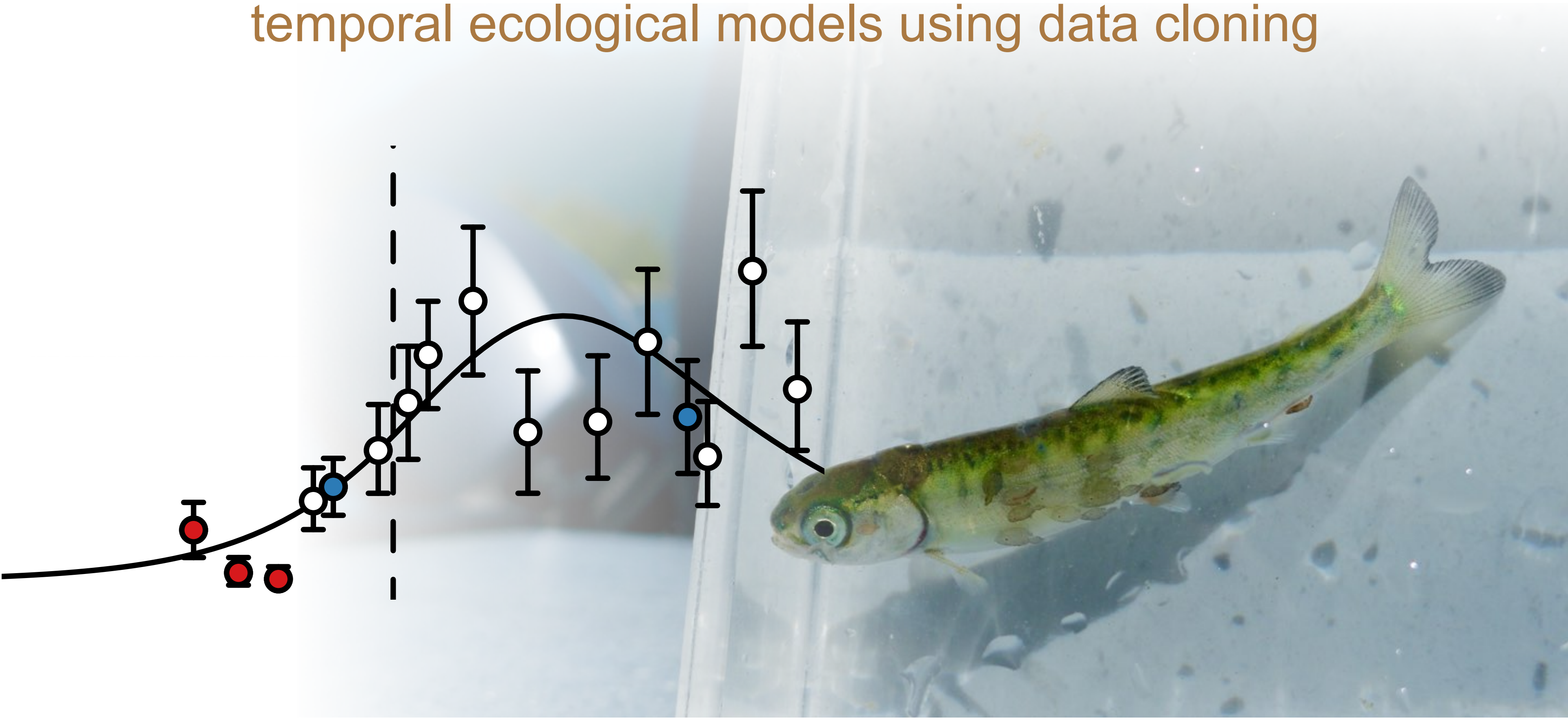
Lousy lessons learned

Study design and parameter estimability for spatial and temporal ecological models using data cloning



Mark Lewis

Stephanie Peacock, Martin Krkošek, Andrew Bateman, and Subhash Lele

UNIVERSITY OF CALGARY
UNIVERSITY OF TORONTO
UNIVERSITY OF ALBERTA

Parameter estimability

Non-indentifiability

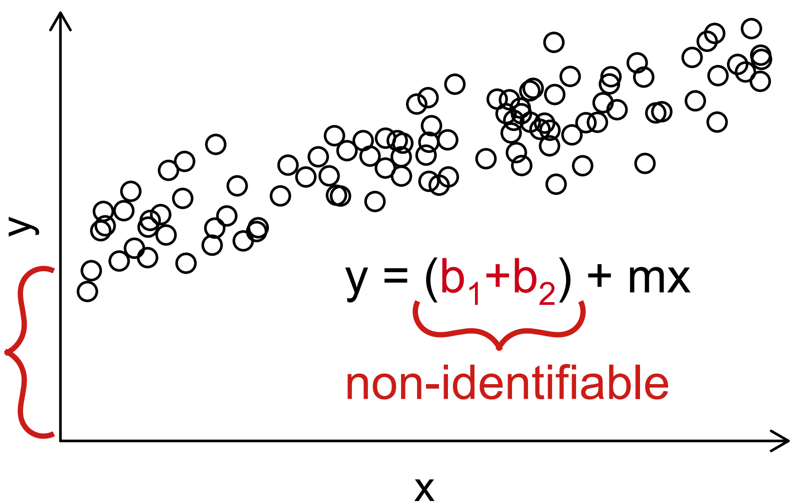
Parameter can not be estimated, no matter how many data

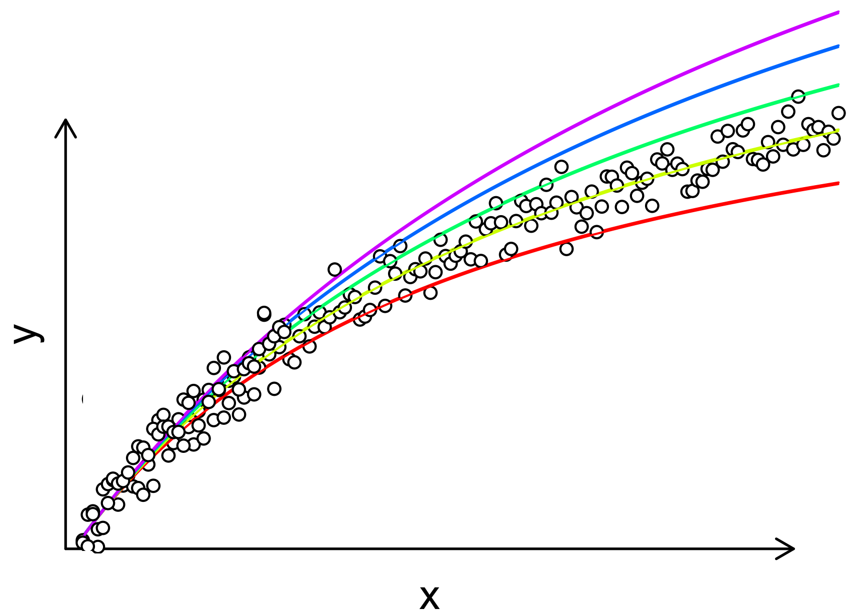
$\Rightarrow$

$\nLeftarrow$

Non-estimability

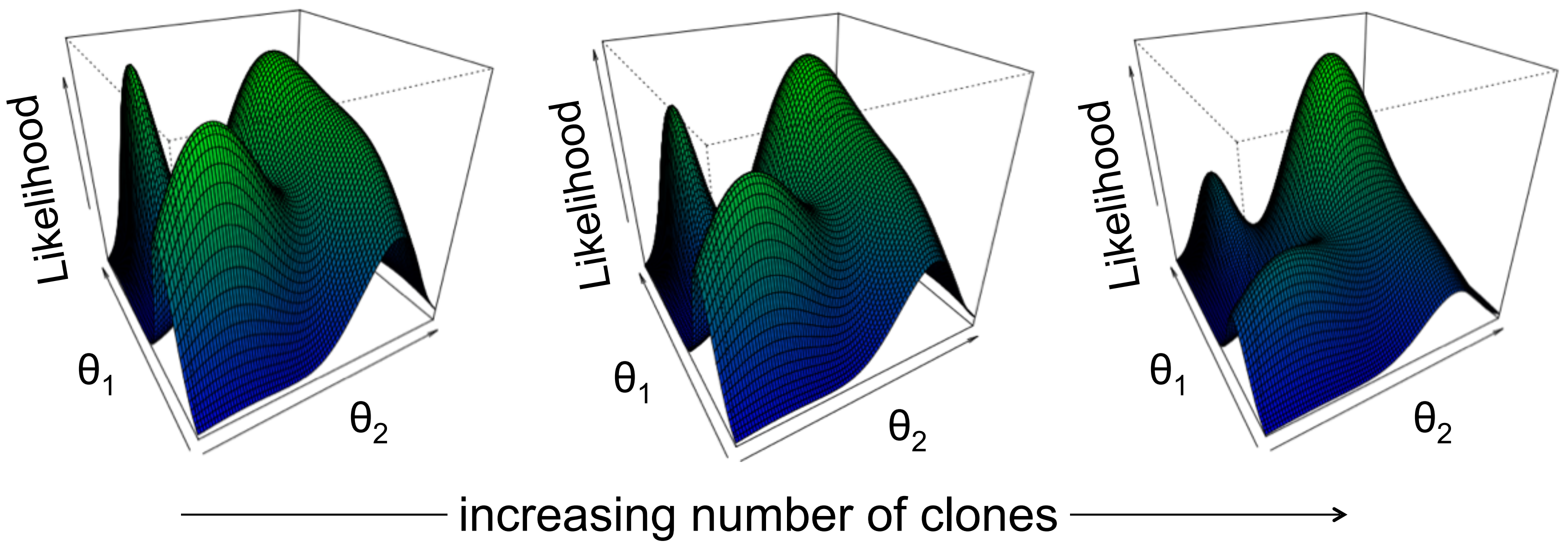
Parameter can not be estimated, given the data you have





Data cloning

- Maximum Likelihood Estimates (MLEs) using **MCMC** in a **Bayesian framework** by *overwhelming* the prior
- Global MLEs when your likelihood surface may be flat or multi-modal
- Estimability of parameters in your model



Lele et al. 2007 *Ecology Letters*

How does data cloning work?

1. Create a *K* cloned data set

2. Using MCMC, generate draws from your posterior, based on some (proper) prior and the likelihood of the cloned data vector

3. Compute means and sample variances from the marginal posteriors from the MCMC output

$$\mathcal{D}^k = \underbrace{\mathcal{D}, \mathcal{D}, \ldots, \mathcal{D}}_{k \text{ times}}$$

$\mathcal{D}$: data

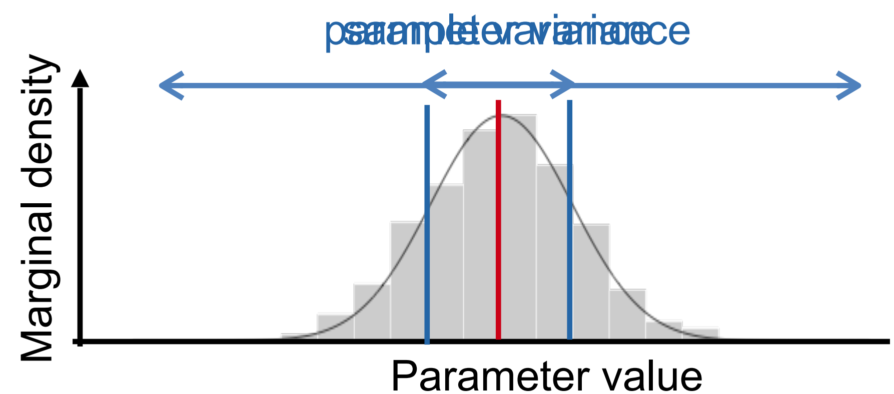
k : number of clones

$\mathcal{M}$: model

$$Pr(\mathcal{M}|\mathcal{D}^k) \propto Pr(\mathcal{D}^k|\mathcal{M})Pr(\mathcal{M})$$

Given enough clones….

4. The MLE is the mean of the posterior, and the variance is ***K* times** the MCMC sample variance

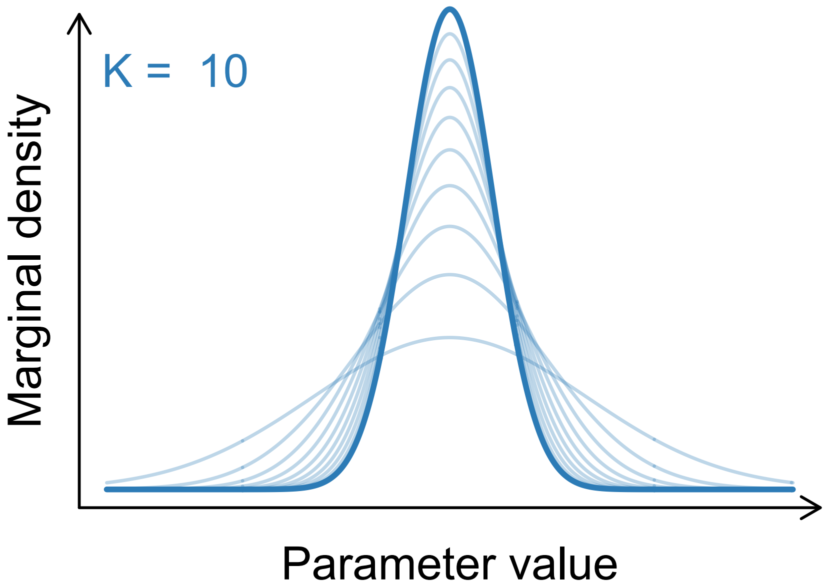


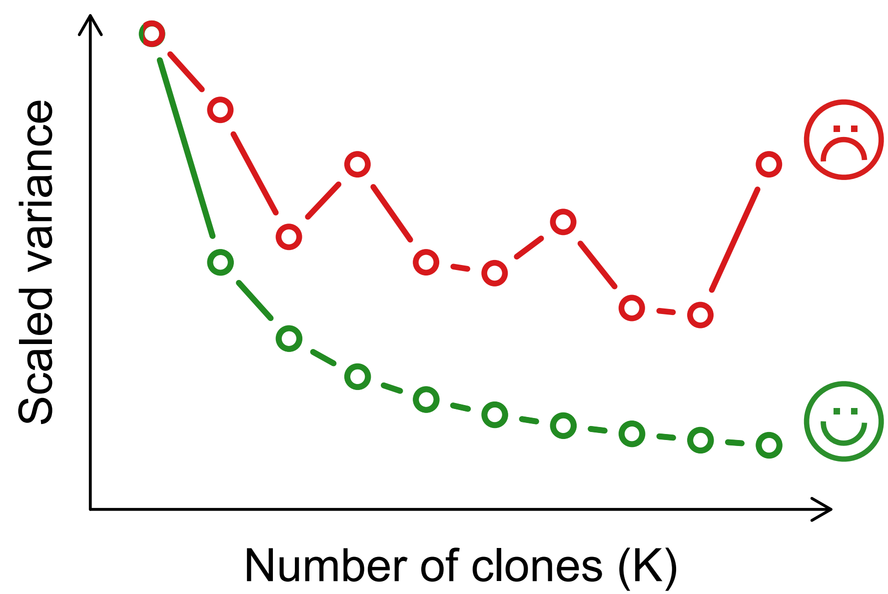

```
> install.packages("dclone")
```

Lele et al. 2007 *Ecology Letters*

Data cloning and estimability

Variance in posterior should decrease at a rate of $1/K$



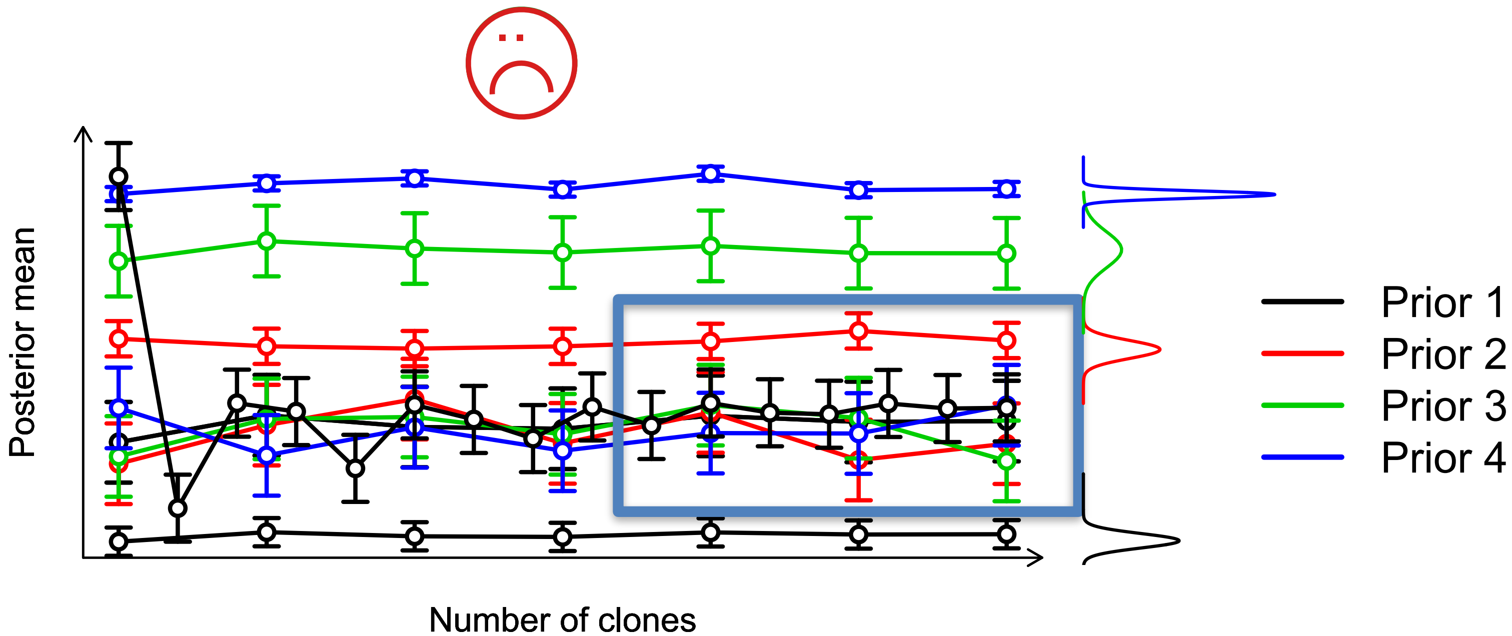


Lele et al. 2010 *J. Amer. Stat. Assoc.*

Data cloning and estimability

Variance in posterior should decrease at a rate of 1/*K*

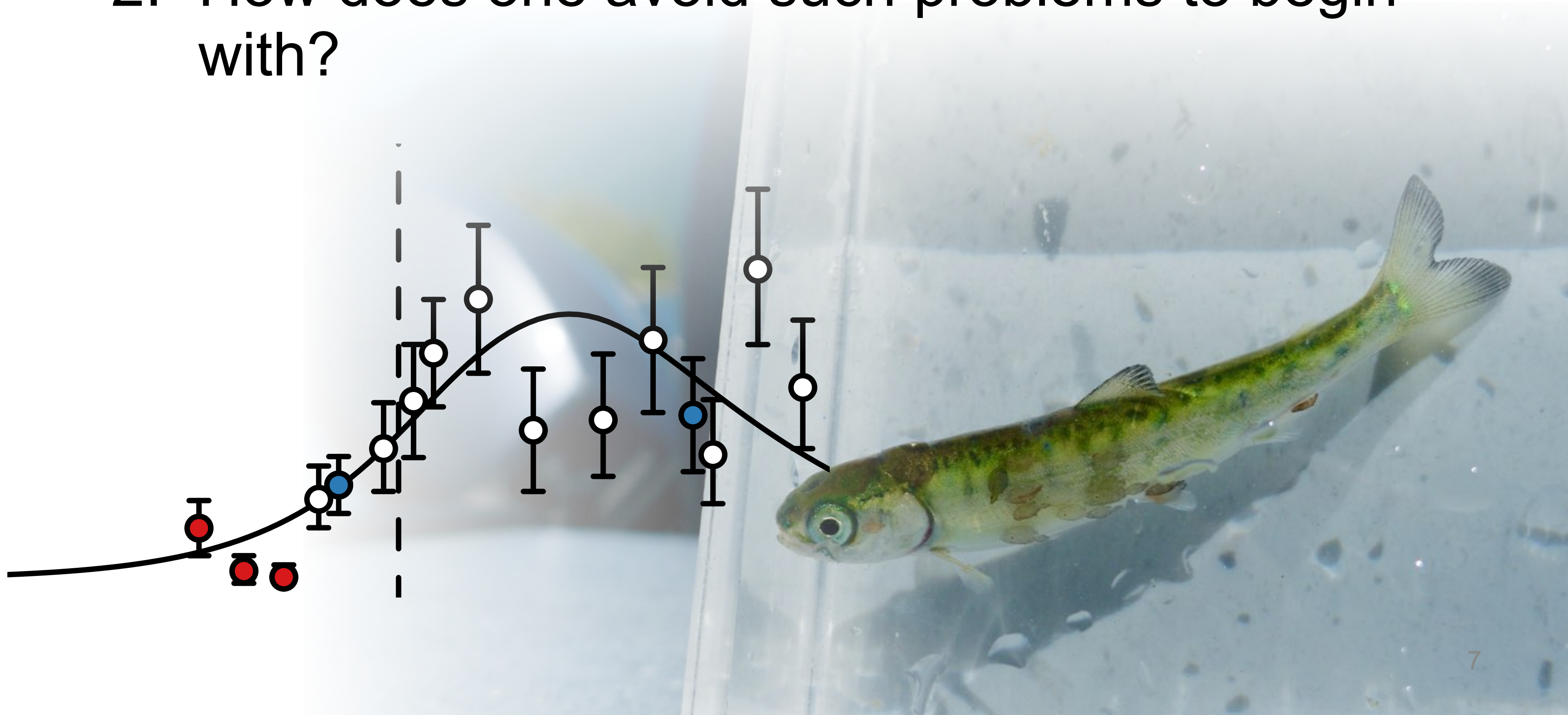
Posterior mean should be invariant to the choice of prior



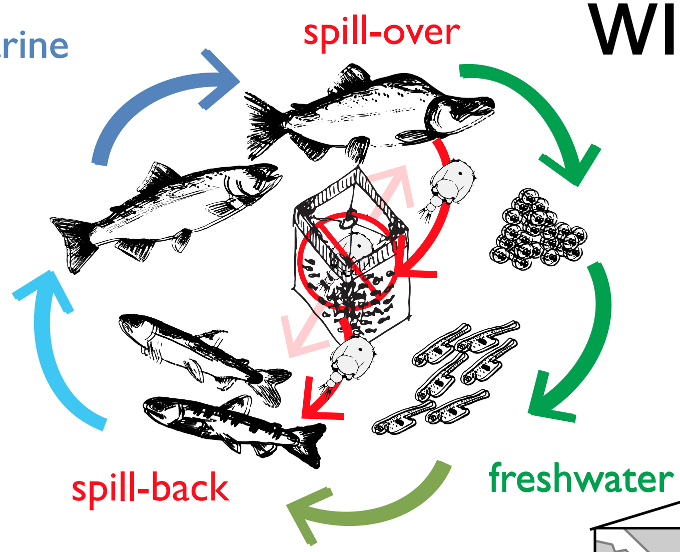
Campbell and Lele 2013 *Comput. Stat. & Data Analysis*

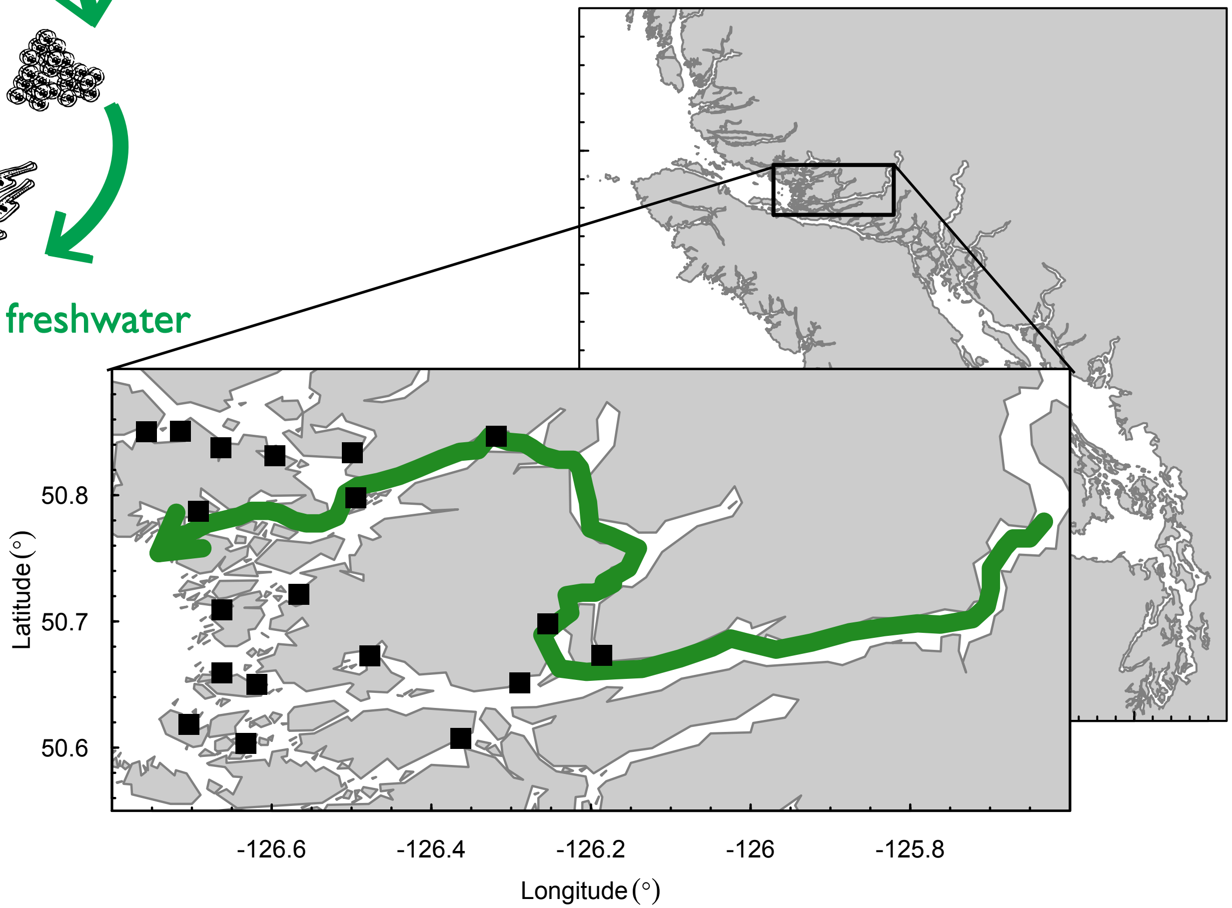
Case study: sea lice on juvenile salmon

1. How does one address problems of parameter non-estimability?
2. How does one avoid such problems to begin with?



Sea louse transmission from farmed to wild salmon



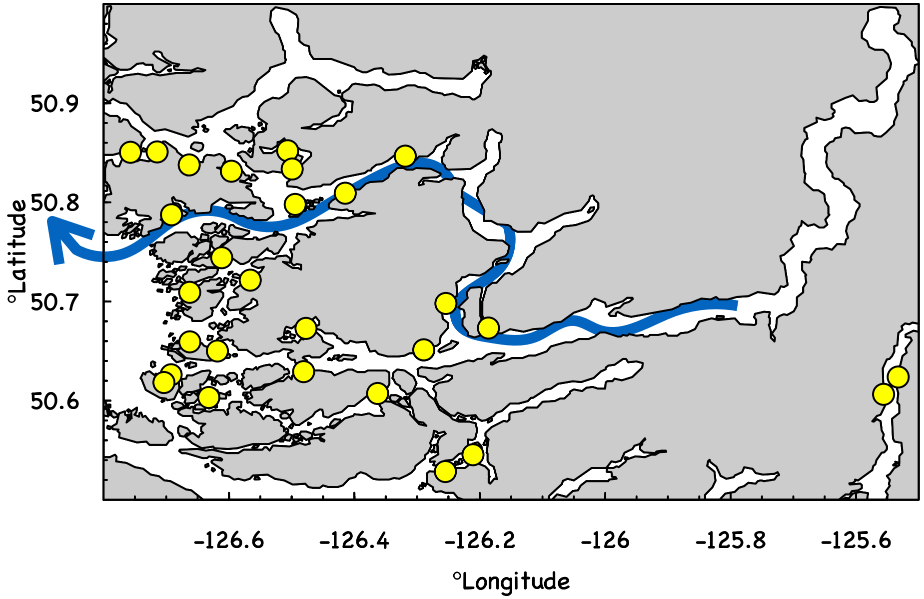


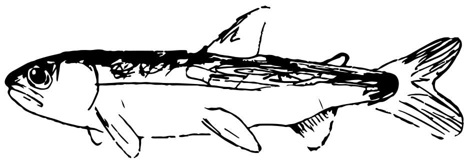
Krkošek et al. 2005 *Proc Roy Soc B: Biol. Sci.*

Data:
Spatial surveys of sea louse abundance

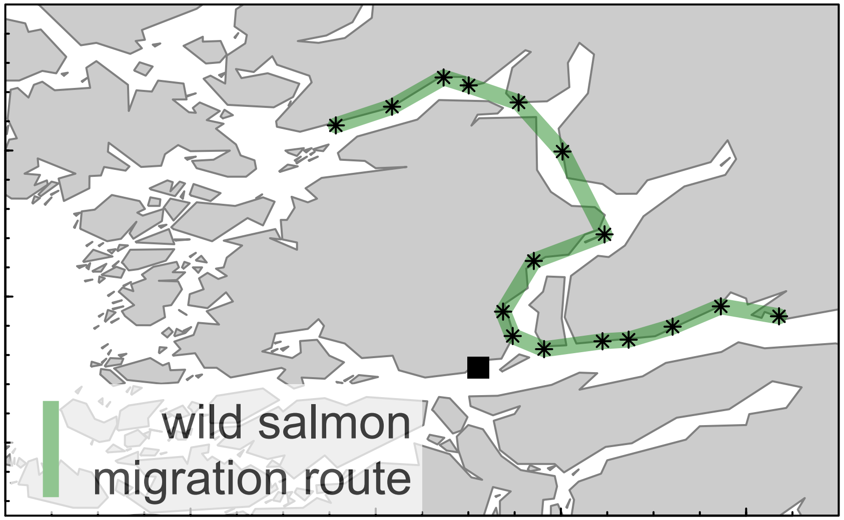
location of salmon farms

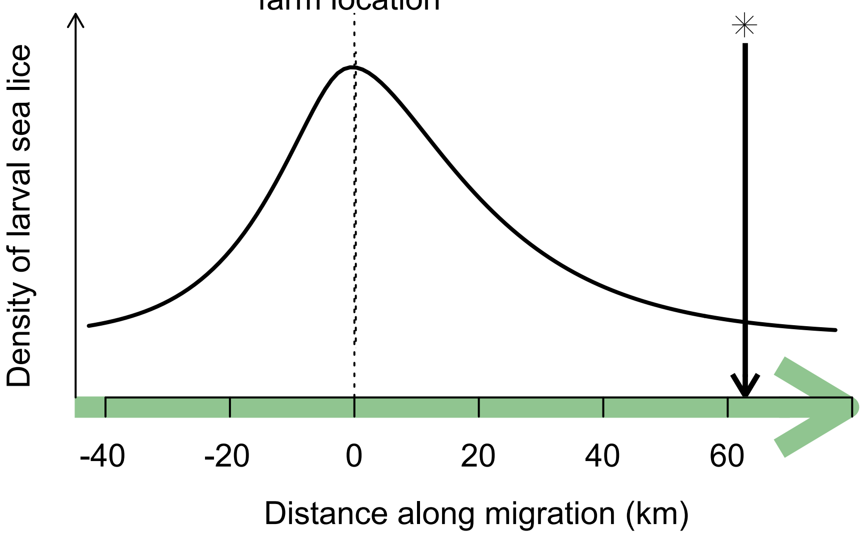
wild salmon migration route



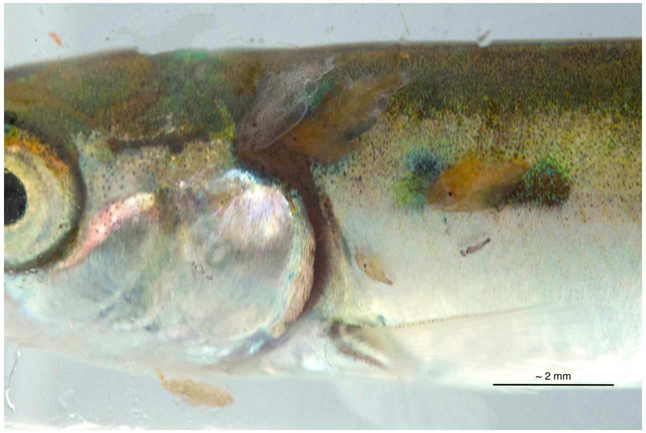


The model





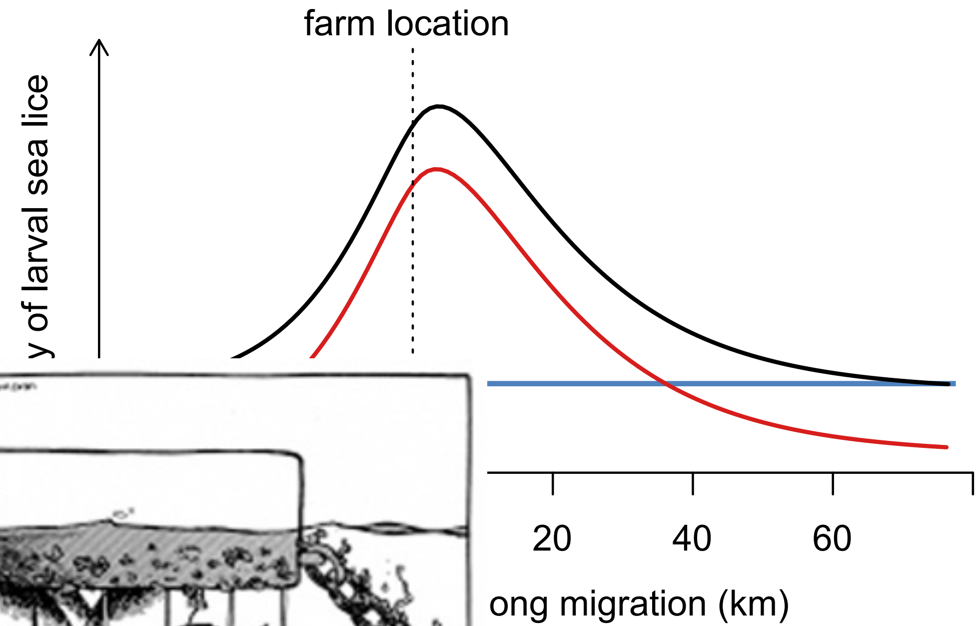
What is the relative contribution of farm and ambient sources of sea lice?

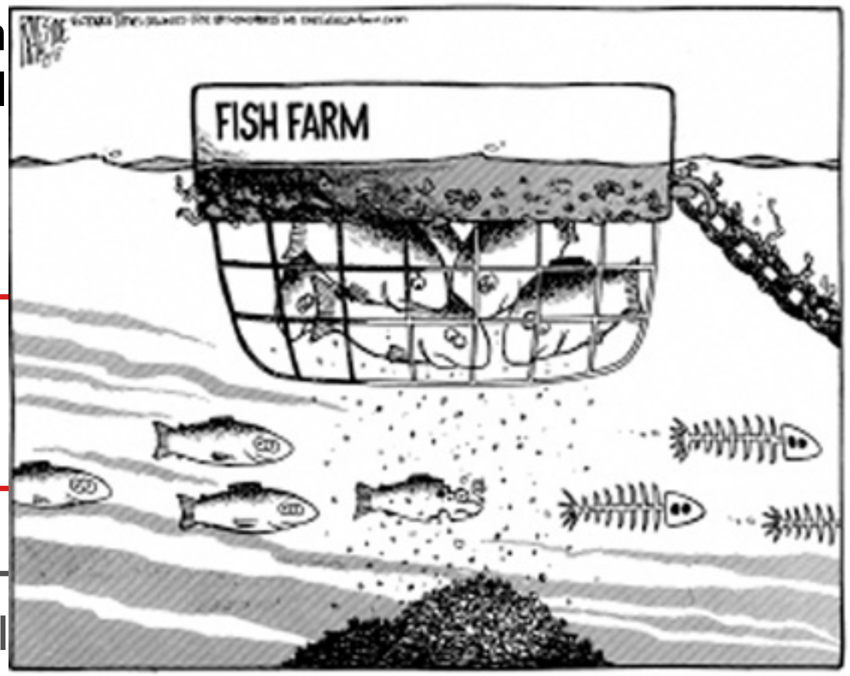


Krkošek et al. 2005 *Proc Roy Soc B: Biol. Sci.*

Three hypotheses:

1. No effect of farms
2. Point source of sea lice at location of farm
3. Both backgroun sources of sea l





Constants
D, g, m_n, m_p, q_n

und sources of
arasites

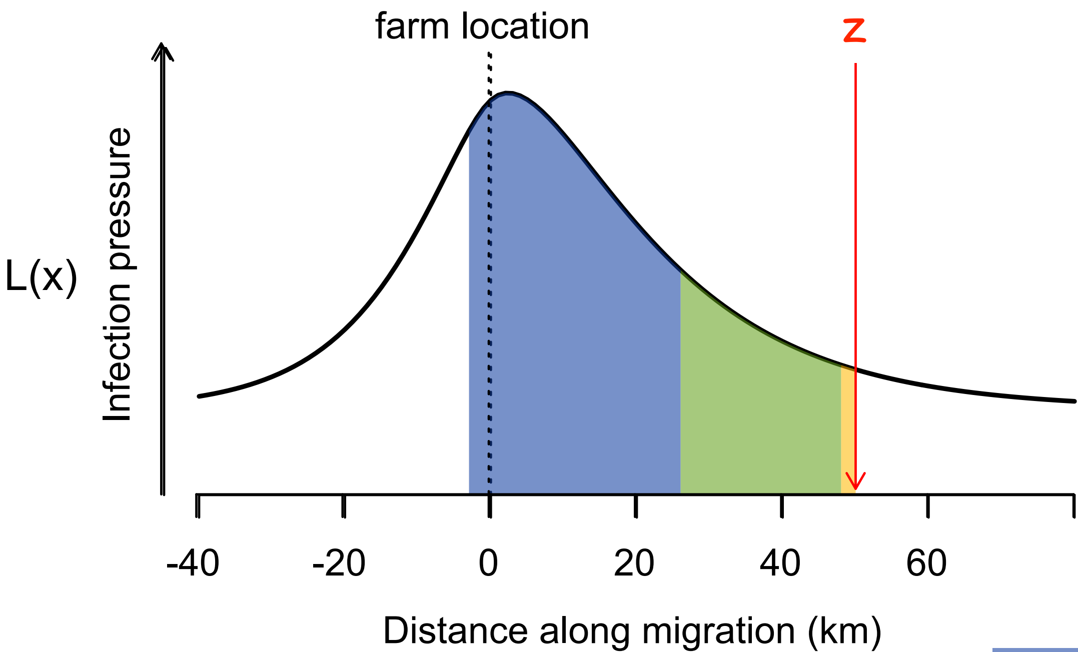
Wild data
c_i, h_i, m_i

l
on wild salmon
infection → host migration + development + survival (s_c, s_h)

Estimated parameters
f, k, l_c, l_h, l_m, s_c, s_h

Krkošek et al. 2005 *Proc Roy Soc B: Biol. Sci.*

The math



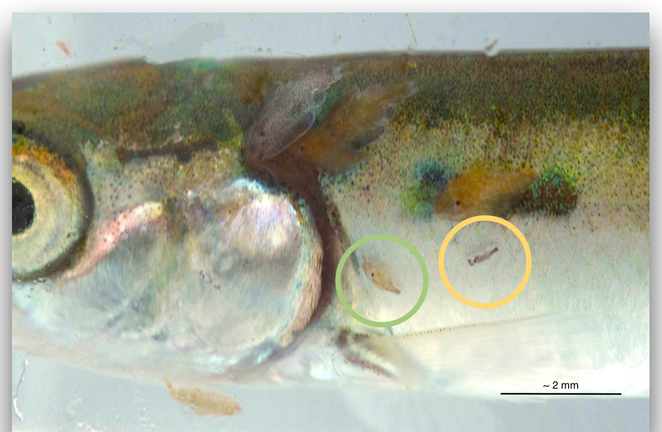
$$L(x) = k + f\, f(x)$$

Expected number of lice of each stage:

$$C(z) = \beta \int_{z-\lambda_c}^{z} L(x)dx$$

$$H(z) = \beta s_c \int_{z-\lambda_c-\lambda_h}^{z-\lambda_c} L(x)dx$$

$$M(z) = \beta s_h s_c \int_{z-\lambda_c-\lambda_h-\lambda_m}^{z-\lambda_c-\lambda_h} L(x)dx$$

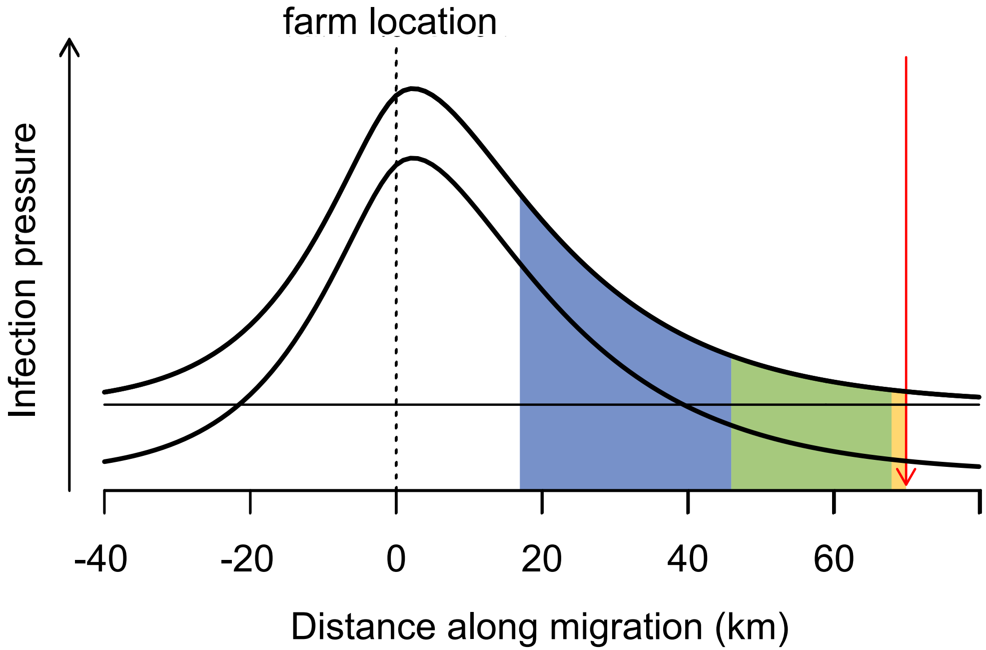


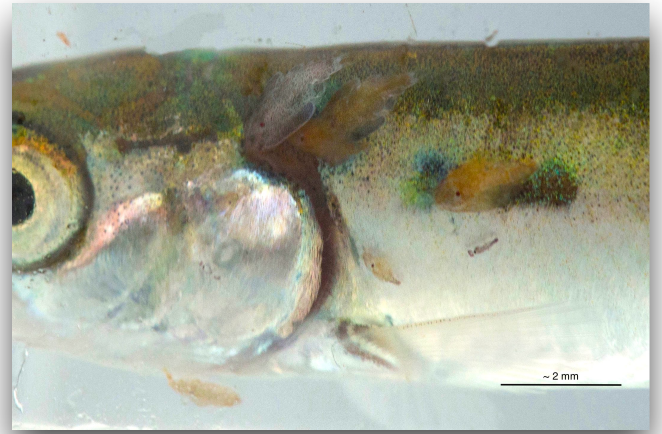
b transmission coefficient (unknown)

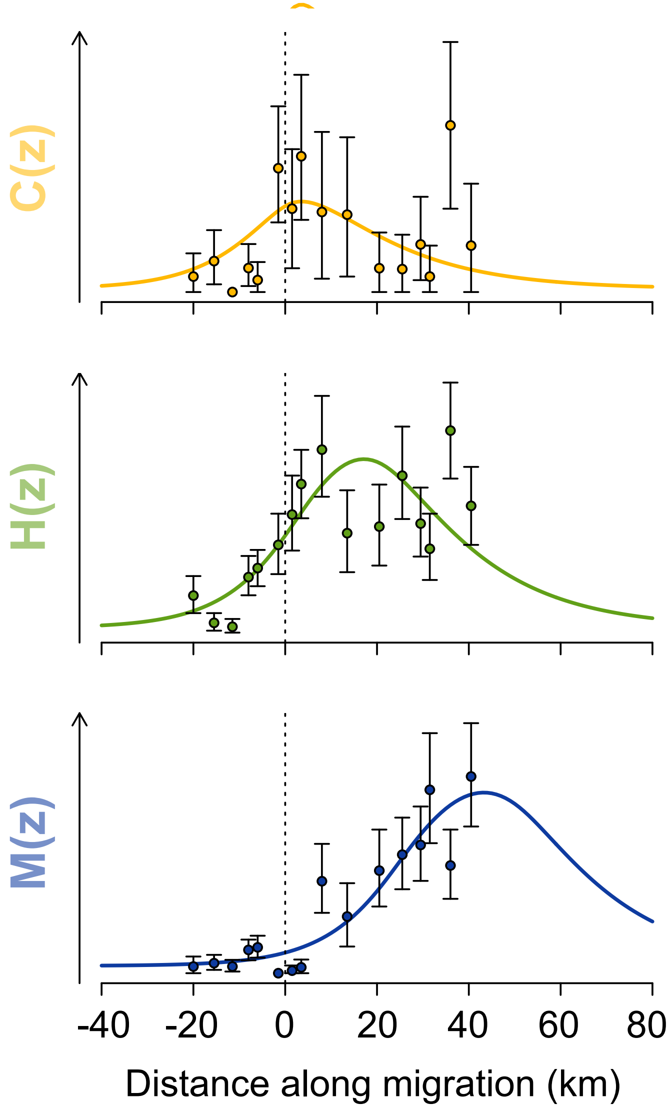
l_i distance migrated during stage *i* (days)

s_i survival of lice from stage i to i+1

The math

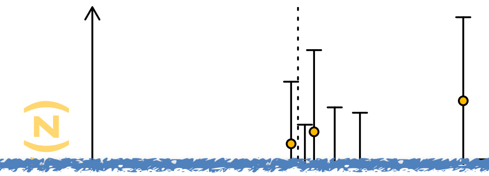

farm location
Infection pressure
-40
-20
0
20
40
60
Distance along migration (km)


~ 2 mm


C(z)
H(z)
M(z)
-40
-20
0
20
40
60
80
Distance along migration (km)

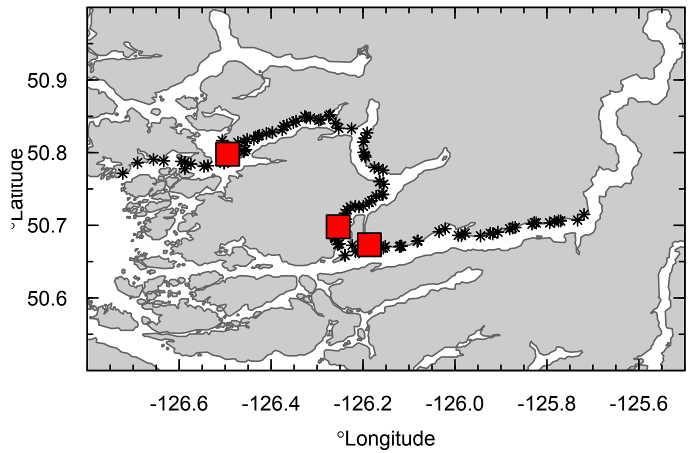
Model results

"Farm salmon were the primary source of lice, raising the density of infective parasite larvae



Can we apply the model to test if we can measure effects of multiple farm sources of lice (f)?

2006:



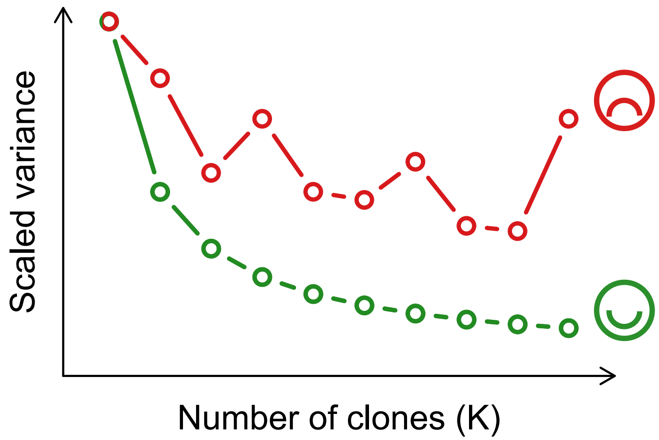
Ecological Applications, 23(3), 2013, pp. 606–620

Cessation of a salmon decline with control of parasites

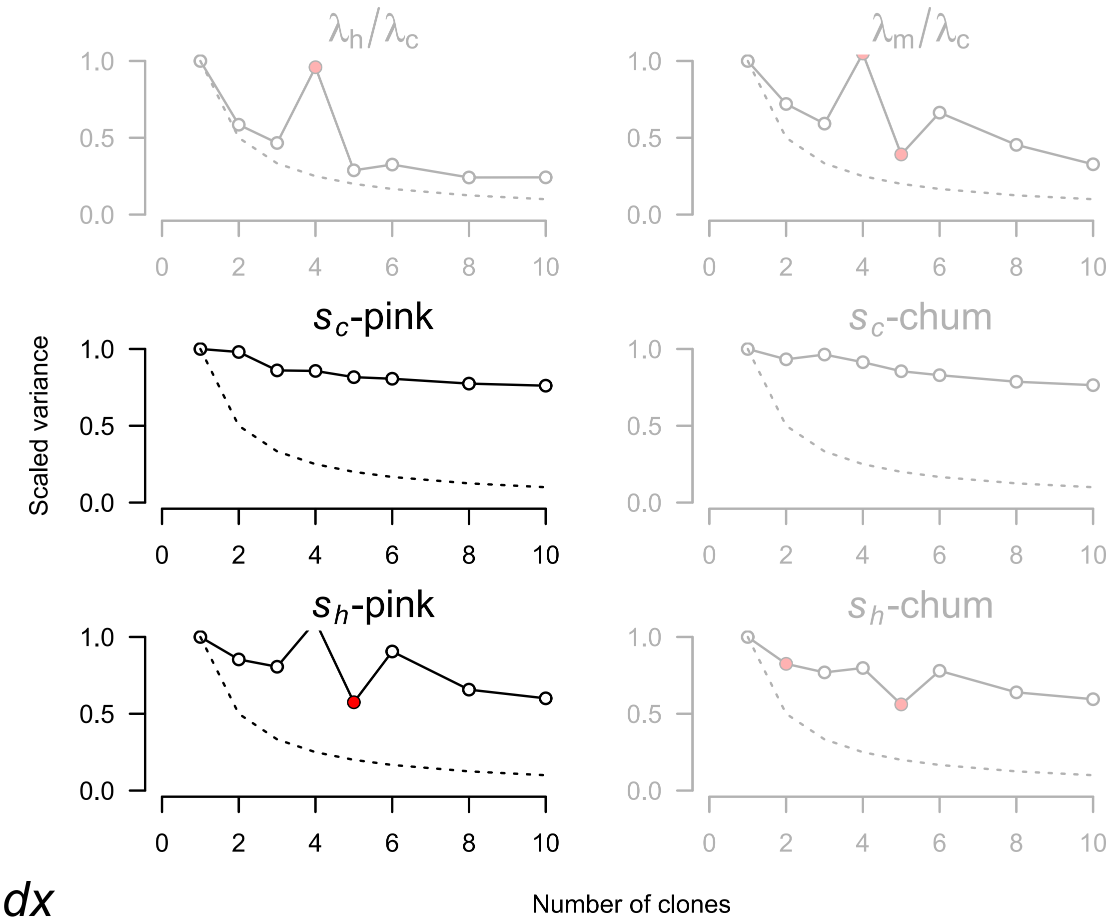
Stephanie J. Peacock,[1,2,6] Martin Krkošek,[3,4] Stan Proboszcz,[5] Craig Orr,[5] and Mark A. Lewis[1,2]

[1]Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2G1 Canada
[2]Centre for Mathematical Biology, Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta T6G 2G1 Canada
[3]Department of Zoology, University of Otago, Dunedin, New Zealand

Non-estimability of parameters



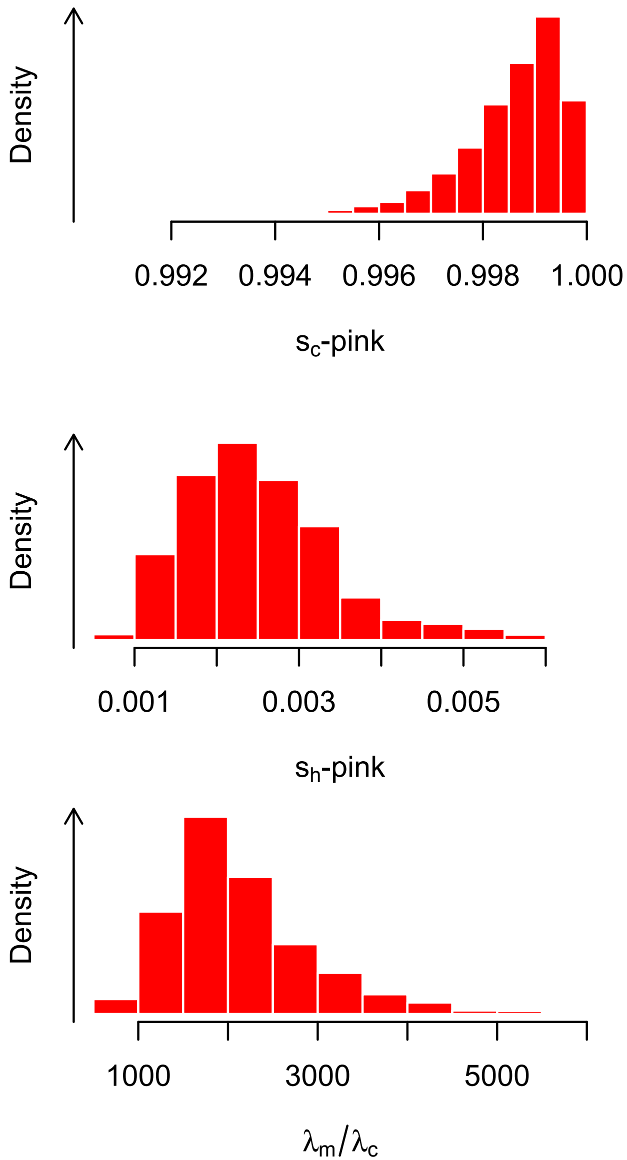
2006 data

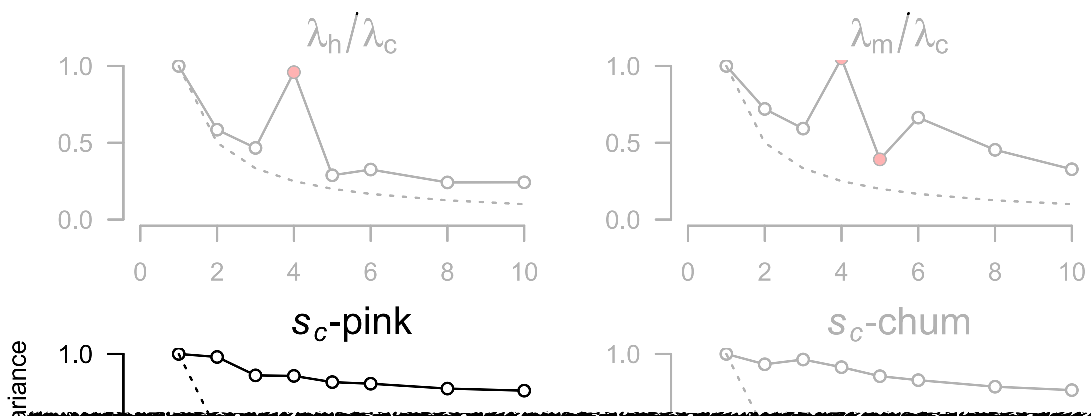


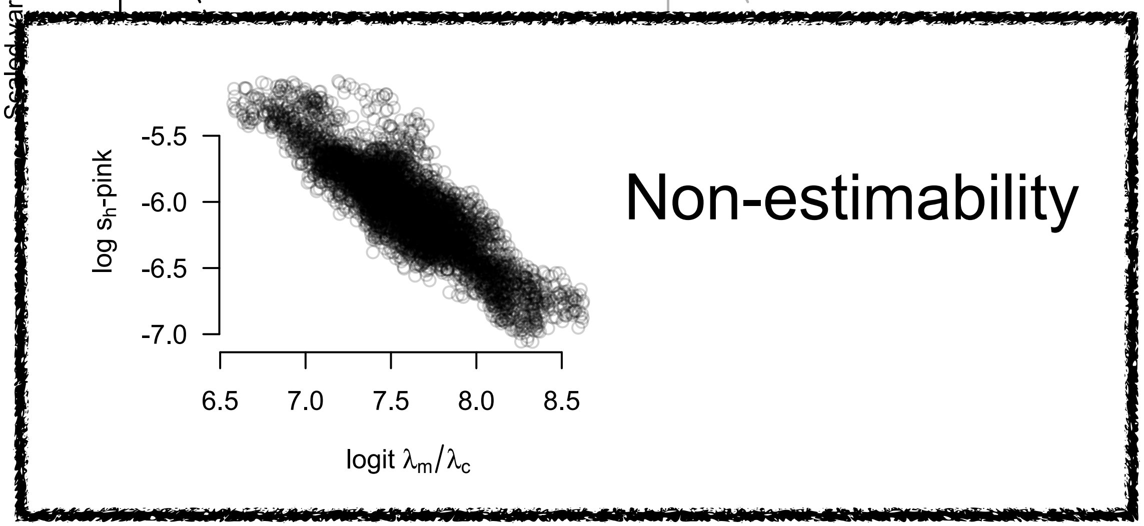
$$H(z) = \beta s_c \int_{z-\lambda_c-\lambda_h}^{z-\lambda_c} L(x)dx$$

$$M(z) = \beta s_h s_c \int_{z-\lambda_c-\lambda_h-\lambda_m}^{z-\lambda_c-\lambda_h} L(x)dx$$

Non-estimability of parameters

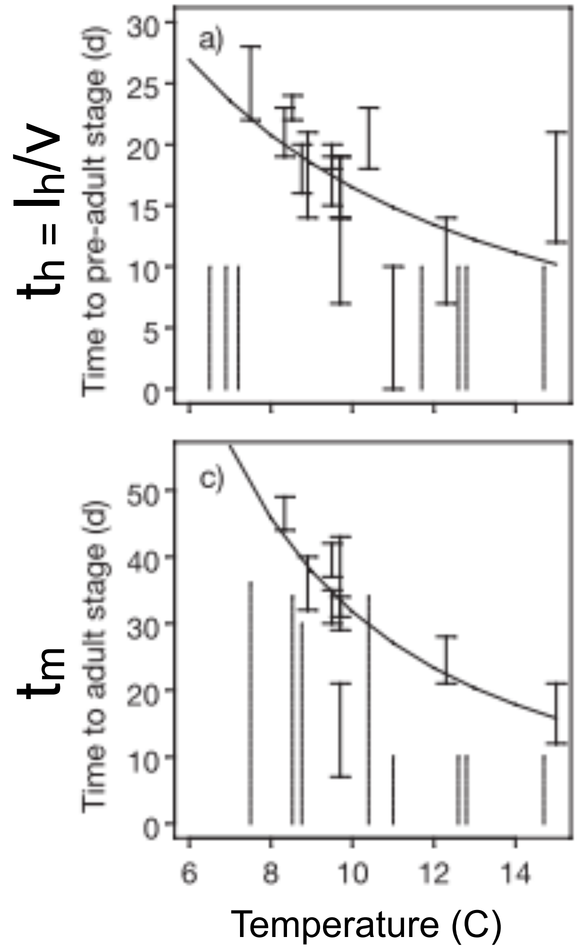






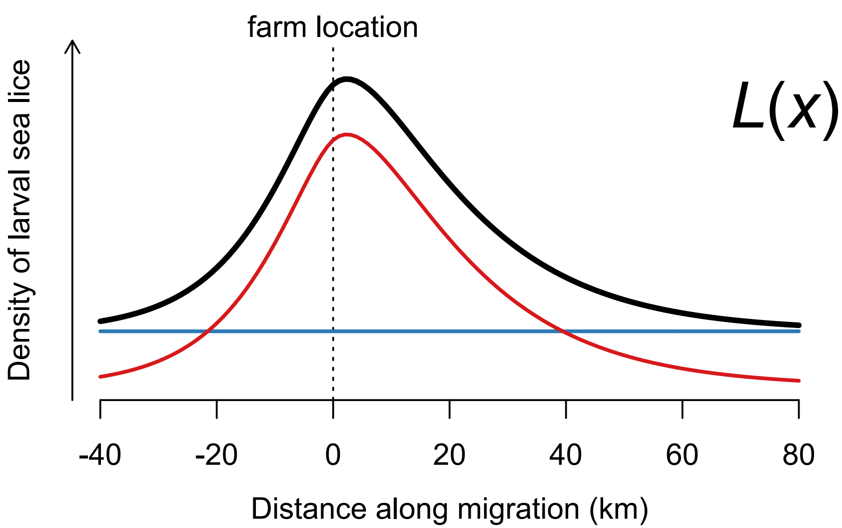
What to do about non-estimability?

- Fix parameters if information is available
  - Previous study
  - Parameters for different host species not significantly different
- Revisit model str
  - Is something
  - Can additiona included?



Stien et al. 2005 *Mar. Ecol. Prog. Ser.*

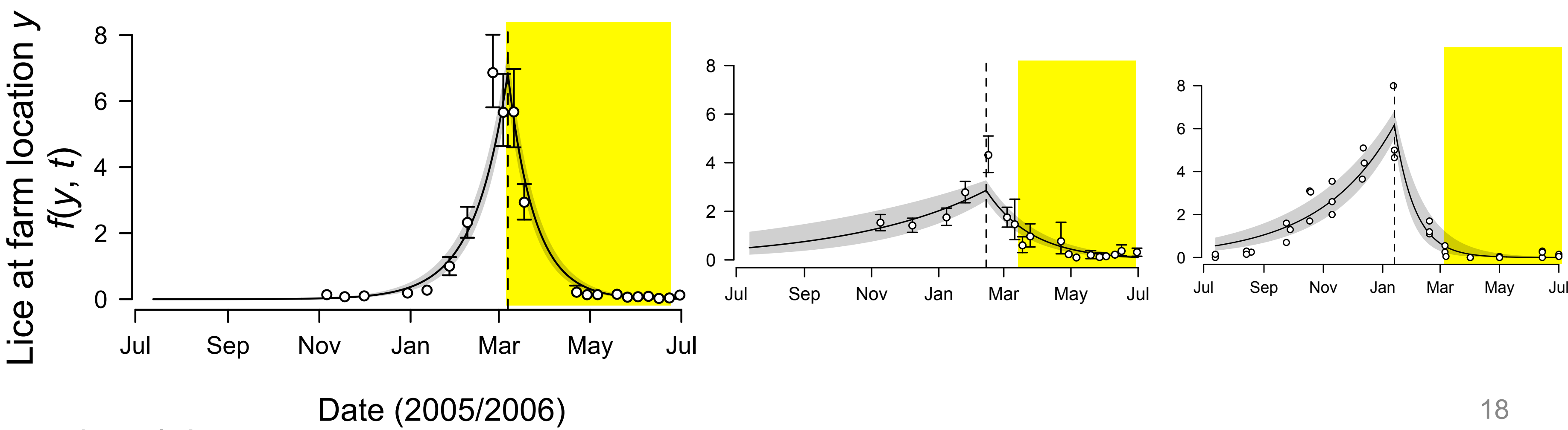
Revisiting model structure



$L(x) = k + f\, f(x)$

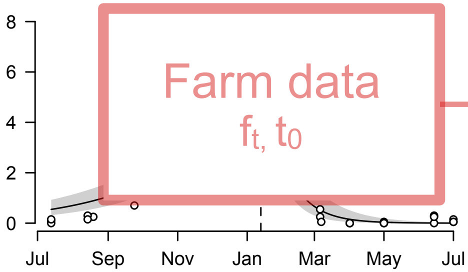
↑

- Assumed to be a constant source through time
  - steady-state solution to advection-diffusion-decay equations



Peacock et al. *In prep.*

Revisiting model structure



Farm data
f_t, t_0

Temporal dynamics on farms
growth (r_1) → treatment (f_0) → decay (r_2)

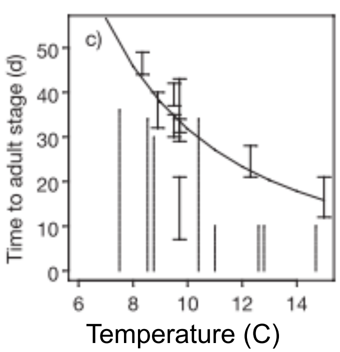
Estimated parameters
r_1, f_0, r_2

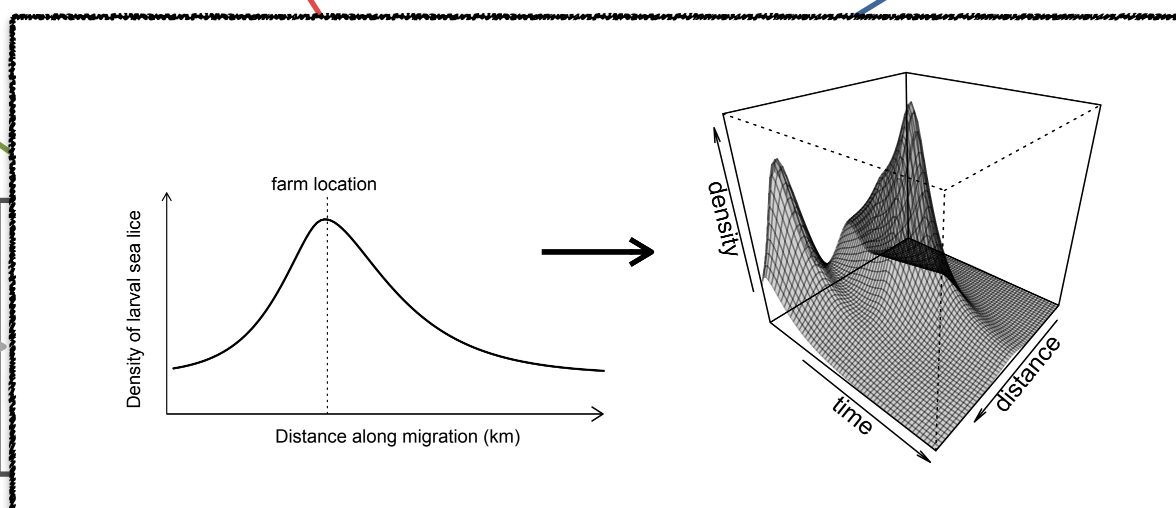
Constants
D, g, m_n, m_p, q_n

Advection, diffusion of parasites from farms

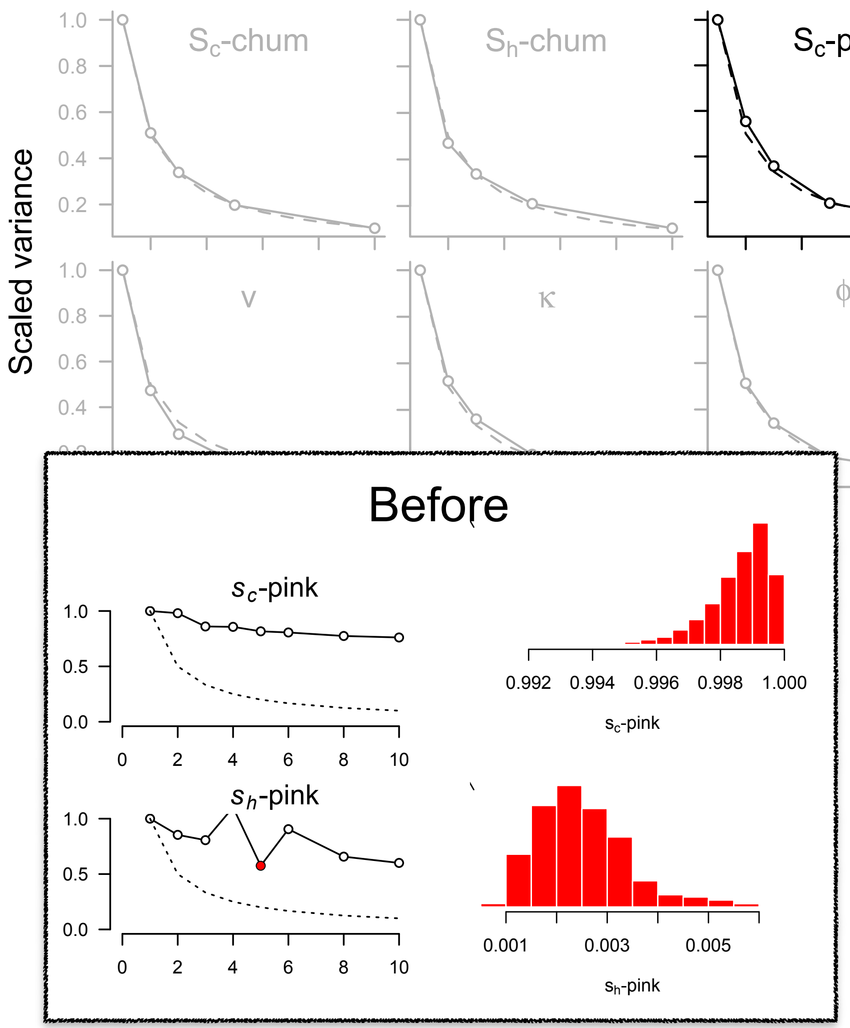
Background sources of parasites

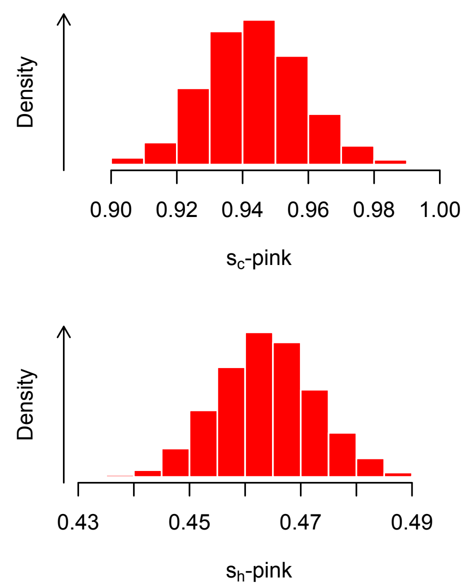
Constants
t_c, t_h, t_m

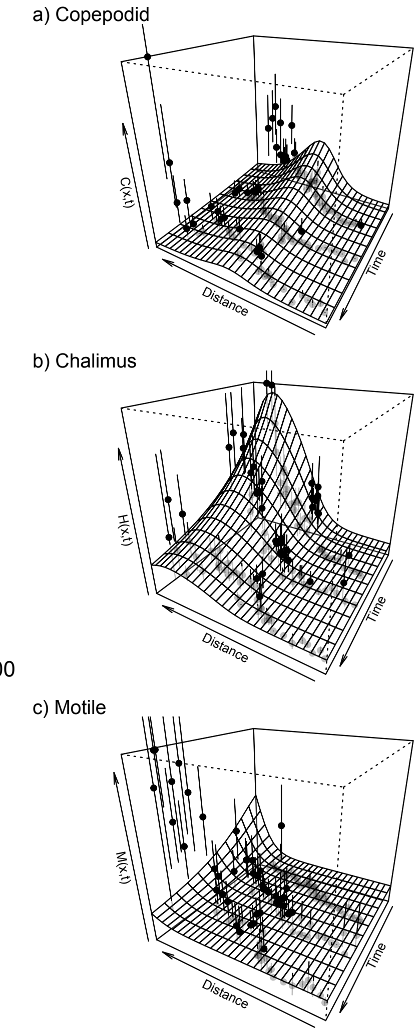




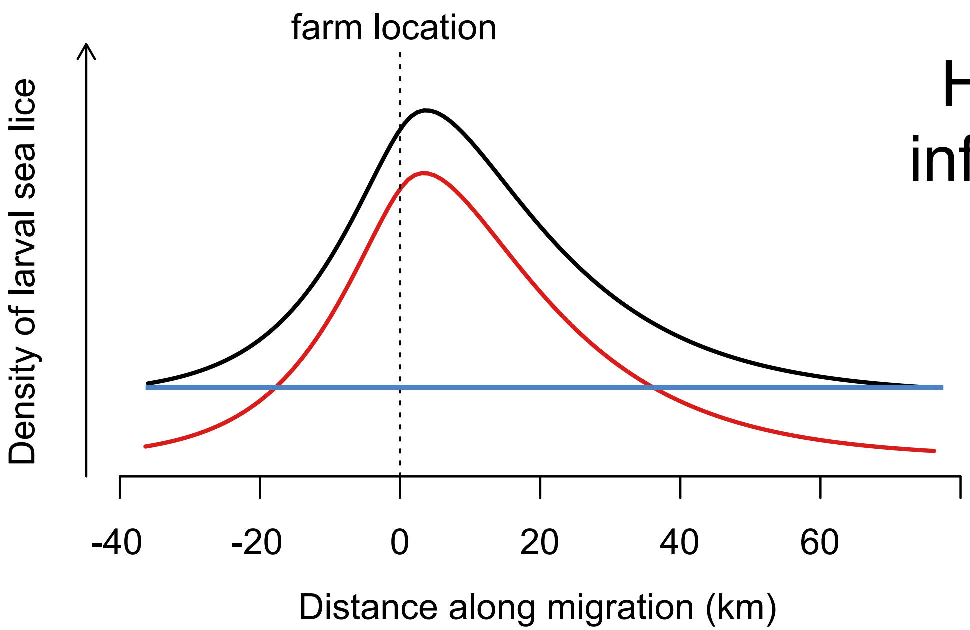
Non-estimability of parameters
Estimability of parameters







Revisiting our hypotheses



How did non-estimability influence our conclusions?

1. Ambient sources - k
2. Farm sources - f
3. Both - k & f

Relative strength of farm vs. ambient sources of lice:

f / k

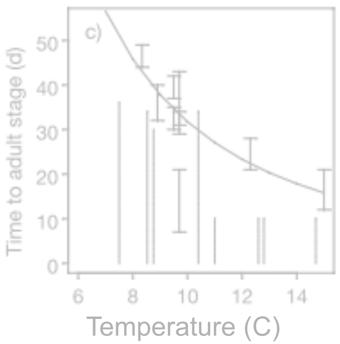
Original 2006 model (non-estimable)	29 - 16,165
Revised spatiotemporal 2006 model (estimable)	578
Published 2004 estimates	28 - 21,445

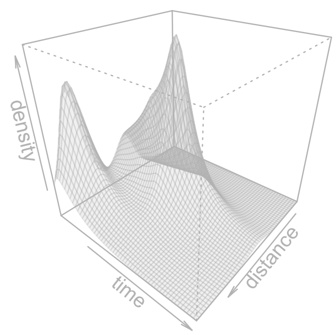
Krkosek et al. 2006 *PNAS*

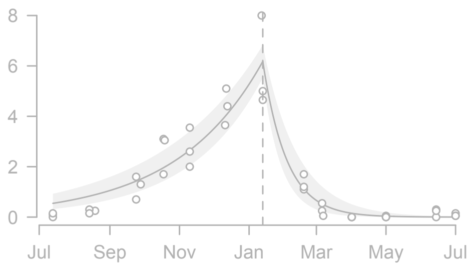
Case study: sea lice on juvenile salmon

1. How does one address problems of parameter non-estimability?

✓ Fixing parameters ✓ Revisiting model ✓ Collecting more/ different data*

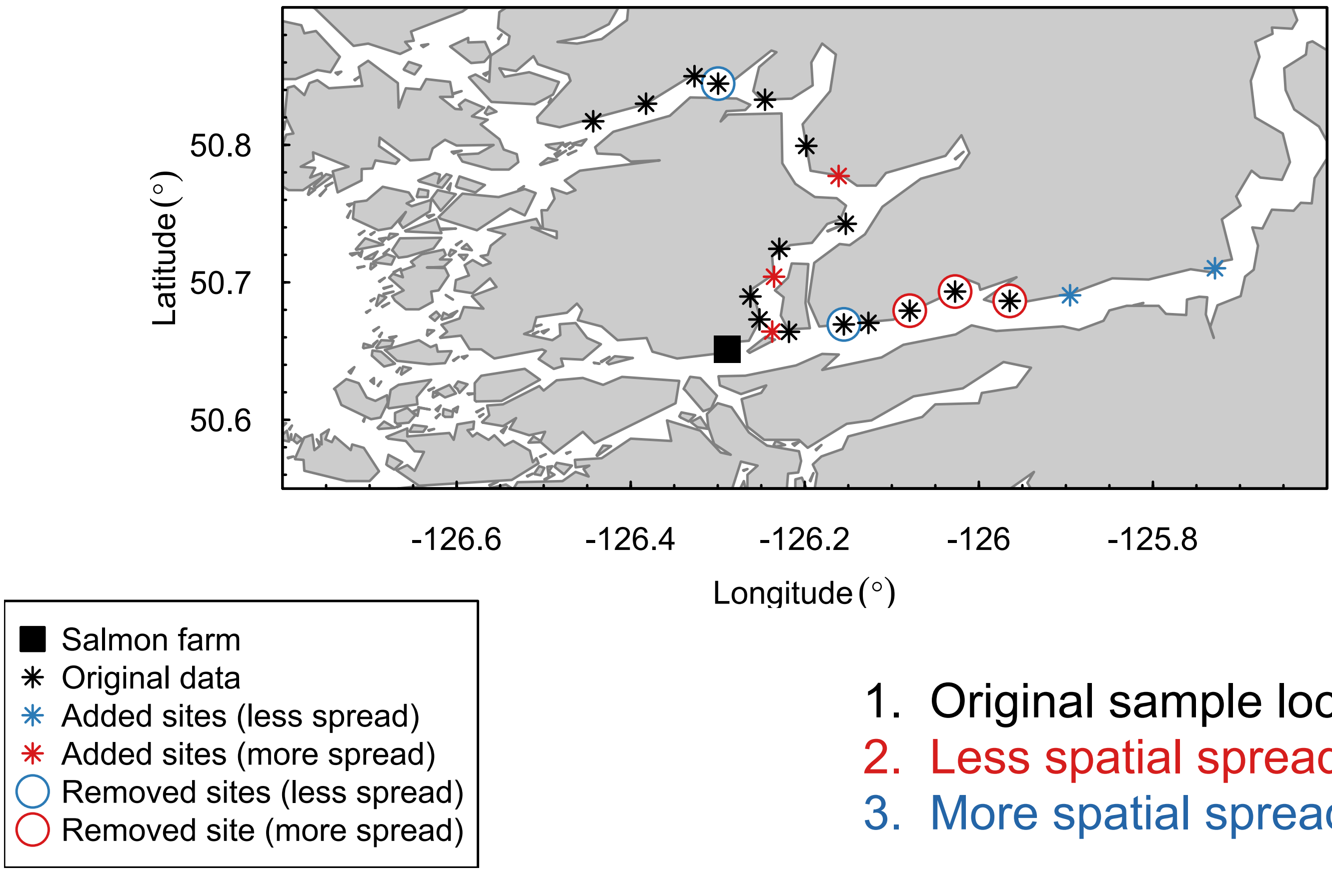






2. How does one avoid such problems to begin with!?

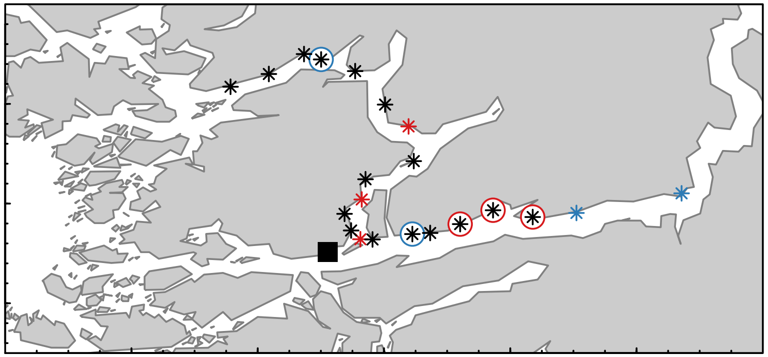
Which sampling design should be adopted to ensure *k* and *f* are estimable?

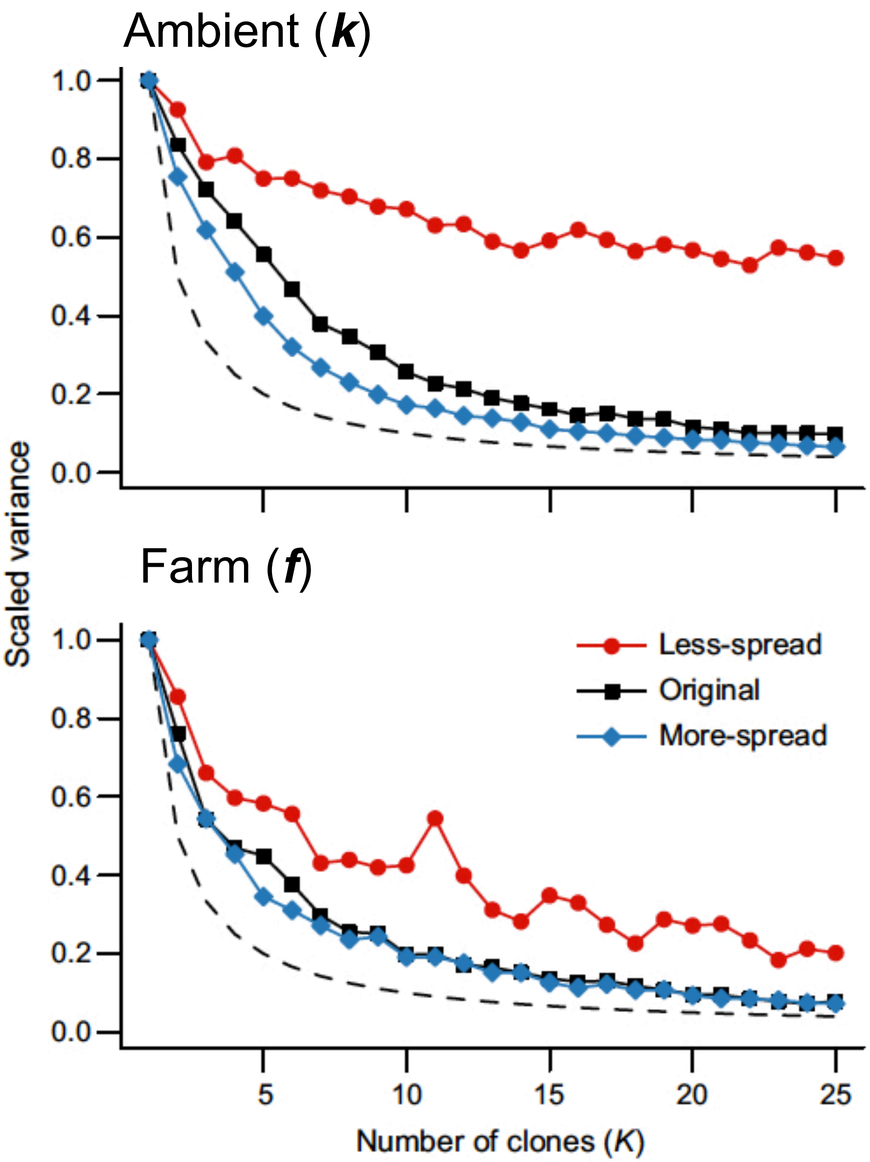


1. Original sample locations (2004)
2. Less spatial spread
3. More spatial spread

Peacock et al. Ecology & Evolution (2017) **7**:762–770

Estimability of ambient and farm sources of sea lice





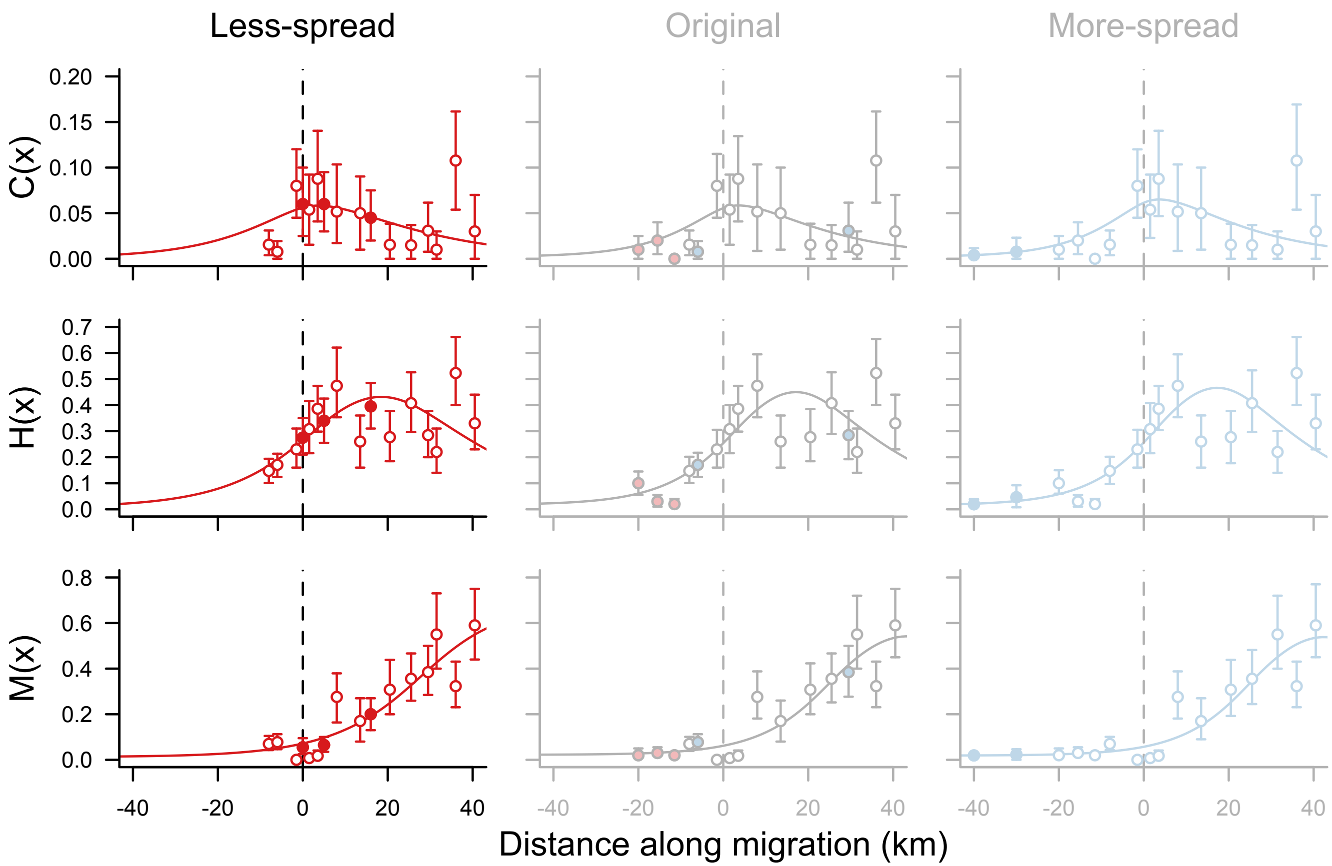
Peacock et al. Ecology & Evolution (2017) **7**:762–770

Fits to data

pattern ☺

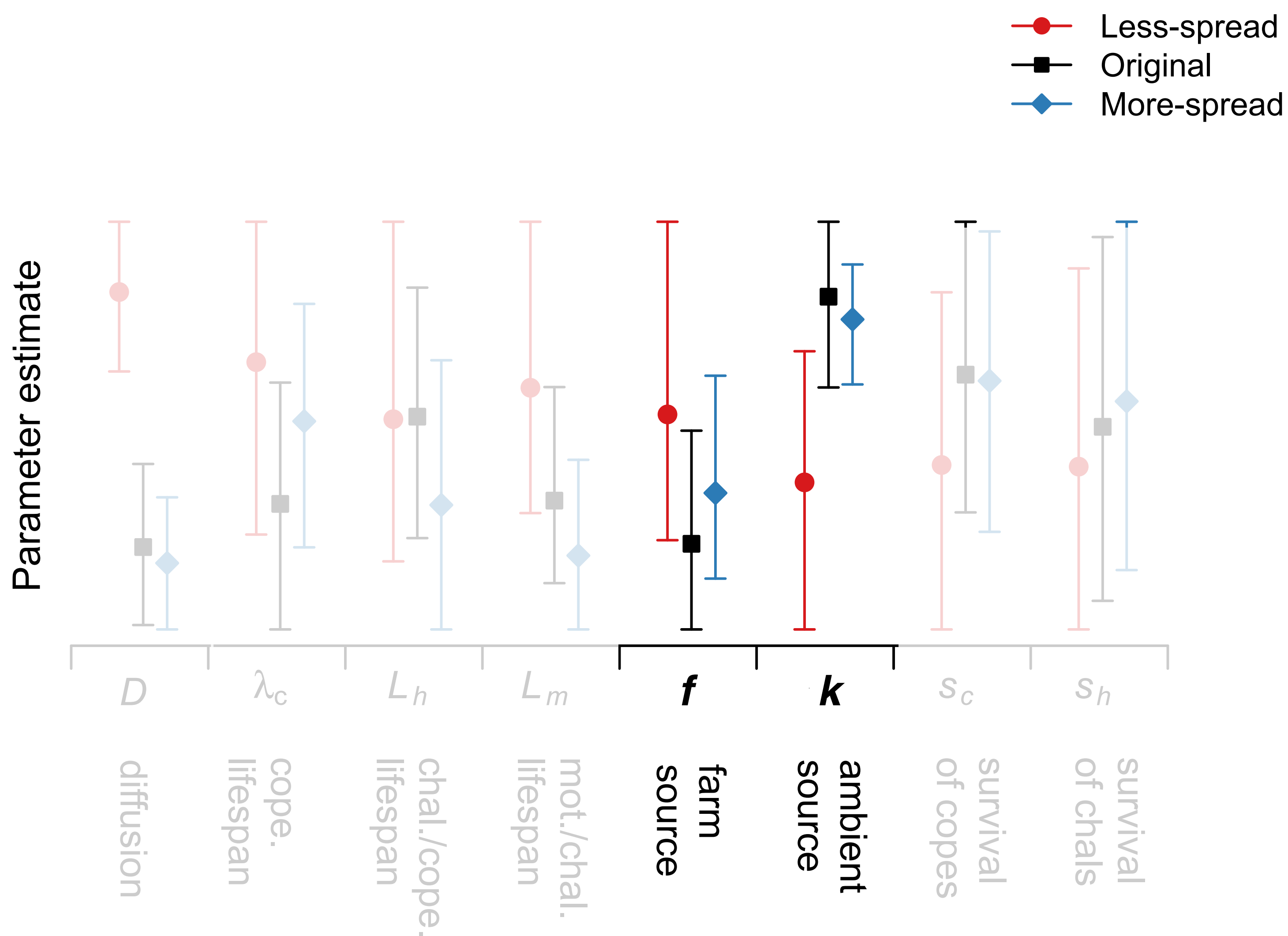
↑

process ☹



Peacock et al. Ecology & Evolution (2017) **7**:762–770

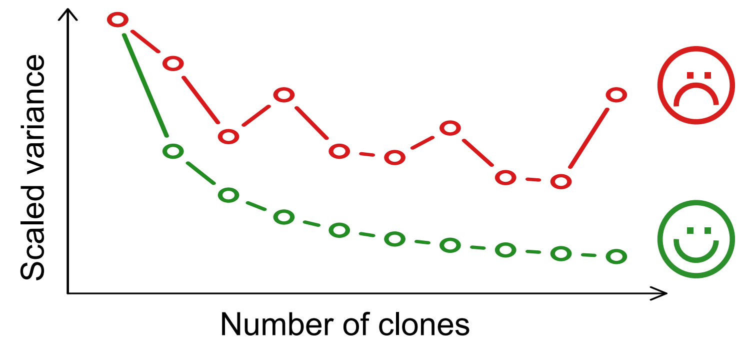
Parameter estimates



Peacock et al. Ecology & Evolution (2017) **7**:762–770

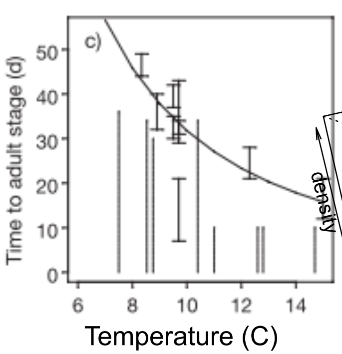
Lousy lessons learned

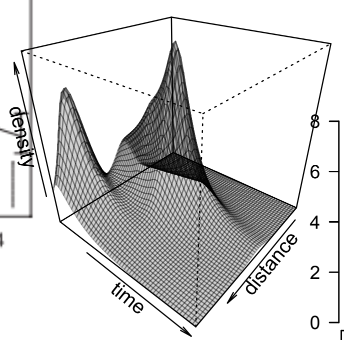
Check parameter estimability!

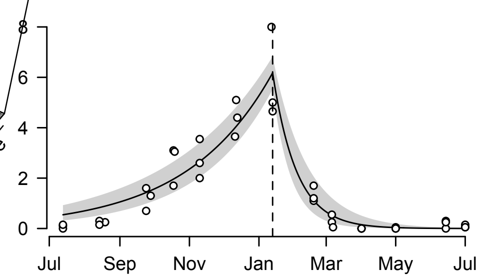


In our case estimability was fixable by:

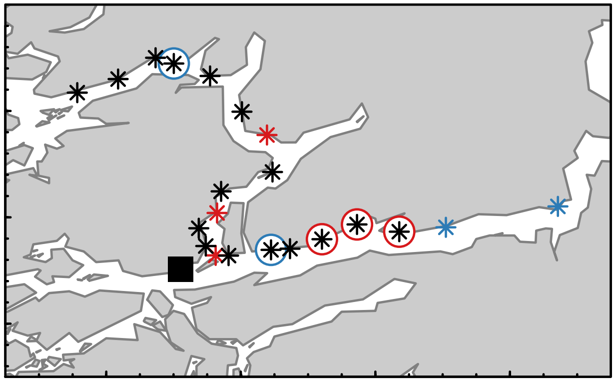
- ✓ Fixing parameters
- ✓ Revisiting model structure
- ✓ Collecting more/ different data*







OR **avoided** by investigating different spatial/temporal designs






Coauthors: Stephanie Peacock (Calgary), Subhash Lele (Alberta), Martin Krkosek (Toronto), Andrew Bateman (UVic)

Salmon Coast Field Station

pims

NSERC CRSNG

Alberta Innovates Technology Futures

Max Wyman - Term Assistant Professor - Statistics, University of Alberta

The Department of Mathematical and Statistical Sciences at the University of Alberta invites applications for a Max Wyman Term Assistant Professorship in the area of Statistics. This is a non-tenure-track, three-year fixed-term position. The position offers a stimulating research and teaching environment with a reduced teaching load.

https://www.careers.ualberta.ca/Competition/A113337476/