Estimate RSV onset and peak timing

Tutorial 2 for transition

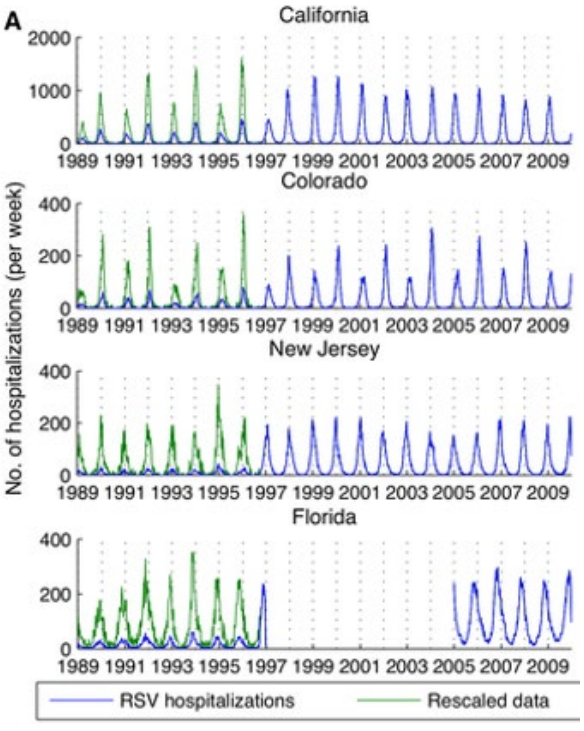
Gigi (Zhe Zheng)[1] Dan (Daniel Weinberger) Ginny (Virginia Pitzer)

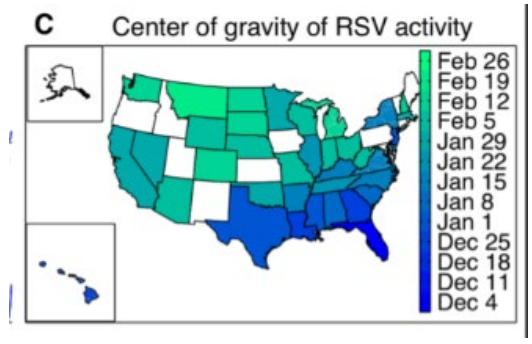
2023-01-04

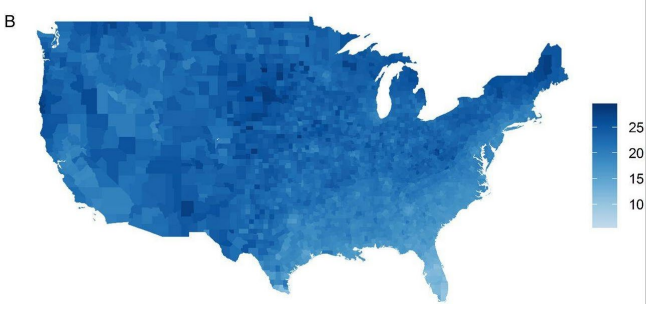
[1]zhe.zheng@yale.edu; zhe.zheng@aya.yale.edu; gigi.zhe.zheng@gmail.com

Background

Before the COVID-19 pandemic, the timing of RSV seasonal epidemics exhibited notable spatial patterns







Virginia E. Pitzer, 2015, PLoS Pathog
Daniel Weinberger, 2015, CID

Outline

- In section 1, we will first introduce how to find the peak timing of RSV epidemics using harmonic regression (given regular annual/biennial seasonality). We will learn how to use R to identify peak timing of periodic RSV epidemics.
- In section 2, we will then talk about identifying the onset of RSV epidemics using second derivative method, regardless of the seasonality of RSV.
- In section 3, we will incorporate the spatial component of RSV epidemics. We will first introduce the concept of spatial autocorrelation and then lean to use R to account for spatial autocorrelation.

References

Relevant readings:

- RSV onset timing at county level

- RSV peak timing on state level

- Comparing RSV onset timing before and during the COVID-19 pandemic

- RSV peak timing at ZIP code level and the drivers of RSV spread

- Assessment and optimization of respiratory syncytial virus prophylaxis in Connecticut, 1996–2013

- Disease outbreak outcome estimation using penalized splines

Section 1: Harmonic regression to estimate the peak timing of RSV epidemics

Note: Most of the following materials came from Dan and Ginny's Lecture notes and Harmonic Regression by NCSS[2].

Please check out:

- Dan's class: Public Health Surveillance
- Ginny's class: Quantitative Methods in Infectious Diseases

$$X_t = \mu + R\cos(2\pi ft - d) + e_t$$

Where
X_t is the time-series contains a periodic (cyclic) component.
μ is mean of the series.
R is the amplitude of seasonality.
$f = \frac{1}{period}$ is the frequency of the periodic.
d is the phase or horizontal offset.
e_t is the random error (noise) of the series.
t is the time step

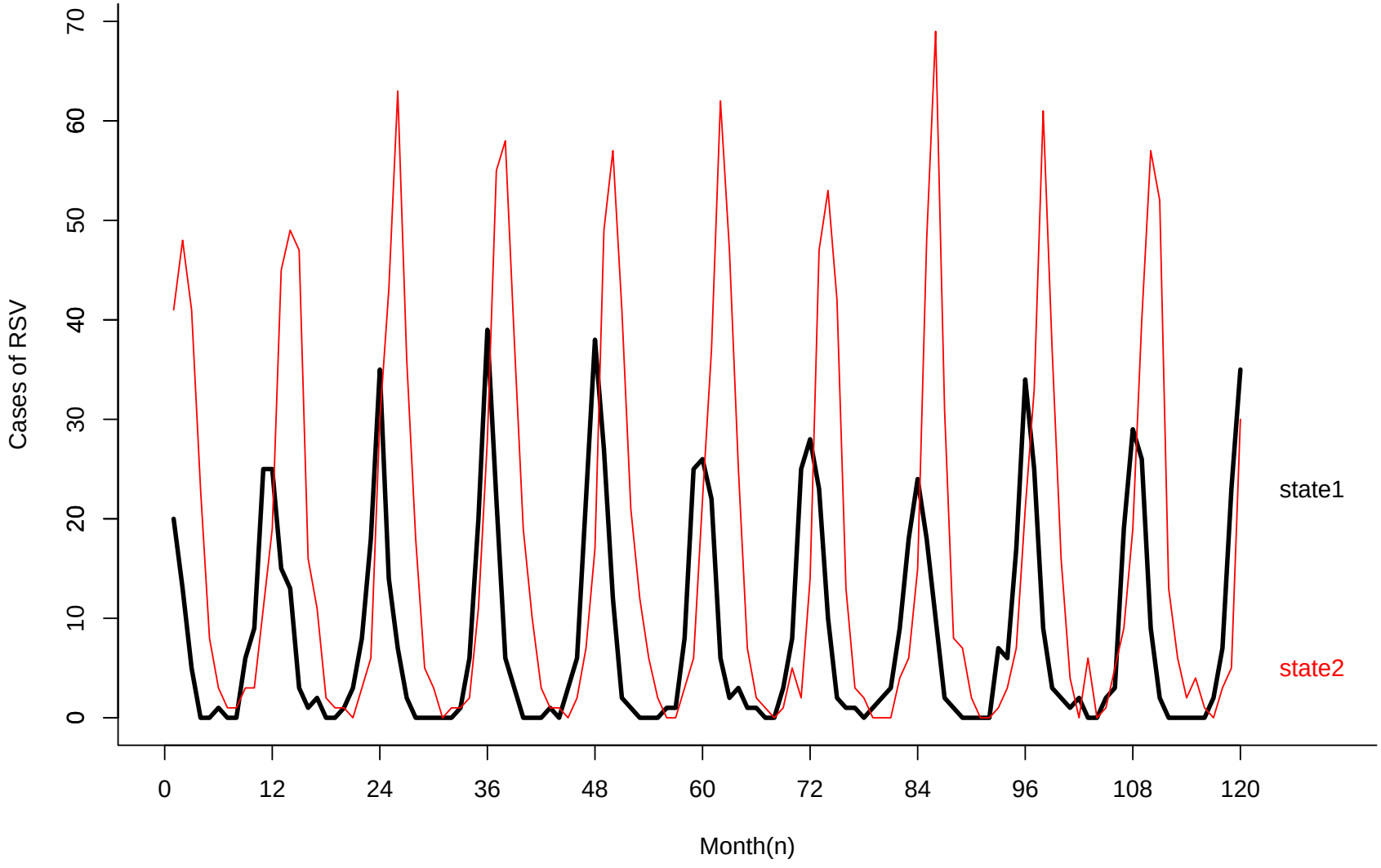
[2] https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Harmonic_Regression.pdf

Pseudo-RSV data: Simulate time series with a 12 month period

Imagine this is RSV case data from 2 states, and we want to investigate the epidemic characteristics in these states and the lag between states.

```
set.seed(123)
n=120 # 10 years
t <- seq(1,n)
amp1=2.5 # high amplitude
freq=1/12 # frequency = 1/period
amp2=2 # low amplitude

xt1a=amp1*cos(2*3.14159*t*freq)

#other series shifted by 2 months
xt2a=amp2*cos(2*3.14159*t*freq-1)

#Simulate some poisson count data
xt1=rpois(n,exp((1+xt1a)))
xt2=rpois(n,exp(2+xt2a))
```

The observed pseudo-RSV cases over time in two states



Investigate the epidemic characteristics of the pseudo-RSV time-serieses

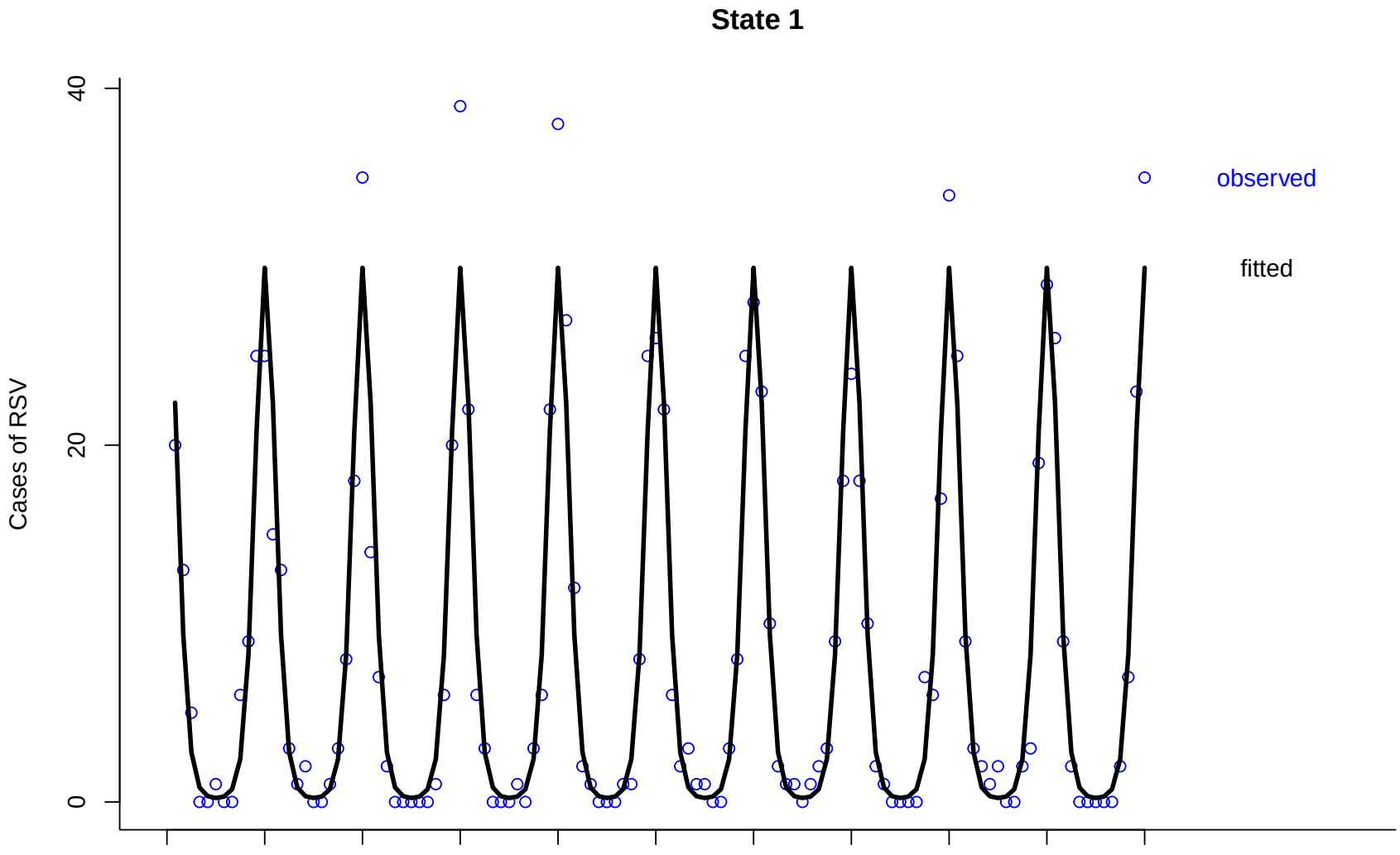
This is based on the prior knowledge that RSV has annual cycle in temperate regions and biannual cycle in high latitude regions. For other viruses or RSV circulation in other climate, you should consider using wavelet analysis to identify the periodicity first.

```
# Create the needed harmonic variables with
# 6 month, 12 month, and 24 month periodicities
t<-1:120
#Create harmonic variables
sin6=sin(2*pi*t/6)
cos6=cos(2*pi*t/6)
sin12=sin(2*pi*t/12)
cos12=cos(2*pi*t/12)
sin24=sin(2*pi*t/24)
cos24=cos(2*pi*t/24)
```

When you write the equations with sin and cos terms, you will need to include them both in an equation. For example, if the coefficient of cos12 is significant, you will need to include sin12 even though the coefficient is not significant in the summary.

Fit a simple poisson regression with 12 month period for state 1

```
fit1a <- glm(xt1~sin12+cos12, family='poisson')
pred1a<- fitted(fit1a)
#summary(fit1a)
```



Add in 24 month periodicity

```
fit2a <- glm(xt1~sin12+cos12+sin24+cos24,family='poisson')
pred2a<- fitted(fit2a)
summary(fit2a)
```

```
##
## Call:
## glm(formula = xt1 ~ sin12 + cos12 + sin24 + cos24, family = "poisson")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1406  -0.7825  -0.2894   0.3672   2.3956
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.94989    0.07873  12.065   <2e-16 ***
## sin12        0.07283    0.05737   1.269    0.204
## cos12        2.44820    0.09474  25.840   <2e-16 ***
## sin24        0.04088    0.09204   0.444    0.657
## cos24        0.04591    0.03407   1.348    0.178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1495.23  on 119  degrees of freedom
## Residual deviance:  101.72  on 115  degrees of freedom
## AIC: 436.78
##
## Number of Fisher Scoring iterations: 5
```

Add in 6 month periodicity

```
fit3a<-glm(xt1~sin12+cos12+sin24+cos24+sin6+cos6,family='poisson' )
pred3a<-fitted(fit3a)
summary(fit3a)
```

```
##
## Call:
## glm(formula = xt1 ~ sin12 + cos12 + sin24 + cos24 + sin6 + cos6,
##     family = "poisson")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0585  -0.8416  -0.2304   0.4924   2.5479
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.00529    0.09101  11.045   <2e-16 ***
## sin12        0.11920    0.09601   1.242    0.214
## cos12        2.32944    0.13971  16.673   <2e-16 ***
## sin24        0.04088    0.09204   0.444    0.657
## cos24        0.04591    0.03407   1.348    0.178
## sin6        -0.04481    0.07659  -0.585    0.558
## cos6         0.08943    0.07949   1.125    0.261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1495.23  on 119  degrees of freedom
## Residual deviance:  100.13  on 113  degrees of freedom
## AIC: 439.19
##
## Number of Fisher Scoring iterations: 5
```

Determine best model with AIC (smaller=better)

Winner is Model 1 (12 period). Also, when models have similar AIC score (within 2 points), we prefer a simpler model.

```
 AIC(fit1a)
```

```
## [1] 434.8372
```

```
 AIC(fit2a)
```

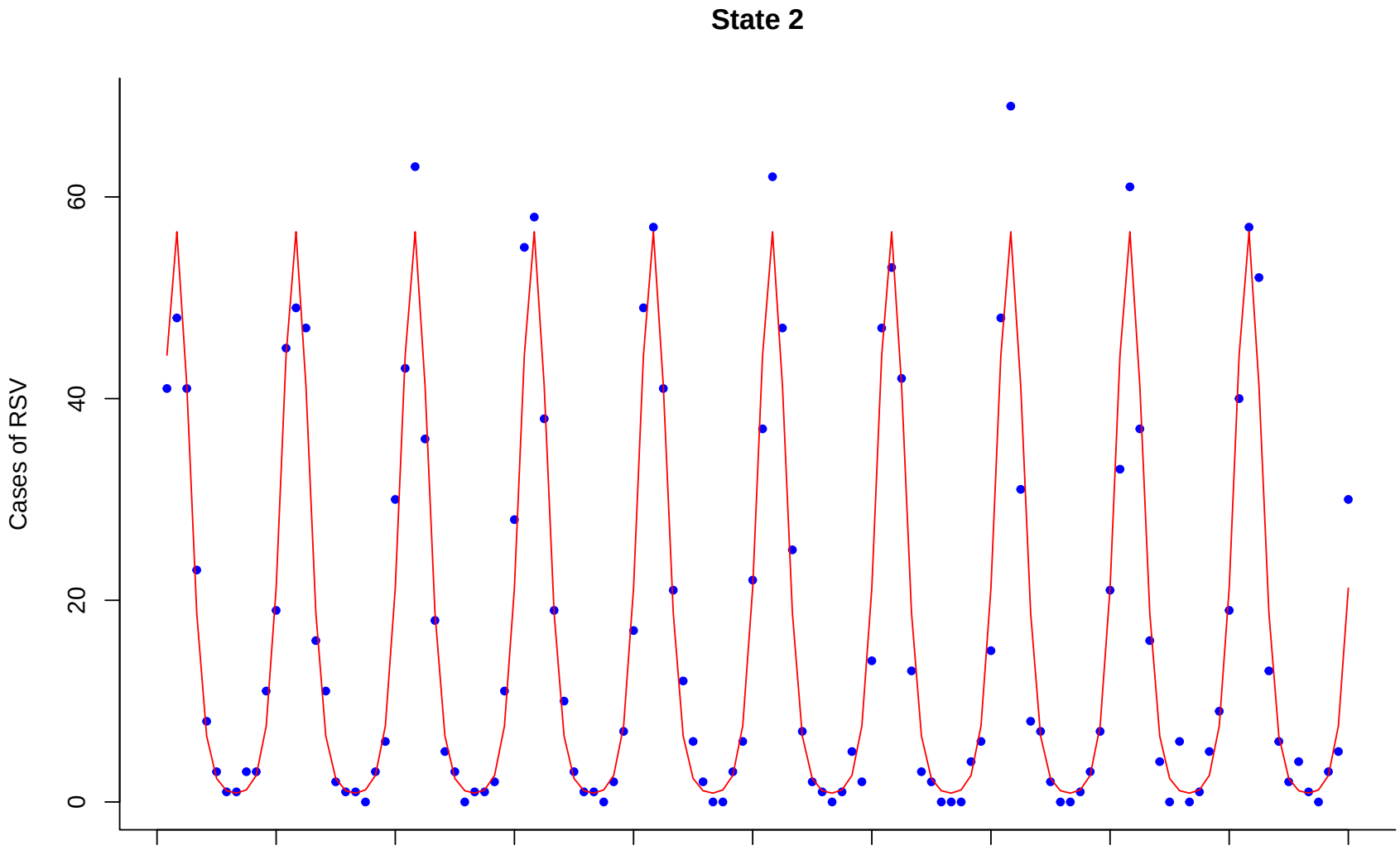
```
## [1] 436.7778
```

```
 AIC(fit3a)
```

```
## [1] 439.1851
```

Fit a simple poisson regression with 12 month period for state 2

```
fit1b <- glm(xt2~sin12+cos12, family='poisson')
pred1b<- fitted(fit1b)
#summary(fit1a)
```



Calculate amplitudes of the 12 periods in state1 and state2

```
beta_sin12_xt1<-coef(fit2a)['sin12']
beta_cos12_xt1<-coef(fit2a)['cos12']

amp12_xt1<-sqrt(beta_sin12_xt1^2+
                  beta_cos12_xt1^2)

amp12_xt1 #True value=2.5
```

```
##    sin12
## 2.449281
```

```
beta_sin12_xt2<-coef(fit1b)['sin12']
beta_cos12_xt2<-coef(fit1b)['cos12']

amp12_xt2<-sqrt(beta_sin12_xt2^2+
                  beta_cos12_xt2^2)

amp12_xt2 #True value=2
```

```
##    sin12
## 2.086493
```

Calculate phase of the 12 month period in state1 and state2

```
# Phase angle
phase12_xt1 <- -atan(beta_sin12_xt1/beta_cos12_xt1)
phase12_xt1
```

```
##       sin12
## -0.02973933
```

```
# True value = 0
phase12_xt2 <- -atan(beta_sin12_xt2/beta_cos12_xt2)
phase12_xt2
```

```
##     sin12
## -1.013021
```

```
# True value = 1
```

Calculate peak timing in month in state1 and state2

```
# Average peak timing
12*(1-phase12_xt1/(2*pi)) # True value = 0 month
```

```
##   sin12
## 12.0568
```

```
12*(1-phase12_xt2/(2*pi)) # True value = 2 month
```

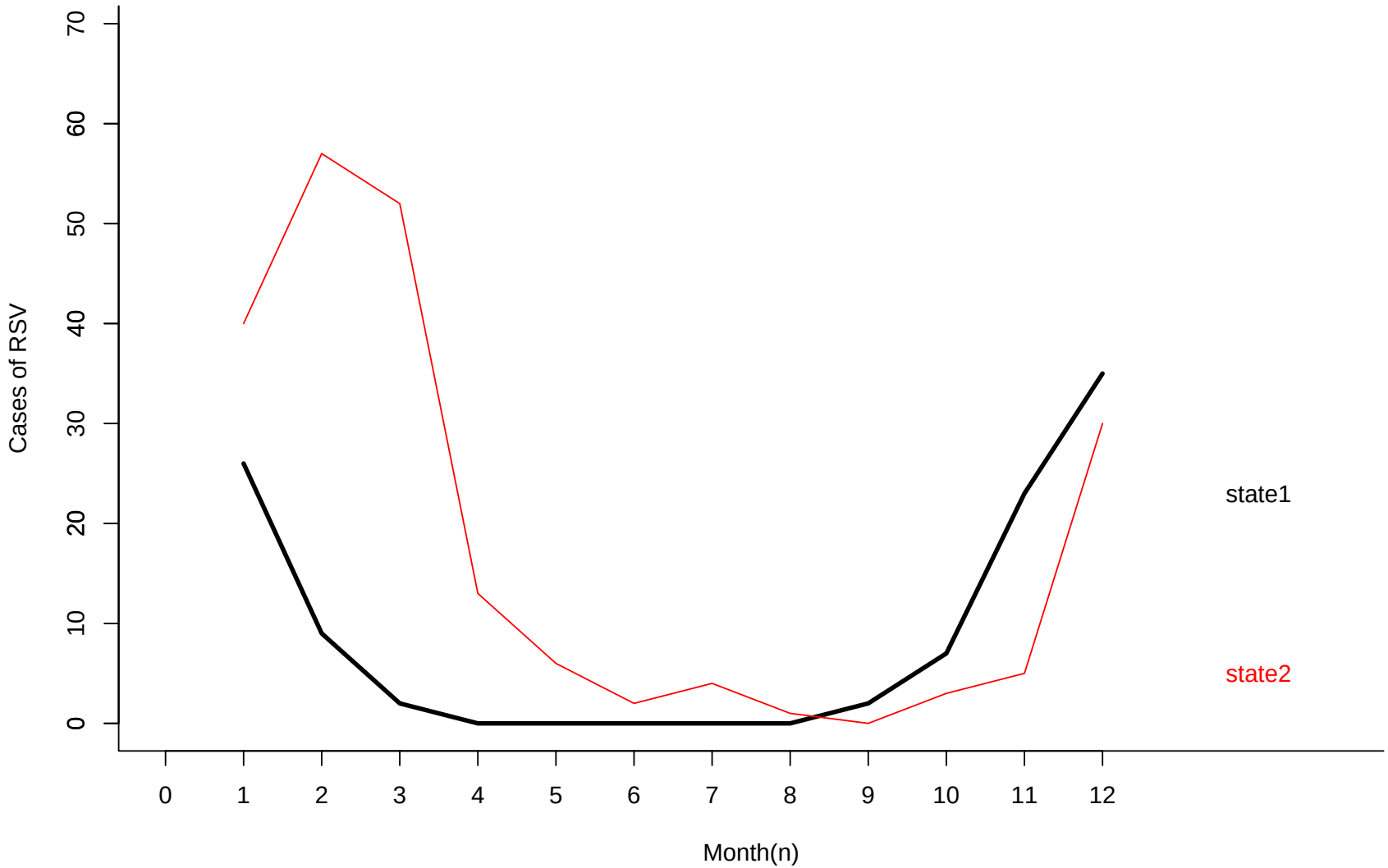
```
##    sin12
## 13.93473
```

Note: the period is 12 month - Therefore, peak timing of 12.06 month (year 1) equal to peak timing of 0.06 month (year 0)

- peak timing of 13.93 month (year 1) equals to 1.93 month (year 0)

Section 2: Identify the onset of an epidemic using the second derivative method

When we look at the pseudo-RSV cases in the last 12 months, it is hard to determine when is the onset by eyes.



So what are the methods to identify epidemic onsets?

The most common way is fixed-value thresholds based on the number of cases defined by researchers. For example, in Dr. Rachel Baker's paper[3], they defined onset as when the normalized mean RSV incidence per week exceeds 0.2. However, this type of threshold is based on the assumption that the threshold "... will be low enough to constitute onset but high enough to exceed random fluctuations in the data...". It generally works well but it requires extensive knowledge of the epidemic curves in different locations to define this value. Therefore, we introduce a new method, second derivative, to estimate the onset of an outbreak.

[3]Baker, R.E., Mahmud, A.S., Wagner, C.E. et al. Epidemic dynamics of respiratory syncytial virus in current and future climates. Nat Commun 10, 5512 (2019).

Theoretical background on first and second derivative

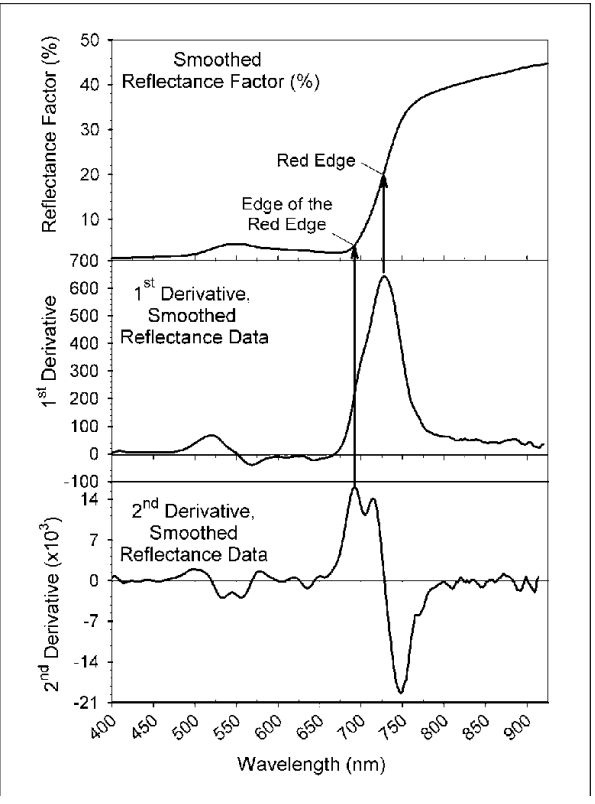


Figure 1: Observed curve, first derivative, and second derivative

The figure shows the relationship among the observed curve, the first derivative of the observed curve, and the second derivative of the observe curve."Peak in the first derivative curve identify the location of the red edge. Peak in the second derivative identify the edge of the red edge."[a]

When the first derivative is positive, the original curve is increasing. In epidemic, this corresponds to the early stage of an outbreak as cases are increasing. When the second derivative reach its maximum in the segment that the first derivative is positive, it means that the growth rate of the increasing trend reach its maximum. This fits for the definition of the starting point of an disease outbreak.

---

[a] Randy M Hamilton et al. (2009). Pre-visible Detection of Grub Feeding in Turfgrass using Remote Se Engineering & Remote Sensing. 75. 179-192. 10.14358/PERS.75.2.179.

Fit a regression to the observed RSV cases

Because now we only observe part of the epidemic, we cannot use the harmonic regression to fit what we observe. Instead, we use p-spline regression, a generalized additive model. Please check out Smooth terms in GAM, P-splines in GAMs and Iris's work on p-spline inference for introduction.

```
require(mgcv)
```

```
## Loading required package: mgcv
```

```
## Loading required package: nlme
```

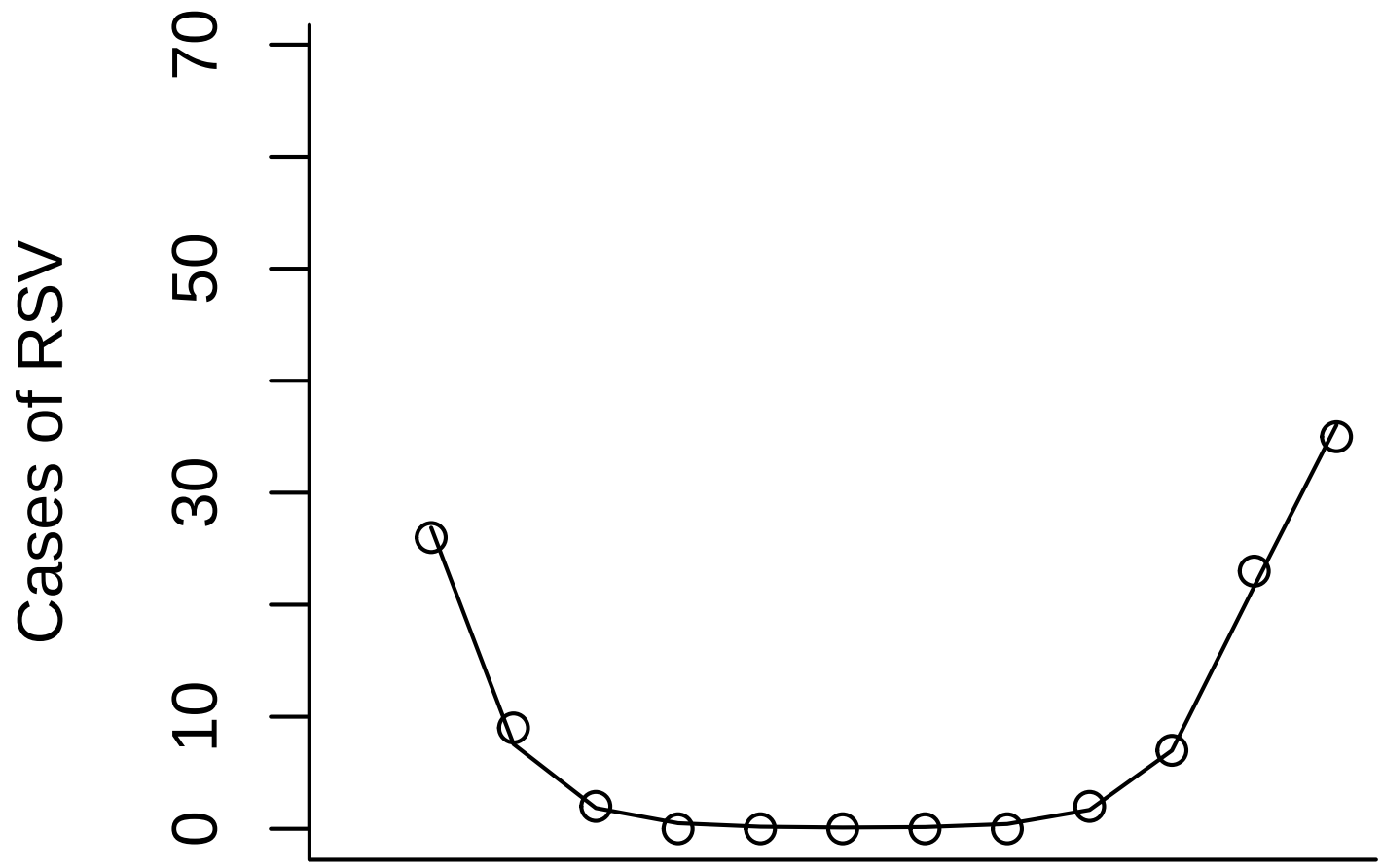
```
## This is mgcv 1.8-40. For overview type 'help("mgcv-package")'
```

```
irregular_model <- gam(cases ~ s(x=time, k=5, bs="ps"),
                       family=poisson, method="REML",
                       data=data.frame(time=seq(1:12),
                                       cases=tail(xt1,12)))
# k is knots
# The value for k set the upper limit on the
# wiggliness of the smooth function
# You can choose the value based on AIC scores as well
```

Comparing the model fit and what we observed

The dots are what we observed and the line is the model fit

```
plot(1:12, tail(xt1,12),type="p", xaxt="n",bty="l",
      ylab="Cases of RSV", xlab="Month(n)",
      ylim=range(c(xt1,xt2)),
     xlim=c(0,12))
lines(1:12,fitted(irregular_model))
```



Function to calculate the first and second derivatives

```
deriv <- function(x, y) diff(y) / diff(x)
# function to calculate derivative

middle_pts <- function(x) x[-1] - diff(x) / 2
# function to assist derivative calculation
```

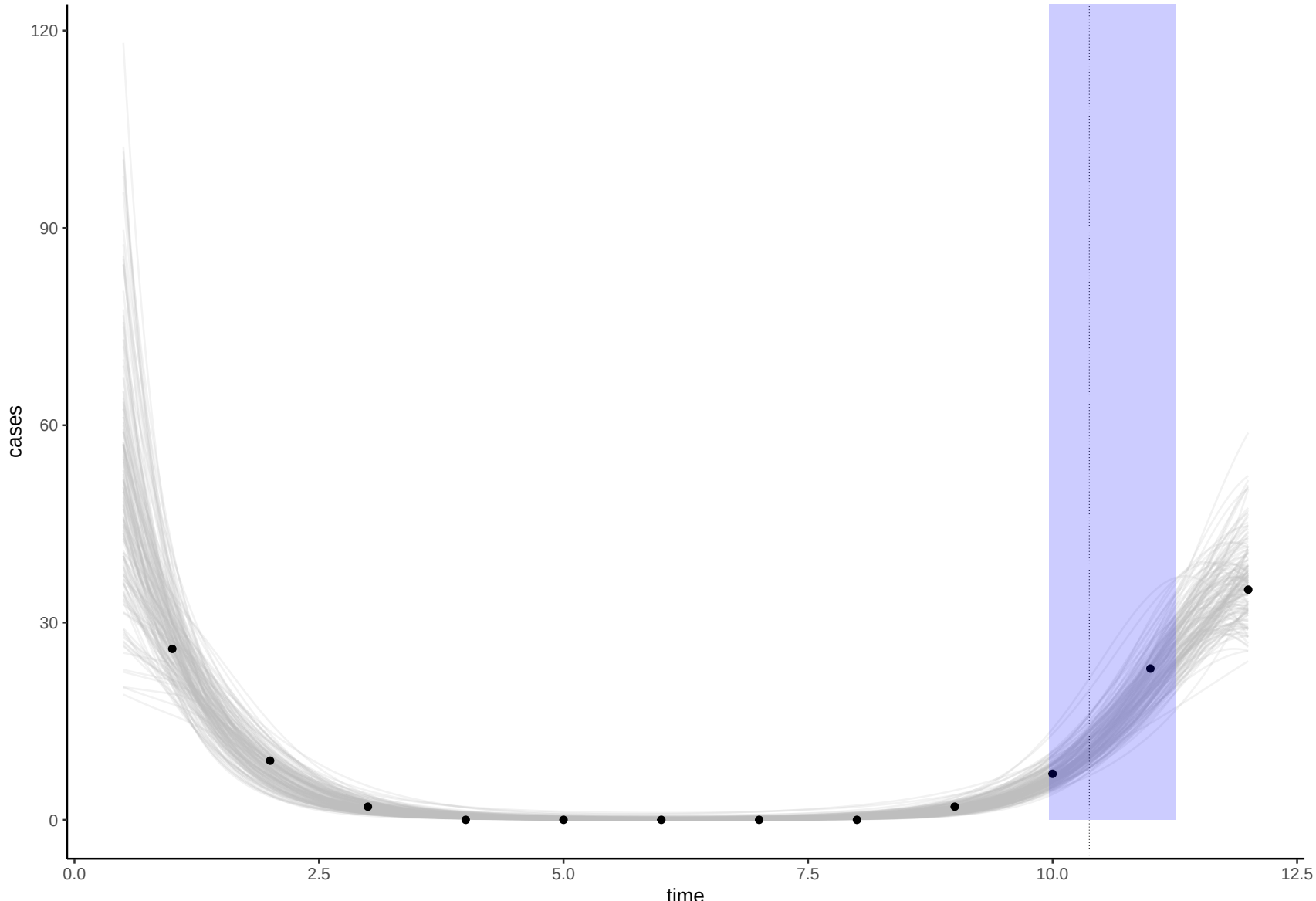
```
#install.packages("pspline.inference")
# require did not work for this package
library(pspline.inference)
require(dplyr)

t <- seq(0.5,12,0.01)
dtime <- seq(0.5,12,0.01)
```

Find uncertainty interval of disease trajectory and onset timing

```
# we use Iris's package to get the uncertainty interval of disease trajectory.
cases.samples=pspline.sample.timeseries(
  irregular_model, data.frame(time=t),
  pspline.outbreak.cases, samples=150)

onset.samples = cases.samples %>%
  group_by(pspline.sample) %>% # for each sample do the following
  do((function(data){
    deriv.pred = data.frame(deriv=diff(data$cases)/diff(data$time),
                            time=c(1:length(diff(t))))
    # calculate the first derivative

    second_d <- data.frame(second.deriv=deriv(middle_pts(dtime),deriv(dtime,data$cases)),
      time=c(1:(length(diff(t))-1))) # calculate the second derivative

    indicator = deriv.pred[which(deriv.pred$deriv>0),]
    # only look at second derivatives in the increasing segment (first derivative>0 )
    second_d_test <- second_d[second_d$time%in%indicator$time,]

    onset = dtime[second_d_test$time[second_d_test$second.deriv==
        max(second_d_test$second.deriv)]]
# find when the second derivative of  the smooth functions reached its maximum
    data.frame(
      onset = onset,
      cases = data$cases[which(data$time==onset)])
    # find the number of RSV
    # when the second derivative reach its maximum
  })(.))
```

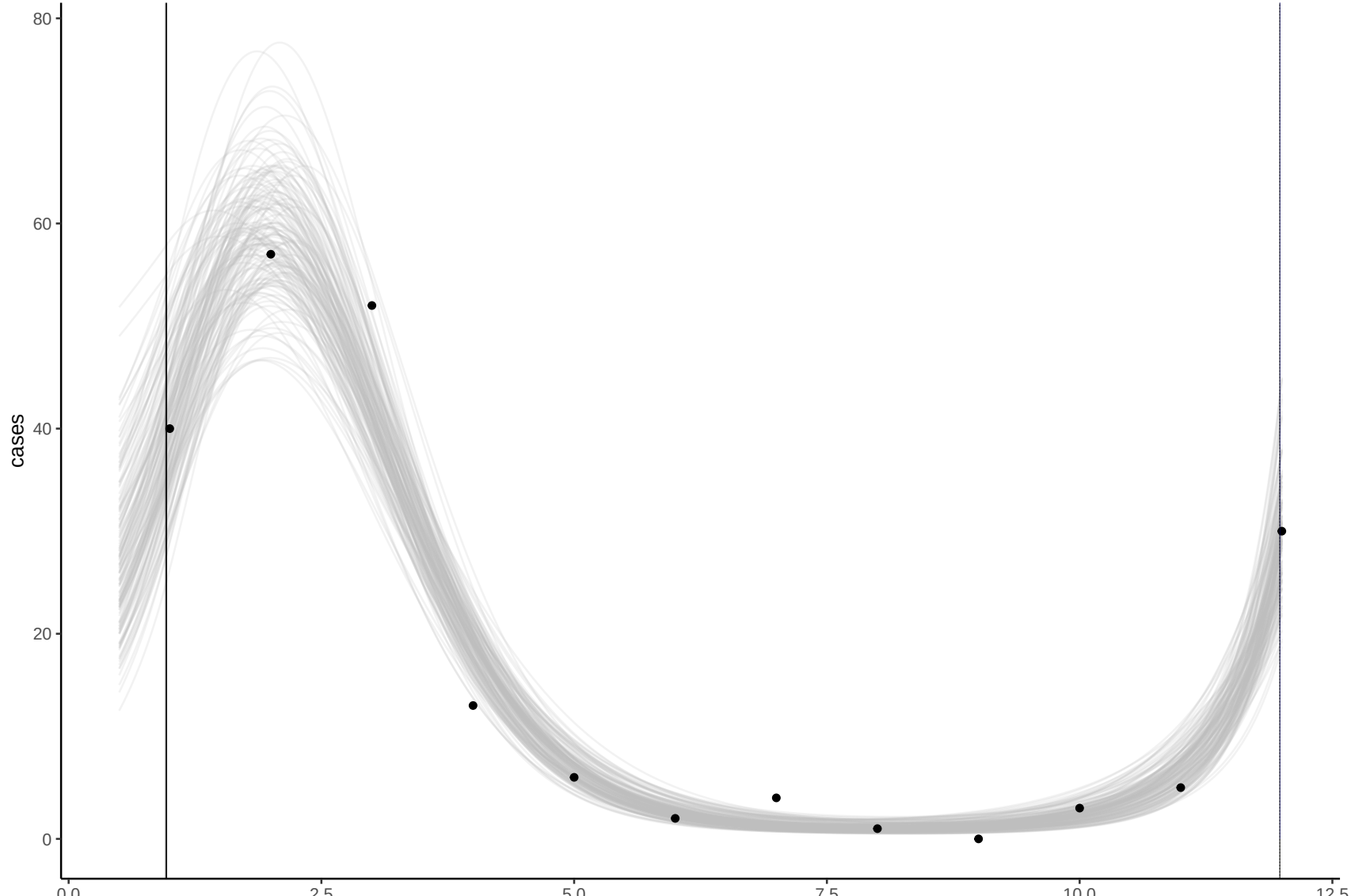
Plot the range of the onset on top of the epidemic curves

The dots are what we observed and the grey lines are the fitted epidemic curves. The blue area are the 95% confidence interval of epidemic onsets and the vertical dotted line shows the mean of the epidemic onset estimate.



Plot the onset estimates of state 2.

The estimates show bimodal distribution because we cut through the epidemic season. To correctly analyze the data, we need to limit the time to after the peak of previous outbreak.



Retry the analysis for the onset estimates of state 2.

This analysis will work better with weekly data and after the epidemic take off as you will have more data points.