Inference in stochastic process based models

Model-based-proposals

Propose large change in parameter space
Then create a history to match this based on model dynamics



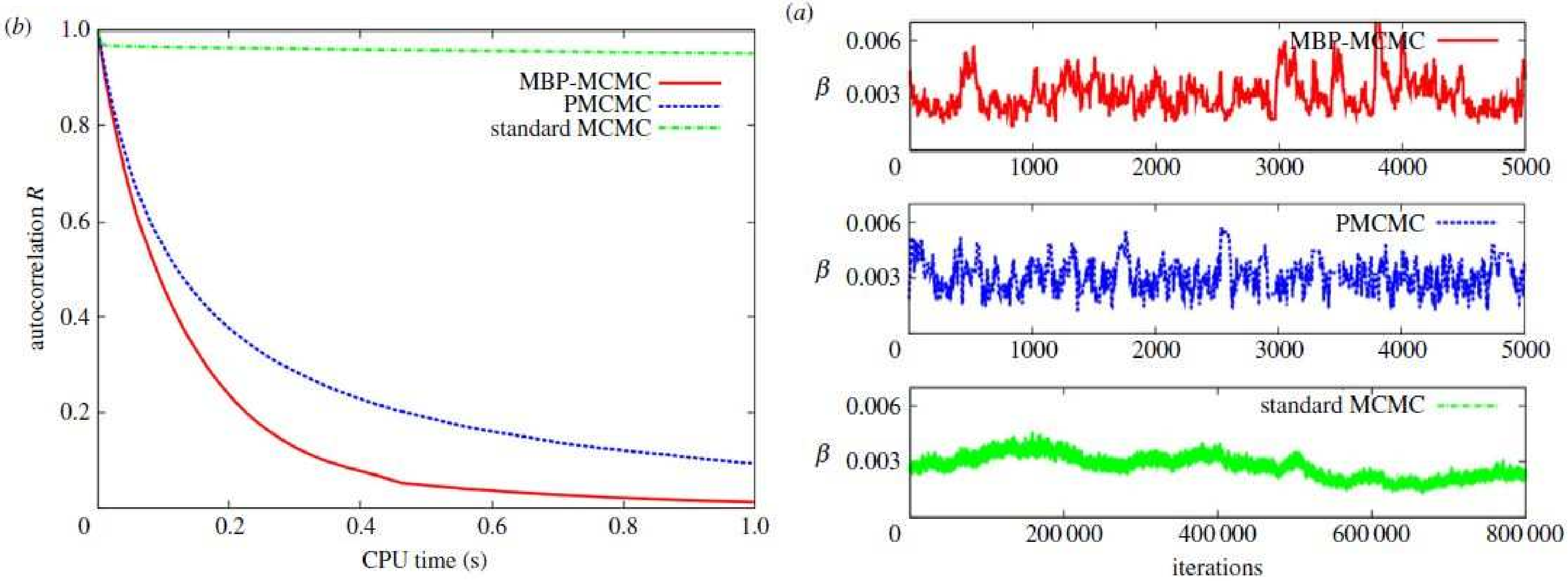
Add events based on difference in rates $\Delta r_e(t) = \max\{0, r_e(\theta^f, n^f(t)) - r_e(\theta^i, n^i(t))\}.$

Also keep some old events $P_k^{\text{keep}} = \max\left\{1, \frac{r_{\varepsilon_k}(\theta^f, n^f)}{r_{\varepsilon_k}(\theta^i, n^i)}\right\}$

Use M-H accept/reject to keep on track $P_{\text{MH}} = \min\left\{1, \frac{\pi(y|\xi^f)\pi(\theta^f)}{\pi(y|\xi^i)\pi(\theta^i)}\right\}$

Inference in stochastic process based models
Model-based-proposals improve performance

Auto-correlation

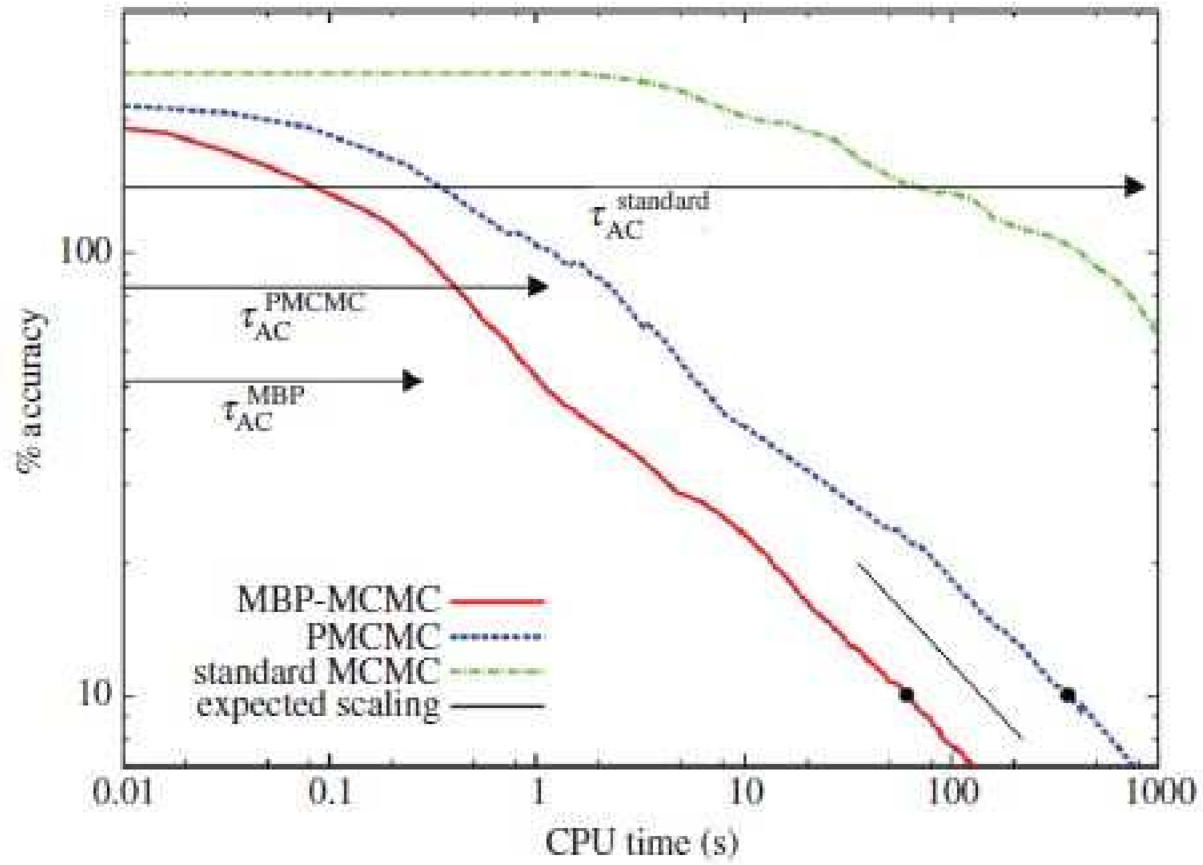


Compared with Standard MCMC and particle MCMC

Inference in stochastic process based models

Model-based-proposals improve performance

Computation time



Compared with Standard MCMC and particle MCMC

Inference in stochastic process based models

New methodology

Current research focussed on two main areas

- Model assessment/model selection
    - Latent residual approach: based on Sellke thresholds
    - Designed to assess fit for different sub-processes in model
    - e.g. spatial spread modelled via kernel density
    - Can be computed cheaply from augmented data
    - Useful as can identify areas where model is deficient

- The need for speed!
    - Increasing efficiency with which we can draw samples from the posterior
    - e.g. non-centred parameterisations/schemes
    - Model-based-proposals