MAC-MIGS Sandpit Nov 22nd 2019

glenn.marion@bioss.ac.uk

Head of Research

Biomathematics & Statistics Scotland is:

- A research institute, partly funded by Scottish Government, scientifically independent but formally part of the James Hutton Institute
- A member of the Scottish Environment Food & Agriculture Research Institutes (SEFARI) collective
- Eligible for UKRI funding since June 2018 (happy to discuss potential bids)

BioSS

The James Hutton Institute

Royal Botanic Garden Edinburgh

SEFARI

LEADING IDEAS FOR BETTER LIVES

Moredun Research Institute

The Rowett Institute

SRUC

BioSS Research

We aim to be

- A world leader in developing and applying mathematical and statistical techniques to improve agriculture, food and the environment
- Applying & helping develop cutting edge methodology for:
  - statistical analysis of large and/or complex datasets;
  - **mathematical modelling of processes and systems**;
  - statistical / computational bioinformatics and genetics

Biomathematics & Statistics Scotland

BioSS Research

Collaboration, Collaboration, Collaboration

- Co-construction of data analysis and modelling
- Statistical design of field trials, experiments, and studies
- **Applying & advancing methods to extract maximum value from data and from models**
- Estimating quantities that are impractical to measure directly
- Using modelling and data analytics to understand complex systems
- Co-constructing actionable insights to manage complex systems

BioSS Research

Trans-disciplinary success sustained through tension

Applied quantitative research: close links including embedding of BioSS staff with applications focussed research institutes ensures research is applications focussed and responsive to changing priorities.

Applied methodological research: strong links with methodological community to ensure continued application and development of cutting edge methods.

Research continuum

Biomathematics & Statistics Scotland

BioSS research

Methodological Research Themes

- Statistical methodology
- Bioinformatics and Statistical Genetics
- **Process and Systems modelling**

This research is inspired by stakeholder problems

Methodological research

Process and systems modelling

embed process and systems models within statistical framework

methods for simplification, analysis and approximation of complex models

co-construction of models for scientific discovery

Qualitative mathematical model

	C	P	SE	GE	ED	NCH
C	-1	-1	0	0	0	0
P	1	0	-1	-1	0	0
SE	0	1	0	0	0	0
GE	0	0	0	-1	1	1
ED	0	0	0	0	-1	0
NCH	0	0	0	0	0	-1

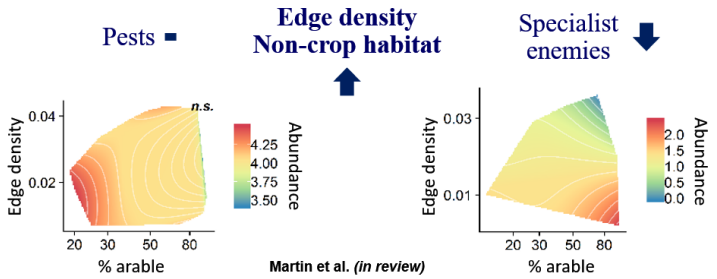
System of differential equations

Pests

$$\frac{dP}{dt} = aP - \mu_P P - \frac{f_s PN}{P + K_s} - \frac{f_g PG}{P + K_g} + \delta_P$$

Specialist enemies

$$\frac{dN}{dt} = -\mu_N N + \frac{\varepsilon_s f_s PN}{P + K_s} + \delta_N$$



Collaboration within MAC-MIGS

4 areas of potential interest

1. Predicting spread of disease, pests and invasive species on landscapes

Statistical methods to infer and test spread mechanisms both when population at risk is known e.g. farm type/location (HPAI) and unknown e.g. wildlife hosts (ASF)

2. Combine process understanding and data for scientific discovery

Using mathematical and statistical methods to develop data-driven tests of alternative hypotheses e.g. understanding within-host disease progression Tb in badgers

3. Phenotype from genotype

Current applications to estimating genetics effects on disease resistance, infectivity and recoverability phenotype

4. Systems modelling tools for better outcomes

or How to escape the law of unintended disaster (LOUD)! e.g. natural and chemical pest control, unintended effects of culling to control wildlife disease

Collaboration within MAC-MIGS

4 areas of potential interest

1. Predicting spread of disease, pests and invasive species on landscapes

Statistical methods to infer and test spread mechanisms both when population at risk is known e.g. farm type/location (HPAI) and unknown e.g. wildlife hosts (ASF)

2. Combine process understanding and data for scientific discovery

Using mathematical and statistical methods to develop data-driven tests of alternative hypotheses e.g. understanding within-host disease progression Tb in badgers

3. Phenotype from genotype

Current applications to estimating genetics effects on disease resistance, infectivity and recoverability phenotype

4. Systems modelling tools for better outcomes

or How to escape the law of unintended disaster (LOUD)! e.g. natural and chemical pest control, unintended effects of culling to control wildlife disease

Inference for complex dynamic models

Focus on continuous time discrete state-space Markov processes

The problem:

- ❑ Know model structure but …
- ❑ Don't know **model parameters $\boldsymbol{\theta}$**
- ❑ In reality do not know **full history $\boldsymbol{H}$** – e.g. all births, deaths etc
- ❑ Instead have **partial observations given by data $\boldsymbol{D}$**

How to do inference from data $\boldsymbol{D}$?

- ❑ i.e. how to estimate **model parameters $\boldsymbol{\theta}$**
- ❑ or reconstruct the **history $\boldsymbol{H}$**

Bayesian inference

Combines

- Process model $P(H \mid \theta)$ - just the model definition
- Observation model $P(D \mid H)$
- Prior information about parameters $P(\theta)$

To give:

Posterior distribution of unknowns in terms of knowns

$$P(\theta, H \mid D) \propto P(D \mid H)P(H \mid \theta)P(\theta)$$

Inference of unknowns $\boldsymbol{H}, \boldsymbol{\theta}$ from the data $\boldsymbol{D}$

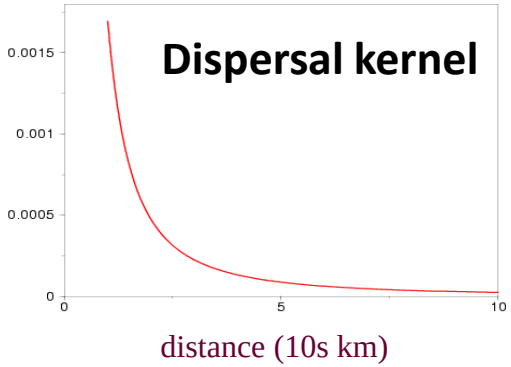
- **model parameters $\boldsymbol{\theta}$**
- **full history $\boldsymbol{H}$**
- **observed data $\boldsymbol{D}$**

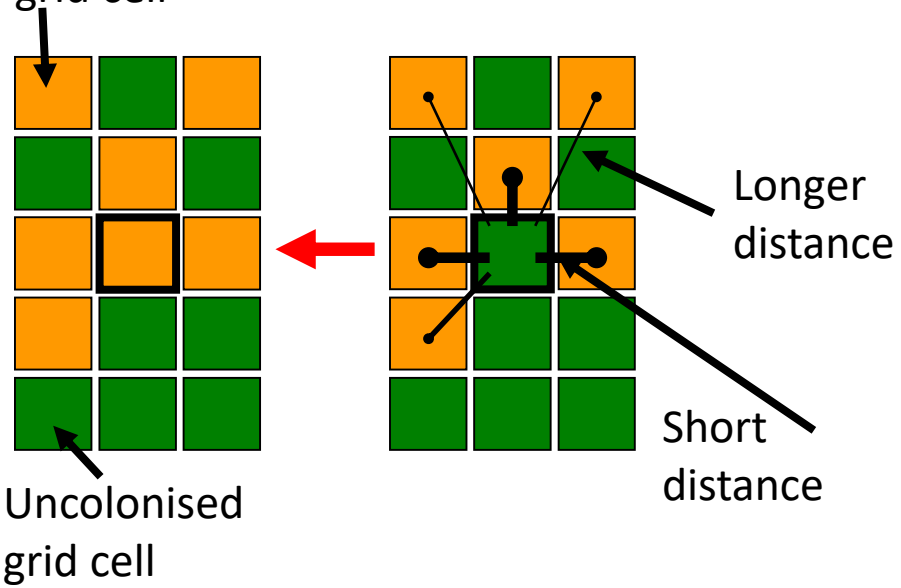
Computational methods e.g. Markov chain Monte Carlo (MCMC) needed to work with this distribution (high-dimensional) even when don't know constant of proportionality (normalisation factor)

Spatial spread models

Dispersal between grid cells in landscape

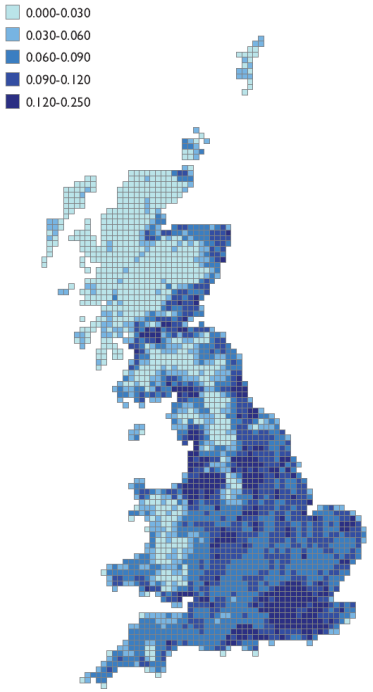
- reduces with distance from colonised locations
- increases with number of colonised locations





Suitability of grid cells in landscape

- function of local characteristics
  e.g. landuse, climate etc.
- varies across landscape
- represents info on pop at risk



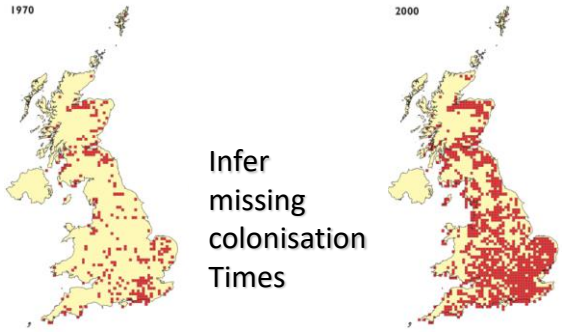
Inference for complex dynamic models

Apply Bayesian inference to

- ❑ Estimate dispersal model parameters
- ❑ Estimate suitability model parameters
- ❑ And estimate colonisation times

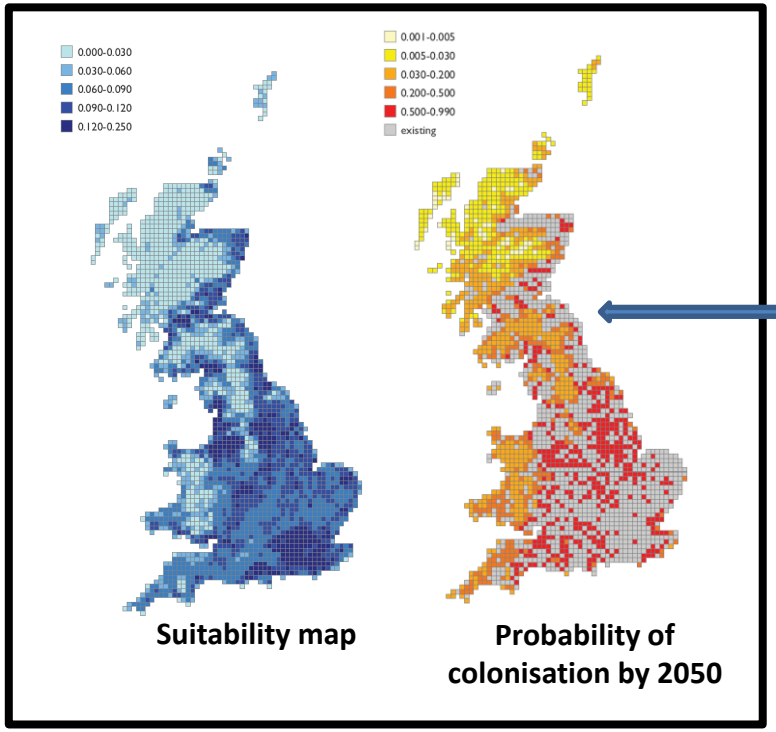
From observed species distribution maps (atlas data) at two time points

Inferring the spread of invasive aliens



From Data – maps at 2 times
Infer: suitability
Infer: spatial spread kernel

Dynamics crucial to correct inference

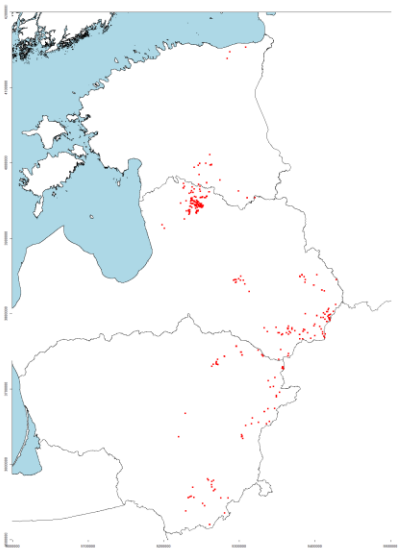


Repeatedly simulate from fitted stochastic model (the posterior) to sample uncertainty in model parameters

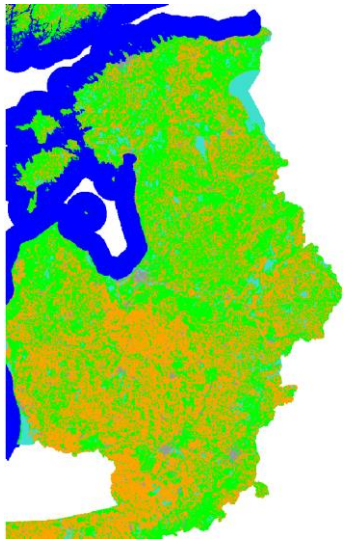
Data-driven risk assessment: population at risk unknown

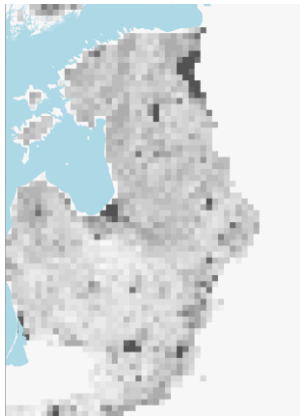
- Outbreak data typically comprise ONLY cases with times/locations
- BUT most analysis methods require knowledge of the host distribution
- These methods allow us to model the outbreak data when the host distribution is uncertain
- Can perform prediction of spread and control as before

ASF cases Apr 2015 (wild boar)

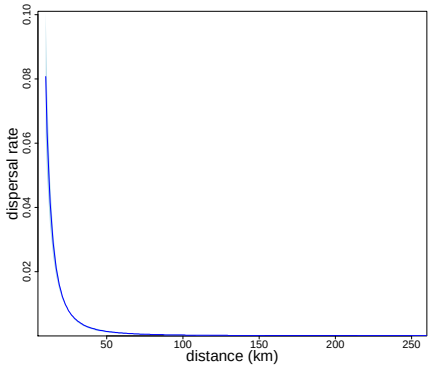


Land use data





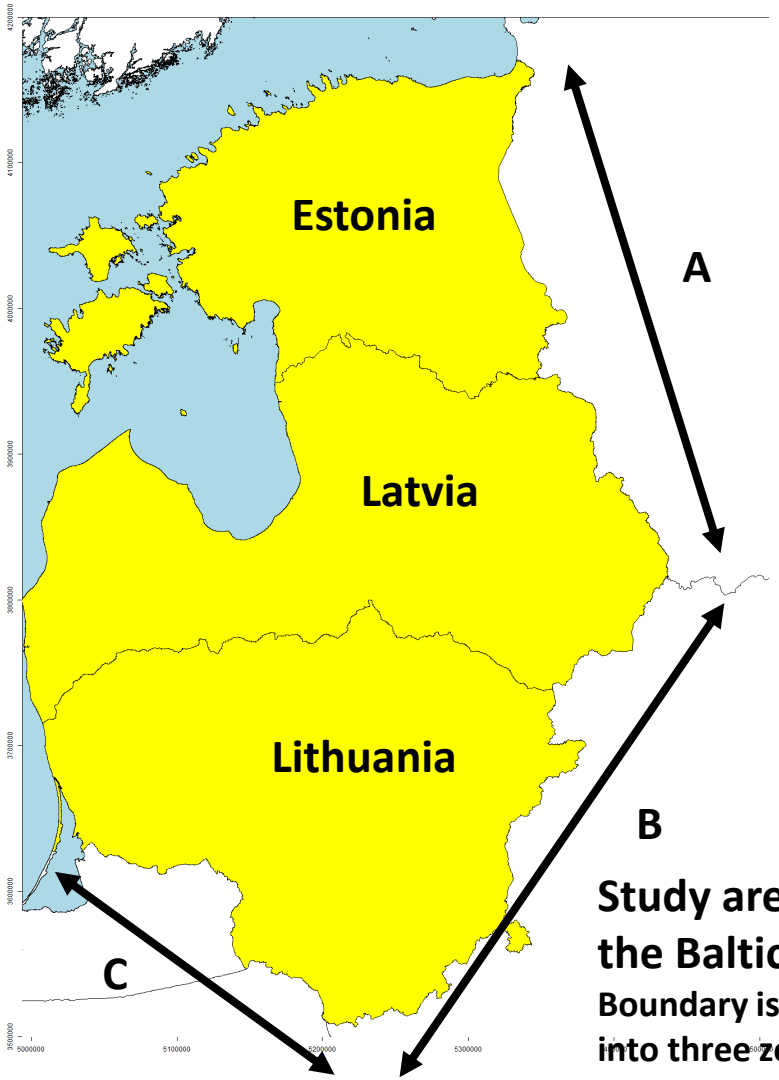
Wild boar density estimate (susceptibility)



Estimated transmission kernel

Application to African Swine Fever

Refine analysis to account for lack of data outside EU

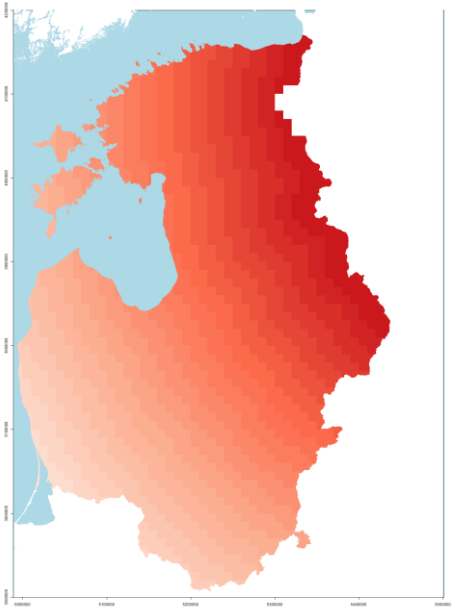


Study area:
the Baltic states
Boundary is split into three zones

Account better for

- **Transmission from out with study area e.g. treat zones A, B & C separately**
- **Differences between countries background infection/introduction**

Distance from border zone A



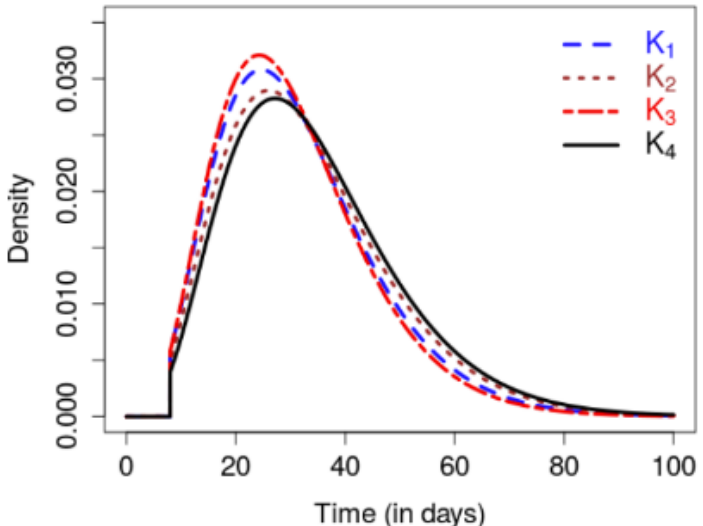
Between-farm outbreak: model

Disease progression on farm: S -> I ->R

S – susceptible I – infectious

R – detected and depopulated

- S -> I – disease transmission (see next slide)
- I ->R – disease detection



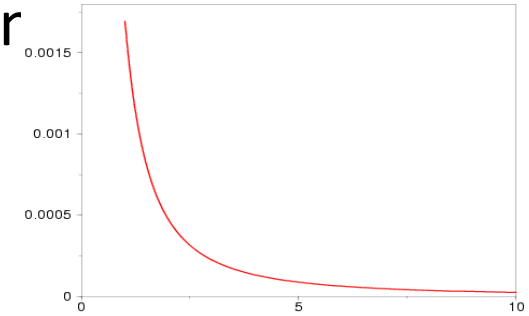
Time to detection

Between-farm outbreak: model

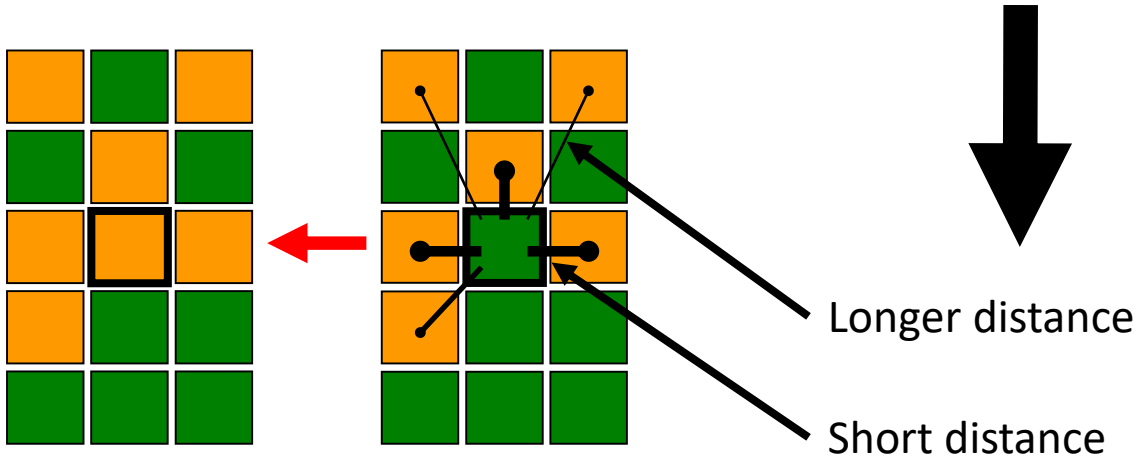
Spread of disease between farms

- Transmission reduces with distance between infected and susceptible farms
- Increases with number of infected farms

Transmission kernel



distance (10s km)



Note: farms can be arranged anyhow

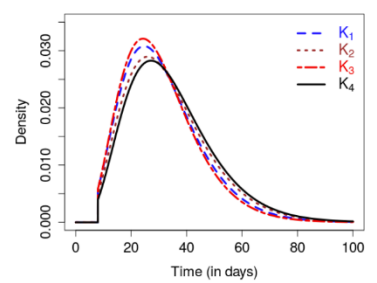
epic
Centre of Expertise on Animal Disease Outbreaks

Rapid inference to support emergency outbreak response

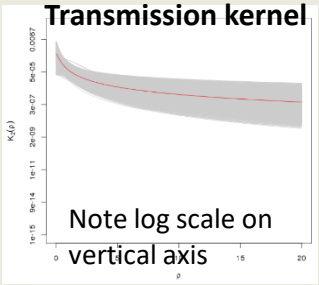
Application to CSF data from 2000 Norfolk outbreak

Here can infer
- Times of infection
- Time to detection I->R
- Transmission kernel

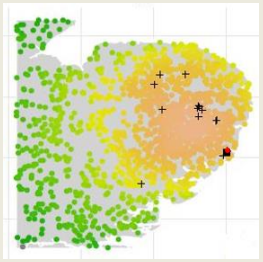
Time to detection



Transmission kernel



Data-driven risk assessment



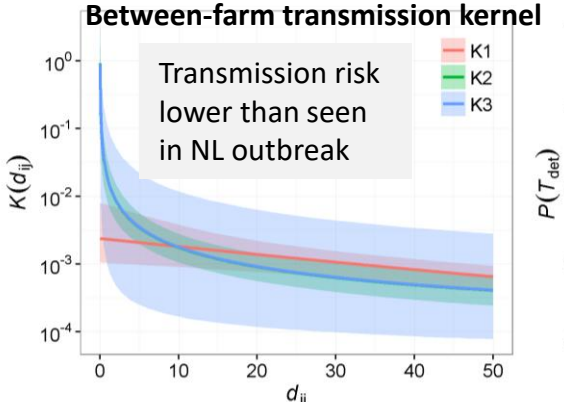
Use Bayesian latent residuals to assess models of transmission kernels
Lau et al 2014 Journal of the Royal Society Interface 11, 20131093.

Gamado, K.M., Marion, G. and Porphyre, T. (2017). *Frontiers in Veterinary Science* 4, 16.

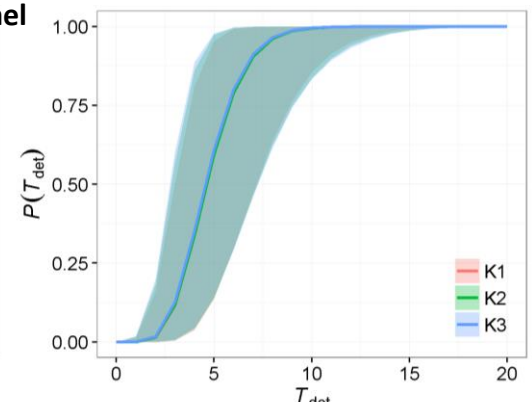
Methods extended to understand 2015 US outbreak of HPAI (H5N2)

- Explore the role of local transmission between farms of known location in Iowa (77 detected cases)
- **Infer the time taken by surveillance to detect infection in farms – <u>corroborated by analysis of within farm mortality</u>**
- Estimate non-local disease incursions (95% Cr. I. 0.76, 17.0)

Between-farm transmission kernel



Cumulative prob. of time to detection



Can also apply model assessment tools to select between competing models

Model assessment via latent residuals for transmission kernel functions

Kernels	Pr(p < 5%)
$K_1(d_{ij})$	0.1435
$K_2(d_{ij})$	0.0674
$K_3(d_{ij})$	0.0598

Assess transmission kernels: K_1 poor fit

Porphyre (Roslin), Gamado, Marion (BioSS), Delgado, Schoenbaum, Torchetti (USDA-VS)

Collaboration within MAC-MIGS

4 areas of potential interest

1. Predicting spread of disease, pests and invasive species on landscapes

Statistical methods to infer and test spread mechanisms both when population at risk is known e.g. farm type/location (HPAI) and unknown e.g. wildlife hosts (ASF)

2. Combine process understanding and data for scientific discovery

Using mathematical and statistical methods to develop data-driven tests of alternative hypotheses e.g. understanding within-host disease progression Tb in badgers

3. Phenotype from genotype

Current applications to estimating genetics effects on disease resistance, infectivity and recoverability phenotype

4. Systems modelling tools for better outcomes

or How to escape the law of unintended disaster (LOUD)! e.g. natural and chemical pest control, unintended effects of culling to control wildlife disease

Data on badger population and disease dynamics

Woodchester Park

- Long term study site for badgers and bTB ~ 40 years
- Very detailed data on individuals and groups
- Capture-mark-recapture data on ~2700 individuals
  - 4 capture campaigns per year
  - Individual data on approx. age, sex, location captured
  - disease status as measured by several different tests
  - Unidentified badgers recorded and 'marked'

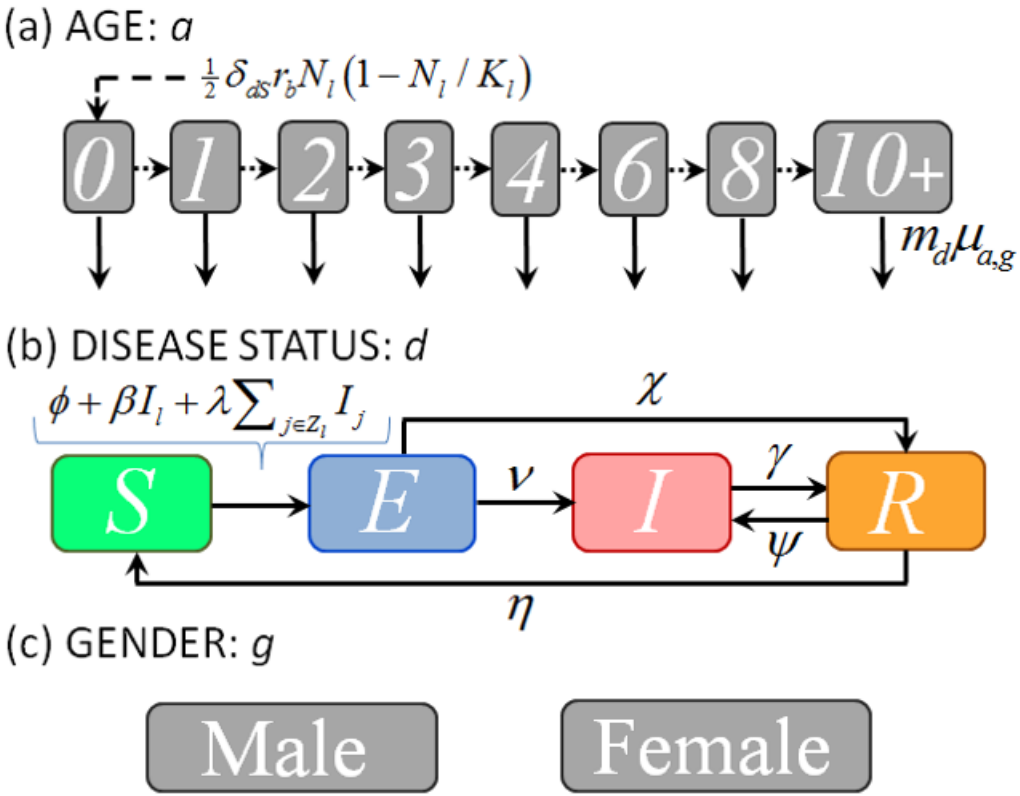
Aim: **understand TB dynamics in badgers**
+ our ability to monitor them

Individual-based model: demography and disease

Individual based stochastic model accounts for

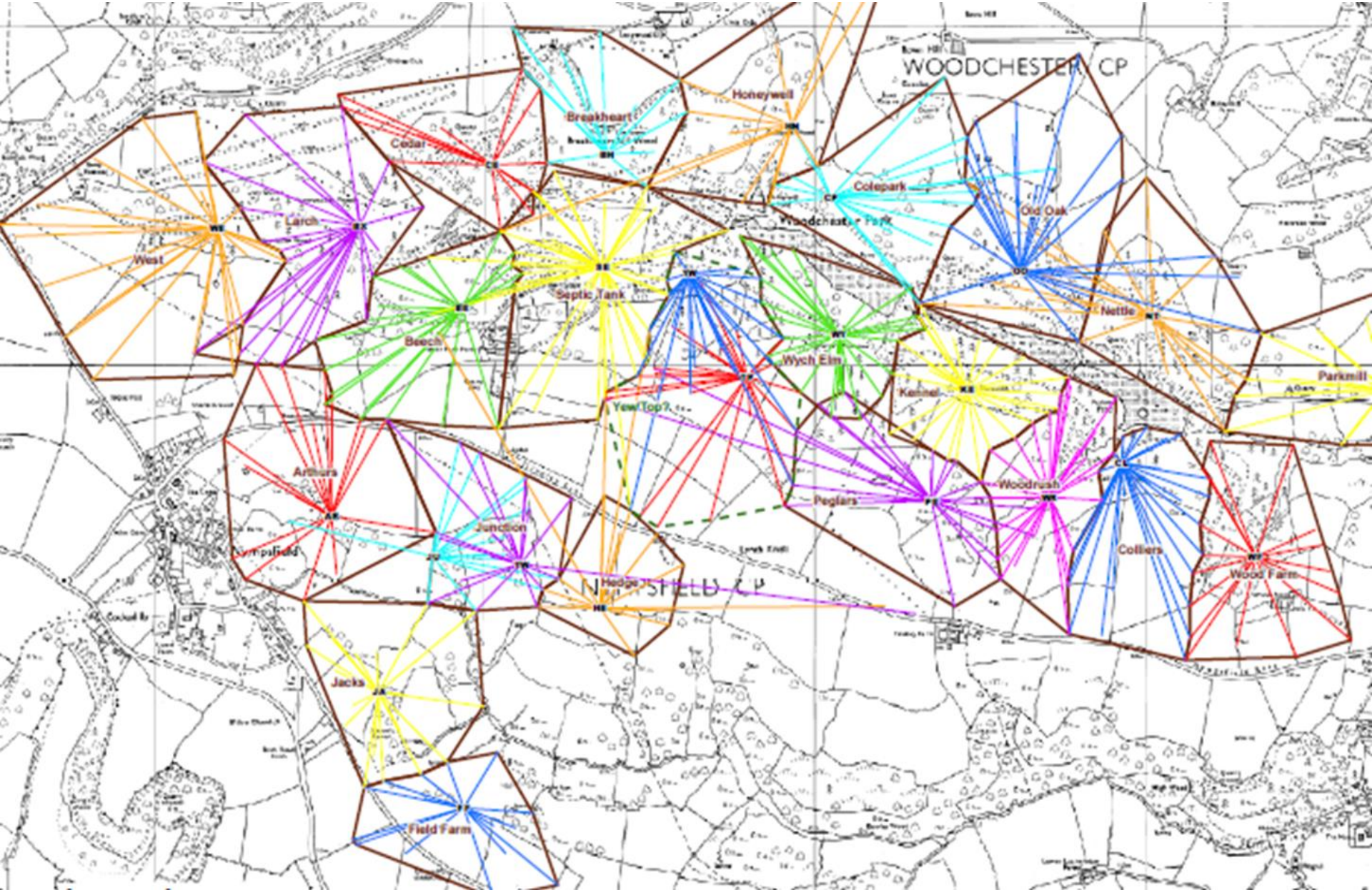
- *Demography including births, deaths and dispersal*
- *Age and sex*
- *Disease induced mortality* m_d

- *Disease progression*
- *Transmission both within* β *and between* λ *social groups*
- *External transmission* ϕ



Individuals live in social groups

Badgers live in social groups with stable territories



Individual-based model: spatially explicit

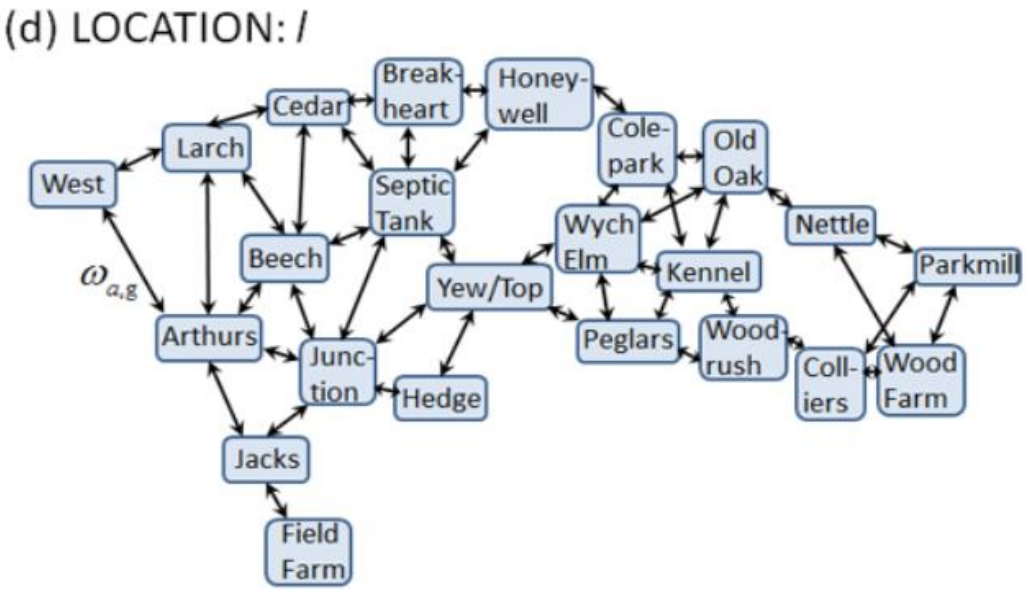
Individual based stochastic model accounts for

- *Demography including births, deaths and* ***dispersal***
- *Age and sex*
- *Disease induced mortality* m_d

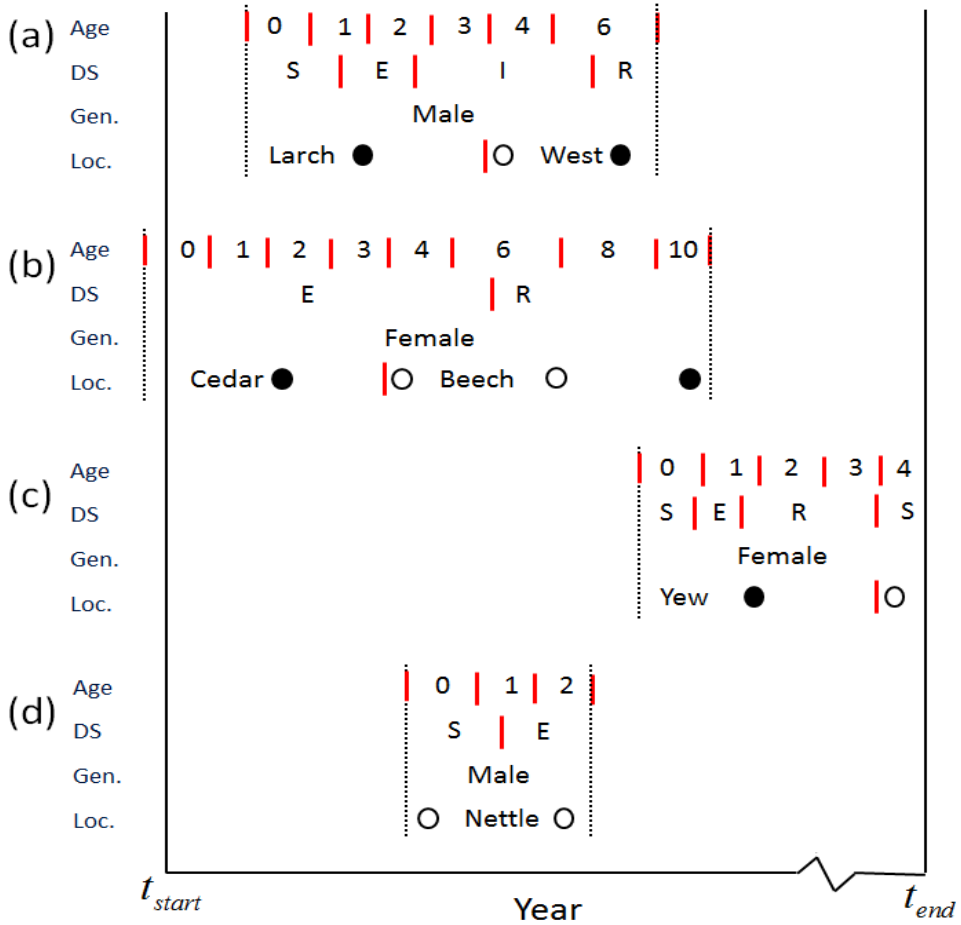
- *Disease progression*
- ***Transmission*** *both within* β
- *and* ***between*** λ ***social groups***
- *External transmission* ϕ

Individual dispersal and between social group disease transmission occurs on neighbourhood structure determined from bait marking maps

Longer range interactions negligible



1. Process model represents underlying life history events including disease dynamics



Life history event sequence *H*

2. Observation model $P(D \mid H)$ determines probability of observed data given complete life history event sequence ***H***

It accounts for

- *Individual trapping*
- *Sex and age differences in trapability*
- *Seasonal variation in capture rates*
- *Se & Sp of diagnostic tests*
  - *State R undetectable by diagnostics*
  - *states E and I detectable by Brock-ELISA, Stat-Pak, and IFN-γ,*
  - *M. bovis culture detects only individuals in the infectious state*

$$P(D \mid H) = \prod_{obs.} p_D \prod_{not\ obs.} (1-p_D) \prod_k Se_k^{N_k^{+,+}} (1-Se_k)^{N_k^{+,-}} Sp_k^{N_k^{-,-}} (1-Sp_k^{N_k^{-,+}}),$$

$$p_D = \max\left\{ z_d\, p_{a,g} \left(1 + \Delta p_D \cos\left(2\pi(t+\theta_D)\right)\right), 1 \right\}$$

Bayesian inference

Combines

- Process model $P(H \mid \theta)$ - just the model definition
- Observation model $P(D \mid H)$
- Prior information about parameters $P(\theta)$

To give:

Posterior distribution of unknowns in terms of knowns

$$P(\theta, H \mid D) \propto P(D \mid H)P(H \mid \theta)P(\theta)$$

Inference of unknowns $\boldsymbol{H}, \boldsymbol{\theta}$ from the data $\boldsymbol{D}$

- **model parameters $\boldsymbol{\theta}$**
- **full history $\boldsymbol{H}$**
- **observed data $\boldsymbol{D}$**

Computational methods e.g. Markov chain Monte Carlo (MCMC) needed to work with this distribution (high-dimensional) even when don't know constant of proportionality (normalisation factor)

Optimising MCMC proposals

Proposals: approximate sampling of individual event sequences
Corrected as usual using Metropolis-Hastings accept/reject step

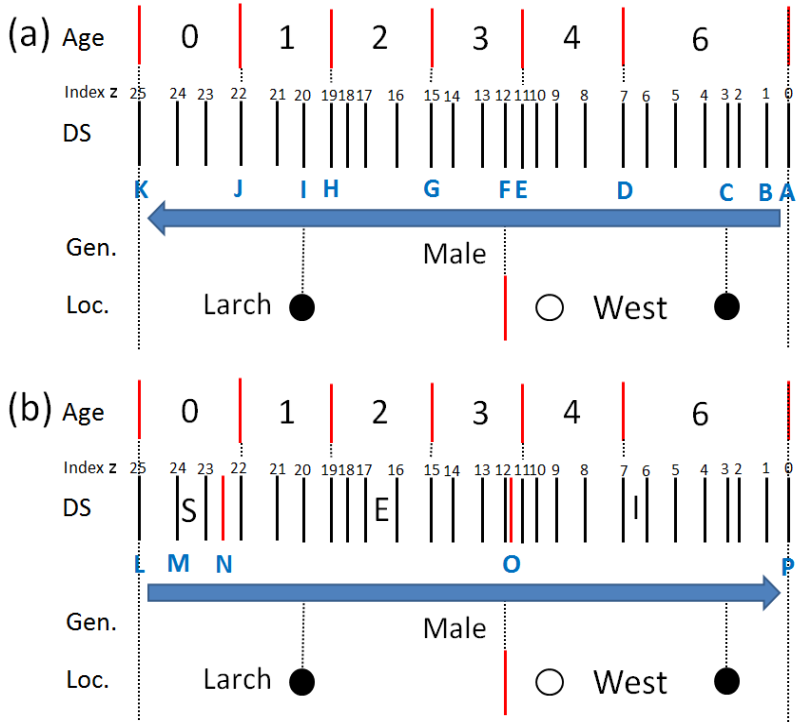
Approximate sampling scheme based on using discrete time (black vertical lines):

Two stage process which involves

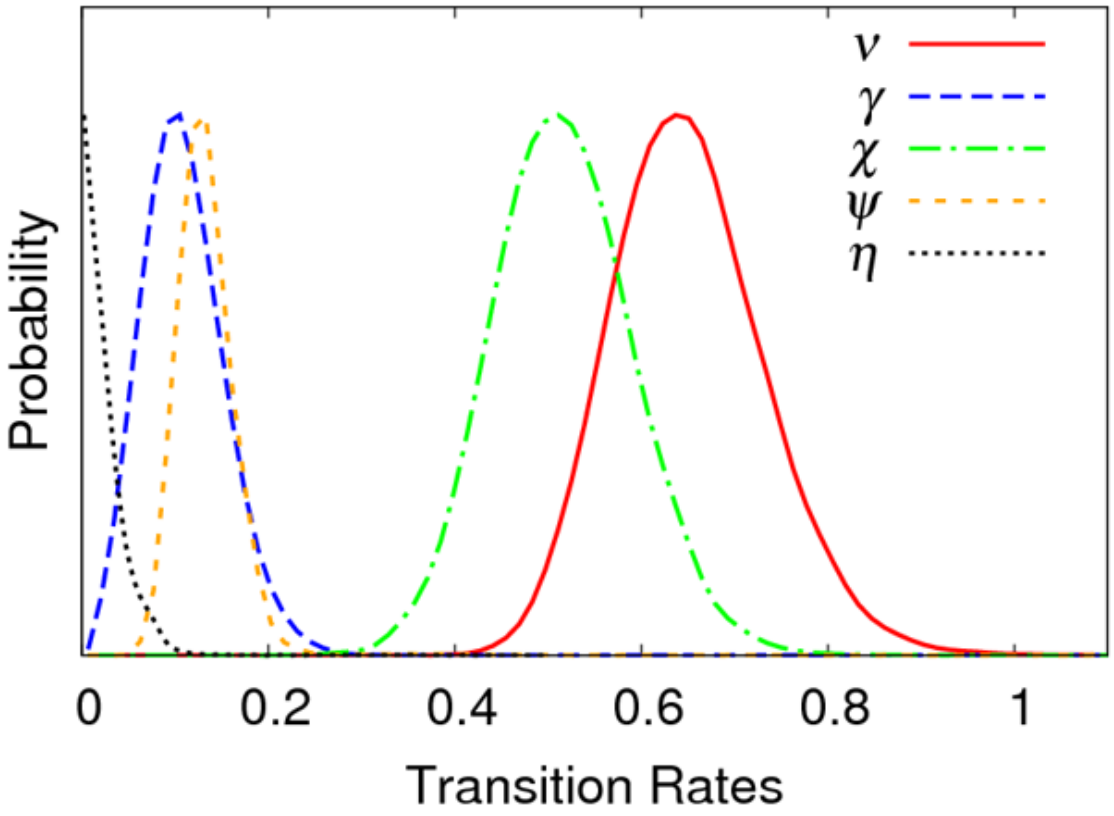
a) back propagation from death of individual to birth to calculate effective 'observation probability' $P_d(t)$ of being in disease state *d*
   Depends on
   - *events affecting location and age*
   - *obs. on disease states (test results)*

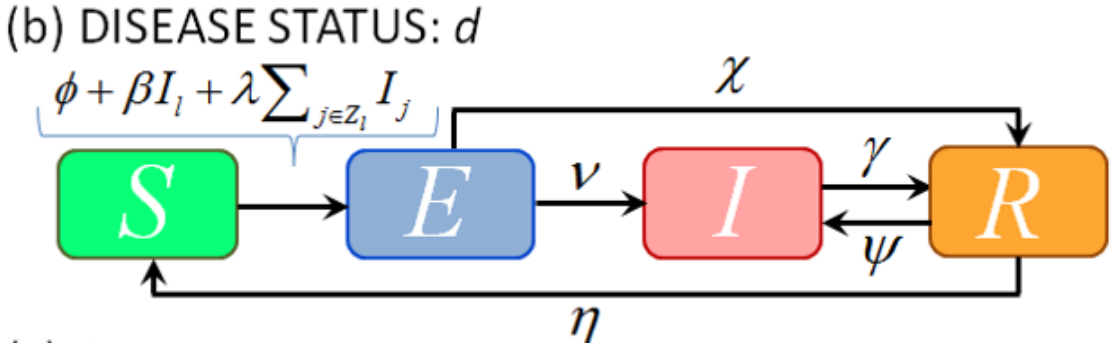
b) forward simulation scheme to generate the new proposed sequence of events based on model (Gillespie algorithm) and $P_d(t)$
   - *modified* Gillespie algorithm

Note: Apply to each individual classification (disease state, age, location) separately



Infer biological mechanism: disease dynamics





SEI model is the standard view of TB dynamics

Disease transitions non-zero suggesting:

- R state (I -> R)
- E -> R transition
- R -> I transition

But not

- R -> S transition

Dormant state

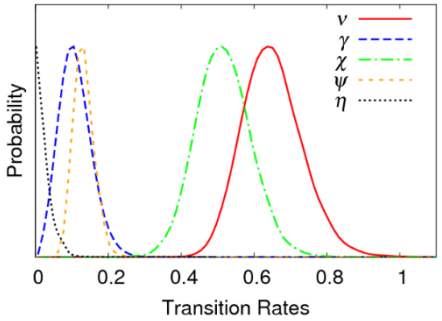
- **average duration 7.6 years** a significant fraction of a typical badger lifespan
- whereas the exposed state lasts on average for only 10 months
- 9%, 6% and **12% of individuals** in states E, I and **R** resp. for WP parameters

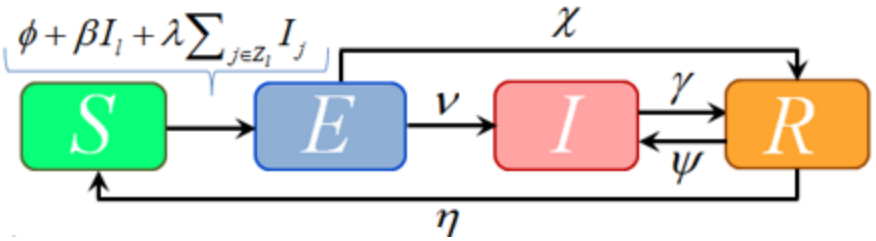
Dormancy recognised in human TB and as a evolutionary adaptation to persistence in small groups of humans – Chisholm & Tanaka, *Proc. R. Soc. B-Biological Sci.* **283,** 20160499 (2016).

Previous study suggested pathology of TB in badgers consistent with dormant state – Gallagher et al *Vet. Rec.* **142,** 710–714 (1998)

But standard models of TB in badgers have not included such a state

Evidence of a long lasting inactive state R

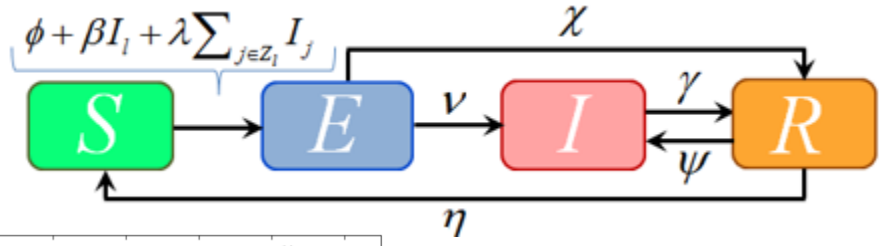


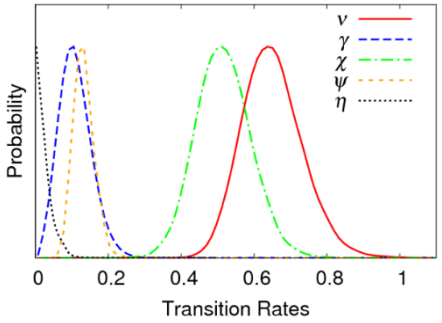


Combine process understanding and data for scientific discovery

- **TBMI consortium** led by Glasgow aims to provide next generation of models to inform UK policy for bovine TB.
- **BioSS led** state-of-the-art parametrisation of models characterising local dynamics of badgers and TB
  - enables, for the first time, simultaneous estimation of operational, diagnostic, demographic and epidemiological parameters
  - informs national modelling of TB in cattle led by Glasgow
  - **yields novel insights into badger-TB system**

Evidence of a temporarily inactive state R





Standard models over many years have ignored the non-infectious inactive state R

Field performance of diagnostic tests

Test sensitivity	Field estimate – full method	Field estimate - raw data	Previous studies*
Se_{ELISA}	0.74 – 0.81	0.79	0.50
$Se_{StatPak}$	0.79 – 0.89	0.78	0.58
$Se_{IFN-\gamma}$	0.45 – 0.55	0.58	0.81
Se_{Cult}	0.24 – 0.30	--	0.275

- Estimates typically based on post-mortem confirmation
- Our results suggest this does not reflect infectiousness in the field

Pooley (Roslin), White (York), Hutchings (SRUC) Kao (Roslin), Bishop (Roslin), Smith (APHA), Delahay (APHA) and Marion (BioSS), *In*

Collaboration within MAC-MIGS

4 areas of potential interest

1. Predicting spread of disease, pests and invasive species on landscapes

Statistical methods to infer and test spread mechanisms both when population at risk is known e.g. farm type/location (HPAI) and unknown e.g. wildlife hosts (ASF)

2. Combine process understanding and data for scientific discovery

Using mathematical and statistical methods to develop data-driven tests of alternative hypotheses e.g. understanding within-host disease progression Tb in badgers

3. Phenotype from genotype

Current applications to estimating genetics effects on disease resistance, infectivity and recoverability phenotype

4. Systems modelling tools for better outcomes

or How to escape the law of unintended disaster (LOUD)! e.g. natural and chemical pest control, unintended effects of culling to control wildlife disease

Phenotype from genotype

Adapting disease dynamics models to account for host genetics

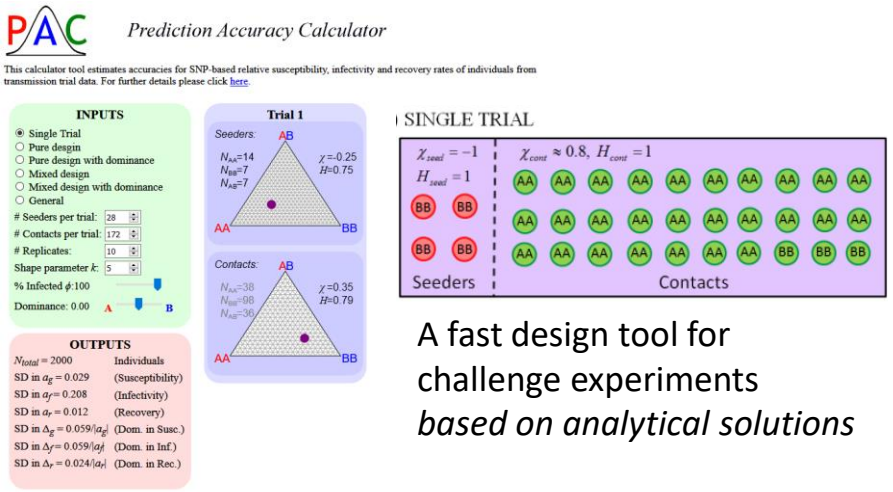
- Quantifying host SNPs in terms of impact on susceptibility, infectivity and recovery

$$g_j^{SNP} = \begin{cases} a_g \\ 0 \\ -a_g \end{cases}, \quad f_j^{SNP} = \begin{cases} a_f \\ 0 \\ -a_f \end{cases}, \quad r_j^{SNP} = \begin{cases} a_r \\ 0 \\ -a_r \end{cases} \begin{matrix} \text{if } j \text{ is } AA \\ \text{if } j \text{ is } AB. \\ \text{if } j \text{ is } BB \end{matrix}$$

- Estimate genetic effects from field observations or disease challenge experiments where SNP data are available
- Potential future application to GWAS type studies

Design

- Use models to assess empirical design
  A general approach – not just for disease!
  simulate data and see how good inferences would be



A fast design tool for challenge experiments
based on analytical solutions

Pooley (Roslin), Marion (BioSS), Bishop (Roslin), and Doeschl-Wilson (Roslin),*In preparation*

Collaboration within MAC-MIGS

4 areas of potential interest

1. Predicting spread of disease, pests and invasive species on landscapes

Statistical methods to infer and test spread mechanisms both when population at risk is known e.g. farm type/location (HPAI) and unknown e.g. wildlife hosts (ASF)

2. Combine process understanding and data for scientific discovery

Using mathematical and statistical methods to develop data-driven tests of alternative hypotheses e.g. understanding within-host disease progression Tb in badgers

3. Phenotype from genotype

Current applications to estimating genetics effects on disease resistance, infectivity and recoverability phenotype

4. Systems modelling tools for better outcomes

or How to escape the law of unintended disaster (LOUD)! e.g. natural and chemical pest control, unintended effects of culling to control wildlife disease

Systems modelling tools for better outcomes

Understanding natural pest control

$4.5B/year in the USA

Chemical pest control

health effects

impacts on environment

pesticide resistance

Qualitative mathematical model

	C	P	SE	GE	ED	NCH
C	-1	-1	0	0	0	0
P	1	0	-1	-1	0	0
SE	0	1	0	0	0	0
GE	0	0	0	-1	1	1
ED	0	0	0	0	-1	0
NCH	0	0	0	0	0	-1

System of differential equations

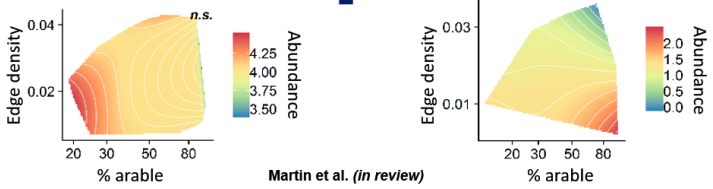
Pests

$$\frac{dP}{dt} = aP - \mu_P P - \frac{f_s PN}{P + K_s} - \frac{f_g PG}{P + K_g} + \delta_P$$

Specialist enemies

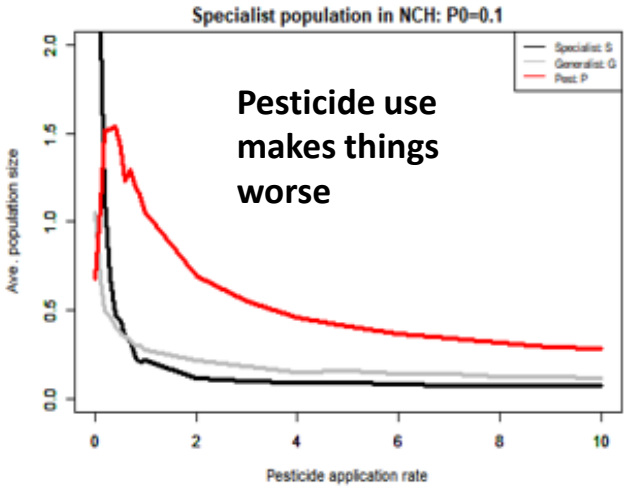
$$\frac{dN}{dt} = -\mu_N N + \frac{\varepsilon_s f_s PN}{P + K_s} + \delta_N$$

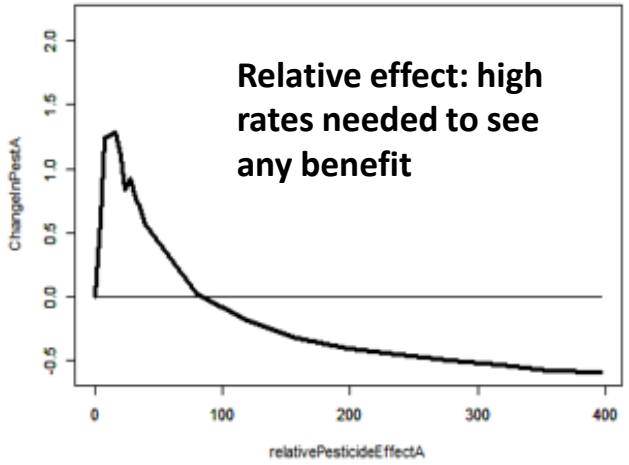
Pests ▪ **Edge density Non-crop habitat** Specialist enemies ⬇



Martin et al. *(in review)*

Unintended impact of pesticide use





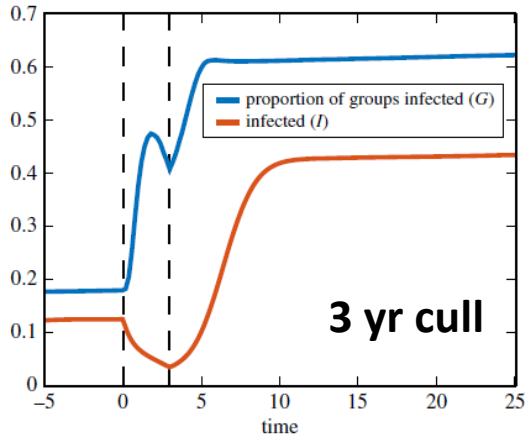
Serious issue for rice production

Systems modelling tools for better outcomes

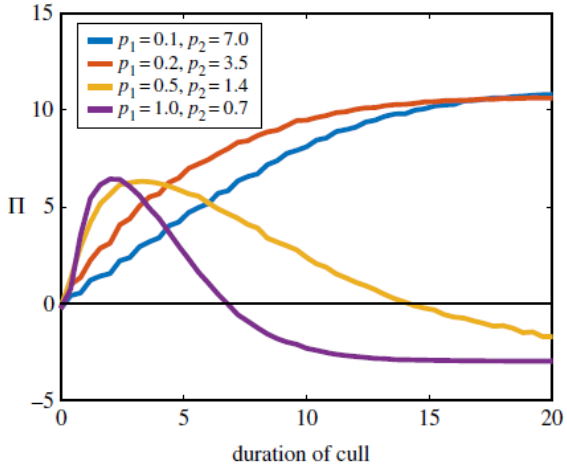
Perturbation effect: Culling of wildlife to control disease is sometimes observed to lead to increases in disease levels

Use models to explore mechanisms, design strategies and when it is not worth it

Culling spreads disease between groups



Spatial distribution and duration of cull of effort impacts perturbation effect size



Model results

Collaboration within MAC-MIGS

Applications need new methodology

Modelling methods for **highly heterogeneous systems**.

Current applications: modelling microbial communities with applications to human gut-health, methane production in ruminants and anti-microbial resistance

Quantitative tools for **modelling dynamics of and on networks**

Current applications: animal trading networks and disease spread on them

Parameter inference for stochastic processes which aims to be fast, rigorous and readily applicable to a range of models and data.

Current applications in epidemiology, ecology and genetics of disease resistance

Efficient model selection methods to drive data-driven hypothesis testing via comparison of models that represent different sub-system processes.

Current applications in epidemiology and ecology

glenn.marion@bioss.ac.uk