A brief survey of statistical models to analyze the transmission of infectious diseases

Shannon Gallagher, PhD
Post-doctoral Fellow
Biostatistics Research Branch, Division of Clinical Research
National Institute of Allergy and Infectious Disease

Studying infectious disease is important because…

- You are getting credit for this class (utility)
- You or someone you have known has experience being sick (empathy)
- Infectious disease has a huge impact on the world around us
  - Lives
  - Time
  - Money
  - Media

So why wouldn’t we make a statistical model?

A series of unfortunate events (to be avoided)

1. Ask the wrong question
    - E.g. Ask what is R0 when we really want Rt
2. Use the wrong model
    - E.g. modeling spread of disease assuming everyone acts approximately the same
3. Make the wrong (statistical) assumptions
    - E.g. Assuming the Central Limit Theorem applies when it does not
4. Take the wrong (optimization) paths
    - E.g. STAN vs. optim() vs. PRISM vs. SASS vs. next big thing
5. Infer the wrong conclusions
    - E.g. vaccines cause autism

We need ***reliable*** epidemic models

- We want to **predict** who, what, where, and when a disease will spread
- We want to **infer** how and why disease spreads
- What is **reliable**?
- What is **reliable enough**?

Some criteria for reliable models

1. Reproducibility
2. Plausibility
3. Diagnostics and model interrogation
4. Fitting data to model vs. fitting model to data

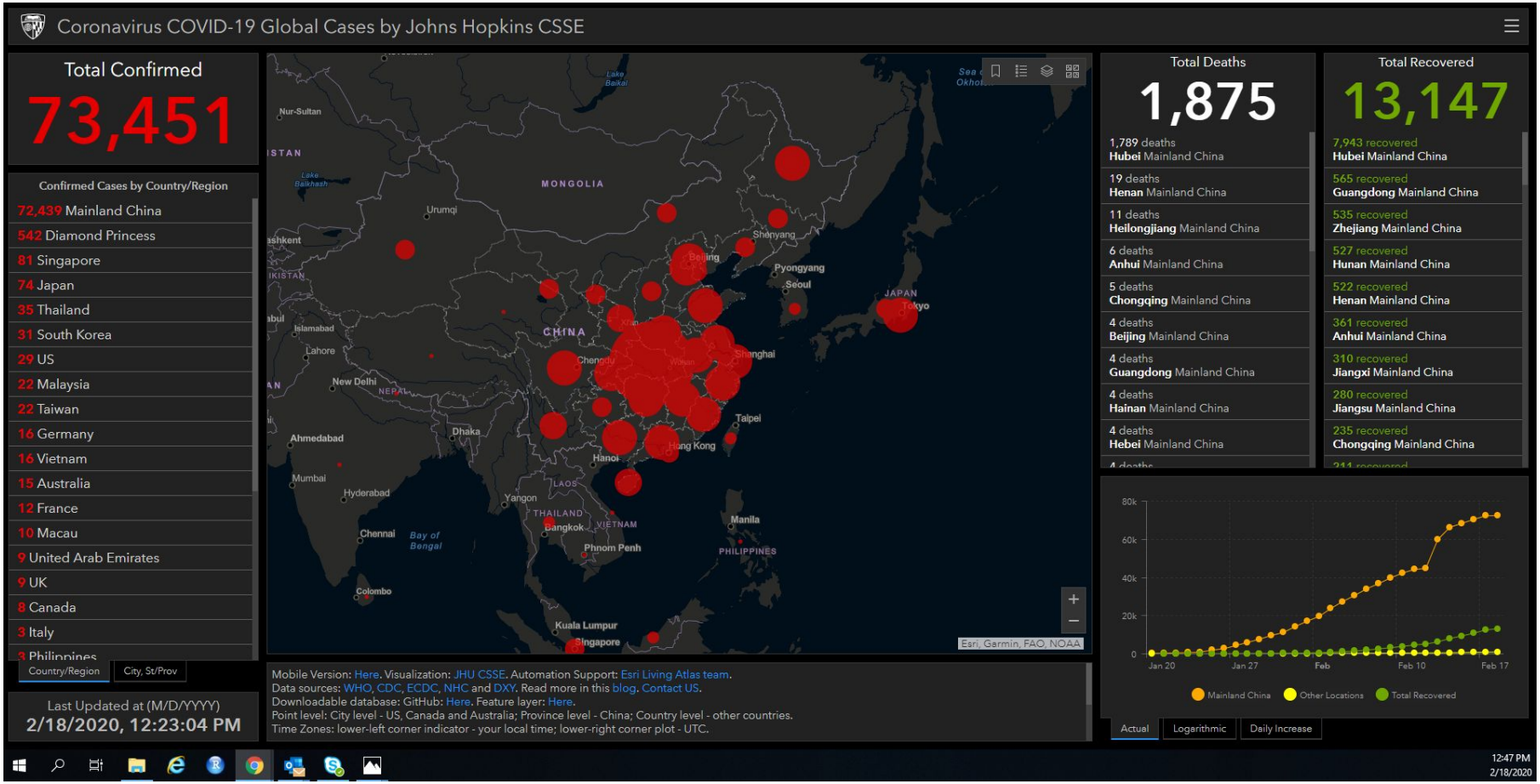
Exploratory data analysis is important

- What does your data look like?
  - Does the data tell the story?
  - (Are we sure it’s the true story?)
- What are we looking for?
- What factors can be skewing our analysis?

Ex. 2019-nCoV - Coronavirus.

- What would you like to see?

How people are visualizing spread of 2019-nCoV


Coronavirus COVID-19 Global Cases by Johns Hopkins CSSE
Total Confirmed
73,451
Confirmed Cases by Country/Region
72,439 Mainland China
542 Diamond Princess
81 Singapore
74 Japan
35 Thailand
31 South Korea
29 US
22 Malaysia
22 Taiwan
16 Germany
16 Vietnam
15 Australia
12 France
10 Macau
9 United Arab Emirates
9 UK
8 Canada
3 Italy
Country/Region
City, St/Prov
Last Updated at (M/D/YYYY)
2/18/2020, 12:23:04 PM
MONGOLIA
CHINA
Nur-Sultan
Urumqi
Beijing
Shenyang
Pyongyang
Seoul
JAPAN
Tokyo
Shanghai
Islamabad
Lahore
New Delhi
NEPAL
Ahmedabad
Dhaka
Hong Kong
Taipei
Hanoi
Mumbai
Hyderabad
Yangon
LAOS
THAILAND
Bangkok
VIETNAM
Manila
PHILIPPINES
Chennai
Bay of Bengal
Phnom Penh
Colombo
Kuala Lumpur
Singapore
Esri, Garmin, FAO, NOAA
Mobile Version: Here. Visualization: JHU CSSE. Automation Support: Esri Living Atlas team.
Data sources: WHO, CDC, ECDC, NHC and DXY. Read more in this blog. Contact US.
Downloadable database: GitHub: Here. Feature layer: Here.
Point level: City level - US, Canada and Australia; Province level - China; Country level - other countries.
Time Zones: lower-left corner indicator - your local time; lower-right corner plot - UTC.
Total Deaths
1,875
1,789 deaths Hubei Mainland China
19 deaths Henan Mainland China
11 deaths Heilongjiang Mainland China
6 deaths Anhui Mainland China
5 deaths Chongqing Mainland China
4 deaths Beijing Mainland China
4 deaths Guangdong Mainland China
4 deaths Hainan Mainland China
4 deaths Hebei Mainland China
Total Recovered
13,147
7,943 recovered Hubei Mainland China
565 recovered Guangdong Mainland China
535 recovered Zhejiang Mainland China
527 recovered Hunan Mainland China
522 recovered Henan Mainland China
361 recovered Anhui Mainland China
310 recovered Jiangxi Mainland China
280 recovered Jiangsu Mainland China
235 recovered Chongqing Mainland China
80k
60k
40k
20k
0
Jan 20
Jan 27
Feb
Feb 10
Feb 17
Mainland China
Other Locations
Total Recovered
Actual
Logarithmic
Daily Increase
12:47 PM
2/18/2020

Prediction: CDC Yearly Influenza Forecasting Contest

- Statistical methods to predict the following
  - Peak incidence
  - Peak day
  - Next three time points percent influenza like illness (ILI)
  - For 50 states and DC
  - For 28 military bases

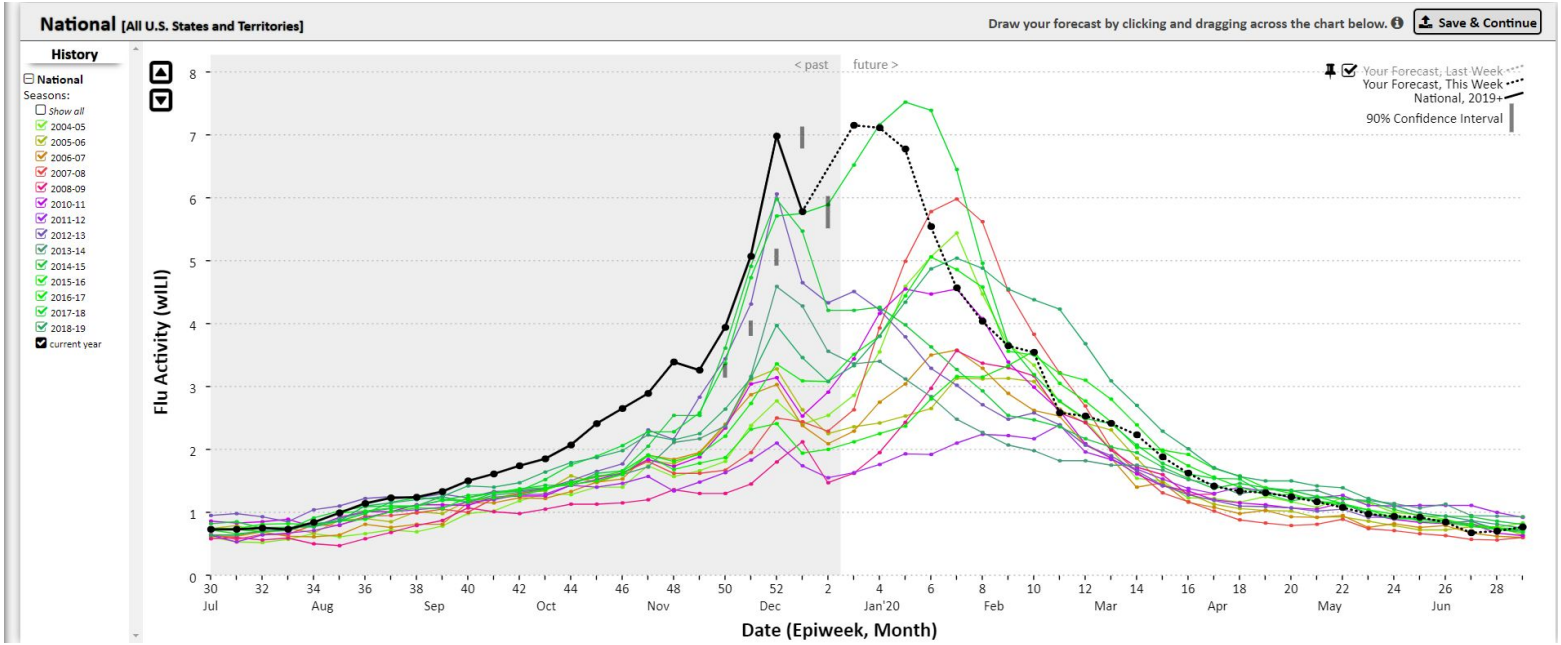
- Over 30 teams contribute
  - Carnegie Mellon University's DELPHI group led by Dr. Roni Rosenfeld has had leading forecasts now for 5 years

Available Data to make forecasts include

- CDC's reports of ILI for each state up to the given time period
  - The data is constantly changing!
- Google Flu Trends
  - (requires special permission)
- Wikipedia and Twitter data
  - Manually scraped and cleaned from the web
- Weather

Methods used for prediction

- EpiCast.org (from DELPHI group)



Some methods of prediction

- Kalman Filters
- Empirical Bayes using past flu seasons
- Time series
- Climate-based models

How are prediction models validated?

- In linear regression, e.g. we have a data that are independent of one another
  - Training and test sets of data, easy cross-validation (CV), or bootstrapping

- But for infectious disease,
  - Data set is small (e.g. only 52 points for flu in a year)
  - Very dependent on previous time steps

How can we assess our model performance?

Some possible solutions for validation

- Fit model to past/current data. Test on future
- Leave one-season out cross-validation
- A good plot is hard to find

Leave one season out cross-validation

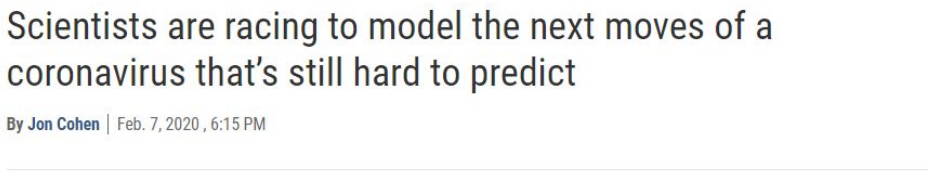
- Great for seasonal/cyclical/periodic diseases
- Assumes independence between seasons
- Requires a number of seasons to already be recorded

Do prediction models have issues?

- Overfitting!
- Many (non-parametric) methods are uninterpretable
  - With respect to the mechanistic processes of disease transmission
- Prediction is really, really hard
- Where does “machine learning” fit in?
  - Sometimes associated with very large N (observations) or very large P (parameters)
  - Neither of which we have
    - Not yet anyway

Predicting coronavirus: ongoing work

- Scientists are racing to model the next moves of a coronavirus that's still hard to predict


Scientists are racing to model the next moves of a coronavirus that's still hard to predict

By Jon Cohen | Feb. 7, 2020 , 6:15 PM

- https://www.scientificamerican.com/article/heres-how-computer-models-simulate-the-future-spread-of-new-coronavirus/


COMPUTING

Here's How Computer Models Simulate the Future Spread of New Coronavirus

They aim for clarity amid confusion surrounding the outbreak

By Jeremy Hsu on February 13, 2020

Which model(s) do we trust??

- If more models agree, do we trust these more?
- What if one model predicted a certain point better than the others?
- What would you do?

Parametric models and inference

- Using models to **explain** the process of transmission, not just predict

- Common parameters of interest
  - R0 - the initial reproduction number
  - Beta - average rate of infection
  - Gamma - average rate of recovery
  - Omega - the serial interval
  - K - number of sub-groups that behave differently
  - Final size
  - Outbreak duration
  - Vaccination threshold

The biggest challenge is making the model

- Specify how and why a disease moves through a population

- “A Thousand and One Epidemic Models” (Hethcote 1994)
  - Epidemiological compartment structure, i.e. states (susceptible, infectious, recovered)
  - Incidence and distribution of waiting times (transition eqs.)
  - Demographic structure (heterogeneity of population)
  - Epidemiological-interactions (vectors, mutations, interventions, etc.)

Kermack and McKendrick SIR Model (discrete time)

S - number of susceptible at time t

I - number of infectious at time t

R - number of recovered at time t

N = total number of people (fixed)

□ - average infection rate

γ - average recovery rate

Typically, (S(0), I(0), R(0)) are known

$$\begin{cases} \frac{\Delta S}{\Delta t} = -S \times \beta \frac{I}{N} \\ \frac{\Delta I}{\Delta t} = S \times \beta \frac{I}{N} - I \times \gamma \\ \frac{\Delta R}{\Delta t} = I \times \gamma \end{cases}$$

Epidemiological states

- S(t) - # Susceptible individuals at t
- I(t) - # Infectious individuals at t
- R(t) - # Recovered individuals at t

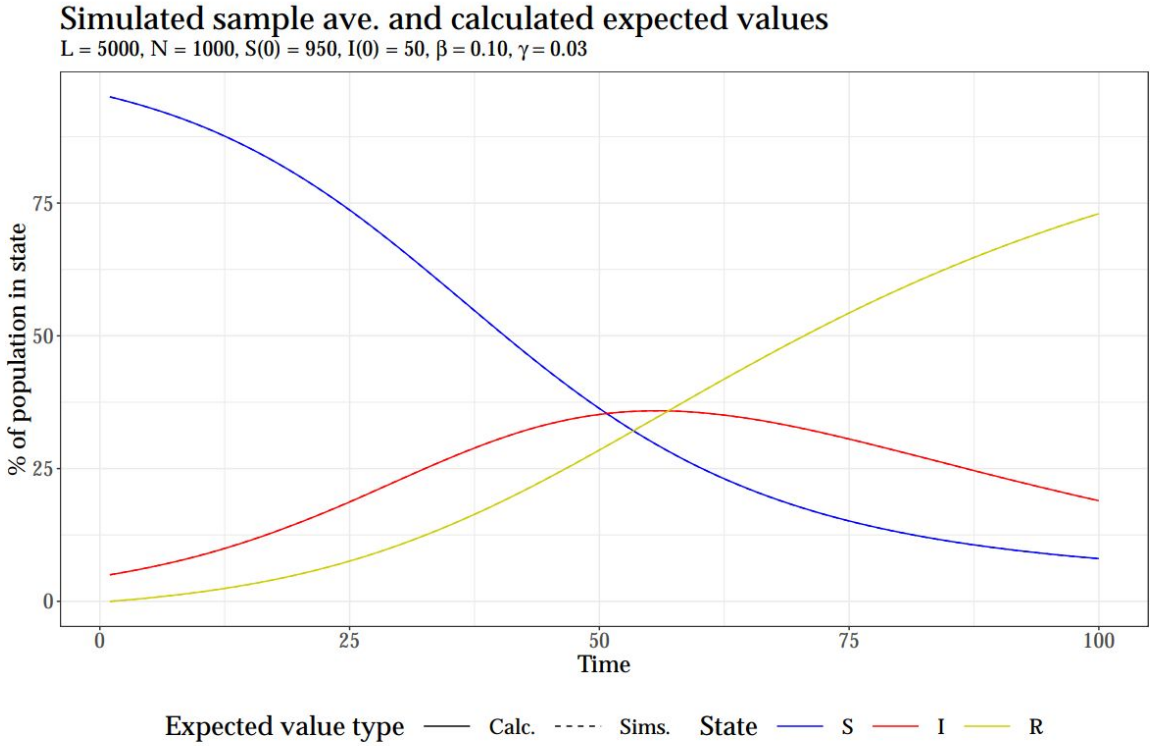
Demographics and interactions

- β – rate of infection
- γ – rate of recovery
- N – fixed population size
- $S(0), I(0), R(0)$ known

Incidence and distributions

$$\begin{cases} \frac{\Delta S}{\Delta t} & = -S \times \beta \frac{I}{N} \\ \frac{\Delta I}{\Delta t} & = S \times \beta \frac{I}{N} - I \times \gamma \\ \frac{\Delta R}{\Delta t} & = I \times \gamma \end{cases}$$

An example of a K&M SIR Model



Critical value?

R0 = □ / **γ** = 3.33
("R-naught")

R0 - the initial reproduction number

DEF: Number of expected infections when a single infector is introduced to an entirely susceptible population (Anderson and May 1992)

Adding stochasticity - one example

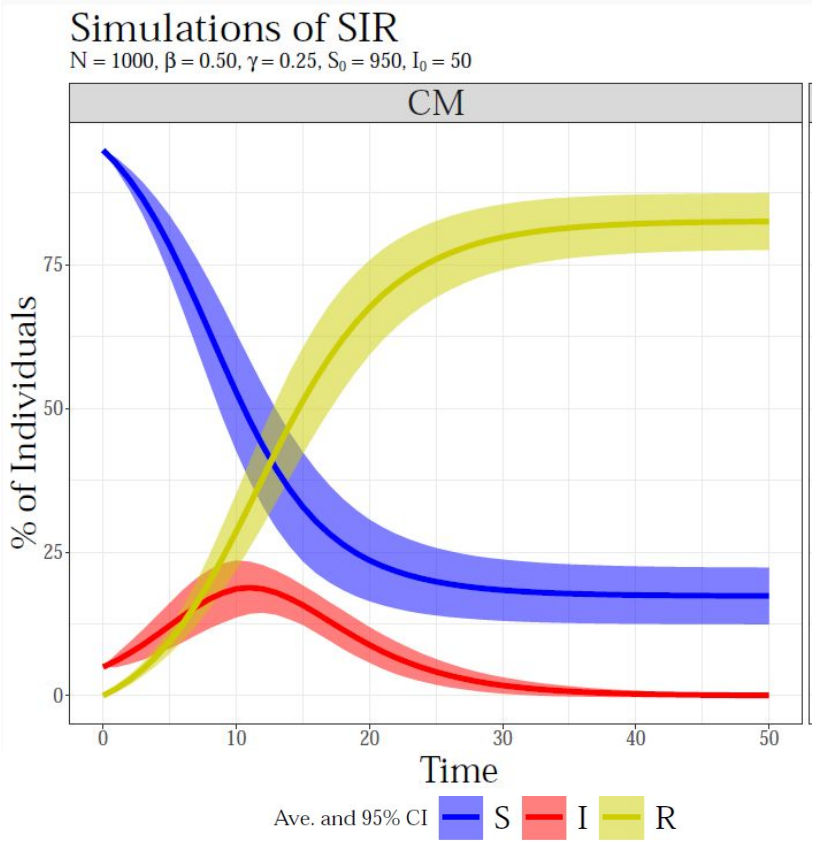
$$Z_{t-1,S}|S_{t-1}, I_{t-1} \sim \text{Binomial}\left(S_{t-1}, \beta\frac{I(t-1)}{N}\right)$$

$$Z_{t-1,R}|S_{t-1}, I_{t-1} \sim \text{Binomial}\left(I_{t-1}, \gamma\right)$$

$$S_t\,|S_{t-1}, I_{t-1} = S_{t-1} - Z_{t-1,S}$$

$$I_t|S_{t-1}, I_{t-1} = N - S_t - R_t$$

$$R_t\,|S_{t-1}, I_{t-1} = R_{t-1} + Z_{t-1,R},$$



What can R0 tell us?

Pros

- “Arguably the most important quantity in the study of epidemics” (Heesterbeek 2002)
- Tells us proportion to vaccinate 1 - 1/R0
- Tells us chance of an outbreak
- Can compare diseases to one another
- One number summary

Cons

- Plagued by recent criticism
  - Does not tell ‘full’ story
  - Does not account for dynamic parameters
  - Model dependent
- No standard way to estimate
- Confusion of whether it contains preventive measures

- Property of the Model!!!

R0 is a property of the model! SEIR vs. SIR

A disease passes through a population,

Creates **(S(t), E(t), I(t), R(t))**

Scientist **A** observes (S(t), E(t), I(t), R(t)) (**correct** SEIR)

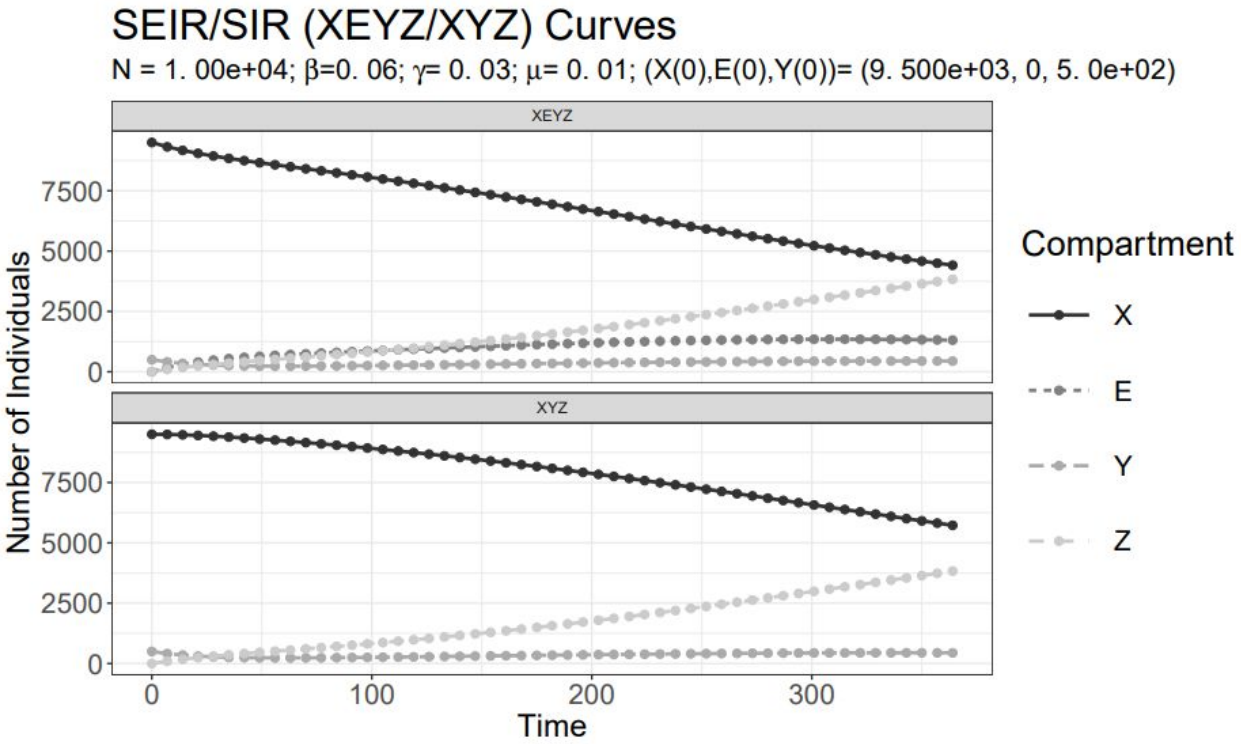
Scientist **B** observes (S(t) + E(t), I(t), R(t)) (**incorrect** SIR)

Both estimate R0

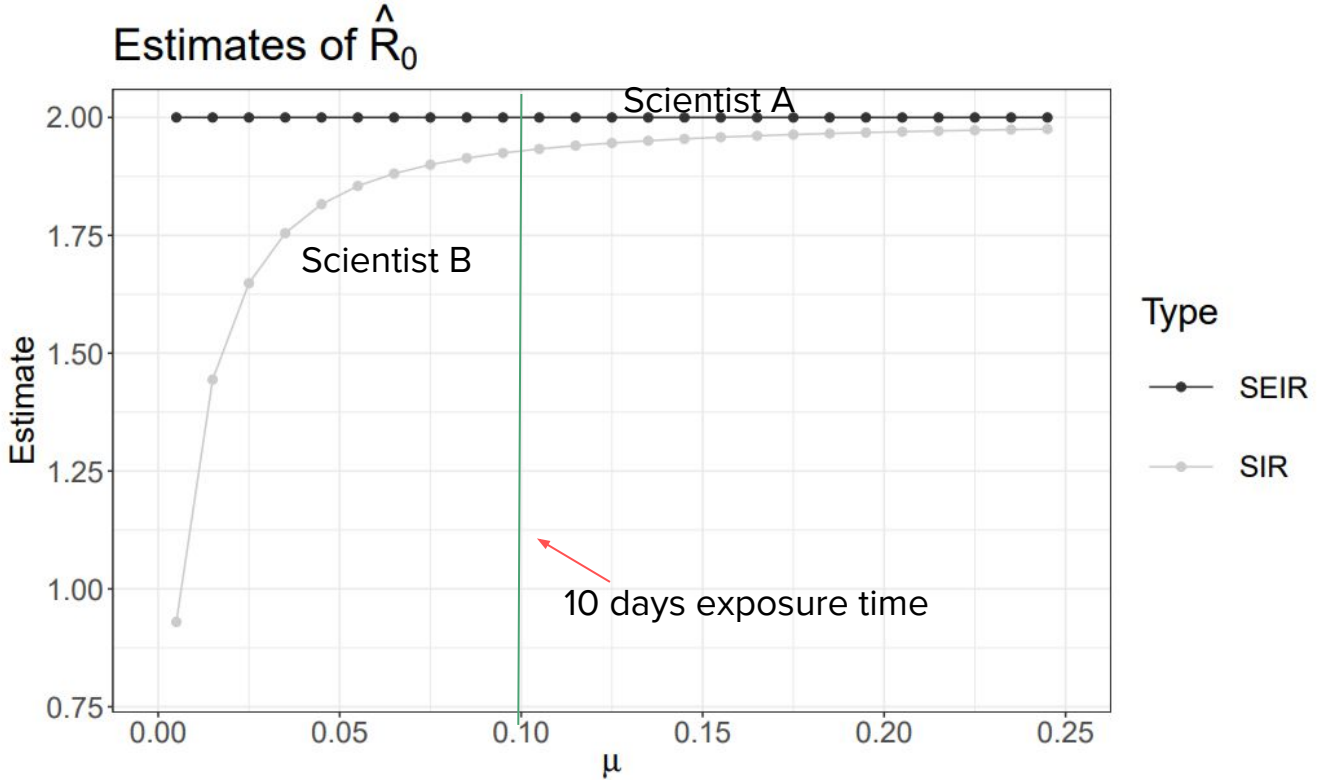
R0 is a property of the model! SEIR vs. SIR

Scientist A

Scientist B



R0 is a property of the model! SEIR vs. SIR



1/(Latent/Exposed time)

R0 estimates for common diseases

Disease	R0	Sources
Coronavirus	**??**	
Measles	(6-12)	(Guerra et al. 2017)
Zika	3.80 (2.40, 5.60)	(Towers et al. 2016)
Spanish Influenza 1918	1.32 (1.29-1.36) 1.80 (1.47-2.27)	(Camara et al. 2009) (Biggerstaff et al. 2014)
Seasonal Influenza	1.28 (1.19-1.37)	(Biggerstaff et al. 2014)
Pandemic Influenza 2009	1.46 (1.30-1.70)	(Biggerstaff et al. 2014)
Ebola (Guinea)	1.51 (1.50-1.52)	(Althaus 2014)

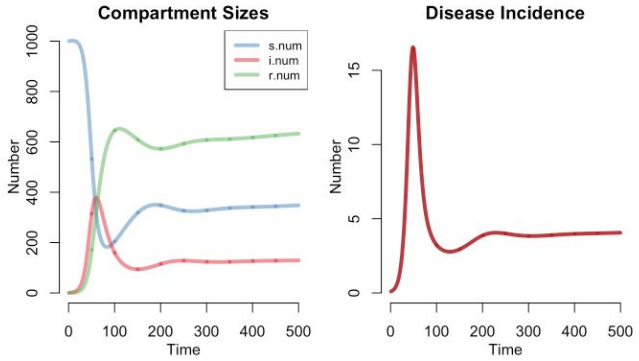
Accessible software for infectious disease modelling

- Partially observed Markov processes (An R package)
  - kingaa.github.io/pomp
  - https://kingaa.github.io/pomp/vignettes/pompjss.pdf
- EpiModel
  - https://www.epimodel.org/

Plotting

Next we plot the results of the model to demonstrate several plot arguments. First, the `par` function is used to change some default graphical options. In the left plot, the `popfrac=FALSE` argument plots the compartment size (rather than prevalence) and `alpha` increases the transparency of the lines for better visualization. By default, the `plot` function will plot the prevalences for all compartments in the model, but in the right plot we override that using the `y` argument to specify that disease incidence (the `si.flow` element of the model object) should be plotted.

```
par(mar = c(3.2, 3, 2, 1), mgp = c(2, 1, 0), mfrow = c(1, 2))
plot(mod, popfrac = FALSE, alpha = 0.5,
     lwd = 4, main = "Compartment Sizes")
plot(mod, y = "si.flow", lwd = 4, col = "firebrick",
     main = "Disease Incidence", legend = "n")
```



It is possible to specify a single line color, a vector of colors, or a color palette using the `col` argument, and the legend options are set using the `legend` argument.

Takeaways from R0

- Way to compare severity of diseases with a one number summary
- Take with a grain of salt

Compartment models

- Origins from pharmacokinetics - tracked blood flow through different *compartments* of heart

- Describes how objects in discrete compartments/states move from one state to the next

- Essential parts
  - Disease states
  - births/deaths (population dynamics)
  - Sub-groups? (females vs. males, children vs. adults)
  - Transitions between states

Some CM examples

SI: susceptible-infectious.

SIR: susceptible-infectious-recovered

SIS: susceptible-infectious-susceptible

SEIR: susceptible-exposed-infectious-recovered

SEIFHR: susceptible-exposed-infectious-funeral-hospitalized-recovered

The Fundamental Assumption of CMs

Individuals in the same compartment at time t are **indistinguishable** from one another

Implications:

- Cannot track individuals through a disease
- Need individuals in the same state to act approximately in the same manner
- Don’t quite need independence of individuals but close to it
  - ‘exchangeability’

Fitting a CM to data

- I have an SIR model
  - But do I have SIR data?

Data: What I want

Time	# S	# I	# R
0	99	1	0
1	94	5	1
2	84	7	9

Data: What I have

Time	# New cases
0	1
1	5
2	10

Agent-based models

- Like the SIMs but less fun

ID	Country	Year	Occupation	Age	Gender	Environment	ID	Environment	Capacity	Latitude	Longitude
P1	U.S.	2010	Statistician	30	M	E1	E1	PPG Paints Arena	100	40.4396	−79.9893
P2	U.S.	2010	Data scientist	54	F	E2, E3	E2	Ballard High School	4000	40.4474	−79.9498
P3	U.S.	2010	Bagpiper	56	M	E1, E2, E3	E3	Carnegie Mellon	50	40.4428	−79.9430

Example of agents and their environment

Agent-based models: essential parts

- Agents
- Environment
- Interactions
- You make the rules!

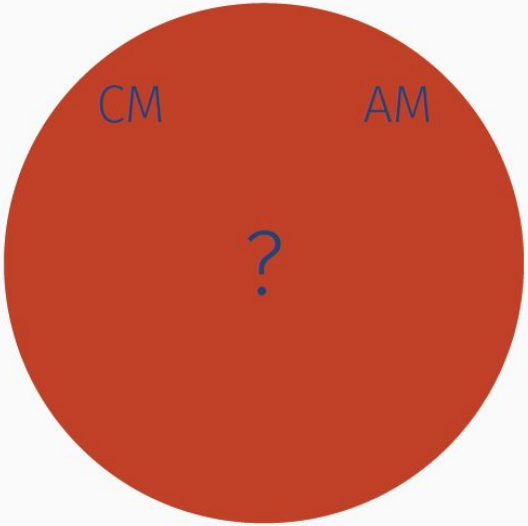
Roller Coaster Tycoon. An agent-based model?

Limitations and examples of agent-based models

- They can be glacially slow
- Model calibration?
  - Do our simulations mean anything?
  - The problem of ‘docking’ (Epstein 2007)
- There are some very interesting ABMs and researchers out there
  - Los Alamos National Lab (TRANSIMS and more)
  - University of Virginia Biocomplexity Institute (See this article)
  - FRED

CMs vs AMs

Quality	CM	AM
1. Interpretable	✓	✓
2. Accessible	✓	✓
3. Modular	✓	✓
4. Individual info	✓	✓
5. Fast computer run time	✓	✓
6. Low computer memory	✓	✓
7. Theory	✓	✓
8. Parameter estimation	✓	✓
9. Statistical software	✓	✓



Are these classes statistically the same?

Theorems - yes, they are the same in some ways

Theorem 1: Given deterministic transition matrix $D(t)$ of size $K \times K$, there exists a stochastic CM-AM pair such that $X^{CM} \overset{d}{=} X^{AM}$ and the models are unbiased w.r.t $D(t)$

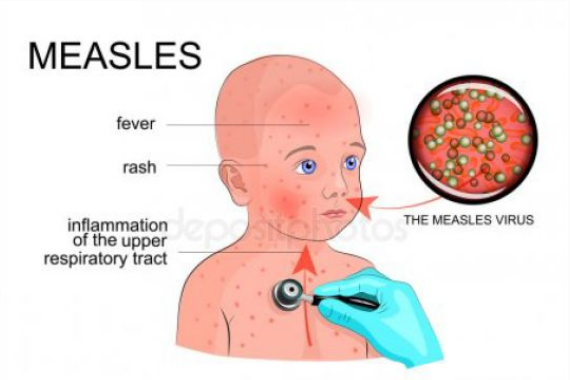
- K is the number of states
- $D_{ij}(t)$ is the non-negative # of individuals moving from state i to j from time t to $t+1$
- Row sums are total number individuals moving out of state i
- Column sums are total number of individuals moving into state j
- $D(t) - D^T(t)$ gives back the original difference equations

Ex. SIR $D(t)$:

$$D(t) = \begin{pmatrix} S(t) - \beta S(t)\frac{I(t)}{N} & \beta S(t)\frac{I(t)}{N} & 0 \\ 0 & I(t) - I(t)\gamma & I(t)\gamma \\ 0 & 0 & R(t) \end{pmatrix}$$

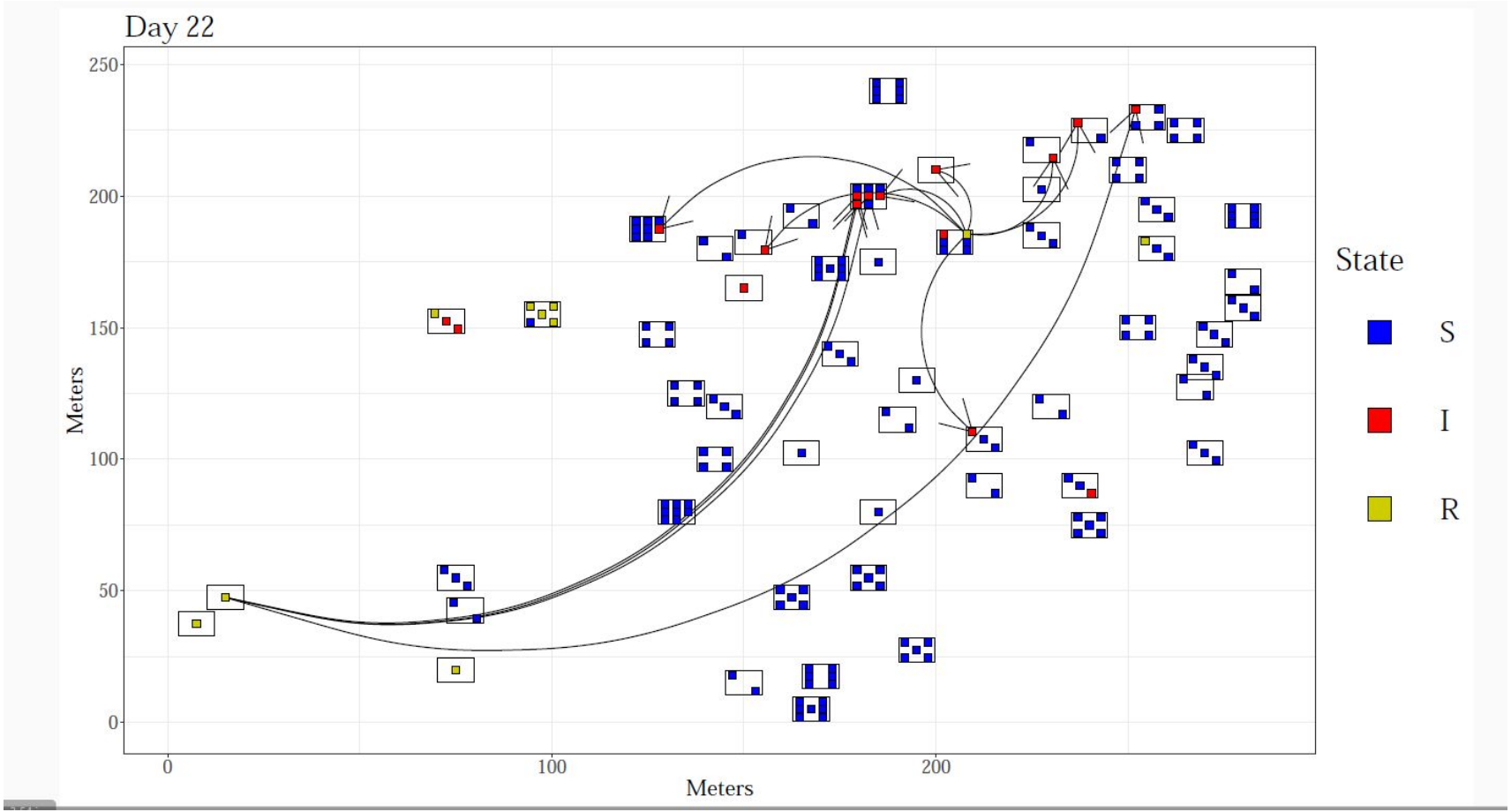
Hagelloch -- measles outbreak in 1861

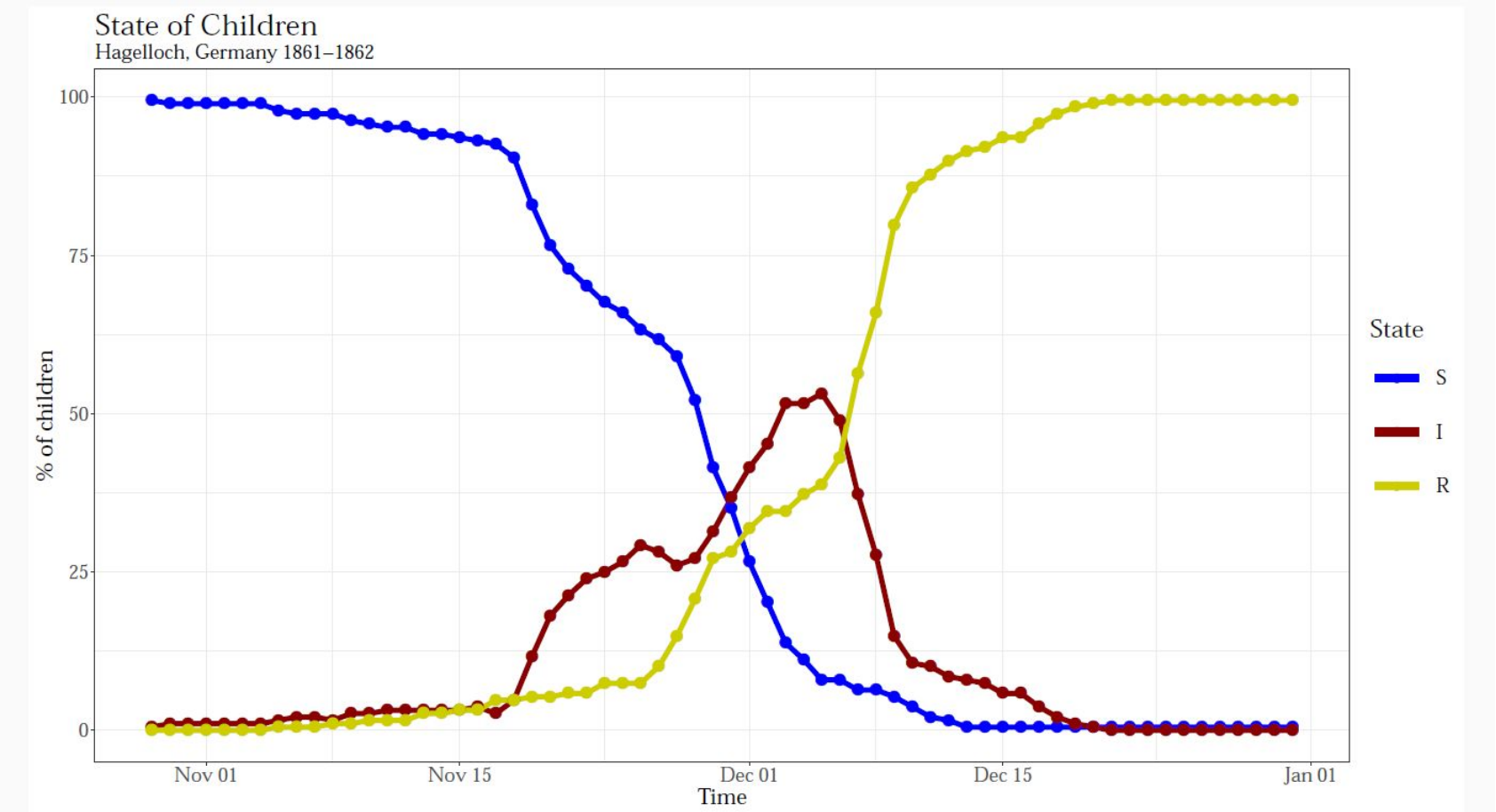
- Highly infectious childhood disease ($\mathcal{R}_0 = 19$) (Anderson & May, 1992)
  - Influenza $\mathcal{R}_0 \approx 1.2$
- Prodromes – initial symptoms
  - high fever, cough, runny nose, red, watery eyes
  - 2-3 days after, tiny white spots in mouth
- Measles rash and high fever: 3-5 days after symptoms begin
- 2-3 days after rash, child recovers
- CDC reports person is infectious ± 4 days after rash appearance
- Lifelong immunity after infection



Measles: data from R surveillance package

ID	Household	Class	Age	Sex	I	R	Infector
1	61	1st	7	F	22	29	45
2	61	1st	6	F	23	32	45
3	62	pre-K	4	F	29	37	NA
4	63	2nd	13	M	27	32	180
5	63	1st	8	F	22	31	45


Day 22
250
200
150
100
50
0
Meters
0
100
200
Meters
State
S
I
R

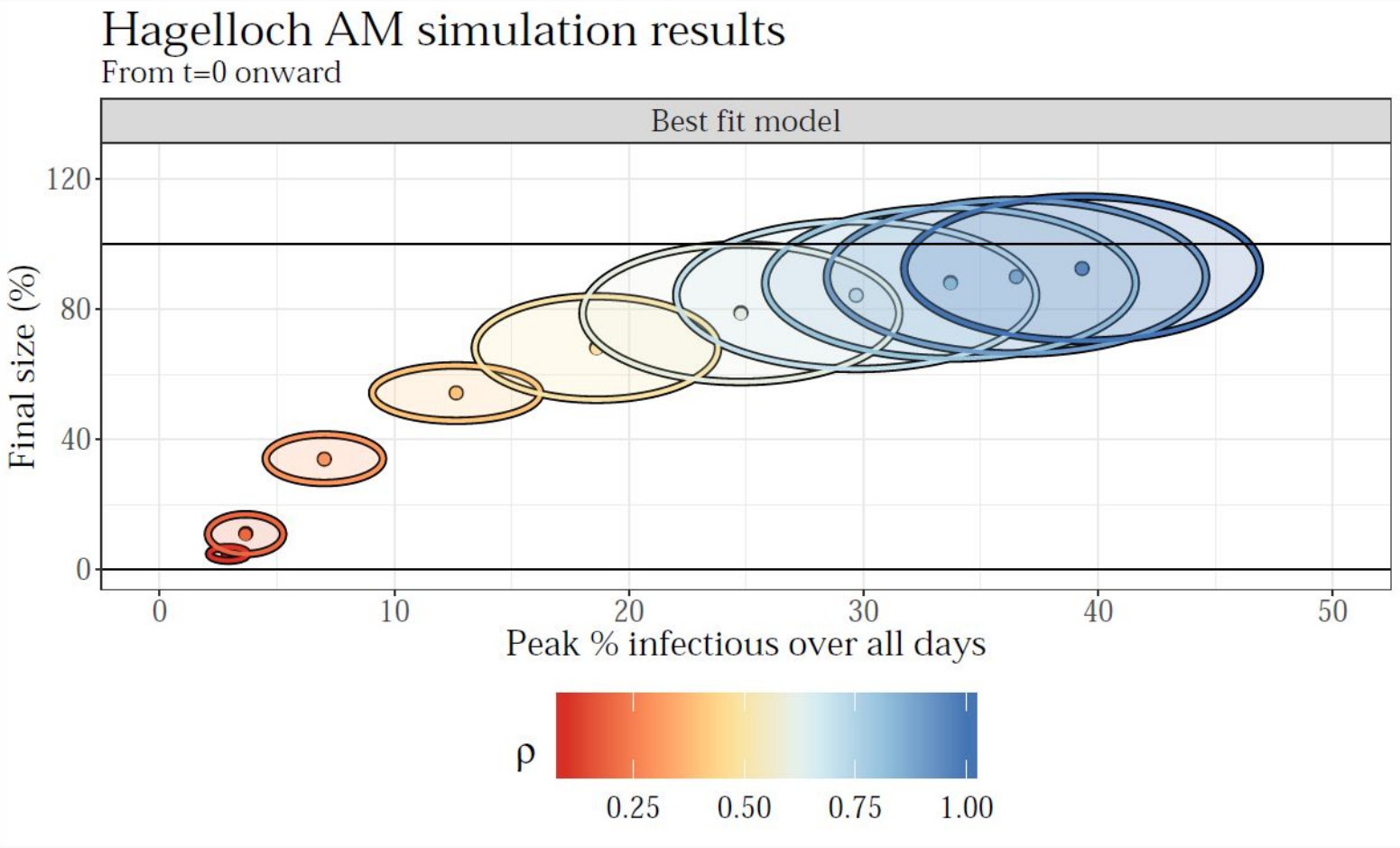

State of Children
Hagelloch, Germany 1861–1862
100
75
50
25
0
% of children
Nov 01
Nov 15
Dec 01
Dec 15
Jan 01
Time
State
S
I
R

Scenario 1

1. We have estimate(s) of $\hat{\beta}$, the infection parameter
2. Assume we can reduce infectivity to $\rho \cdot \hat{\beta}$
3. How would outbreak have changed?

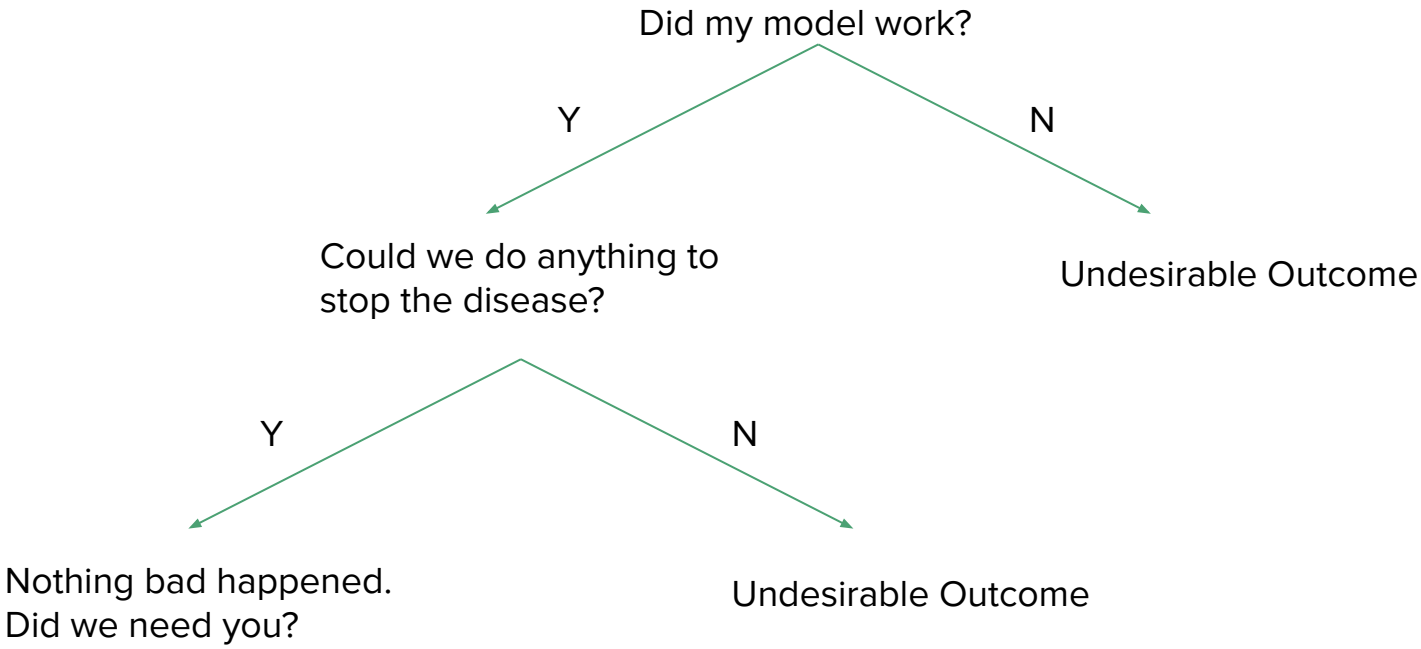
Analysis 1

1. Initialize our CM-AM pair with estimates
2. Vary ρ in our simulations
3. Analyze resulting outbreaks


Hagelloch AM simulation results
From t=0 onward
Best fit model
Final size (%)
120
80
40
0
0
10
20
30
40
50
Peak % infectious over all days
ρ
0.25
0.50
0.75
1.00

- What is K^*, the minimal number of states?
  - $K^* = 6$
- What is $\mathcal{R}_0$?
  - Between 4-5.
- What is the associated CM-AM pair?
  - $S^2I^2R^2$ with groups before and after $t = 25$
- What would have happened...
  - if we reduced the infectivity of the disease?
    - Want to reduce $\hat{\beta}$ by about half
  - if we isolated infectious individuals?
    - Reduce size of epidemic, even if isolated 8 days after initial infection
  - if we shut down the school?
    - Inconclusive results due to assumptions of model

Epidemic disease modelling is a thankless task



Upshot of our brief tour of all of infectious diseases

- There are many good resources/classes/software out there
  - But the individual effort is more important than ever
- Make good decisions
- Sensitivity analysis and uncertainty analysis are vital

How would you like to model coronavirus?

Some references

EDA

Carroll, L. N., Au, A. P., Detwiler, L. T., Fu, T. C., Painter, I. S., & Abernethy, N. F. (2014). Visualization and analytics tools for infectious disease epidemiology: a systematic review. *Journal of biomedical informatics*, *51*, 287-298.

Prediction

Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., & Rosenfeld, R. (2015). Flexible modeling of epidemics with an empirical Bayes framework. *PLoS computational biology*, *11*(8).

Shaman, J., Pitzer, V. E., Viboud, C., Grenfell, B. T., & Lipsitch, M. (2010). Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biol*, *8*(2), e1000316.

Inference

Abbey, H. (1952). An examination of the Reed-Frost theory of epidemics. *Human biology*, *24*(3), 201.

Anderson, R. M., Anderson, B., & May, R. M. (1992). Infectious diseases of humans: dynamics and control. Oxford university press.

Heesterbeek, H., Anderson, R. M., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., ... & Hollingsworth, T. D. (2015). Modeling infectious disease dynamics in the complex landscape of global health. *Science*, *347*(6227), aaa4339.

Daley, D. J., & Gani, J. (2001). Epidemic modelling: an introduction (Vol. 15). Cambridge University Press.

Mishra, S., Fisman, D. N., & Boily, M. C. (2011). The ABC of terms used in mathematical models of infectious diseases. *Journal of Epidemiology & Community Health*, *65*(1), 87-94.

R0

Hethcote, H. W. (2009). The basic epidemiology models: models, expressions for R0, parameter estimation, and applications. In Mathematical understanding of infectious disease dynamics (pp. 1-61).

Van den Driessche, P., & Watmough, J. (2008). Further notes on the basic reproduction number. In Mathematical epidemiology (pp. 159-178). Springer, Berlin, Heidelberg.

Agent-based models

Epstein, J. M. (2006). Generative social science: Studies in agent-based computational modeling. Princeton University Press.

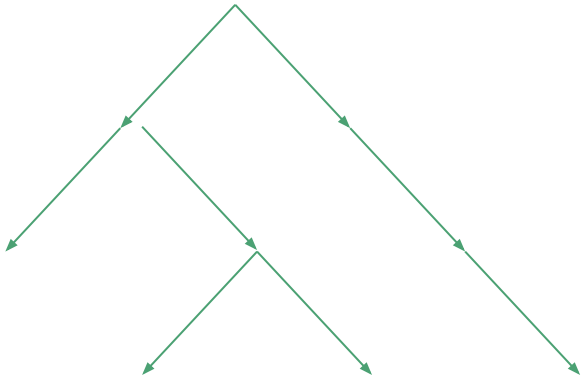
Eubank, S., Guclu, H., Kumar, V. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., & Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. Nature, 429(6988), 180-184.

Grefenstette, J. J., Brown, S. T., Rosenfeld, R., DePasse, J., Stone, N. T., Cooley, P. C., ... & Guclu, H. (2013). FRED (A Framework for Reconstructing Epidemic Dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. BMC public health, 13(1), 940.

Reconstructing disease transmission chains

Branching Processes:

- Concept of generations
- time is non-factor



How we do model this?

- Classically assumes infinite population-
- geometric distribution of number of infections

Chain Binomials - taking into account finite population

Chain Binomials:

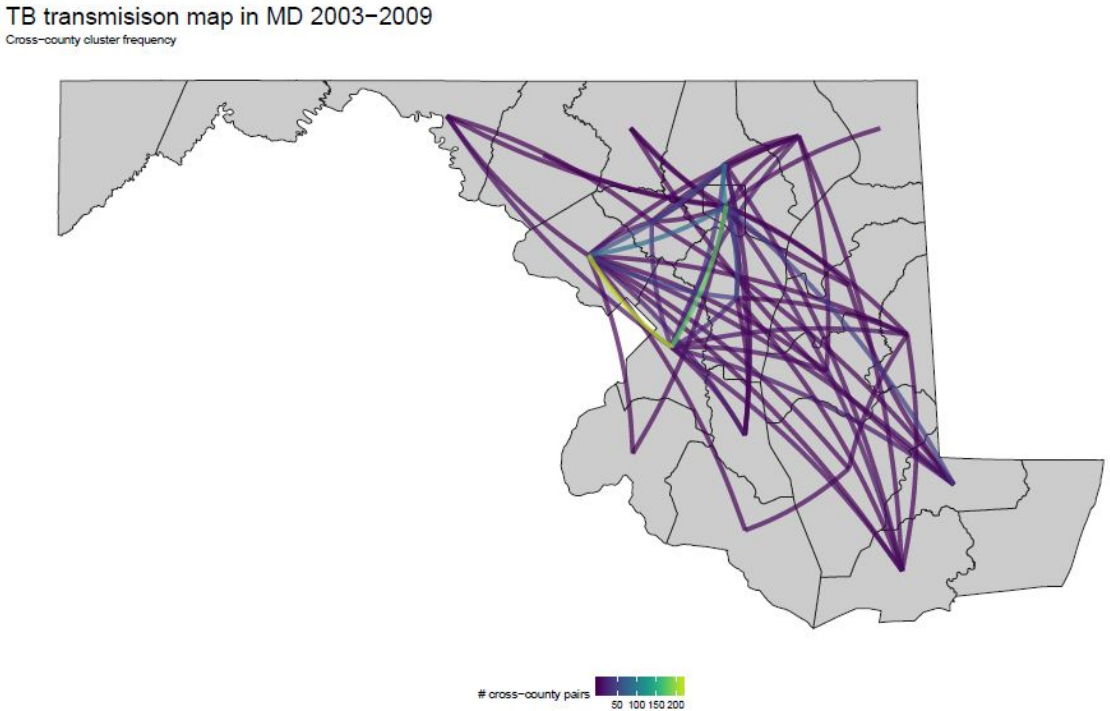
P(#new inf = y) = Binomial(N, 1 - (1-p)^I)

N = number of susceptibles

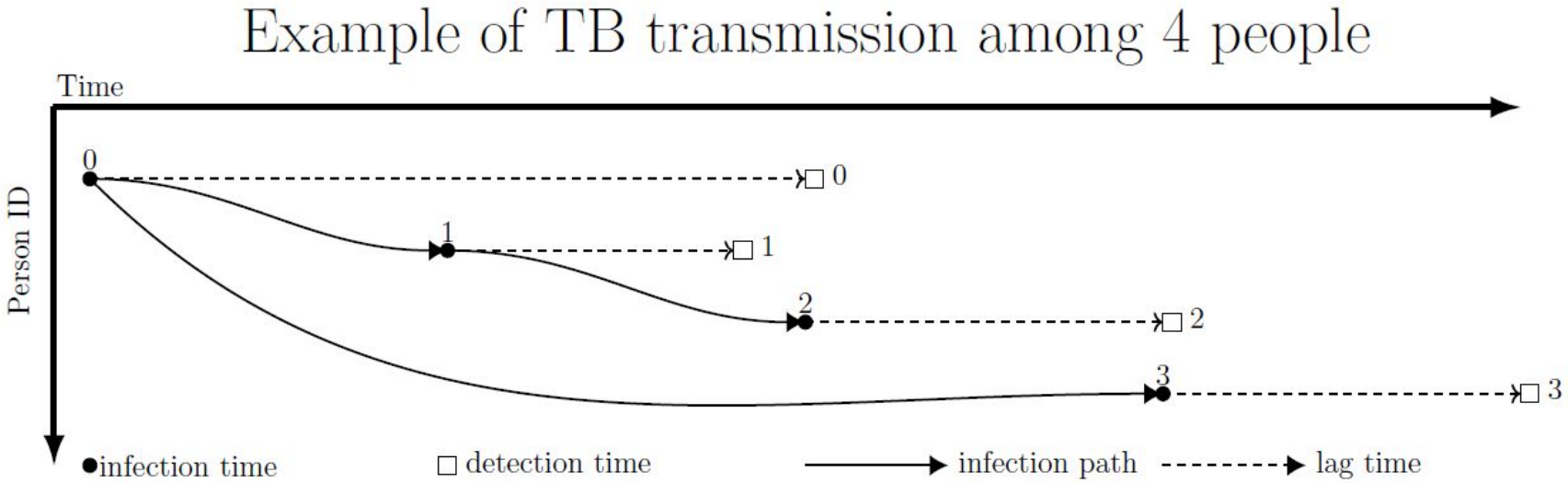
p = probability of infection from one infectious contact

I = number of infectious contacts

TB in MD


TB transmisison map in MD 2003−2009
Cross−county cluster frequency
cross−county pairs
50 100 150 200

TB transmission example



Overarching issues with transmission trees

- Computational tractability: number of trees of size n is n^(n-2) - yikes!
  - 10^(10-2) = 100 million!!
  - 25^(23) ≍ 10^32 (forget about it)
  - Need to do approximate sampling, MCMC, ABC, or other

- Time vs. generations
  - Could have infections occurring at same time but in completely different generations
  - Treatment date != infection date
  - Latent/exposure periods?
  - Underreporting
  -