Combining Mathematical and Statistical Models: a Disease Ecology Perspective

Jennifer Hoeting
Department of Statistics
Colorado State University
www.stat.colostate.edu/∼jah

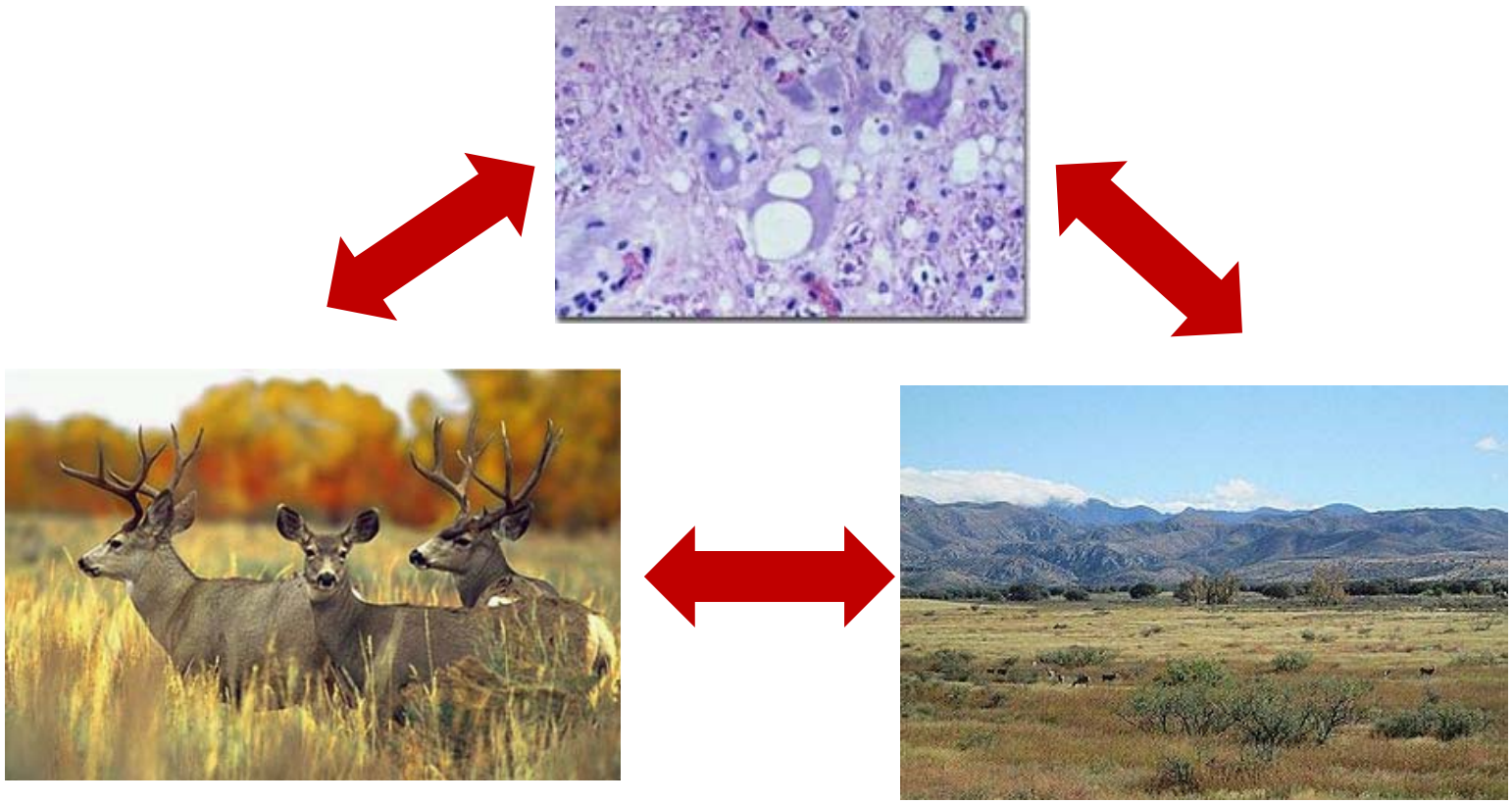
International Statistical Ecology Conference

July 2018

Ecology of Infectious Diseases

Explores the relationship between

1. Pathogen
2. Host: animal and human
3. Environment



Ecology of Infectious Diseases
Tasmanian Devil Facial Tumor Disease

SAVE THE TASMANIAN DEVIL PROGRAM

Ecology of Infectious Diseases
Zika virus

Overview: Motivation

Focus of talk

Challenges of statistical parameter estimation and inference for dynamical models of infectious disease

Why are infectious diseases challenging to model?

1. Disease transmission is not observable
2. Much uncertainty about what is observed
3. Highly nonlinear dynamics

Overview: Motivation

Why are infectious diseases challenging to model?

You need to synthesize a broad range of ideas from
biology,
mathematics,
and
statistics

Overview: Process versus pattern

Classification of modeling approaches

Mathematical	versus	Statistical
Deterministic	versus	Stochastic
Theoretical	versus	Phenomenological
Process	versus	Pattern

Traditional approach:

choose between a mathematical or a statistical model

Alternative approach:

combine the advantages of mathematical and statistical models

Overview: Process versus pattern

Mathematical model (process)

Model structure reflects explicit hypotheses about the biological mechanisms that drive infection dynamics

Example: Susceptible-Infectious-Recovered (SIR) Model

$$\frac{dS}{dt} = -\beta SI \qquad \frac{dI}{dt} = \beta SI - \gamma I \qquad \frac{dR}{dt} = \gamma I$$

Advantages:

- Explicitly model nonlinearities in the process
- Useful for prediction

Disadvantages:

- Represent the average behavior
- Often focus is on model form, not parameter estimation for observed data

Overview: Process versus pattern

Statistical model (phenomenological)

Model describes the observed relationship between variables

Example: Linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + \epsilon$$

Advantages:

- Data can inform the model
- Rich characterization of different types of errors

Disadvantages:

- May only describe the observed data
- Gives little information about the mechanism
- Interactions not always well captured

Overview: history

Why not combine the advantages of mathematical biology and statistics?

- Long history of combining mathematical and statistical models (e.g., Berliner 1991)
- Many sessions at this conference include these ideas

References and this talk:

- See the end of the talk for complete list of references
- These slides are available online: www.stat.colostate.edu/∼jah

Motivating example: CWD Transmission

Chronic Wasting Disease

Deer (female) with
Chronic Wasting Disease

Healthy deer (male)

Motivating example: CWD Transmission

Chronic wasting disease (CWD)

- 100% fatal contagious disease that affects cervids (deer family)
- CWD is a prion disease
- Important to understand the transmission mechanisms of CWD
- Several deterministic epidemic models were proposed by Miller, Hobbs & Tavener (2006)

Motivating example: CWD Transmission

Mathematical model for disease transmission
Susceptible-Infectious-Recovered (SIR) Model

time	# animals S	I	R
t_1	100	0	0
t_2	90	10	0
t_3	80	10	10
t_4	70	10	20

$N = 100$

Motivating example: CWD Transmission

We develop a type of Susceptible-Infectious-Recovered (SIR) model for disease transmission where the state variables are described by a set of differential equations.

Consider the state vector $\mathbf{X}(t) = (S(t), I(t), R(t))^T$, where

- S is the number of susceptible animals
- I is the number of infectious animal
- R is the number of deaths from CWD

Motivating example: : CWD Transmission

Direct transmission ODE model for CWD

$$\begin{aligned}\frac{dS}{dt} &= a - S(\beta I + m)\\ \frac{dI}{dt} &= \beta SI - I(\mu + m)\\ \frac{dR}{dt} &= \mu I\end{aligned}$$

where

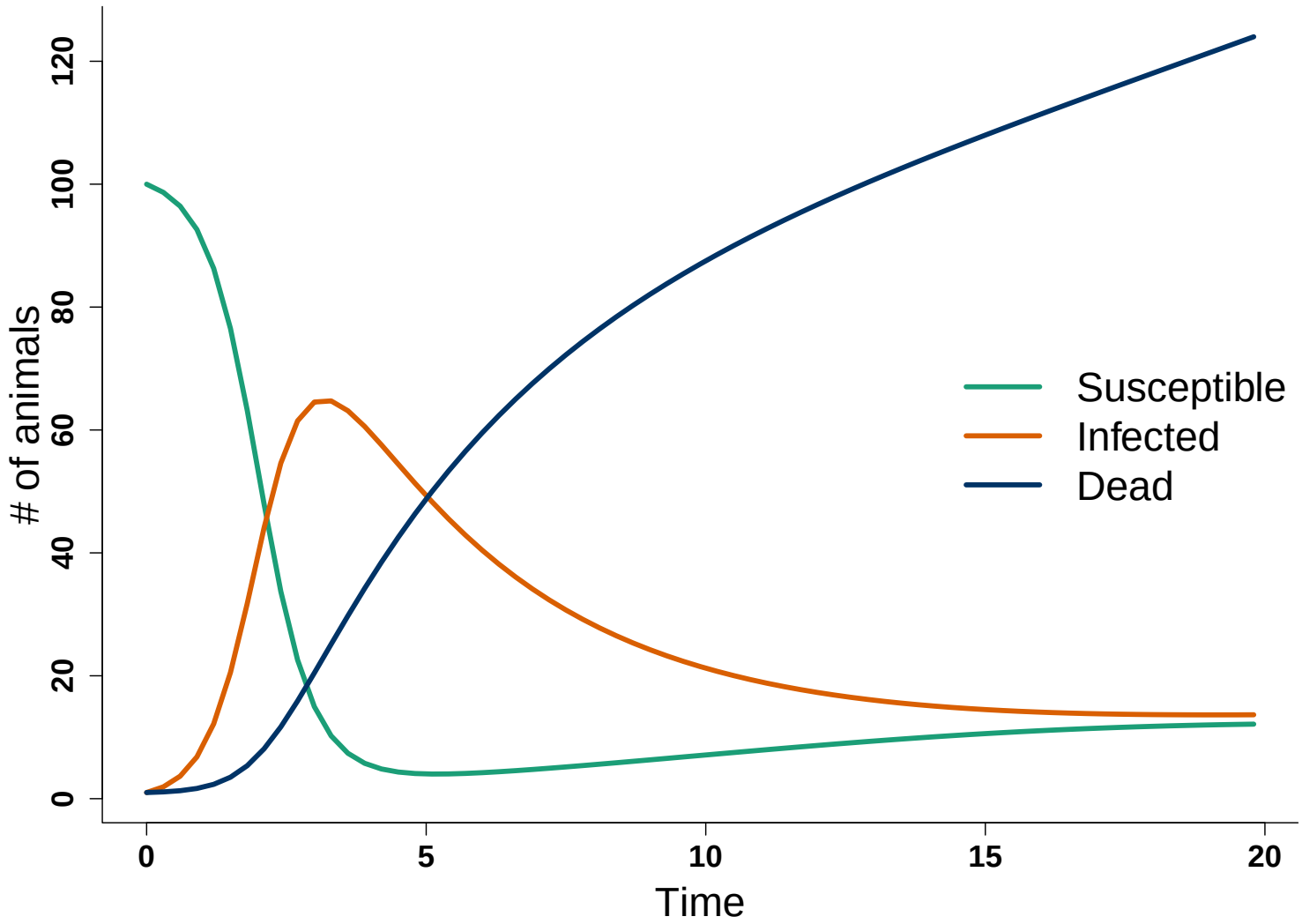
- β is the transmission rate
- μ is the per capita CWD mortality rate

} unknown

- a is the number of susceptible animals annually added to the population via births or importation
- m is the per capita natural mortality rate

} known

Motivating example: CWD Transmission



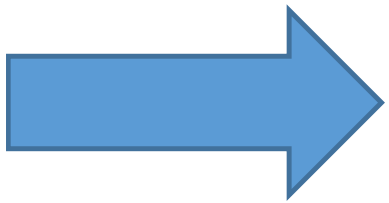
Motivating example: CWD Transmission

There are always challenges with observed data:

Complete data

Time	# of animals S	I	R
t_1	100	0	0
t_2	90	10	0
t_3	80	10	10
t_4	70	10	20

N=100

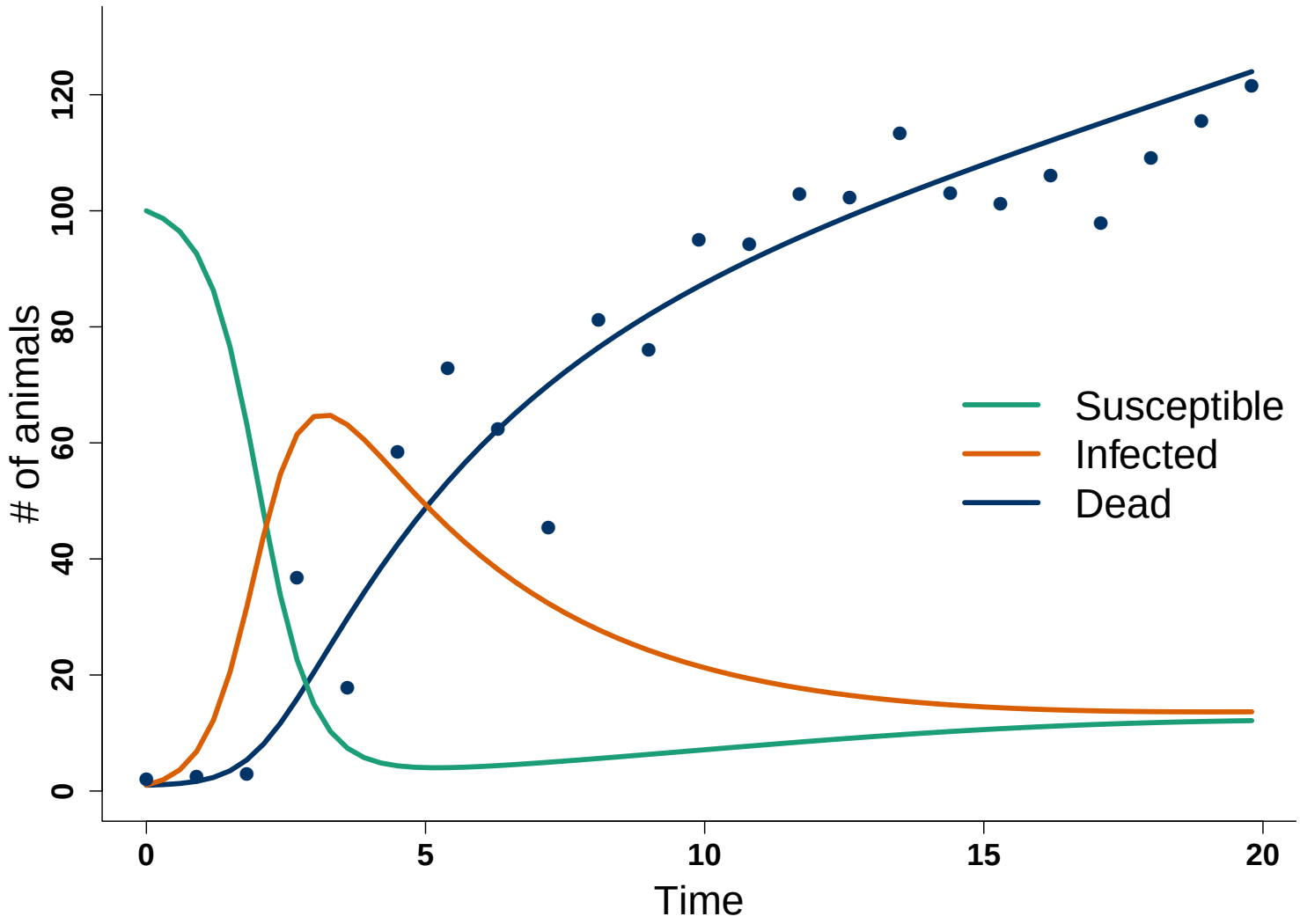


Observed data

Time	# of animals S	I	R
t_1			0
t_2			0
t_3			10
t_4			20

N=100

Motivating example: CWD Transmission



Challenges

We explore challenges in

1. Observed data
2. Model development:
   - Which class of dynamical model? (Mathematics)
   - Which mechanisms to include? (Biology)
3. Statistical inference:
   - Which statistical model/paradigm?
   - Which computational method?
   - How to select models?

Challenge 1: Data challenges

Challenges in the observed data:

1. Missing data
   - You have to find the animals, they don't visit the nearest health clinic
   - Some states are unobserved
2. Sparse data
   - Interval between observations can be long and irregular
3. Uncertain data
   - Uncertainties about the testing procedure (false negatives and/or false positives)
4. Initial conditions unknown
   - Usually don't observe the population before the disease outbreak

Challenge 2: Model development

Challenges in model development:

1. Which class of dynamical model? (Mathematics)
2. Which mechanisms to include? (Biology)

Challenge 2: Model development

Which class of dynamical model?

What is a dynamical model?
A dynamical model describes a system that evolves in time.

A dynamical model includes:

- A description of the state(s) of the system
- A time index
- A rule by which the state(s) evolves forward in time

Some classifications of dynamical models

- continuous or discrete
- stochastic or deterministic

Challenge 2: Model development

Which class of dynamical model?

Ordinary Differential Equation (ODE) model:

- ODE models can be used to determine whether or not disease transmission will occur.
- Dynamical model classifications: deterministic, continuous time, continuous state space model

Example:

$$\frac{dS}{dt} = a - S(\beta I + m)$$
$$\frac{dI}{dt} = \beta SI - I(\mu + m)$$
$$\frac{dR}{dt} = \mu I$$

Challenge 2: Model development

Which class of dynamical model?

Stochastic Differential Equation (SDE) model:

- SDE models be used to determine the probability of disease transmission between two individuals
- Natural extension of ODE models
- Dynamical model classifications: stochastic, continuous time, continuous state space model

Challenge 2: Model development

Which class dynamical model?

A SDE model for direct transmission of CWD is given by

$$\begin{aligned} dS &= [a - S(\beta I + m)]dt + B_{11}dW_1 + B_{12}dW_2 + B_{13}dW_3, \\ dI &= [\beta SI - I(\mu + m)]dt + B_{21}dW_1 + B_{22}dW_2 + B_{23}dW_3, \\ dC &= \mu I dt + B_{31}dW_1 + B_{32}dW_2 + B_{33}dW_3, \end{aligned}$$

where

- **W** is a k-dimensional standard Wiener process.
- $B = (B_{ij}) = \sqrt{\Sigma}$ with

$$\Sigma = \begin{bmatrix} a + S(\beta I + m) & -\beta SI & 0 \\ -\beta SI & \beta SI + I(\mu + m) & -\mu I \\ 0 & -\mu I & \mu I \end{bmatrix}.$$

Challenge 2: Model development

Which class of dynamical model?

Continuous time Markov chain (CTMC) model:

- CTMC models be used to determine the probability of disease transmission between two individuals
- May be more complicated to derive than SDE models
- Dynamical model classifications: stochastic, continuous time, discrete state space model

Challenge 2: Model development

Which class of dynamical model?

A CTMC model for direct transmission of CWD is given by

$$P\left(\begin{array}{l|l} S(t+\delta)=i+1 & S(t)=i \\ I(t+\delta)=j & I(t)=j \\ R(t+\delta)=k & R(t)=k \end{array}\right) = a\delta + o(\delta),$$

$$P\left(\begin{array}{l|l} S(t+\delta)=i-1 & S(t)=i \\ I(t+\delta)=j & I(t)=j \\ R(t+\delta)=k & R(t)=k \end{array}\right) = im\delta + o(\delta),$$

$$P\left(\begin{array}{l|l} S(t+\delta)=i-1 & S(t)=i \\ I(t+\delta)=j+1 & I(t)=j \\ R(t+\delta)=k & R(t)=k \end{array}\right) = \beta ij\delta + o(\delta),$$

$$P\left(\begin{array}{l|l} S(t+\delta)=i & S(t)=i \\ I(t+\delta)=j-1 & I(t)=j \\ R(t+\delta)=k & R(t)=k \end{array}\right) = jm\delta + o(\delta),$$

$$P\left(\begin{array}{l|l} S(t+\delta)=i & S(t)=i \\ I(t+\delta)=j-1 & I(t)=j \\ R(t+\delta)=k+1 & R(t)=k \end{array}\right) = j\mu\delta + o(\delta),$$

where $o(\delta) \to 0$ as the time interval $\delta \to 0$.

Each probability statement in the CTMC model corresponds to a component of the ODE model.

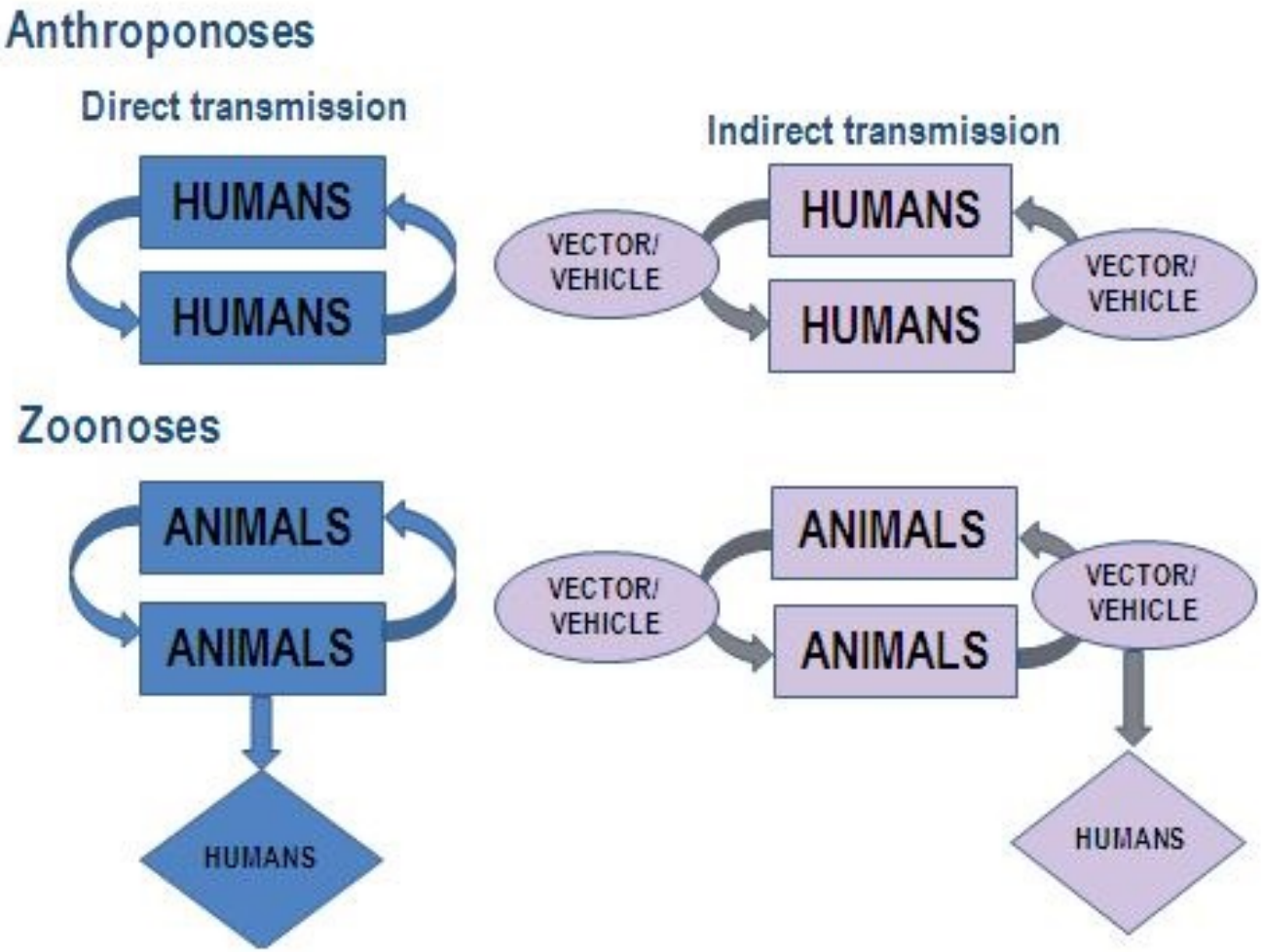
Challenge 2: Model development

Model development:

1. Which class of dynamical model? (Mathematics)
2. Which mechanisms to include? (Biology)

Challenge 2: Model development

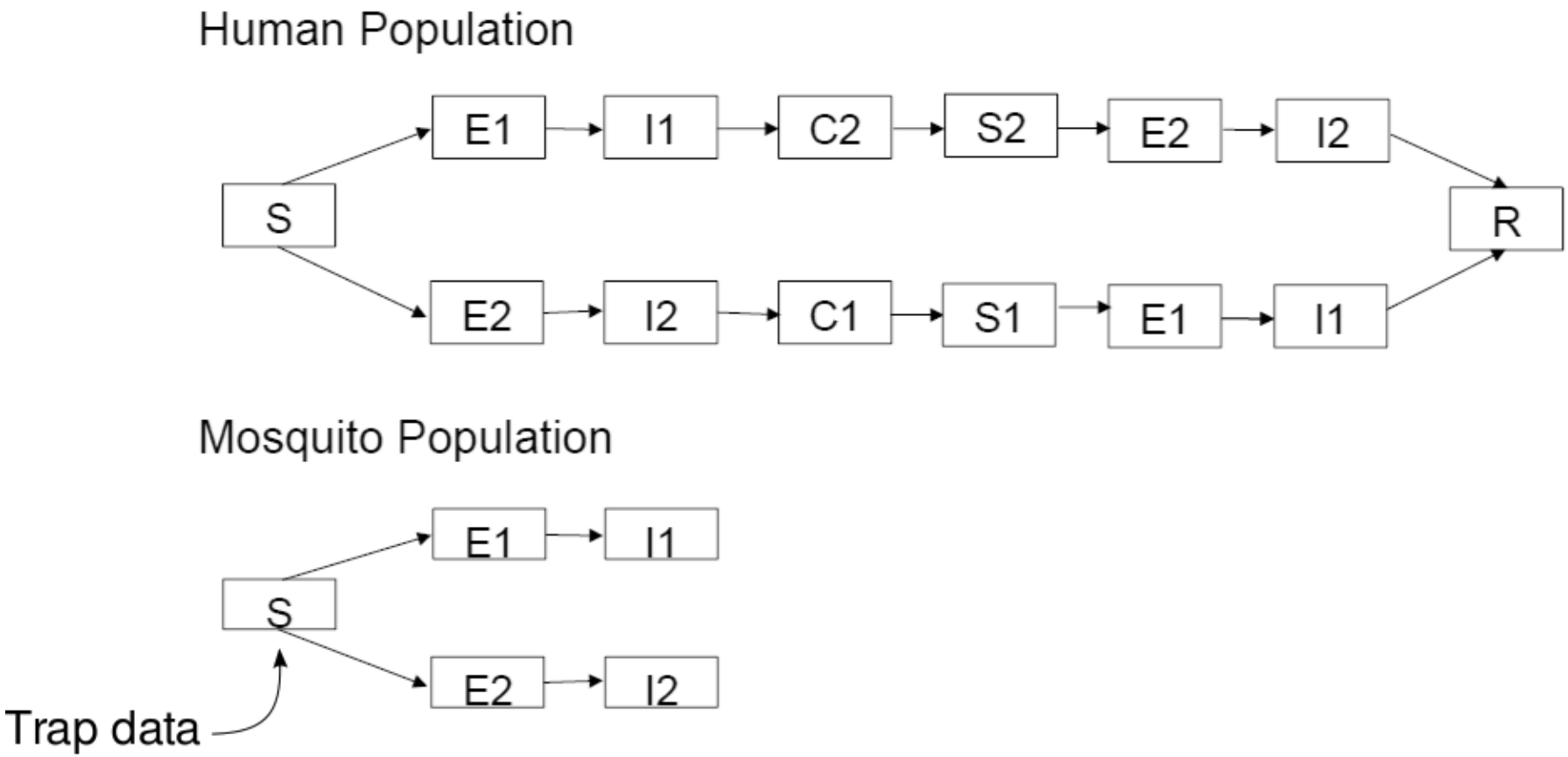
Which mechanisms to include?



Challenge 2: Model development

Which mechanisms to include?

A two serotype model for dengue fever



C. Leach, 2015

Challenge 3: Statistical inference

Challenges in statistical inference

1. Which statistical model/paradigm?
2. Which computational method for inference?
3. Which model do the data support?

Challenge 3: Statistical inference

Statistical model/paradigm

The statistical method you select for inference will depend on the

- shortcomings of your data
- dynamical model you developed

Challenge 3: Statistical inference

Statistical model/paradigm

Example: For the Chronic Wasting Disease data

- It is reasonable to allow for errors in the observed number of deaths
- We use a hierarchical model with a dynamical model at the process level

Hierarchical model consists of

Stage 1: Data model

Stage 2: Process model

Stage 3: Parameter model

Challenge 3: Statistical inference

Computational methods for inference

Some possible computational methods to enable statistical inference for dynamical models

1. Bayesian approaches
   - Markov chain Monte Carlo (MCMC)
   - Approximate Bayesian Computation (ABC)
2. Maximum likelihood approaches
   - Iterated filtering
   - Penalized simulated maximum likelihood
3. Least squares approaches
   - Gradient matching
   - Trajectory matching

Many other options available. See talk references.

Challenge 3: Statistical inference

Computational methods for inference

Most computational methods can't be used 'out of the box' for modeling infectious diseases due to

- The sparse nature of the data
- Small changes in the parameters can lead to very different dynamic behavior

Finding good starting values for the computational statistical algorithms can be particularly challenging.

Latin hypercube sampling (Marino et al., 2008) can be useful

Challenge 3: Statistical inference

Model selection

Choice of model selection methods will depend on the inference paradigm you choose.

Some options:

1. Bayesian
   - Compare models via their posterior model probabilities.
     For model $\mathcal{M}_k$ the posterior model probability is given by $P(\mathcal{M}_k|D)$.
   - Compare models using Bayes factors (Kass & Raftery 1995)
2. Maximum likelihood: AIC
3. Other options for both paradigms: see references

Just to be precise: model selection isn't statistical inference

Example: Chronic Wasting Disease

Putting all the pieces together:

Parameter inference and model selection in deterministic and stochastic dynamical models via approximate Bayesian computation (Sun, Lee, Hoeting, 2015)

Challenges in

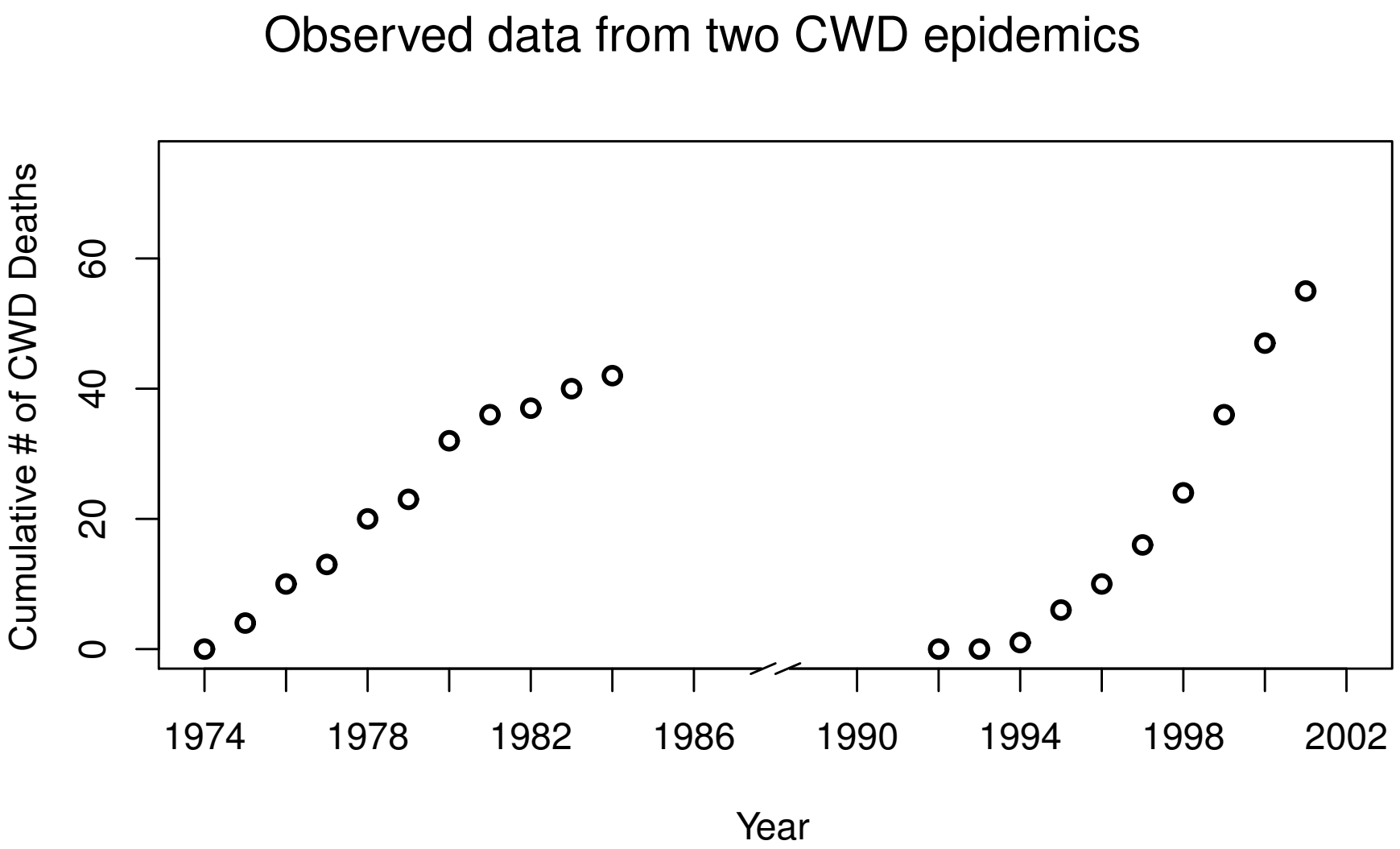
1. Observed data
2. Model development: We'll consider several disease models
3. Statistical inference:
   - Which model/paradigm? Bayesian hierarchical model
   - Which computational method? ABC
   - How to select models? Posterior model probabilities and Bayes factors

Motivating example: CWD Transmission

The observed data:

- Annual observations of cumulative mortality from two CWD epidemics in captive mule deer
- No live-animal test, vaccine, or treatment for CWD existed prior to 2008.
- Epidemic 1: 1974 to 1985
  Epidemic 2: 1992 to 2001 (in a new deer herd)
- 21 observations over time
- The dataset also includes
  - annual number of new deer added to the herd
  - per capita losses due to natural deaths and removals

Example: Chronic Wasting Disease



Example: Chronic Wasting Disease

For our chronic wasting disease example:

Stage 1: Data Model

At time t let $\tilde{R}(t)$ = observed cumulative number of deaths from CWD where

$$\tilde{R}(t) \sim \text{Binomial}\left(N(t); \frac{R(t)}{N(t)}\right)$$

where

- $N(t) = S(t) + I(t) + R(t)$ is the total # of animals at time t
- Only $\tilde{R}(t)$ and $N(t)$ observed at discrete time $t = t_0, t_1, \ldots, t_n$

Example: model set-up

Stage 2: Process Model: Direct transmission ODE

$$\begin{aligned}\frac{dS}{dt} &= a - S(\beta I + m)\\ \frac{dI}{dt} &= \beta SI - I(\mu + m)\\ \frac{dR}{dt} &= \mu I\end{aligned}$$

Stage 3: Parameter Model

Prior distributions for all model parameters

Inference: We can't write the likelihood in closed form so we certainly don't have the posterior distribution in closed form

Example: Model selection

Biologists and statisticians have proposed multiple reasonable models. Which model should we use?

Goal 1: Choose a data model
We consider Binomial or Poisson

Goal 2: Choose the disease transmission model

1. Direct (basic SIR)
2. Indirect (environmental transmission)
3. Both direct and indirect disease transmission

Goal 3: Choose a class of dynamical model

1. Ordinary differential equation (ODE) model
2. Stochastic differential equation (SDE) model
3. Continuous time Markov chain (CTMC) model

Example: Results for CWD

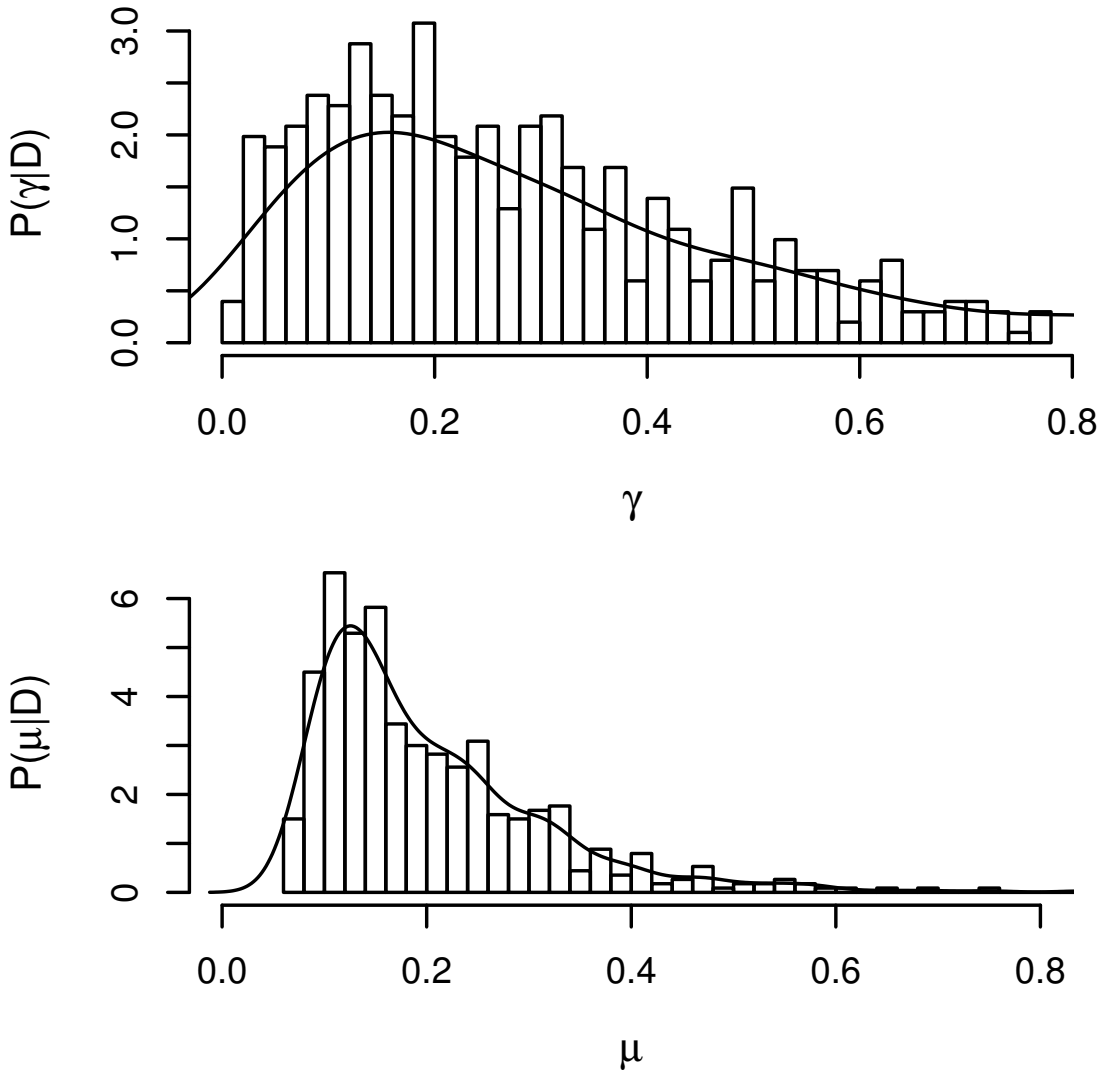
Goal: compare models for Chronic Wasting Disease.

Posterior model probabilities for each model $P(\mathcal{M}|D)$

Data Model	CWD Transmission	Process Model	Informative prior set $P(\mathcal{M}|D)$	Bayes factor
Binom	Direct/Indirect	SDE	0.21	1.00
Binom	Direct	SDE	0.18	1.15
Binom	Direct	ODE	0.13	1.55
Binom	Direct	CTMC	0.11	1.87
Binom	Direct/Indirect	ODE	0.09	2.43
Pois	Direct/Indirect	SDE	0.09	2.27
Pois	Direct	ODE	0.06	3.48
Pois	Direct	SDE	0.05	3.87
Pois	Direct/Indirect	ODE	0.04	4.63
Pois	Direct	CTMC	0.03	6.17

Example: Parameter estimates

The marginal posterior distribution for 2 of the parameters of the indirect transmission SDE model based on the CWD epidemic data.

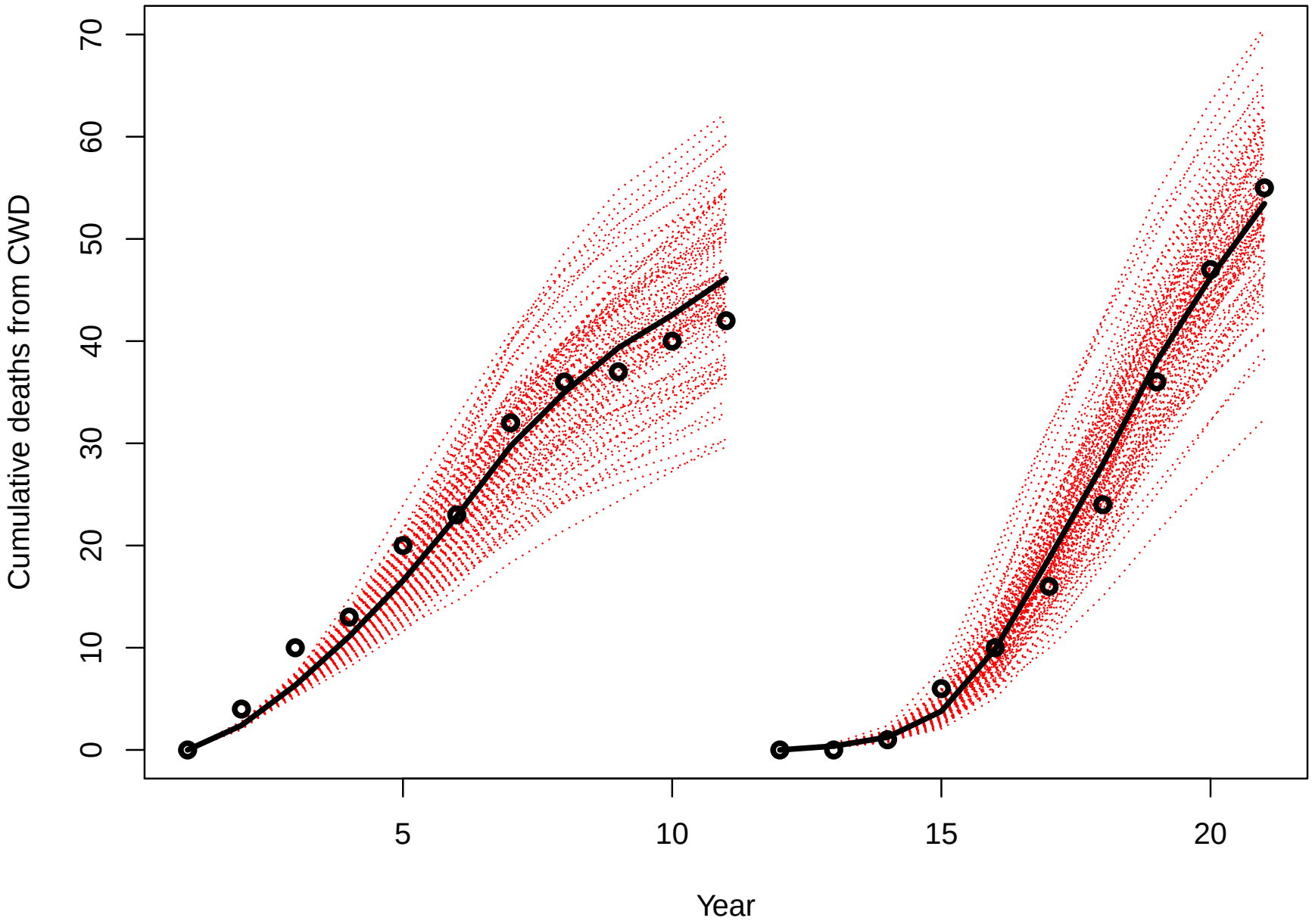


Example: Parameter estimates

Marginal posterior modes and 95% HPD intervals of the parameters of the indirect transmission SDE process model with the Binomial data model based on the CWD epidemic data.

Parameter	Informative prior set Mode	95% HPD
γ = Indirect transmission rate (mass^{-1}yr^{-1})	0.05	(0.01, 0.36)
μ = CWD mortality rate (yr^{-1})	0.20	(0.10, 0.59)
ϵ = Per capita rate of excretion of infectious agent (yr^{-1})	0.47	(0.15, 0.91)
τ = Rate of loss of infectious agent (yr^{-1})	0.88	(0.01, 4.52)
$S(0)$ of the first epidemic	18	(10,26)
$I(0)$	10	(5,18)
$E(0)$	1.73	(0.97,5.84)
$S(0)$ of the second epidemic	48	(24,50)
$I(0)$	2	(0,5)
$E(0)$	3.47	(0.24,4.85)

Example: Fitted SDE Model for CWD



Future work

Much work left to do in development of statistical methods and models for the ecology of infectious disease:

As data complexity and model complexity increase, the current methods often fail.

More developments needed in:

1. Develop efficient computational algorithms for estimation
2. Inference for data from multiple sources and across multiple scales

Did this session pique your interest?
Attend the session on Disease Ecology today at 17:00

Acknowledgments

- A special thank you to
  - Colleen Webb, Colorado State University
  - N. Thompson Hobbs, Colorado State University

  for teaching me much about mathematical biology and ecology

- Thank you to my main collaborators on my work on disease dynamics
  - Libo Sun, Johnson and Johnson
  - Chihoon Lee, Stevens Institute of Technology
  - N. Thompson Hobbs, Colorado State University
  - Mike Miller, Colorado Division of Parks and Wildlife

- Thank you to CSU colleagues (faculty, students, post-docs)
  - Mevin Hooten
  - Clint Leach
  - Joshua Hewitt
  - Erin Gorsich
  - Faustina France-Nkansah

Thank you to the ISEC2018 organizers!

Scientific Committee
Rachel McCrea, University of Kent, UK (Chair)
Res Altwegg, University of Cape Town, SA
Richard Barker, University of Otago, NZ
Simon Bonner, U. of Western Ontario, CAN
Rachel Fewster, University of Auckland, NZ
Marc Kry, Swiss Ornithological Institute, SUI
Jose Lahoz-Monfort, U. of Melbourne, AUS
Brett McClintock, NOAA, USA
Shirley Pledger, Victoria U of Wellington, NZ
Eric Rexstad, University of St Andrews, UK
David Warton, U. of New South Wales, AUS

Conference ambassadors:
Rick Camp
Charles Paxton
Catriona Harris
Lindesay Scott-Hayward

Local Organising Committee
Steve Buckland, Univ. St Andrews (co-chair)
Eric Rexstad, Univ. St Andrews (co-chair)
David Borchers, Univ. St Andrews
Gui Bortolotto, Univ. St Andrews
Mark Brewer, BioSS
Louise Burt, Univ. St Andrews
Claudia Faustino, Univ. St Andrews
Danielle Harris, Univ. St Andrews
Janine Illian, Univ. St Andrews
Ruth King, Univ. Edinburgh
Tiago Marques, Univ. St Andrews
Michail Papathomas, Univ. St Andrews
Charles Paxton, Univ. St Andrews
Rhona Rodger, Univ. St Andrews
Lindesay Scott-Hayward, Univ. St Andrews
Len Thomas, Univ. St Andrews

Projectionists:
Louise Burt, Rick Camp, Fanny Empacher,
Claudia Faustino, Chrissy Fell, Andy Seaton

References I

Talk references (web links in blue)

Slide 9: Ray Hilborn & M. Mangel (2013)
The Ecological Detective Confronting Models with Data
Princeton University Press

Slide 10: Classic references by L. M. Berliner:

- "Statistics, Probability and Chaos," *Statistical Science* (1992)
- "Likelihood and Bayesian Prediction of Chaotic Systems"
  JASA (1991).

Slide 22:

- *Stochastic Epidemic Models and Their Statistical Analysis* (2010), H. Andersson & T. Britton, Springer
- *Statistical Science* special issue on stochastic models for infectious disease dynamics (2018) Vol. 33, No. 1, T. Kypraios and V. Minin, editors.
- J. Cresson and S Sonner (2018)
  A note on a derivation method for SDE models: Applications in biology and viability criteria
  Stochastic Analysis and Applications

References II

Slide 23: Linda Allen (2008)
An introduction to stochastic epidemic models,
Chapter 3 in *Mathematical epidemiology*, p 81–130.

Slide 34: Computation

- D He, EL Ionides, AA King (2009)
  Plug-and-play inference for disease dynamics: measles in large and small populations as a case study
  Journal of the Royal Society Interface
- EL Ionides, C Bret, AA King (2006)
  Inference for nonlinear dynamical systems
  Proceedings of the National Academy of Sciences
- L. Sun, C. Lee, and J. A. Hoeting (2015)
  A penalized simulated maximum likelihood approach in parameter estimation for stochastic differential equations *Computational Statistics and Data Analysis*, 84: 54–67
- J. Ramsay and G. Hooker Dynamic Data Analysis
  Chapman & Hall

References III

- Approximate Bayesian Computation Overview
  David Nott (2018)
  Talk and slides available at SIAM Uncertainty Quantification Conference 2018
- M. Fasiolo, N. Pya, S. Wood (2016)
  A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology
  Statistical Science

Slide 35: S. Marino, I. Hogue, C. Ray, D. Kirschner (2008)
A methodology for performing global uncertainty and sensitivity analysis in systems biology, *Journal of Theoretical Biology*

Slide 36: Model selection

- JA Hoeting, D Madigan, AE Raftery, CT Volinsky (1999)
  Bayesian model averaging: a tutorial
  Statistical Science
- RE Kass, AE Raftery (1995)
  Bayes factors
  Journal of the American Statistical Association

References IV

- MB Hooten, NT Hobbs - Ecological Monographs (2015)
  A guide to Bayesian model selection for ecologists
  Ecological Monographs
- GJ Gibson, G Streftaris, D Thong (2018)
  Comparison and Assessment of Epidemic Models
  Statistical Science

Slide 37:

- L. Sun, C. Lee, and J. A. Hoeting (2015)
  Parameter inference and model selection in deterministic and stochastic dynamical models via approximate Bayesian computation: modeling a wildlife epidemic
  Environmetrics, 26: 451–462.
- Two other perspectives on the same disease system:
  1. C Geremia, MW Miller, JA Hoeting, MF Antolin, NT Hobbs
     Bayesian Modeling of Prion Disease Dynamics in Mule Deer Using Population Monitoring and Capture-Recapture Data
     PloS One 10 (10)

References V

2. L. Sun, C. Lee, and J. A. Hoeting (2015)
A penalized simulated maximum likelihood approach in parameter estimation for stochastic differential equations *Computational Statistics and Data Analysis*, 84: 54–67

Other useful references:

- Recent Review (theory): K. McGoff, S. Mukherjee, N. Pillai (2015)
"Statistical inference for dynamical systems: A review"
Statistics Surveys, 9: 209–252.
- K. Newman (2018) Population Demography for Ecology
in *Handbook of Environmental and Ecological Statistics*
Chapman & Hall, to appear.

Photo acknowledgments

Web links in blue.

Slide 3: Save the Tasmanian Devil Program, Tasmanian Government, Dept of Wildlife Management

'Nice' Tasmanian Devil, Fodors

Snarling Tasmanian Devil, Animal Sounds

Slide 4: Zika virus, Hollywood Gazette

Slide 29: Types of disease transmission

Example: ODE model for direct/indirect transmission of CWD

An ODE model for the direct and indirect transmission of CWD (Miller et al. 2006)

$$d\begin{pmatrix} S \\ I \\ E \\ C \end{pmatrix} = \begin{pmatrix} a - S(\gamma E + m) \\ \gamma SE - I(\mu + m) \\ \epsilon I - \tau E \\ \mu I \end{pmatrix} dt,$$

where

- γ is the indirect transmission coefficient
- ϵ is the per capita rate of excretion of infectious material by infectious animals
- τ is the mass-specific rate of loss of infectious material from the environment

The unknown quantities to be estimated are $(\gamma, \mu, \epsilon, \tau, S(t_0), I(t_0), E(t_0))$.

Example: Parameter estimates

The marginal posterior distribution for the parameters of the indirect transmission SDE model based on the CWD epidemic data.
Left column is prior set 1 and right column is prior set 2