Practical II: epidemiological surveillance and modelling (spatial)

30/31th January 2023

Pathogens move between different geographical units because of movements of animals or humans or other hosts. Spatial patterns of infectious disease transmission arise from heterogeneity in the landscape in which transmission occurs. In large and dense cities for example transmission intensity tends to be higher compared to rural areas. Spatial dispersal in ecology and epidemiology is often approximated by an exponential (probability of dispersal at a distance follows dispersal ∝ exp(-d/a), where a is a shape parameter or Gaussian kernel (dispersal $\propto exp(-(d/a)^2)$ and d is the distance between locations of interest.

For human infectious diseases the process of dispersal rarely follows a continuous spatial expansion but rather spread is better approximated by the spatial organisation of cities and towns. For example, emerging infectious diseases appear first in large urban agglomerations that are connected via air travel and then follow human movements from there to other larger cities or rural counties. The SARS-CoV-2 Omicron variant for example first spread in London where most travellers from abroad arrived.

In this practical we will learn about continuous spatial spread using spatial kernels before investigating the impact of human mobility on spatial dispersal of infectious diseases. The code has been adapted from Chapter 11 of the Epidemics Book by Ottar Bjornstad and additional code can be found here: https://github.com/objornstad/epimdr/blob/master/rcode/chapter11code.r

There will be a 15 minute break after ca. 1h ½.

Please load the required packages for this practical

```r
if (!require("pacman")) install.packages("pacman")
pkgs =
  c("knitr", "rmarkdown",
    "ggplot2","epimdr",
    "ncf", "deSolve",
    "plot.matrix","reshape2",
    "dplyr", 'cowplot') # End
pacman::p_load(pkgs, character.only = T)
```

If you are unable to load the epimdr package using the code above please re-install the htmltools package (a dependency of epimdr)

```r
remove.packages('htmltools')
install.packages('htmltools')
```

Practical II Questions

Q1: What determines the spatial spread of infectious diseases?

Contrast spread of plant diseases vs. human infectious diseases. Here a visualisation of human movement patterns in the UK: https://www.science.org/cms/10.1126/science.abj0113/asset/c322ab29-d54d-42ef-8412-4e6789f1fcbd/assets/images/large/science.abj0113-f1.jpg and dynamics of measles diffusion in the UK: https://www.nature.com/articles/s41559-020-1186-6/figures/3

Here a link to more continuous diffusion of Aedes albopictus in the USA and Europe: https://www.nature.com/articles/s41564-019-0376-y/figures/1

Model answer: Plant diseases often can be approximated by a contiguous spatial kernel such as the exponential or Gaussian distribution. Human infectious diseases in contrast follow dispersal along routes of human mobility.

Q2: Plot the coordinates and infection status (0 for non infected, 1 for infected with different colours for the timing of infection) of a fungal rust pathogen (see picture below) on the Filipendula ulmaria wild plant. Summarise the data (number of infected locations in 1994, 1995 and how many remained uninfected).

Figure 1: Plant Infected With Rust

Context: Triphragmium ulmariae is a species of rust fungus in the family Sphaerophragmiaceae. It causes meadowsweet rust gall which develops as a chemically induced swelling, arising from the lower surface of the Filipendula ulmaria leaves. It has implications for the survival of the meadowsweet seedlings.

R code adapted from Chapter 11, Epidemics: Models and Data in R, Ottar N. Bjornstad (ISBN 978-3-319-97487-3), https://www.springer.com/gp/book/9783319974866

```
data(filipendula)

cols <- c('Uninfected' = '#3a506b','Infected in 1994' = '#ffbc42',
          'Infected in 1995' = '#d81159', 'All plants' = "grey")

a1 <- ggplot() +
  geom_point(aes(X, Y,fill = 'All plants', color = 'All plants'),
             data = filipendula, shape = 21, size = 2,
             alpha = 0.5) +
  theme_bw() +
  xlim(0,700) + labs(x= "Distance in meters", y= "Distance in meters") +
  ylim(0,700) +
  scale_fill_manual(name = 'Infection Status',
                    values = cols) +
  scale_color_manual(name = 'Infection Status',
                     values = cols) +
  theme(legend.position='none')

a2 <- ggplot() +
  geom_point(aes(X, Y,fill = 'All plants', color = 'All plants'),
             data = filipendula, shape = 21, size = 2,
             alpha = 0.5) +
  geom_point(aes(X, Y,fill = 'Infected in 1994', color = 'Infected in 1994'),
             data = filipendula[filipendula$y94==1 & filipendula$y95 == 1, ],
             shape = 21, size = 2,
             alpha = 0.5) +
  theme_bw() +
  xlim(0,700) + labs(x= "Distance in meters", y= "Distance in meters") +
  ylim(0,700) +
  scale_fill_manual(name = 'Infection Status',
                    values = cols) +
  scale_color_manual(name = 'Infection Status',
                     values = cols) +
  theme(legend.position='none')

a3 <- ggplot() +
  geom_point(aes(X, Y,fill = 'All plants', color = 'All plants'),
             data = filipendula, shape = 21, size = 2,
             alpha = 0.5) +
  geom_point(aes(X, Y,fill = 'Infected in 1994', color = 'Infected in 1994'),
             data = filipendula[filipendula$y94==1 & filipendula$y95 == 1, ],
             shape = 21, size = 2,
             alpha = 0.5) +
  geom_point(aes(X, Y,fill = 'Infected in 1995', color = 'Infected in 1995'),
             data = filipendula[filipendula$y94==0 & filipendula$y95 == 1, ],
             shape = 21, size = 2,
             alpha = 0.5) +
  theme_bw() +
  xlim(0,700) + labs(x= "Distance in meters", y= "Distance in meters") +
  ylim(0,700) +
  scale_fill_manual(name = 'Infection Status',
                    values = cols) +
  scale_color_manual(name = 'Infection Status',
                     values = cols) +
  theme(legend.position = 'top')

legend_plot <- get_legend(a3)

a3_new <- a3 +
  theme(legend.position='none')

plot_grid(a1,a2,a3_new, legend_plot, ncol = 3,
          rel_heights = c(6, 1))
```

Model answer: New infections in 1995 tend to occur near clusters of plants previously infected in 1995 e.g. at (580,180). There were 86 infected plants in 1994 and 91 infected plants in 1995 with 12 new infections

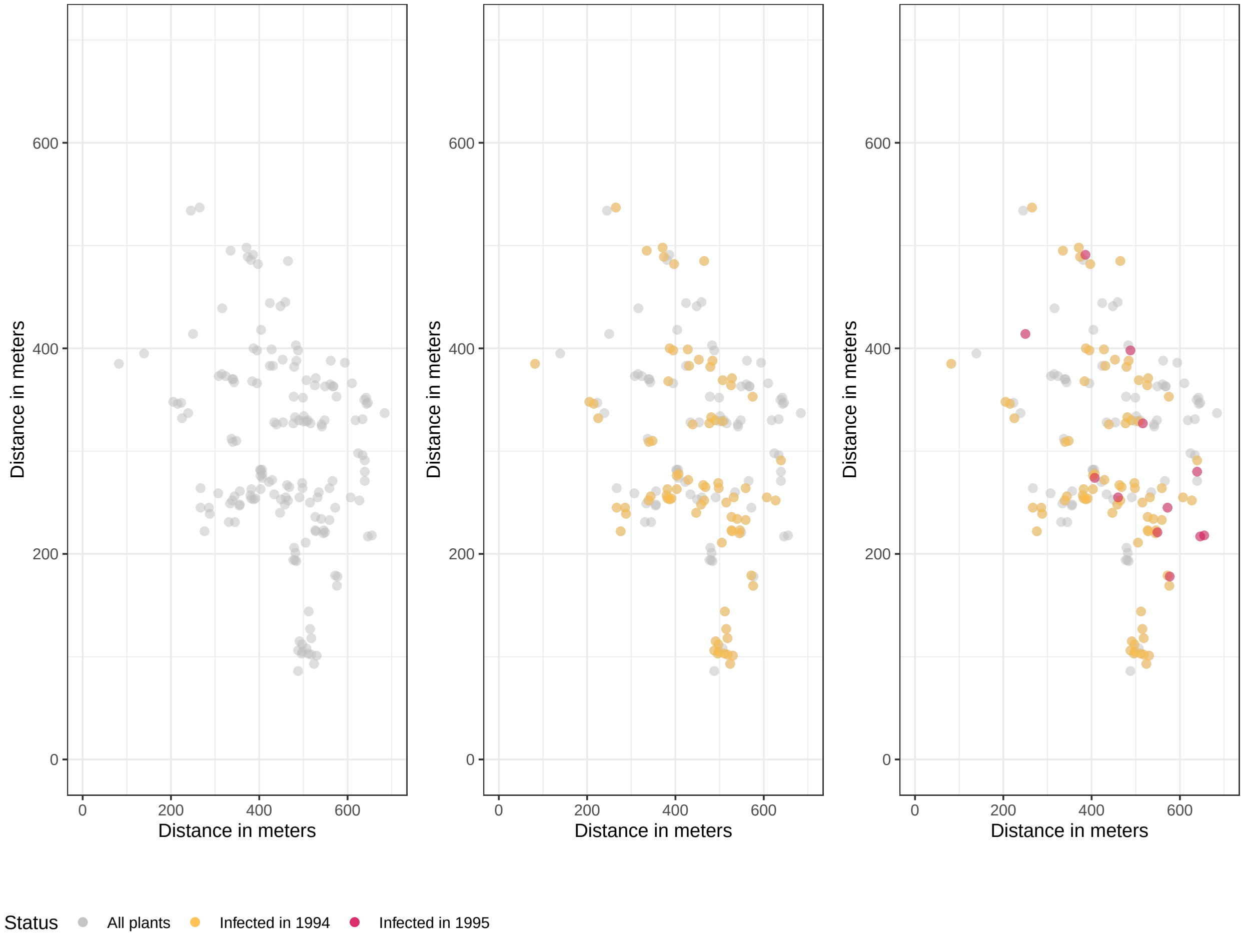


Figure 2: Presence and absence of a pathogen on the Filipendula ulmaria wild plant. Size of dots represent timing of infection in either 1994 or 1995.

Q3: Calculate the distance between each X and Y coordinate using the dist function in R. Visualise the distribution of distances and describe them (unit of distances is meters). Please also explain why we need a distance matrix.

The distance function is given by: The distance between points indicative of their coordinates is calculated according to the formula: $\sqrt{(}(x2\breve{}x1)^2 + (y2\breve{}y1)^2)$

We first have to calculate the distances between each observation of Filipendula ulmaria across the island

```r
dst = as.matrix(dist(filipendula[,c("X","Y")]))
hist(dst)
```

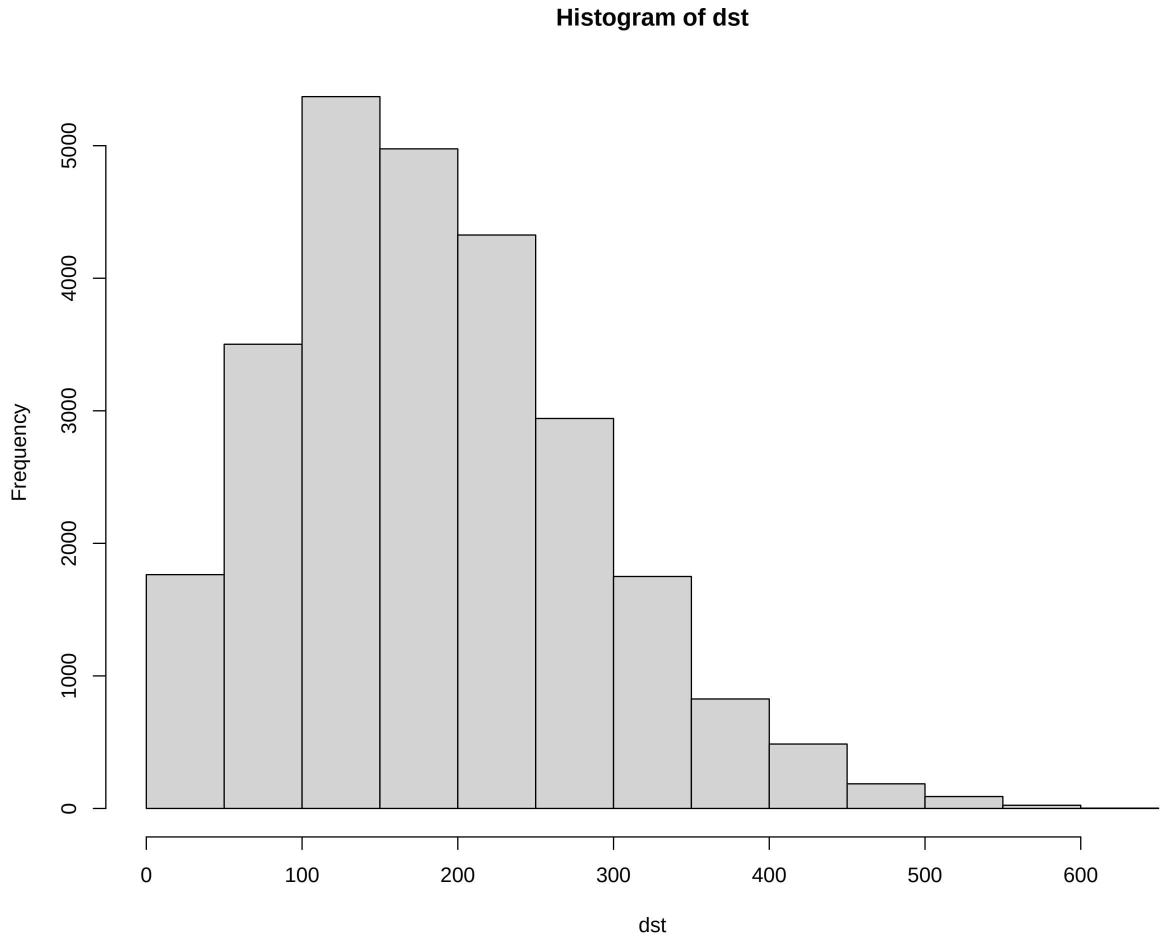


Figure 3: Frequency of distances between each datapoint in meters.

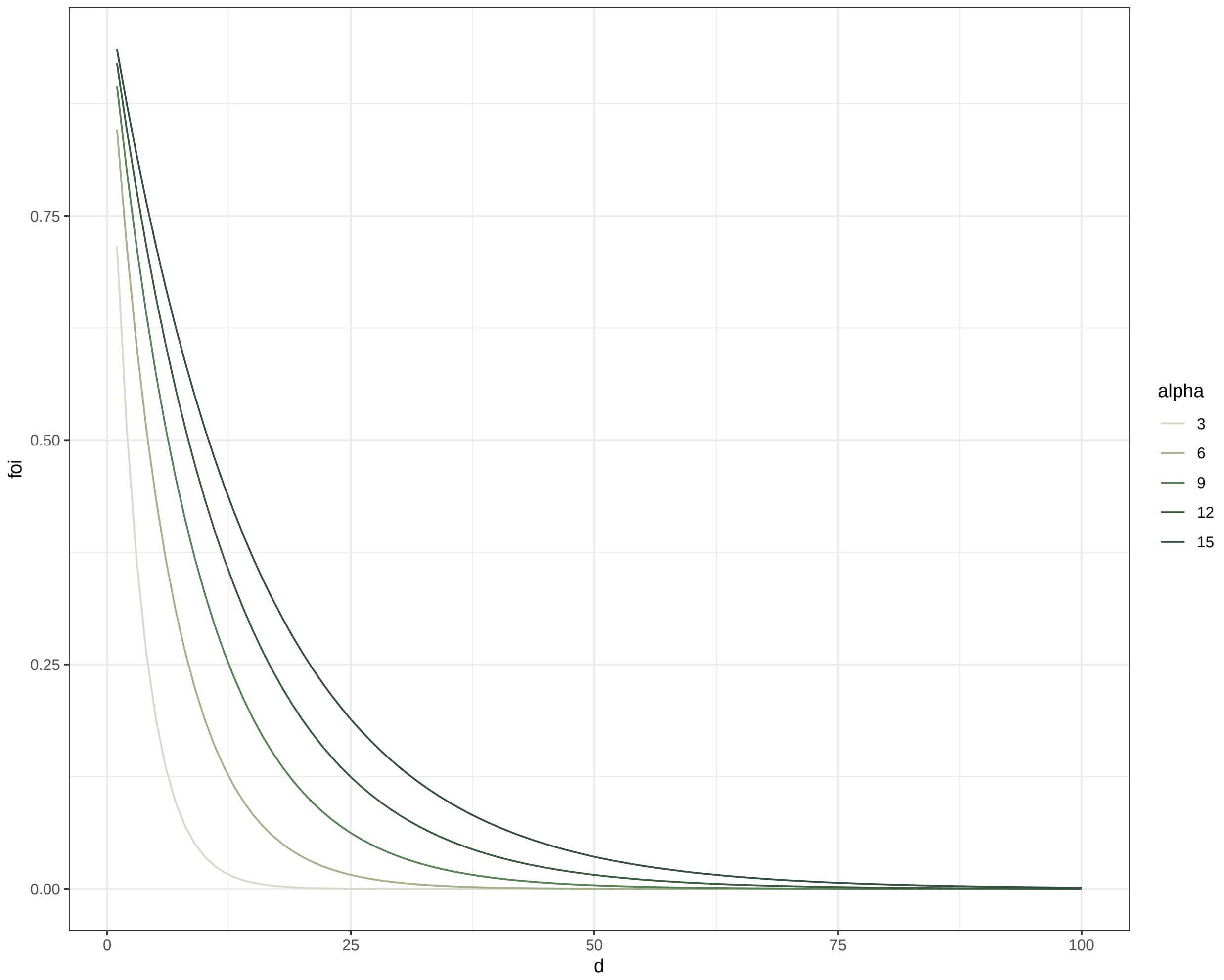
For calculating the force of spatial invasion we need to calculate the distances between locations that are infected vs. those that are not. Predictors of invasion rely on a measure of closeness (here distance).

Q4: Calculate the force of spatial infection based on infected and non-infected plants. The likelihood of becoming infected depends on the spatial force of invasion which is given by the connectivity of an uninfected plant to infected plants:

$\propto \sum_j z_j exp(-d_{ij}/a)$, where z_j is the disease status (0/1) in the previous year and d_{ij} are the distances between locations. a is the shape parameter

Explain in non-technical language how the force of invasion in 1995 is defined in this context and how changing the shape parameter a can affect our estimations of the force of invasion

```
max_dist <- 100
alpha <- rep(seq(3, 15, by = 3), each = max_dist)
d <- rep(seq(1, max_dist), length(unique(alpha)))
df <- cbind(alpha, d) %>% as.data.frame()
df <- df %>%
mutate(foi = exp(-d/alpha))
ggplot(data = df, aes(x = d, y = foi)) +
geom_line(aes(color = factor(alpha)))+
scale_color_manual(values = c("#dad7cd", "#a3b18a", "#588157", "#3a5a40", "#344e41"),
name = "alpha") +
 theme_bw()
```



Define the shape parameter and calculate force of invasion per location for 1995 based on data from 1994

```
a = 10
foi = apply(exp(-dst/a)*filipendula$y94,2,sum)
```

Model answer: Force of infection at any given location in 1995 is defined by how close infected locations are to non-infected locations.By increasing a the force of infection also increases.

Q5: Now we like to compare how a spatial model estimating the dispersal compares to a model which is aspatial (nullmod). To do so we create a second model where risk for infection is uniform across all locations and we then compare this model to the spatial model using the anova function in R (more details available here: https://www.youtube.com/watch?v=wEY1M8Pg0Wg).

Here our outcome is infection status in year 1995 and we want to know which one out of the following is the better predictor, using a logistic regression (because our outcome is binary):

- likelihood of becoming infected which is calculated by the connectivity of an uninfected plant to infected plants (spmod in code)
- No information about connectivity between uninfected and infected plants (nullmod in code)

Explain the results of the anova test below. Please note, the residual deviance would be 0 if the model can perfectly explain the data. Lower residual deviance would mean a better model.

```
# Using a GLM framework (more details on GLM can be found here:
#http://www.simonqueenborough.info/R/statistics/glm-binomial)
lfit = glm(y95 ~ foi, family = binomial(), data = filipendula)
lfit$deviance/2
```

```
## [1] 69.8527
```

```
####################################################
a = seq(1,20, length = 1001)
llik = rep(NA, length(a))
for(i in 1:length(a)){
  foi = apply(exp(-dst/a[i])*filipendula$y94,2,sum)
  lfit = glm(y95~foi, family = binomial(),
             data = filipendula)
  llik[i] = lfit$deviance/2
}

####################################################
ahat = a[which.min(llik)]
foi = apply(exp(-dst/ahat)*filipendula$y94,2,sum)
spmod = glm(y95~foi, family = binomial(), data = filipendula)
nullmod = glm(y95~1, family = binomial(), data = filipendula)
#correct the df of the spmod
spmod$df.residual = spmod$df.residual-1
anova(nullmod, spmod, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: y95 ~ 1
## Model 2: y95 ~ foi
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       161     222.10
## 2       159     109.48  2   112.63 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The spatial model has a lower residual deviance and thus is the better model as compared to the null model (non-spatial).

Q6: We are now interested in whether a Gaussian kernel may be better in approximating the dispersal. To do so we create another model where force of infection is calculated by using the Gaussian shape rather than exponential (see introduction).

Calculate log likelihood for Gaussian kernel and compare both models using AIC (see more details on AIC here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8187274/ section “Traditional null-hypothesis significance testing”, and visualise their kernels. Explain what the plot shows.

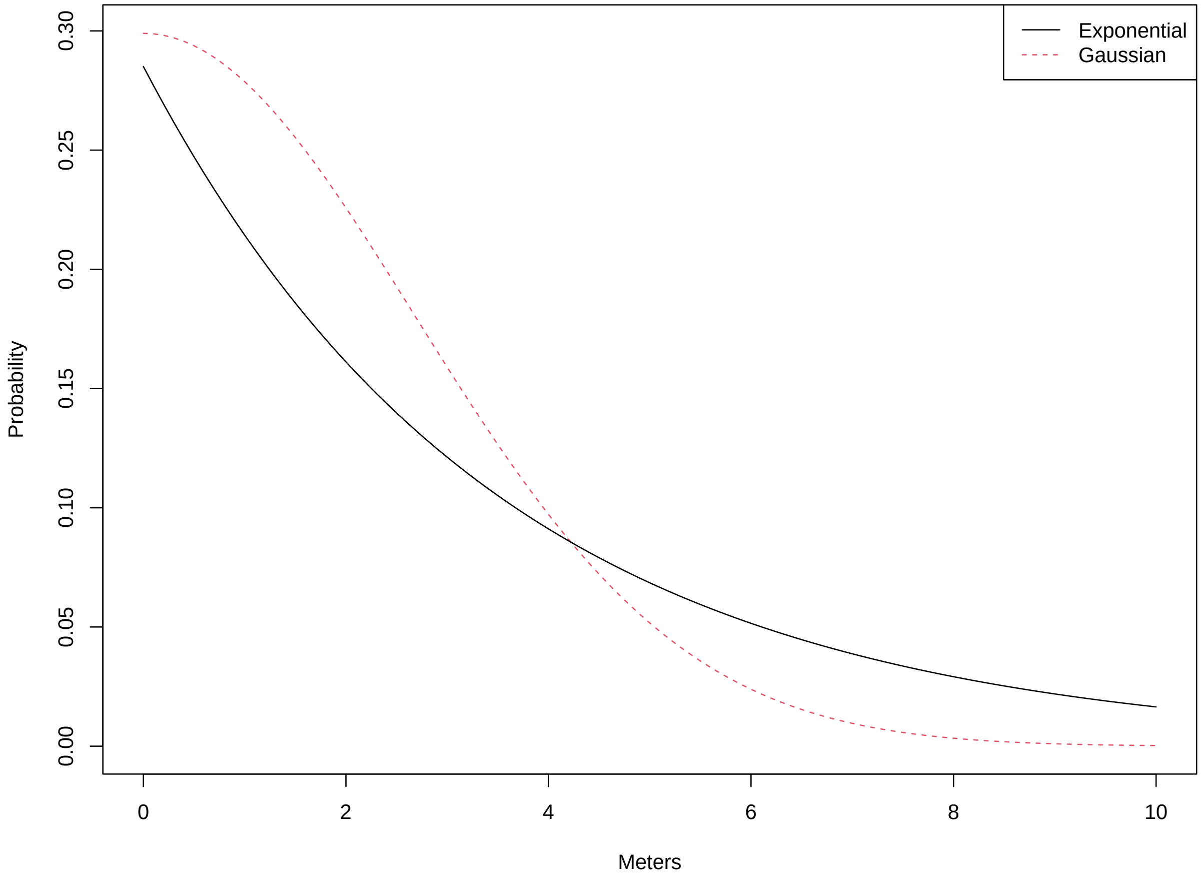
AIC is a function of:

- Model complexity (number of parameters used). It prefers a more parsimonious model
- Likelihood (how well the model reproduces the data)

Hence, the better model according to AIC is the one that explains the greatest amount of variation in the outcome using the fewest possible predictors. AIC = 2*(number of parameters)-2*log(Likelihood). The model with the lower AIC is better.

```
a2 = seq(1,20, length = 1001)
llik2 = rep(NA, length(a2))
for(i in 1:length(a2)){
  foi2 = apply(exp(-(dst/a2[i])^2)*filipendula$y94,2,sum)
  lfit2 = glm(y95~foi2, family = binomial(),
              data = filipendula)
  llik2[i] = lfit2$deviance/2
}
ahat2 = a2[which.min(llik2)]
foi2 = apply(exp(-(dst/ahat2)^2)*filipendula$y94,2,sum)
spmod2 = glm(y95~foi2, family = binomial(),
             data = filipendula)
spmod2$df.residual = spmod2$df.residual-1

###################################################
curve((2/(ahat2*gamma(1/2)))*exp(-((x/ahat2)^2)), 0,
      10, col=2, lty=2, ylab="Probability", xlab="Meters")
curve((1/(ahat)*gamma(1))*exp(-x/ahat), 0, 10, add=TRUE)
legend("topright", c("Exponential", "Gaussian"), lty=c(1,2), col=c(1,2))
```



```
##################################################
spmod$aic
```

```
## [1] 113.4775
```

```
spmod2$aic
```

```
## [1] 116.6538
```

Exponential model (spmod) is favoured over the Gaussian model (spmod2).

Probability of invasion corresponding to distance between infected and uninfected plant with alpha (max range of infection) = 10.

Q7: Human mobility and its impact on spatial dispersal of epidemics.

The diffusion of human pathogens rarely follows simple patterns that can be approximated by the exponential or Gaussian kernels discussed above. The patterns of human mobility are more complex and follow patterns of transportation and human population aggregation. For example, it is more likely to travel to and from major population aggregation than to rural areas, even if the rural areas are closer to an individual's home location. Consider the case of Oxford: on average people living in Oxford are more likely to travel to and from London than travel to let's say, Cirencester. Cirencester is closer to Oxford but travel there via train takes ~2h, vs. 1h to central London. Chichester's population is about 20k vs. London's population is currently estimated at > 8M. So it is no surprise that human mobility patterns today are not strictly determined by distance.

Models approximating these data usually take into account the population of the origin and destination location, distance or travel time between them, and any other variables that may influence travel patterns such as the attractiveness of a population (high degree of shopping opportunities, work opportunities in cities). For the purpose of our practical we will consider the simplest of these models which is the Gravity model (Erlander and Stewart 1990: https://books.google.co.uk/books/about/The_Gravity_Model_in_Transportation_Anal.html?id=tId3PU1leR8C&redir_esc=y). The gravity model posits that movement volume between two communities depends inversely on distance, d, but bilinearly on the size, N, of the communities considered. More generally the model can be written as: $T_{ij} = \frac{N_i^a N_j^b}{d_{ij}^c}$where T_{ij} is the travel volume between locations i and j, N_i^a is the population at origin location i, N_j^b is the population of destination location j and d_{ij}^c is the distance between i and j. d_{ij}^c could also be the travel time between i and j. a,b, c are parameters fitted using empirical data.

Using a previously developed model by Viboud et al. (2006: https://www.science.org/doi/10.1126/science.1125237) for the spatial spread of influenza in and between USA cities we will consider a simple version of the SIR model ignoring any susceptible recruitment (no births or deaths):

$\frac{dS_i}{dt} = -(\beta I_i + \sum_{j \neq i} l_{j,i} I_j) S_i$

$\frac{dI_i}{dt} = (\beta I_i + \sum_{j \neq i} l_{j,i} I_j) S_i - \gamma I_i$

$\frac{dR_i}{dt} = \gamma I_i$

Where $l_{j,i} I_j$ is the gravity weighted force of infection exerted by location j on location i. Beta is the transmission coefficient, I is number of infected individuals, S number of susceptibles. Gamma is the recovery rate.

Please use links provided in the text above and any other online resources to describe in your own words the gravity model and the basic intuition behind it?

Model answer: In analogy to Newton's law of gravity, the gravity law assumes that the number of individuals travelling from i to j (Tij) per unit time is proportional to some power of the population of the source and destination locations and decays with the distance between them. A and b are adjustable exponents and the distance deterrence function is chosen to fit the observed empirical data. The gravity model was derived to describe human mobility patterns in a modern world where travel time was not equal to distance but rather the transport infrastructure between locations.

Q8: We now translate the theory from above into code in R. This requires the R-package 'deSolve'.

Explain in your own words the basics of the SIR model and its key assumptions. Should you be unfamiliar with it there will be another practical going into the theory behind it. Otherwise please use the literature or watch a short section of this video by Prof. Grenfell: https://youtu.be/AzVnN5cCFk4 (from minute ~4 to minute ~7). Some additional details can be found here: https://royalsocietypublishing.org/doi/10.1098/rstb.2020.0263

```r
SIR.space = function(t, y, pars){
  i = c(1:L)
  S = y[i]
  I = y[L+i]
  R = y[2*L+i]
  with(pars,{
    beta = beta[i]
    dS = -(beta*I + m*G%*%I)*S
    dI = (beta*I + m*G%*%I)*S - gamma*I
    dR = gamma*I
    list(c(dS, dI, dR))
  })
}
```

Model answer: The classical model describing outbreaks is the SIR system. The standard SIR model is made up of ordinary differential equations describing the flow of hosts between the Susceptible, Infectious, and Recovered compartments.

Assumptions are:

- The infection circulates in a population of size N with birth and death rates that balance each other out.
- The infection causes acute morbidity but no mortality. In the simple version of the model that means that we can ignore disease-induced mortality. This may not be realistic for example for Ebola.
- Individuals are susceptible from birth (doesn't take into account demographic structure).
- Transmission of infection from infectious to susceptible individuals is controlled by a bilinear contact term. The assumption here is that susceptible and infected individuals mix at random.
- Infectiousness is assumed not to change during the course of the infection in an individual
- Infected individuals become infectious immediately (no latent period)
- Those recovered are immune immediately and indefinitely
- Chance of recovery or death does not change over the course of the outbreak or infection
- Tends to be an oversimplification of complex biological mechanisms

Q9: Now we want to estimate the interaction matrix for all USA states using parameters estimated by Viboud et al. 2006, Science.

Generate matrix using the code below, visualise it, and please explain what it shows (also refer back to Question 7; a = 0.3, b = 0.6, c = 3, and G = T_ij in question 7):

```
require(ncf)
data(usflu)
usdist = gcdist(usflu$Longitude, usflu$Latitude)

###################################################
gravity = function(tau1, tau2, phi, pop, distance){
  gravity = outer(pop^tau1, pop^tau2)/distance^phi
  diag(gravity) = 0
  gravity}
G = gravity(0.3, 0.6, 3, usflu$Pop, usdist)

class(G)
```

```
## [1] "matrix" "array"
```

```
plot(log10(G), xlab="US State", ylab="US State")
```

Spatial interaction matrix here represents the relative connectivity between states in the USA.

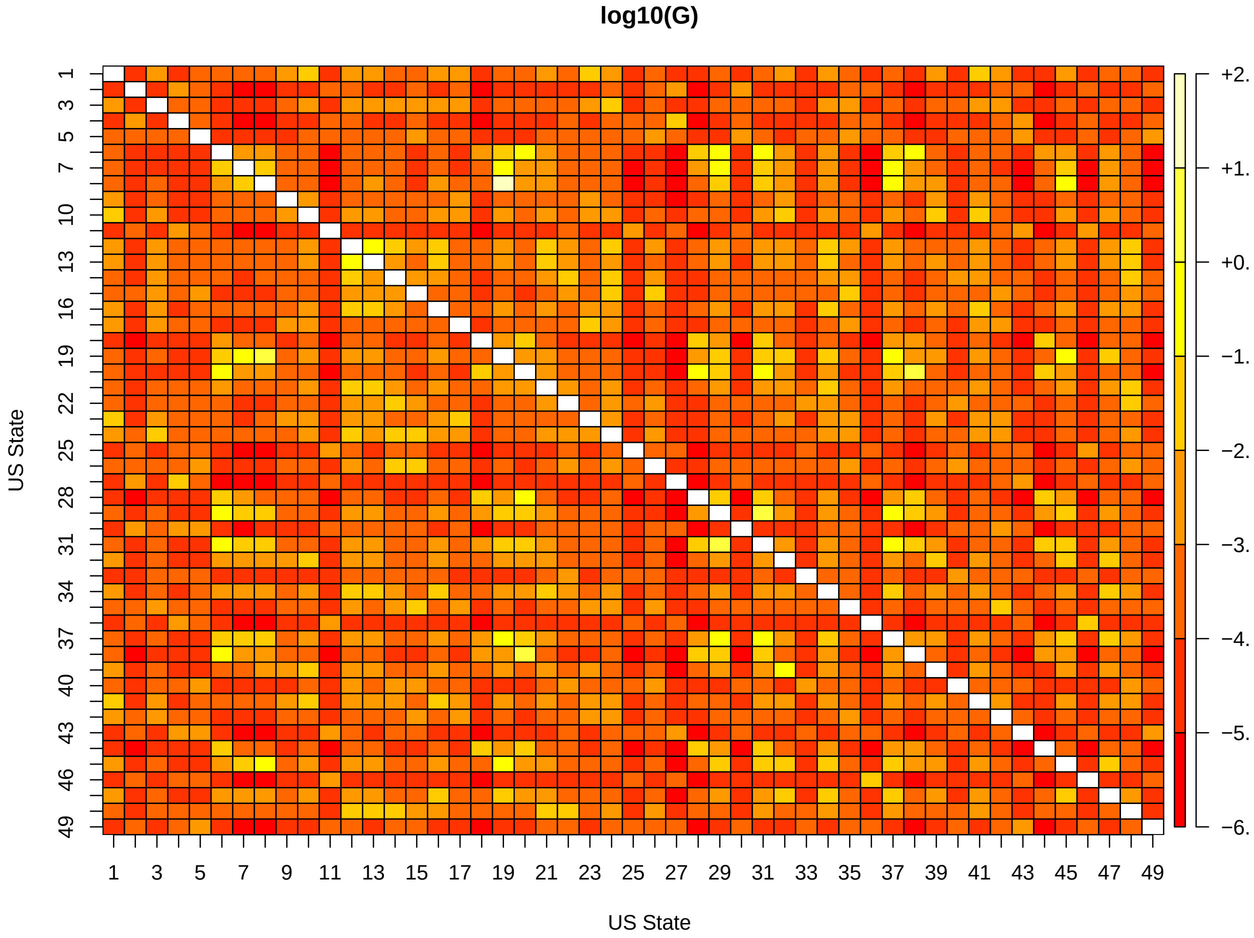


Figure 4: Spatial interaction matrix using parameters from Viboud et al. between all USA states.

Q10: What does G in the model above represent?

Model answer: Spatial interaction matrix.

Q11: In order to estimate the spatial interaction matrix we need multiple data points. What are these?

Model answer:

- Population per location
- Distances between them

Q12: To simulate the epidemic of Flu in the USA we need to define the reproduction number and recovery parameter. We assume all individuals to be susceptible for simplicity.

Generate model and outputs. Explain the Figure output and hypothesise why we observe subsequent waves of flu in different US states.

In the code below, we only visualised a selected few states.

```
gamma = 1/3.5
R0=1.8
beta = R0*gamma/usflu$Pop
m = 1/1000/sum(usflu$Pop)
parms = list(beta = beta, m = m, gamma  =  gamma, G = G)
L = length(usflu$Pop)

head(usflu)
```

```
##   State Acronym      Pop Latitude Longitude Start Peak
## 1     1       AL  3755003  33.0015   -86.766    27   33
## 3     3       AZ  2332966  33.3735  -111.829    25   30
## 4     4       AR  2202789  35.0803   -92.577    28   32
## 5     5       CA 21955156  35.4586  -119.355    28   31
## 6     6       CO  2639909  39.5007  -105.204    28   31
## 7     7       CT  3143652  41.4949   -72.874    26   29
```

```
S = usflu$Pop
R = I = rep(0, length(usflu$Pop))
usflu$State <- 1:nrow(usflu)
I[31] = 1 # State where to initialise the epidemic simulation
```

```
inits = c(S = S, I = I, R = R)


####################################################
times = 0:200
out = ode(inits, times, SIR.space, parms)
infected <- as.data.frame(out[,c(51:99)])
get_state_name <- as.data.frame(usflu)
names(infected) <- as.character(usflu$Acronym)
infected$time <- 1:201
infected_long <- melt(infected, id.vars="time")
state.name <- c('AL','NY','MT','CA','DE')
ggplot(infected_long %>% filter(variable %in% state.name),
       aes(time,value, col=variable)) +
  geom_point() + theme_bw() +
  labs(x = 'Time', y = 'Cases') +
  guides(colour = guide_legend(title="States"))
```

Model answer: Epidemic size depends on population size and their timing on their connectivity to NY state.

Q13: Change initial conditions of the model and initialise epidemic in a rural US state (e.g., Montana (MT), relatively smaller population and further away from high population centres). Describe outputs in relation to an outbreak that originated in NY state. Make use of matrix G and usflu dataframe:

```
G_summary <- as.data.frame(rowSums(G))
colnames(G_summary) <- c('GravitySummary')
usflu <- cbind(usflu, G_summary)
```

Epidemic initialised in Alabama (AL). Smaller epidemic wave there and longer lag between Alaska and onset of epidemics in other states.

```
S = usflu$Pop
R = I = rep(0, length(usflu$Pop))
I[1] = 1
usflu$State <- 1:nrow(usflu)
inits = c(S = S, I = I, R = R)

####################################################
# We are then ready to simulate and visualise the model.
require(deSolve)
times = 0:200
```

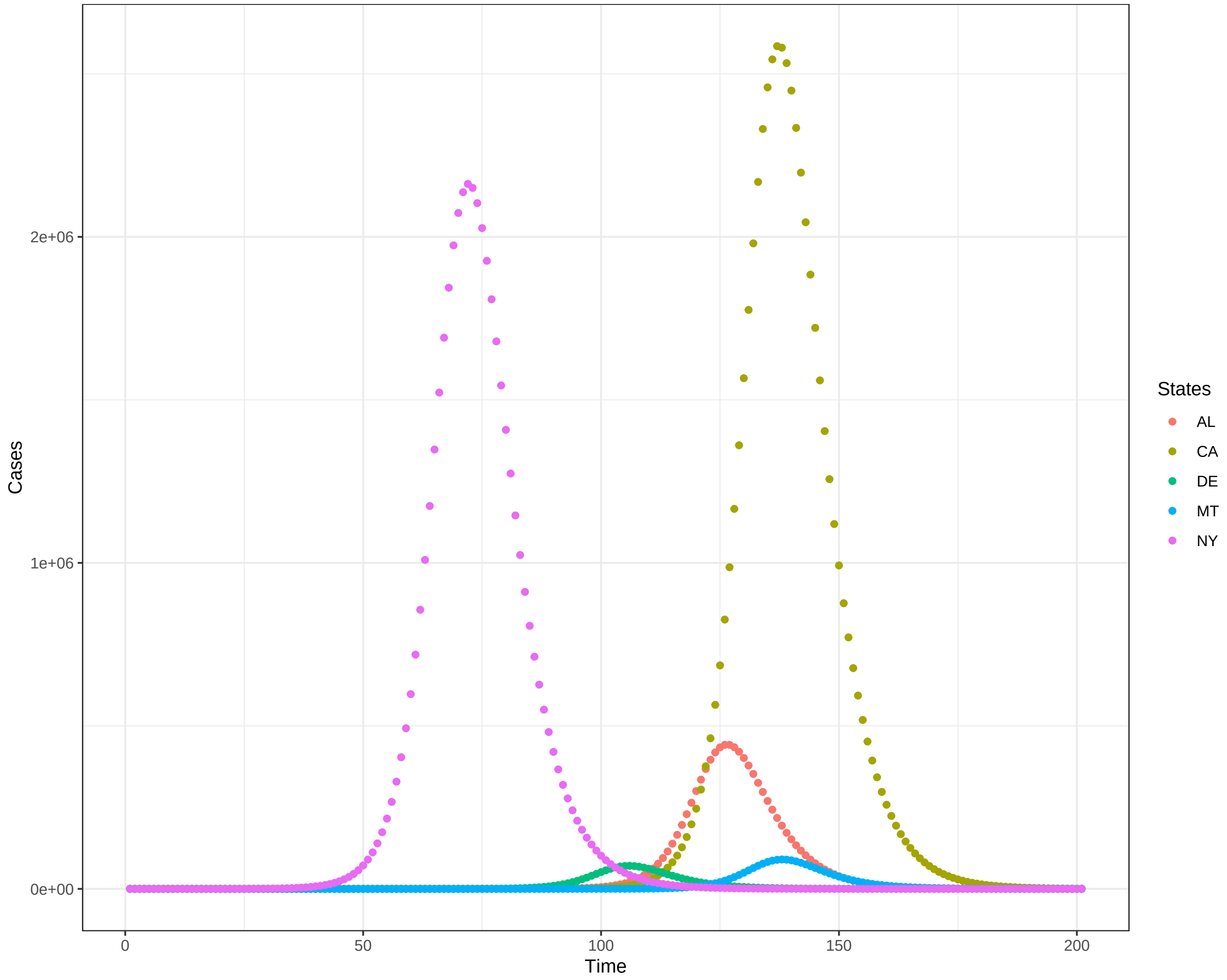


Figure 5: State level epidemics of Flu in the USA using a multipatch SIR model with parameters on spatial interactions based on Viboud et al. 2006. Initial case reported in NY state.

```r
out = ode(inits, times, SIR.space, parms)
infected <- as.data.frame(out[,c(51:99)])
get_state_name <- as.data.frame(usflu)
names(infected) <- as.character(usflu$Acronym)
infected$time <- 1:201
infected_long <- melt(infected, id.vars="time")
state.name <- c('AL','NY','MT','CA','DE')

ggplot(infected_long %>% filter(variable %in% state.name), aes(time,value, col=variable)) +
  geom_point() + theme_bw() +
  labs(x = 'Time', y = 'Cases') +
  guides(colour = guide_legend(title="States"))
```

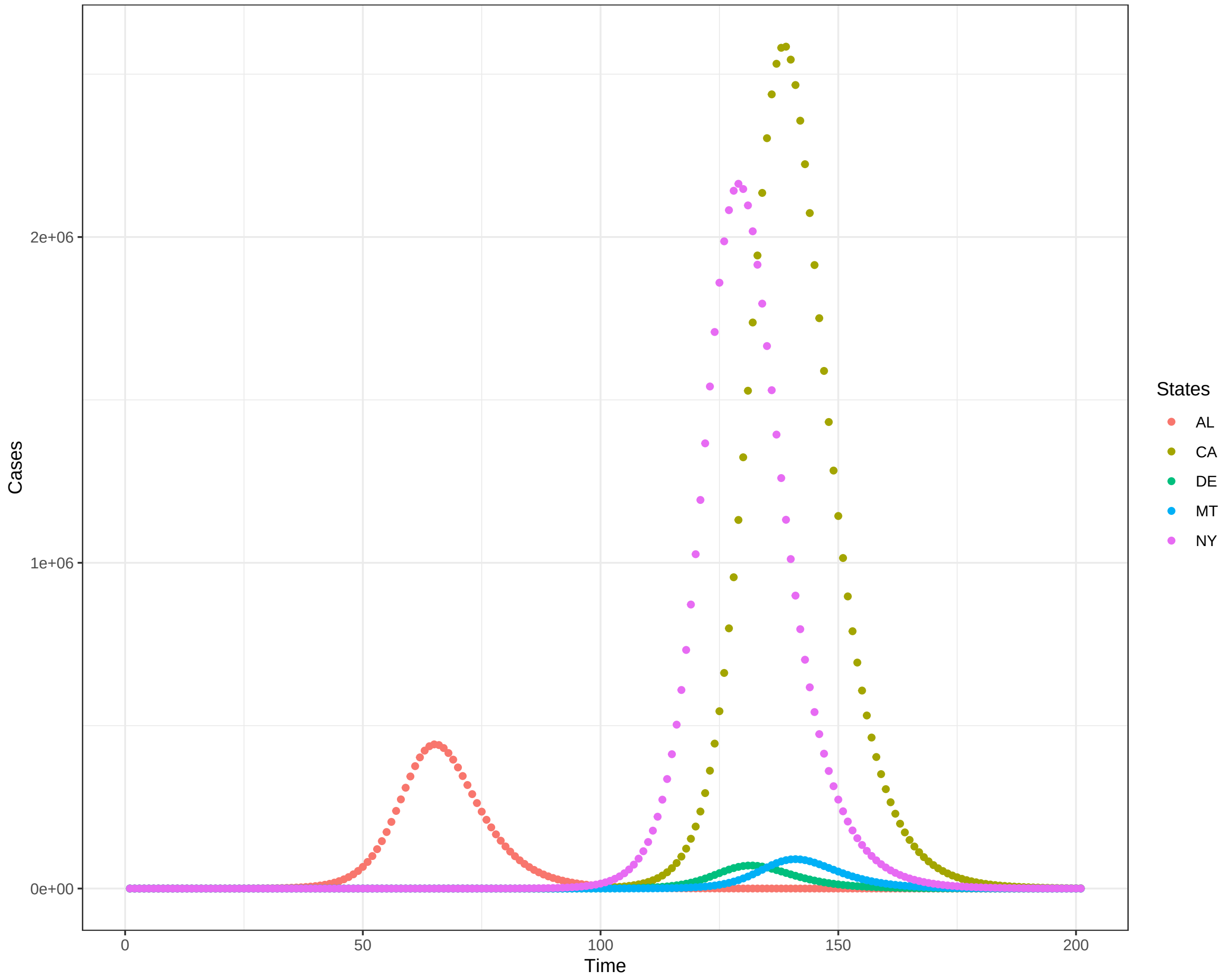


Figure 6: State level epidemics of Flu in the USA using a multipatch SIR model with parameters on spatial interactions based on Viboud et al. 2006. Initial case reported in AL state.

Q14: What are other factors that may explain the spatial synchrony/asynchrony of flu?

- Behavioural factors
- Household size
- Immunity
- Vaccination
- Age distribution
- Climatic factors