Chapter 5. Case study: Measles in large and small towns

Objectives

- To display a published case study using plug-and-play methods with non-trivial model complexities.
- To show how extra-demographic stochasticity can be modeled.
- To demonstrate the use of covariates in **pomp**.
- To demonstrate the use of profile likelihood in scientific inference.
- To discuss the interpretation of parameter estimates.
- To emphasize the potential need for extra sources of stochasticity in modeling.

Motivation: challenges in inference from disease dynamics

- Understanding, forecasting, managing epidemiological systems increasingly depends on models.
- Dynamic models can be used to test causal hypotheses.
- Real epidemiological systems:
  - are nonlinear
  - are stochastic
  - are nonstationary
  - evolve in continuous time
  - have hidden variables
  - can be measured only with (large) error
- Dynamics of infectious disease outbreaks illustrate this well.
- Measles is the paradigm for a nonlinear ecological system that can be well described by low-dimensional nonlinear dynamics.
- A tradition of careful modeling studies have proposed and found evidence for a number of specific mechanisms, including
  - a high value of R_0 (c. 15–20)
  - under-reporting
  - seasonality in transmission rates associated with school terms
  - response to changing birth rates

Outline

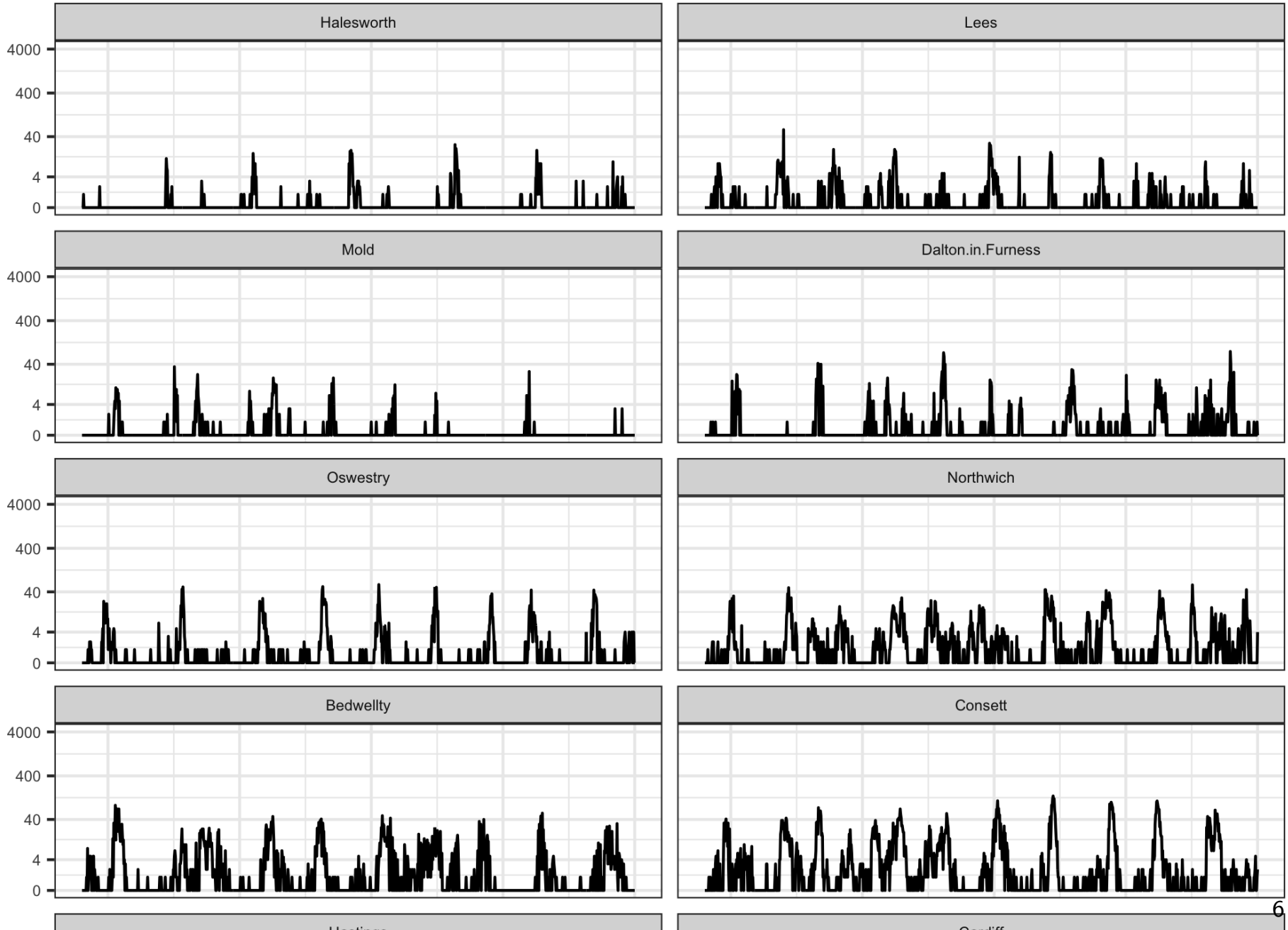
- We revisit a classic measles data set, weekly case reports in 954 urban centers in England and Wales during the pre-vaccine era (1950–1963).
- We examine questions regarding:
- measles extinction and recolonization
- transmission rates
- seasonality
- resupply of susceptibles
- We use a model that
  1. expresses our current understanding of measles dynamics
  2. includes a long list of mechanisms that have been proposed and demonstrated in the literature
  3. cannot be fit by existing likelihood-based methods
- We examine data from large and small towns using the same model, something no existing methods have been able to do.
- We ask: does our perspective on this disease change when we expect the models to explain the data in detail?
- What bigger lessons can we learn regarding inference for dynamical systems?

Data sets

- Twenty towns, including
  - 10 largest
  - 10 smaller, chosen at random
- Population sizes: 2k–3.4M
- Weekly case reports, 1950–1963
- Annual birth records and population sizes, 1944–1963

Map of cities in the analysis

City case counts



Continuous-time Markov process model

```
library(DiagrammeR)
DiagrammeR("digraph SEIR {
  graph [rankdir=TD, overlap=false, fontsize = 10]
  node[shape=egg, label='B'] b;
  subgraph {
    rank=same;
    node[shape=oval, label='S'] S;
    node[shape=oval, label='E'] E;
    node[shape=oval, label='I'] I;
    node[shape=oval, label='R'] R;
    S->E E->I I->R
  }
  node[shape=diamond, label='dead'] d;
  b->S
  {S E I R}->d
   }",type="grViz",engine="dot",height=300,width=800)
```

Continuous-time Markov process model

- Covariates:
  - $B(t) =$ birth rate, from data
  - $N(t) =$ population size, from data
- Entry into susceptible class:

$$\mu_{BS}(t) = (1-c)\,B(t-\tau) + c\,\delta(t-\lfloor t \rfloor)\int_{t-1}^{t} B(t-\tau-s)\,ds$$

  - $c =$ cohort effect
  - $\tau =$ school-entry delay
  - $\lfloor t \rfloor =$ most recent 1 September before t
- Force of infection:

$$\mu_{SE}(t) = \frac{\beta(t)}{N(t)}\,(I+\iota)\,\zeta(t)$$

  - $\iota =$ imported infections
  - $\zeta(t) =$ Gamma white noise with intensity σ_{SE} [@He2010;@bhadra11]
  - school-term transmission:

$$\beta(t) = \begin{cases} \beta_0\left(1+a(1-p)/p\right) & \text{during term} \\ \beta_0\,(1-a) & \text{during vacation} \end{cases}$$

Implementation in **pomp**

- We'll load the packages we'll need, and set the random seed, to allow reproducibility.
- Note that we'll be making heavy use of the **tidyverse** methods.
- Also, we'll be using **ggplot2** for plotting: see this brief tutorial.
- Finally, we'll use the convenient **magrittr** syntax, which is explained here.

```
library(pomp)
library(tidyverse)
theme_set(theme_bw())
set.seed(594709947L)
```

Data and covariates

- Now we'll load the data and covariates.The data are measles reports from 20 cities in England and Wales.
- We also have information on the population sizes and birth-rates in these cities; we'll treat these variables as covariates.

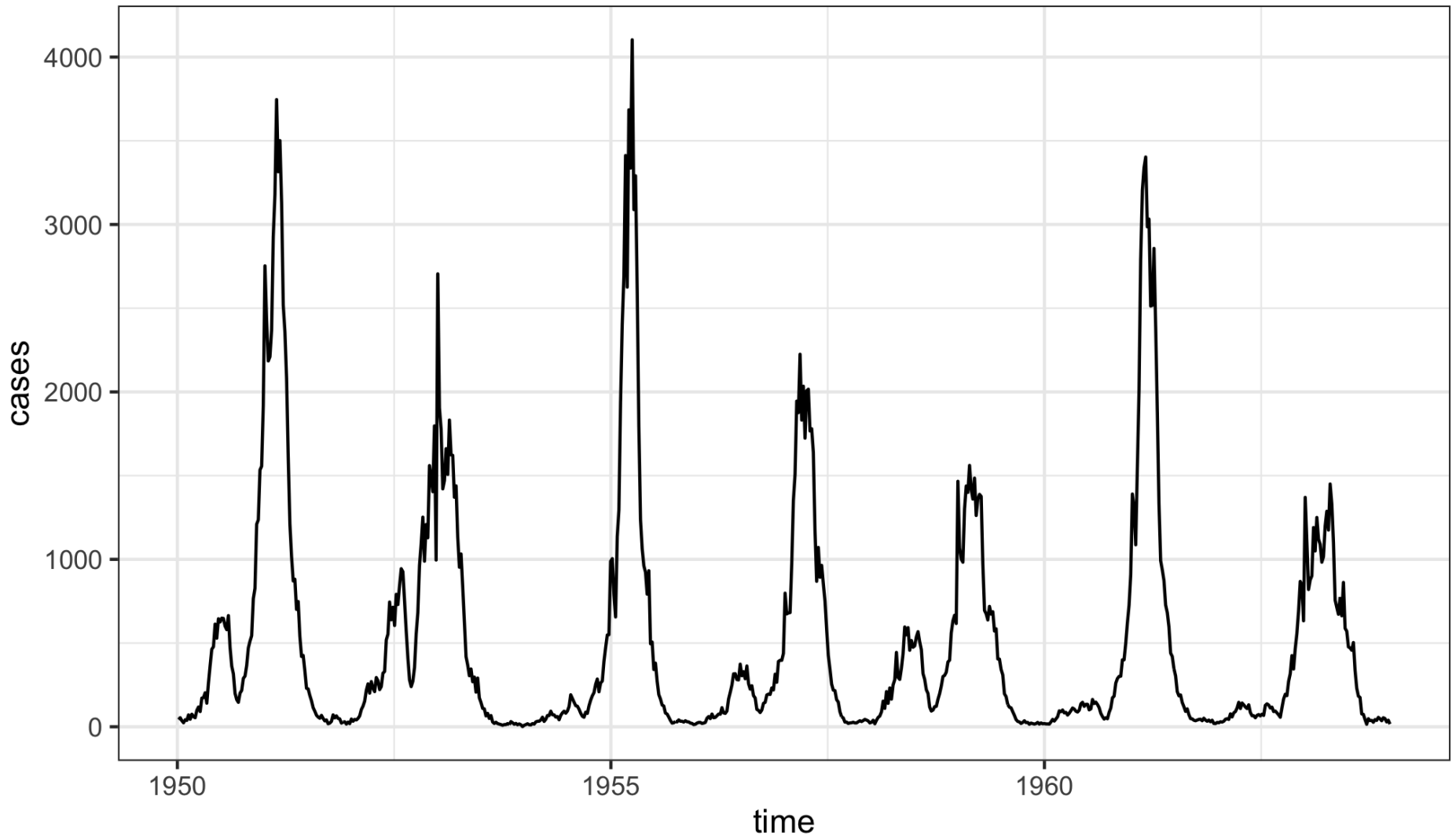
We select the data for London and pre-process the measles and demography data.

```
measles %>%
  mutate(year=as.integer(format(date,"%Y"))) %>%
  filter(town=="London" & year>=1950 & year<1964) %>%
  mutate(
    time=(julian(date,origin=as.Date("1950-01-01")))/365.25+1950
  ) %>%
  filter(time>1950 & time<1964) %>%
  select(time,cases) -> dat

demog %>%
  filter(town=="London") %>%
  select(-town) -> demogLondon
```

Data and covariate plots

```
dat %>% ggplot(aes(x=time,y=cases))+geom_line()
```



The partially observed Markov process model

We require a simulator for the BSEIR model. Notable complexities include:

1. Incorporation of the known birthrate.
2. The birth-cohort effect: a specified fraction ('cohort') of the cohort enter the susceptible pool all at once.
3. Seasonality in the transmission rate: during school terms, the transmission rate is higher than it is during holidays.
4. Extra-demographic stochasticity in the form of a Gamma white-noise term acting multiplicatively on the force of infection.
5. Demographic stochasticity implmented using Euler-multinomial distributions.

Implementation of the process model

```
rproc <- Csnippet([1379 chars quoted with '"'])
```

- In the above, C represents the true incidence, i.e., the number of new infections occurring over an interval.
- Since recognized measles infections are quarantined, we argue that most infection occurs before case recognition so that true incidence is a measure of the number of individuals progressing from the I to the R compartment in a given interval.

State initializations

We complete the process model definition by specifying the distribution of initial unobserved states. The following codes assume that the fraction of the population in each of the four compartments is known.

```
rinit <- Csnippet("
  double m = pop/(S_0+E_0+I_0+R_0);
  S = nearbyint(m*S_0);
  E = nearbyint(m*E_0);
  I = nearbyint(m*I_0);
  R = nearbyint(m*R_0);
  W = 0;
  C = 0;
")
```

The measurement model

- We'll model both under-reporting and measurement error.
- We want $\mathbb{E}[\text{cases}|C] = \rho\, C$, where C is the true incidence and $0 < \rho < 1$ is the reporting efficiency.
- We'll also assume that $\mathrm{Var}[\text{cases}|C] = \rho\,(1-\rho)\,C + (\psi\,\rho\,C)^2$, where ψ quantifies overdispersion.
- Note that when $\psi = 0$, the variance-mean relation is that of the binomial distribution. To be specific, we'll choose cases—C $\sim f(\cdot|\rho, \psi, C)$, where

  $f(c|\rho,\psi,C) = \Phi(c+\frac{1}{2}, \rho\,C, \rho\,(1-\rho)\,C+(\psi\,\rho\,C)^2) - \Phi(c-\frac{1}{2}, \rho\,C, \rho\,(1-\rho)\,C$

  where $\Phi(x, \mu, \sigma^2)$ is the c.d.f. of the normal distribution with mean μ and variance σ^2.

The following computes $\mathbb{P}[\text{cases}|C]$.

```
dmeas <- Csnippet("
  double m = rho*C;
  double v = m*(1.0-rho+psi*psi*m);
  double tol = 1.0e-18;
  if (cases > 0.0) {
```

Case simulations

The following codes simulate cases$|C$.

```
rmeas <- Csnippet("
  double m = rho*C;
  double v = m*(1.0-rho+psi*psi*m);
  double tol = 1.0e-18;
  cases = rnorm(m,sqrt(v)+tol);
  if (cases > 0.0) {
    cases = nearbyint(cases);
  } else {
    cases = 0.0;
  }
")
```

Constructing the ‘pomp‘ object

```
dat %>%
  pomp(t0=with(dat,2*time[1]-time[2]),
    time="time",
    rprocess=euler(rproc,delta.t=1/365.25),
    rinit=rinit,
    dmeasure=dmeas,
    rmeasure=rmeas,
    covar=covariate_table(covar,times="time"),
    accumvars=c("C","W"),
    statenames=c("S","E","I","R","C","W"),
    paramnames=c("R0","mu","sigma","gamma","alpha","iota",
      "rho","sigmaSE","psi","cohort","amplitude",
      "S_0","E_0","I_0","R_0")
  ) -> m1
```

@He2010 estimated the parameters of this model. The full set is included in the **R** code accompanying this document, where they are read into a data frame called 'mles'.
We verify that we get the same likelihood as @He2010. Change eval to T

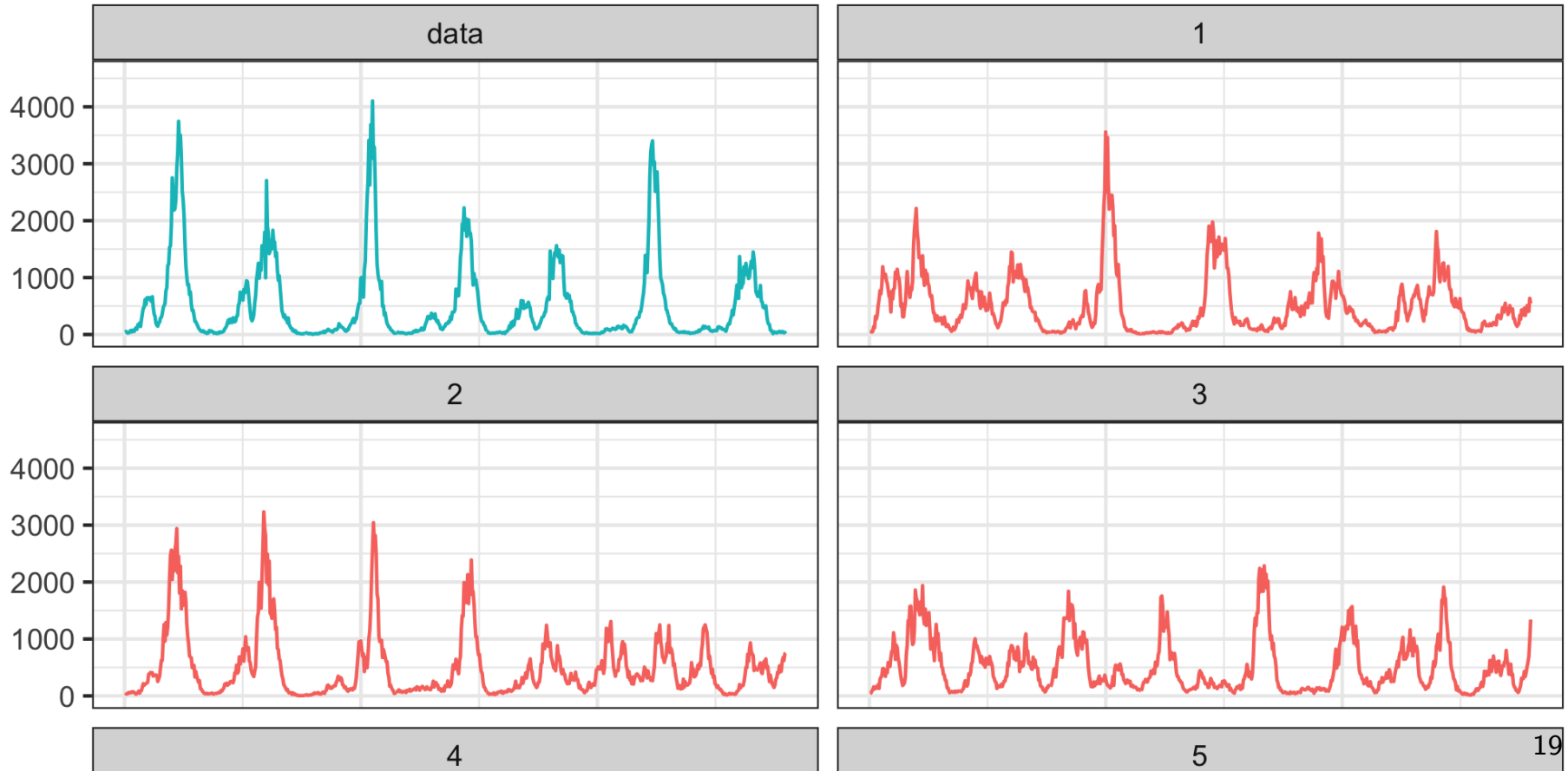
```
library(foreach)
library(doParallel)
library(doRNG)

registerDoParallel()
registerDoRNG(998468235L)

foreach(i=1:4) %dopar% {
  library(pomp)
  pfilter(m1,Np=10000,params=theta)
} -> pfs
logmeanexp(sapply(pfs,logLik),se=TRUE)
```

Simulations at the MLE

```
m1 %>%
  simulate(params=theta,nsim=9,format="d",include.data=TRUE) %>%
  ggplot(aes(x=time,y=cases,group=.id,color=(.id=="data")))+
  guides(color=FALSE)+
  geom_line()+facet_wrap(~.id,ncol=2)
```



Parameter transformations

- The parameters are constrained to be positive, and some of them are constrained to lie between 0 and 1.
- We can turn the likelihood maximization problem into an unconstrained maximization problem by transforming the parameters.
- Specifically, to enforce positivity, we log transform, to constrain parameters to the unit interval, we logit transform, and to confine parameters to the unit simplex, we use the log-barycentric transformation.

```
pt <- parameter_trans(
  log=c("sigma","gamma","sigmaSE","psi","R0"),
  logit=c("cohort","amplitude"),
  barycentric=c("S_0","E_0","I_0","R_0")
)

m1 %>%
  pomp(partrans=pt,
    statenames=c("S","E","I","R","C","W"),
    paramnames=c("R0","mu","sigma","gamma","alpha","iota",
```

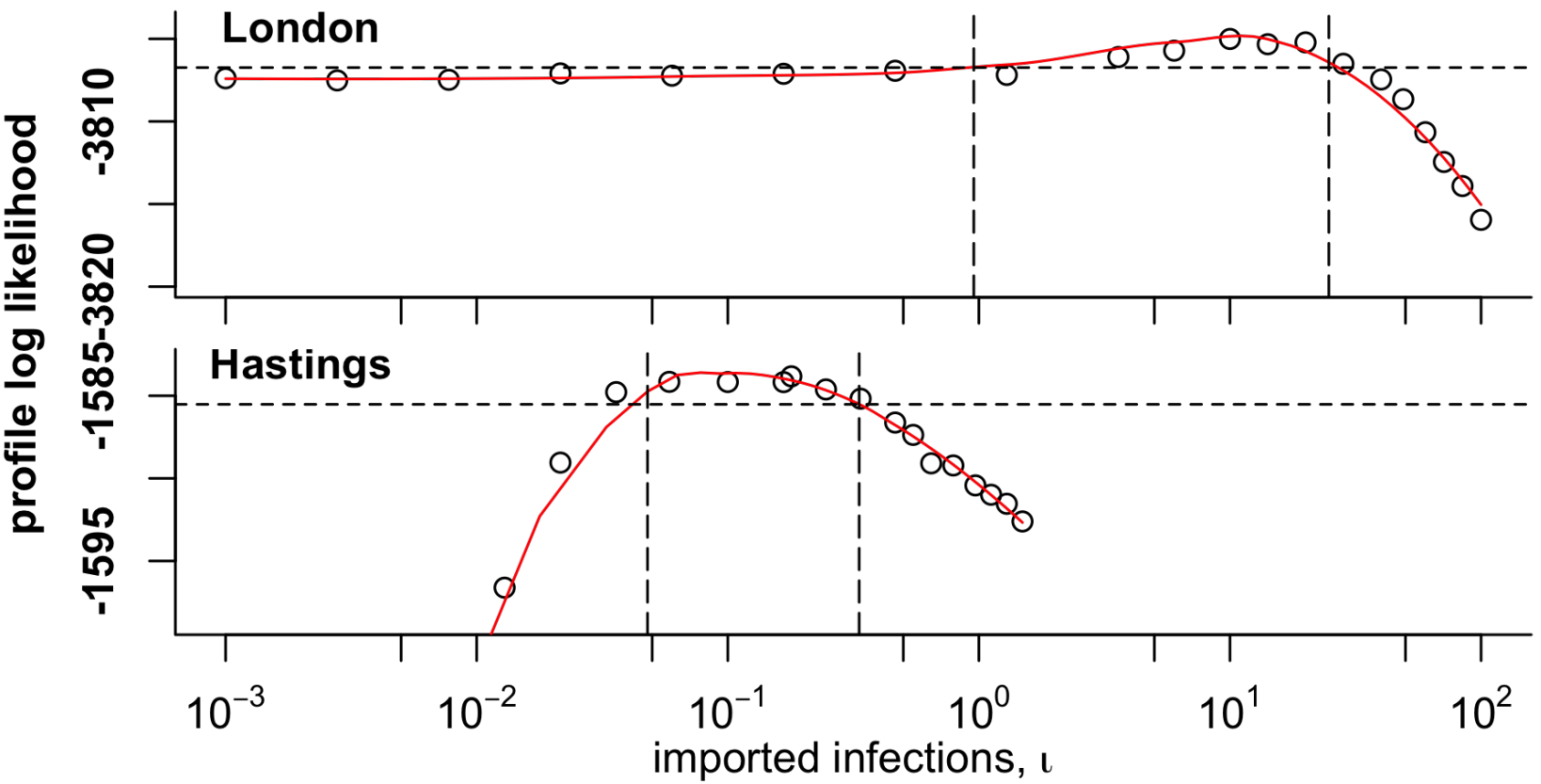
Results from @He2010

The linked document shows how a likelihood profile can be constructed using IF2 The fitting procedure used is as follows:

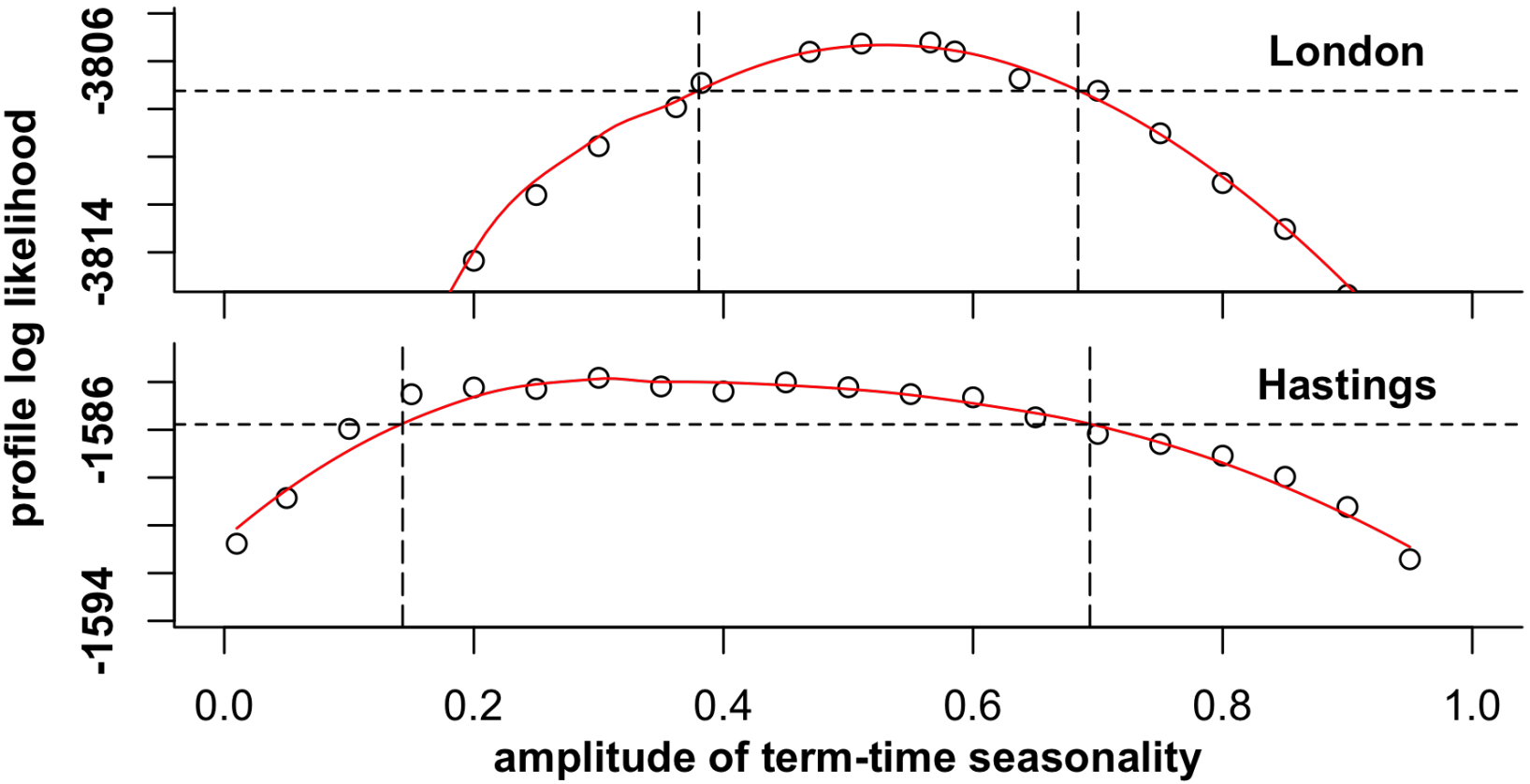
- A large number of searches were started at points across the parameter space.
- Iterated filtering was used to maximize the likelihood.
- We obtained point estimates of all parameters for 20 cities.
- We constructed profile likelihoods to quantify uncertainty in London and Hastings.

Imported infections

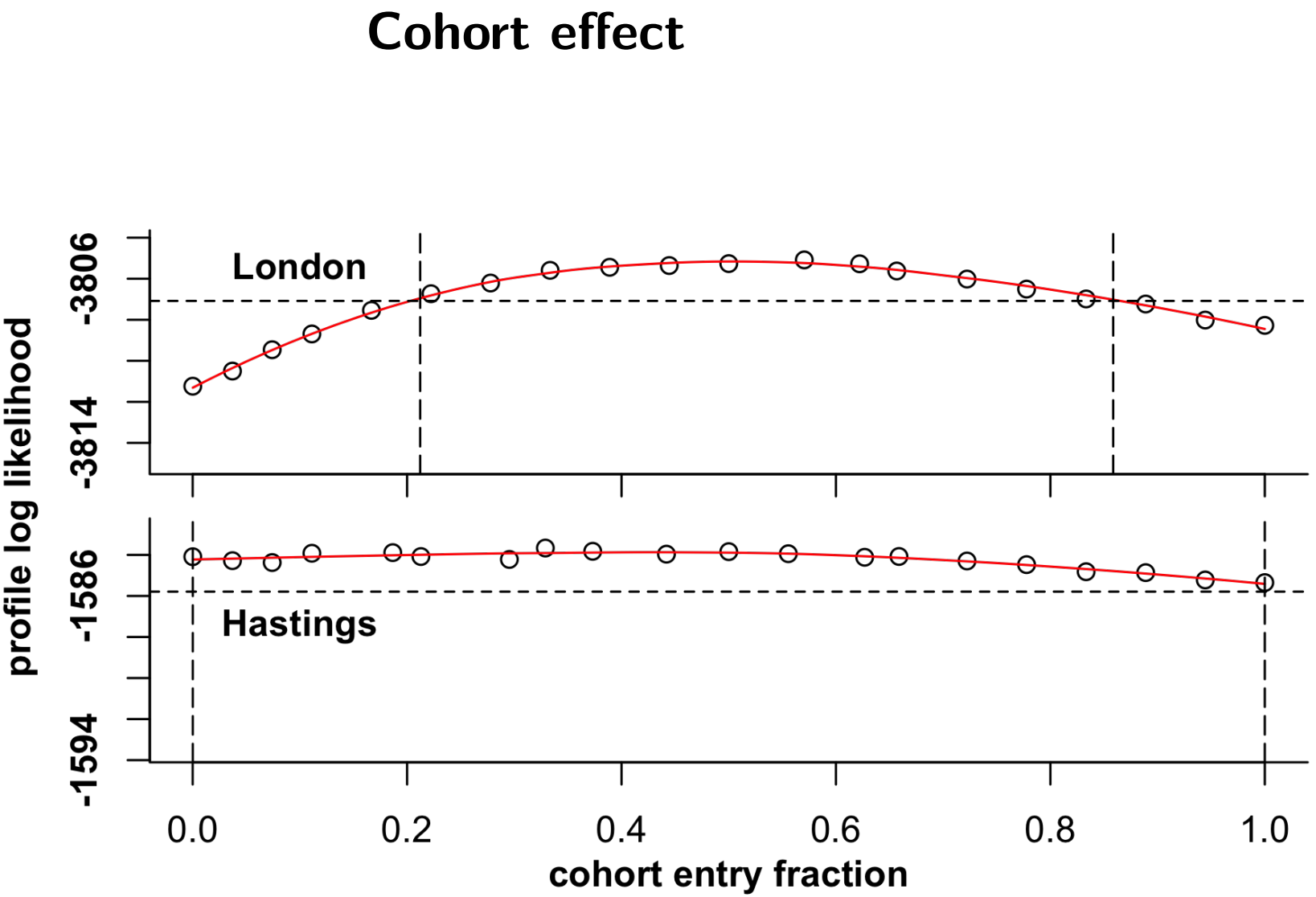
$$\text{force of infection} = \mu_{SE} = \frac{\beta(t)}{N(t)}\,(I+\iota)\,\zeta(t)$$



Seasonality



Notable findings



Notable findings

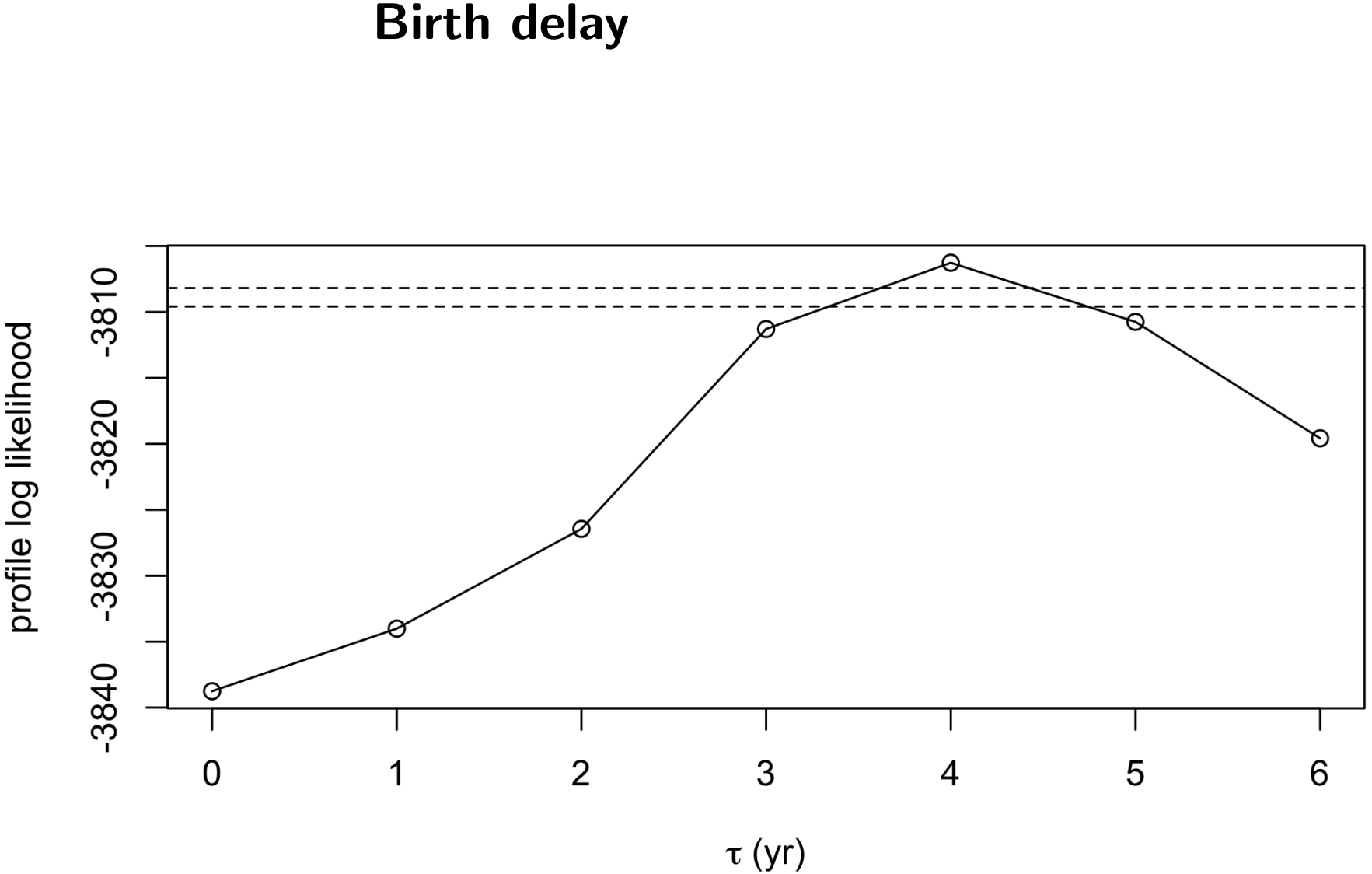
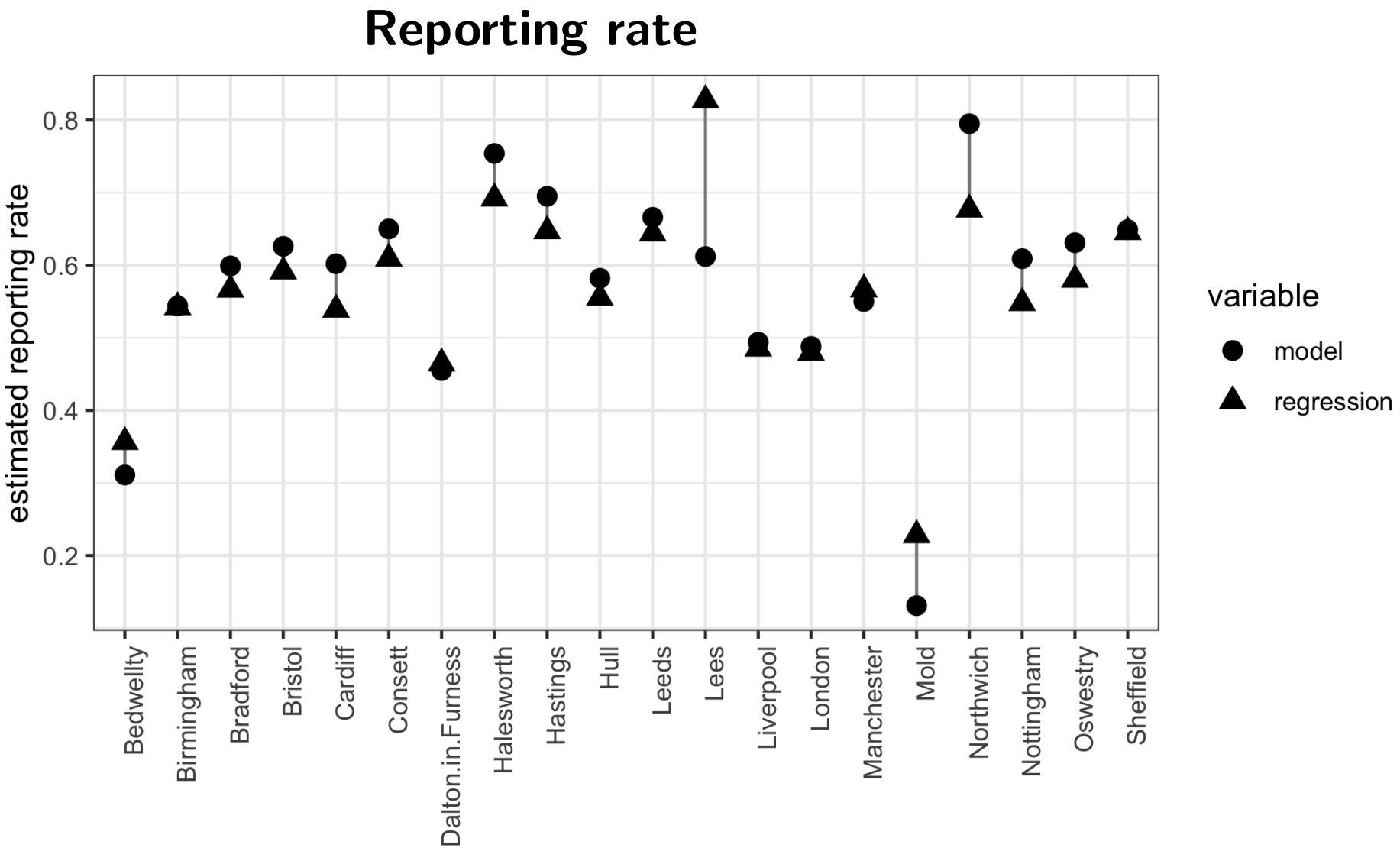
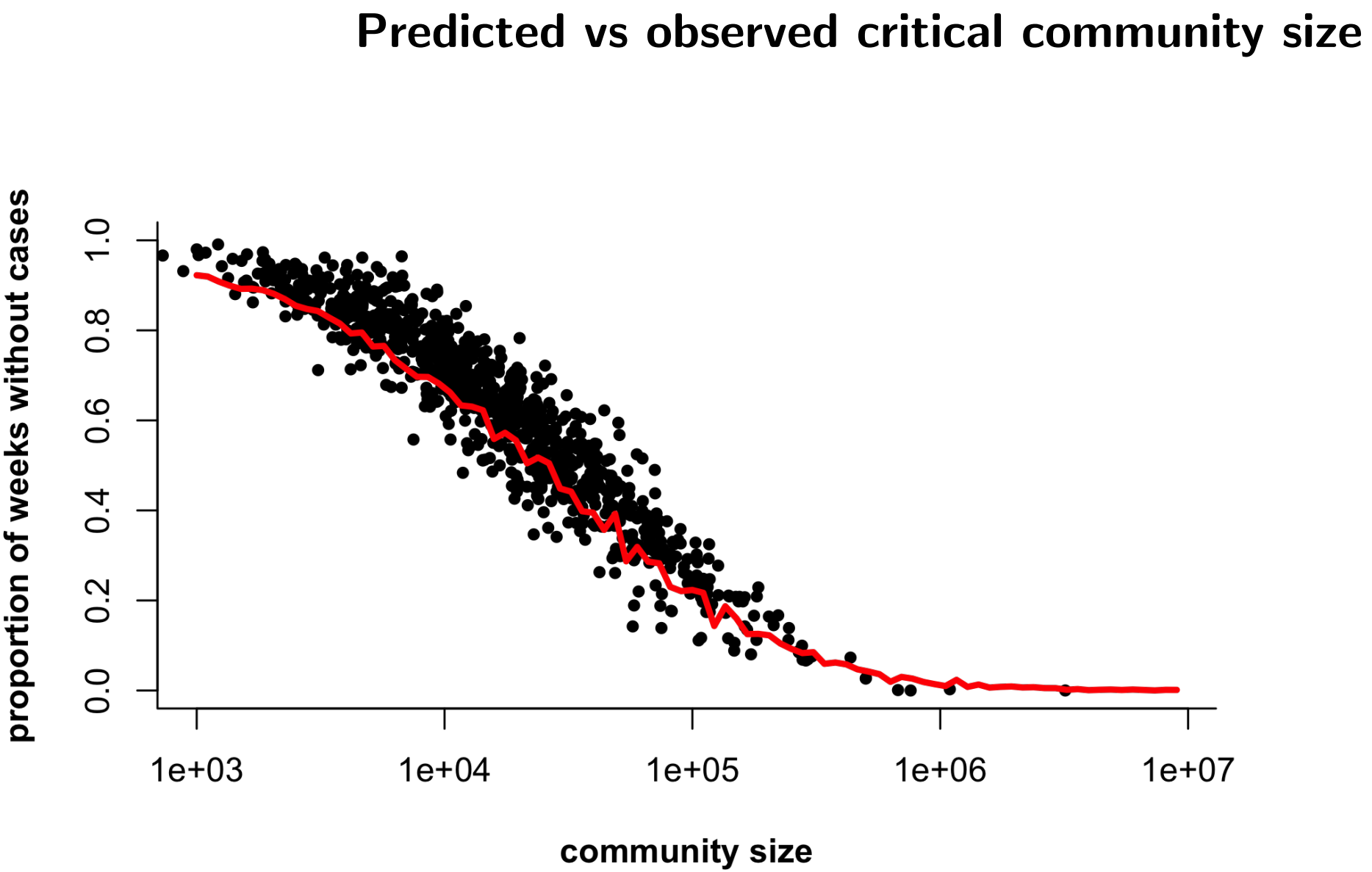


Figure: Profile likelihood for birth-cohort delay, showing 95% and 99% critical

Notable findings



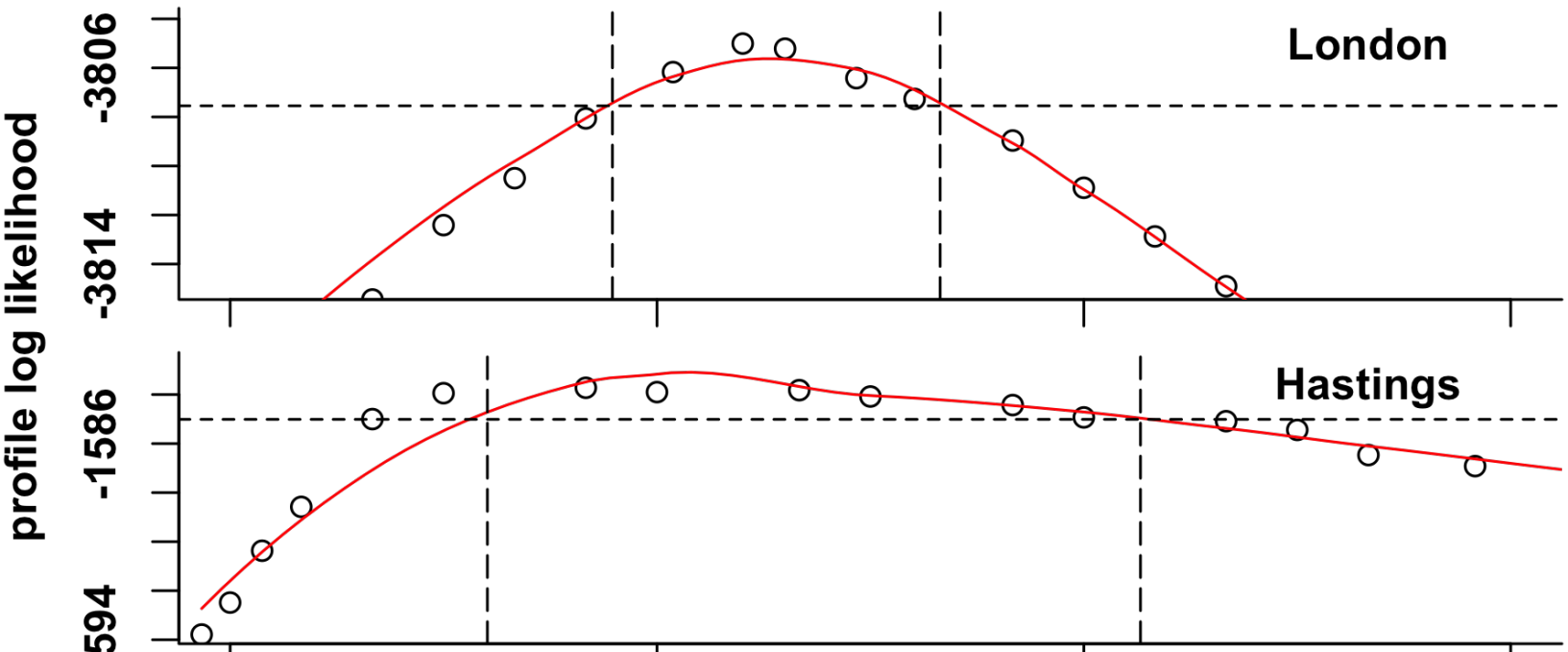
Notable findings



Problematic results

$$R_0$$

- Recall that R_0 is the basic reproduction number: a measure of how communicable an infection is.
- Existing estimates of R_0 (c. 15–20) come from two sources:
  - serology surveys
  - models fit to data using feature-based methods



Problematic results

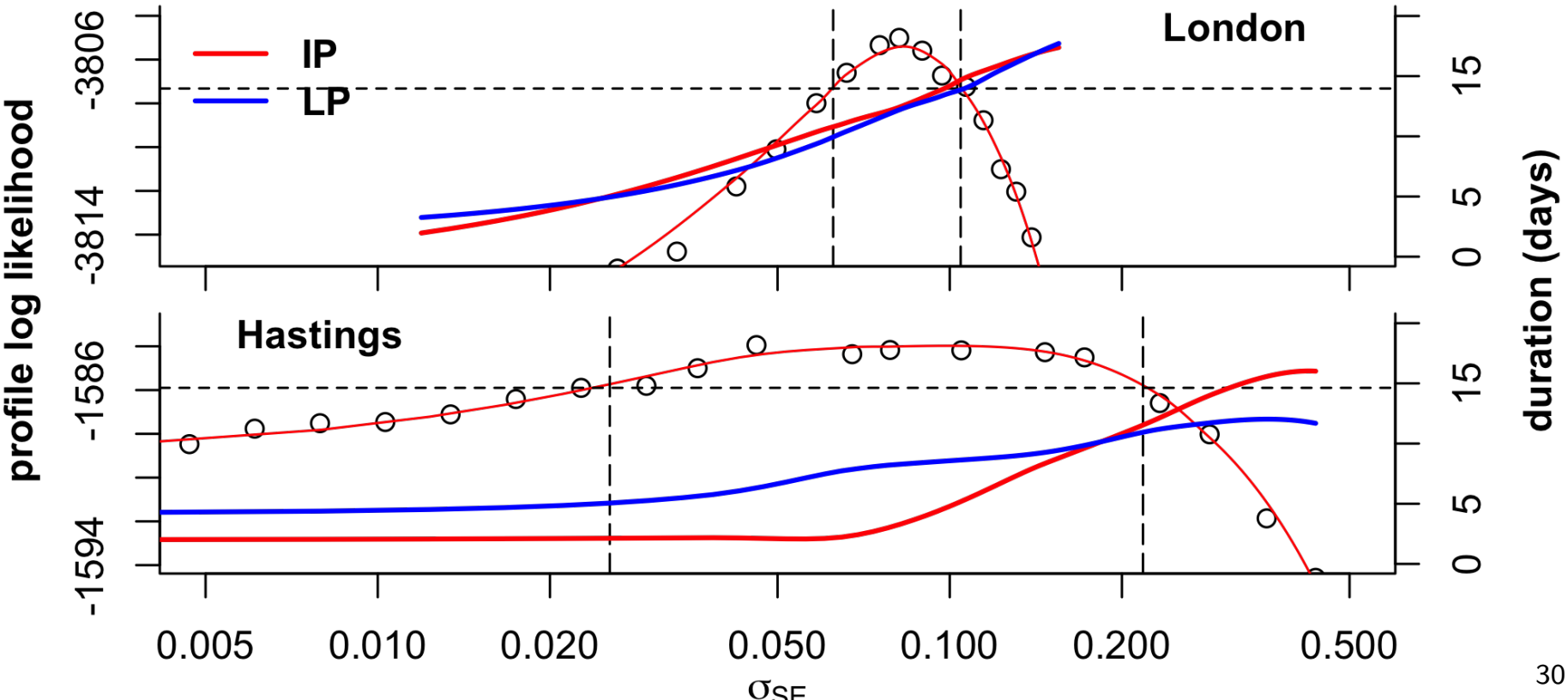
Parameter estimates

	pop	R0	amplitude	LP	IP
Halesworth	2170	33.100	0.381	7.870	2.290
Lees	4250	29.700	0.153	8.510	2.050
Mold	6410	21.400	0.271	5.930	1.780
Dalton.in.Furness	10600	28.300	0.203	5.480	1.980
Oswestry	11000	52.900	0.339	10.300	2.710
Northwich	18300	30.100	0.423	8.510	3.020
Bedwellty	28900	24.700	0.160	6.820	3.030
Consett	39100	35.900	0.200	9.080	2.660
Hastings	65700	34.200	0.299	7.000	5.440
Cardiff	245000	34.400	0.223	9.870	3.090
Bradford	294000	32.100	0.236	8.510	3.360
Hull	302000	38.900	0.221	9.180	5.460
Nottingham	307000	22.600	0.157	5.720	3.700
Bristol	443000	26.800	0.203	6.190	4.940
Leeds	510000	47.800	0.267	9.480	10.900
Sheffield	515000	33.100	0.313	7.230	6.380
Manchester	704000	32.900	0.290	11.100	6.940

Problematic results

Extrademographic stochasticity

$$\mu_{SE} = \frac{\beta(t)}{N(t)} (I + \iota) \zeta(t)$$



Questions

- What does it mean that parameter estimates from the fitting disagree with estimates from other data?
- How can one interpret the correlation between infectious period and city size in the parameter estimates?
- How do we interpret the need for extrademographic stochasticity in this model?

Simulations at the MLE

```
m1 %>%
  simulate(params=theta,nsim=100,format="d",include.data=TRUE) %>%
  select(time,.id,cases) -> simdat

simdat %>%
  mutate(data=.id=="data") %>%
  plyr::ddply(~time+data,plyr::summarize,
    p=c(0.05,0.5,0.95),
    q=quantile(cases,prob=p,names=FALSE)
  ) %>%
  mutate(p=plyr::mapvalues(p,from=c(0.05,0.5,0.95),to=c("lo","med",
         data=plyr::mapvalues(data,from=c(TRUE,FALSE),to=c("data","
  spread(p,q) %>%
  ggplot(aes(x=time,y=med,color=data,fill=data,ymin=lo,ymax=hi))+
  geom_ribbon(alpha=0.2)+
  guides(data=FALSE)
```

Exercises

Exercise 1: Reformulate the model

- Modify the @He2010 model to remove the cohort effect. Run simulations and compute likelihoods to convince yourself that the resulting codes agree with the original ones for 'cohort $= 0$'.
- Now modify the transmission seasonality to use a sinusoidal form. How many parameters must you use? Fixing the other parameters at their MLE values, compute and visualize a profile likelihood over these parameters.

Exercise 2: Extrademographic stochasticity

Set the extrademographic stochasticity parameter $\sigma_{SE} = 0$, set $\alpha = 1$, and fix ρ and ι at their MLE values, then maximize the likelihood over the remaining parameters.

- How do your results compare with those at the MLE? Compare likelihoods but also use simulations to diagnose differences between the models.

Objectives

- To display a published case study using plug-and-play methods with non-trivial model complexities.
- To show how extra-demographic stochasticity can be modeled.
- To demonstrate the use of covariates in **pomp**.
- To demonstrate the use of profile likelihood in scientific inference.
- To discuss the interpretation of parameter estimates.
- To emphasize the potential need for extra sources of stochasticity in modeling.