

Coping with Heterogeneity and Uncertainty of COVID-19 Datasets

Issues, Approaches, and Consequences of the COVID-19 Crisis

Ambuj K Singh
Yu-Xiang Wang

University of California
Santa Barbara

April 28, 2020



UC SANTA BARBARA
Data Science Initiative

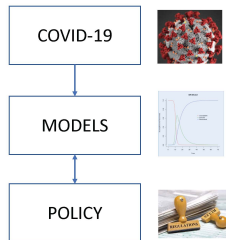


spatial@ucsb
CENTER FOR SPATIAL STUDIES



Previous presentations

- (April 14) SARS-CoV2: The Virus and the Disease (Carolina Gonzalez and Lynn Fitzgibbons)
- (April 21) Current Epidemiological Models: Scientific Basis and Evaluation (Francesco Bullo)



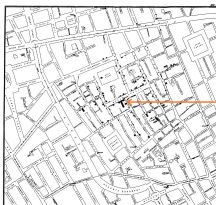
Thanks to doctors, nurses, first responders, and everyone in the hospital/healthcare/biomedical ecosystem. And thanks to everyone else for adapting to the new lifestyle.

Lessons from history

- (1854) Cholera outbreak in London
- (1918–19) Spanish Flu
- (2013–16) Western African Ebola virus epidemic

How have data and models informed decision making in past epidemics?

Tracing the origin of a cholera outbreak




- In August of 1854, Soho, a suburb of London, was hit hard by a terrible outbreak of cholera.
- Dr. John Snow had long believed that water contaminated by sewage was the cause of cholera. A challenge to the theories of disease transmission then.
- Through 'contact-tracing' of water (homes, restaurants, coffee shops, breweries), he established that the source was the pump on Broad Street. Shown is a memorial in his honor.

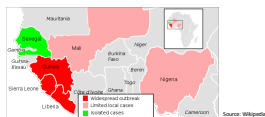
Interventions in Spanish Flu epidemic of 1918–19



- The most severe pandemic in recent history.
- About 500 million people or one-third of the world's population became infected with this virus. The number of deaths was estimated to be at least 50 million worldwide with about 675,000 in the US.
- Hatchett et al analyzed Non-Pharmaceutical Interventions (NPI) in 17 US cities¹.
- Findings support the hypothesis that rapid implementation of multiple NPIs can significantly reduce disease transmission, but that spread will be renewed upon relaxation of such measures.

¹Richard J. Hatchett, Carter E. Mecher, and Marc Lipsitch. “Public health interventions and epidemic intensity during the 1918 influenza pandemic”. In: *Proceedings of the National Academy of Sciences* 104.18 (2007), pp. 7582–7587. 

Western African Ebola virus epidemic (2013–16)



- CDC model predicted 500K–1.4 million infections. A year later there were about 25K Ebola cases, including 10.3K deaths. (Scientific American, Dec 8 2014)
- Prediction was more than 50 times worse. Why?
 - Errors in deterministic modeling²
 - Difficult to model behavioral change.
 - Of course, there were many good predictions post-hoc.
“It is difficult to make predictions, especially about the future.”
- “Ebola defied the prophets of doom. It never went airborne, and its economic effects were less painful than expected. Being wrong rarely feels this good. But it will be harder to catch the world’s attention next time.” Economist, Feb 2015.

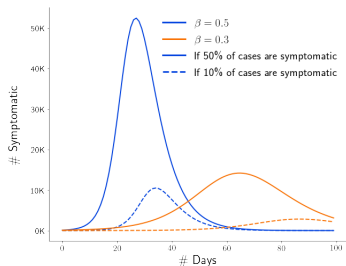
²Aaron A. King et al. “Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola”. In: *Proc. R. Soc. B* (2015).

Talk outline

- Historical context of models and interventions
- COVID-19 data
 - Heterogeneous
 - Uncertain
 - Dynamic
- Measuring the effect of NPIs (Non-Pharmaceutical Interventions)
- How do current models use available data?
- Contact tracing
- Uncertainty and bias in data and models
- Case study: fitting SIR model to Santa Barbara Data
- Challenges and opportunities

COVID-19 data is heterogeneous, uncertain, and dynamic

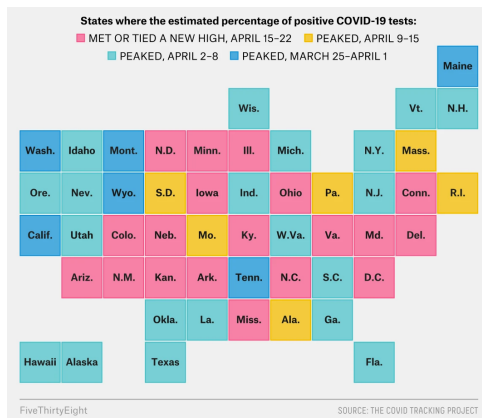
- **Heterogeneous:** Risk factors vary at the individual level (presence of comorbidity). Different countries (and different regions within a country) have different outcomes based on geography, culture, and the human element. Many confounding variables.
- **Uncertain:** Testing error of biomedical tests, uncertainty of transmission fraction, contact rate, recovery rate, 'asymptomatic' vs 'presymptomatic' diagnosis, biased sampling
- **Dynamic:** Tightening and loosening of interventions, behavioral changes, disease dynamics at the individual and population level



Observed *symptomatic* individuals in a homogeneous population of 450K (SIR model), initially with 200 symptomatic and 300 recovered, with $\gamma = 0.2$.

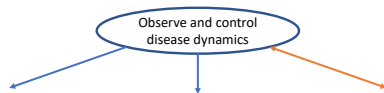
More in the presentation by Yu-Xiang

State of COVID-19 disease



- US is heterogeneous with respect to the disease.
- Disease state can be measured in multiple ways: percent change in infections, death rate, percent change in positive test rate, percent change in immunity, ..

Available data and models



Biomedical

- 1 Transmission fraction
- 2 Incubation period
- 3 Recovery rate
- 4 Death rate
- 5 Asymptomatic fraction
- 6 Immunity to future infections
- 7 Age effects
- 8 Seasonal effects

Population data

- 1 Population density
- 2 Age distribution
- 3 # Hospital admissions
- 4 # ICU admissions
- 5 # Infections
- 6 # Recovery
- 7 # Deaths
- 8 # Immune
- 9 # Doubling rate

Contact data

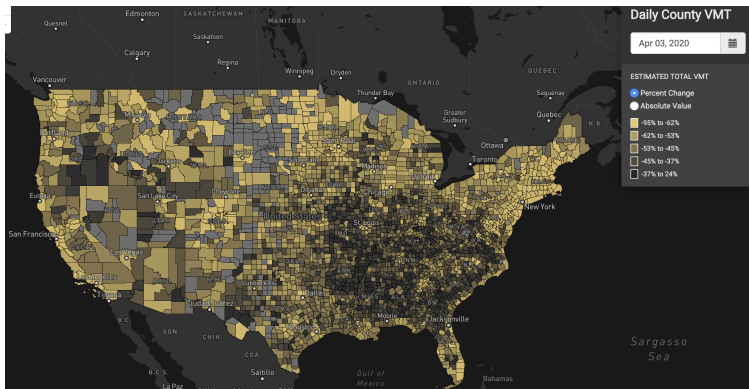
- 1 Stratified contact rates
- 2 Human mobility
- 3 Highway traffic
- 4 Air traffic
- 5 Super-spreaders
- 6 Business closures
- 7 School closures

Sources of data and knowledge

- Data about the epidemic is being made public: nCOv-2019, COVID-19, CORD-19, WHO COVID-2019, COVID-19 Tweet IDs, US Census, Facebook population density map, Covidtracking.com.
 - John Hopkins CSSE curates a worldwide dataset with daily numbers down to the county level³
 - **More in the presentation by Yu-Xiang**
- Since policy-making depends so much on models, we should also make them public.
 - If more data improve models, so does allowing people to look under their bonnets... As well as allowing for expert critique, it is a valuable way of building up public trust. (Economist, April 4, 2020).
- Peer review system is under strain
 - 2319 articles (1823 medRxiv, 496 bioRxiv) since January 19; about 24 per day
 - Challenge of getting right information to public quickly

³<https://github.com/CSSEGISandData/COVID-19>

Measuring the effect of interventions



A number of datasets can be used for measuring the effect of NPIs:

- Facebook social connectedness index
- Google COVID19 community mobility reports
- Streetlightdata (shown above)

Challenge of dealing with uncertainty

- After hearing an economist talk about his forecast's uncertainties and why a range of estimates was needed, President Lyndon B. Johnson reportedly said: "Ranges are for cattle, give me a number."
- "Epidemiology is a science of possibilities and persuasion, not of certainties or hard proof" (New Yorker, April 26, 2020)
- Models typically rely on multiple uncertain inputs whose interactions are difficult to analyze.
- Behavioral changes are especially hard to anticipate (earlier Ebola example).

How do current models use data?

- CHIME model⁴
- Institute for Health Metrics and Evaluation model (IHME)⁵
- Imperial College model⁶
- Multi-level models⁷

⁴COVID-19 Hospital Impact Model for Epidemics (CHIME).

<https://penn-chime.phl.io/>.

⁵Christopher Murray. “Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months”. In: *medRxiv* (2020). DOI: 10.1101/2020.03.27.20043752. eprint: <https://www.medrxiv.org/content/early/2020/03/30/2020.03.27.20043752.full.pdf>. URL:

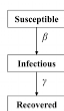
<https://www.medrxiv.org/content/early/2020/03/30/2020.03.27.20043752>.

⁶Neil Ferguson et al. “Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand”. In: (2020).

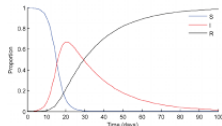
⁷Duygu Balcan et al. “Multiscale mobility networks and the spatial spreading of infectious diseases”. In: *Proceedings of the National Academy of Sciences* 106.51 (2009), pp. 21484–21489. DOI: 10.1073/pnas.0906910106.

CHIME model

- One of two models endorsed to predict hospital capacity in California.
- Based on SIR model
 - Number of current hospitalizations used to predict the number (and rate) of current infections
 - This and observed doubling rate is used to predict the infection rate.
 - Use empirical estimate of recovery rate
- Predict rate of hospital admits, rate of ICU admits, ventilator need
- Can change the dynamics by “social distancing,” realized by a percent reduction in infection rate
- Assumes homogeneity
- Deterministic model



$$\begin{aligned}\frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I\end{aligned}$$



Imperial College model

- Influenced the policy of US and UK governments
- Stochastic simulation based model based on SIR. Parameters learnt from observed data.
 - Age specific variations
 - A fraction of asymptomatic individuals
- Distribute population into age groups
- Contact rates: home, work, school, random
 - Each NPI affects these rates in different ways
- Scalability is an issue
- Can one learn directly from data?

Institute for Health Metrics and Evaluation (IHME) model

- Parameterized matching of time series from Wuhan (and other locations that have recovered) based on the number of reported deaths and the level of NPI.
 - Matching adjusted for age distribution
 - Considers four NPIs: school closures, non-essential business closures, stay-at-home recommendations, and travel restrictions.
 - Three parameters are learned: maximum death rate, inflection point, and a growth parameter.
- Spatial resolution does not allow city-level predictions.
- Sensitive to reported deaths: can be quite different from actual deaths⁸, and can also be affected by the quality of care.
- Difficult to model new interventions.
- Hard to change predictions quickly (over a few days) and accuracy of predictions may not improve as the forecast horizon decreases⁹.

⁸ "Tracking Covid-19 excess deaths across countries". In: *Economist* (Apr. 2020).

⁹ Roman Marchant et al. "Learning as We Go: An Examination of the Statistical Accuracy of COVID19 Daily Death Count Predictions". In: *medRxiv* (2020). DOI: [10.1101/2020.04.11.20062257](https://doi.org/10.1101/2020.04.11.20062257).

Multi-scale mobility models

- Hierarchical model most useful at the level of a country or a large region¹⁰.
- Lowest level dynamics based on SIR (or SEIR with an intermediate 'Exposed' state).
- Middle level dynamics based on traffic between cities: learnt based on actual data or gravitational model.
- Top level dynamics based on air traffic.
- Successfully modeled the effect of air travel on the transmission from Wuhan¹¹.
- **Homogeneous model at the lowest level.**

¹⁰Duygu Balcan et al. "Multiscale mobility networks and the spatial spreading of infectious diseases". In: *Proceedings of the National Academy of Sciences* 106.51 (2009), pp. 21484–21489. DOI: [10.1073/pnas.0906910106](https://doi.org/10.1073/pnas.0906910106).

¹¹Matteo Chinazzi et al. "The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak". In: *Science* 368.6489 (2020), pp. 395–400. DOI: [10.1126/science.aba9757](https://doi.org/10.1126/science.aba9757).

Contact tracing

- A requirement by the CDC for opening up the country¹².
- Manual implementation: in-depth interviews with those who may be infected/exposed. Number of public health employees needed: 100,000–300,000.
- Different combinations of manual and phone-based systems deployed in New Zealand, Taiwan, Singapore (TraceTogether), Australia (COVIDSafe), India (Aarogya Setu), and South Korea.
 - The number of new cases in New Zealand has dropped to single digits.
- Care19 mobile app in North Dakota, South Dakota, Utah.
- Plans for extensive phone-based tracing (Google and Apple).
 - Should contact trace include GPS?
 - Missing data and false positives
 - Privacy

¹²CDC. *Contact Tracing : Part of a Multipronged Approach to Fight the COVID-19 Pandemic*. <https://www.cdc.gov/coronavirus/2019-ncov/php/principles-contact-tracing.html>. 2020.

Can NPIs work on campus?

- Only a personal opinion but do see
 - the opinion piece by Christina Paxson: “College Campuses Must Reopen in the Fall. Here’s How We Do It” (NY Times, April 26, 2020).
- Automated contact tracing and GPS data by opt-in; manual otherwise
- Prior estimate of individual’s disease state based on
 - mobility pattern
 - disease state of contacted individuals
 - disease state of visited places
- Build a contact network of around 35K nodes with attributes ‘at-risk’ and ‘disease state’
- Test individuals on a continuous basis with priority to ‘at-risk’ and ‘hub nodes’ in the network
 - ensure reproduction number < 1 at each node
- Build capacity for
 - healthcare
 - alternate instruction plan
- Ensure privacy of collected information: **More in the presentation by Yu-Xiang.**

