A Spatiotemporal Epidemiological Prediction Model to Inform County-level COVID-19 Risk in the USA

Peter Song and Song Lab

Department of Biostatistics, University of Michigan

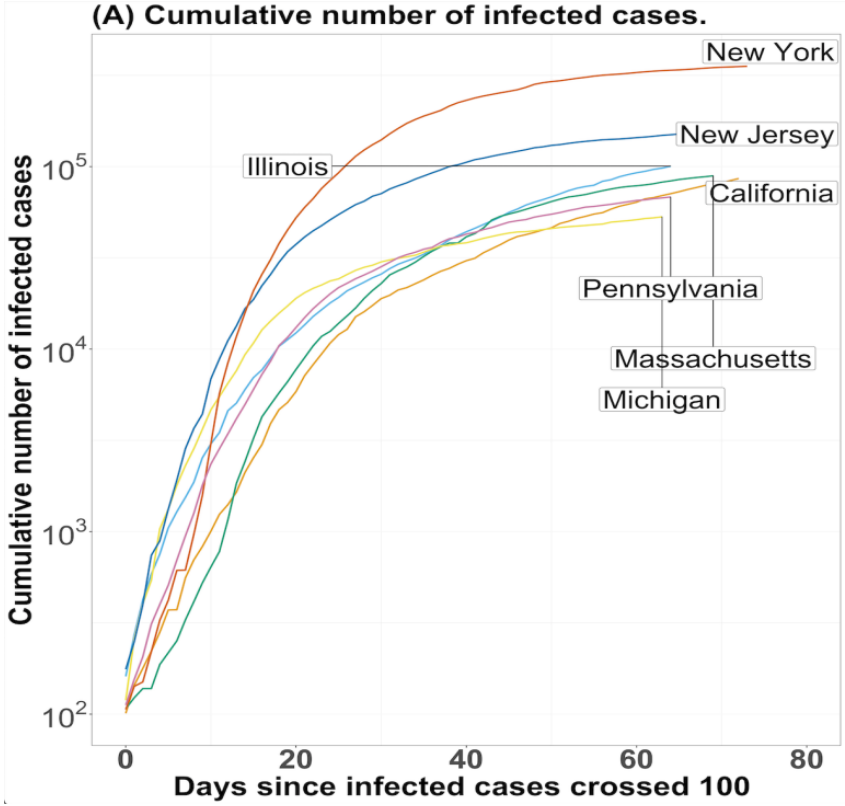
June 5, 2020

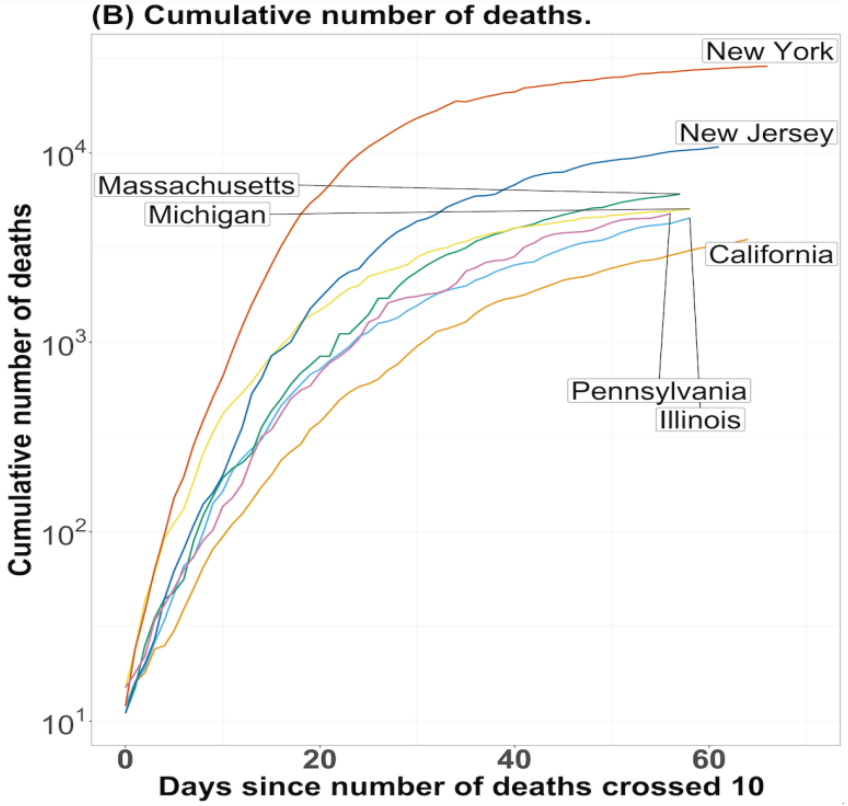
The COVID-19 pandemic in USA continues worsening

- From the CDC data up to May 29, out of 3109 counties in the continental US
- 2,895 (93.1%) have confirmed COVID-19 cases
- 1,666 (53.6%) have reported case fatalities.
- At least 97.1% of the US population are living with their contagious neighbors in the same county.

- Major Data Sources: WHO, US CDC, especially **1Point3Acres**, **Harvard China Data Lab**

Trends of COVID-19 in Heavily Hit States of the US





Opportunities and Challenges:
From macro modeling to micro modeling

- More data become available to the public; daily numbers of confirmed infections, deaths, RT-PCR tests.
- Various data types: surveillance data, lab data, EHR hospital data, antibody survey data, mobile health data.
- High-resolution policy assessment and risk prediction become possible.

- Integrate different data sources to overcome a few key shortcomings in temporal macro models; data underreporting (missing infected asymptomatic cases), assessing effectiveness of social distancing, addressing heterogeneities across geographic locations, etc.
- Need to substantially extend current macro SIR/SEIR model to micro models.
- Overcome prediction latency and improve prediction accuracy

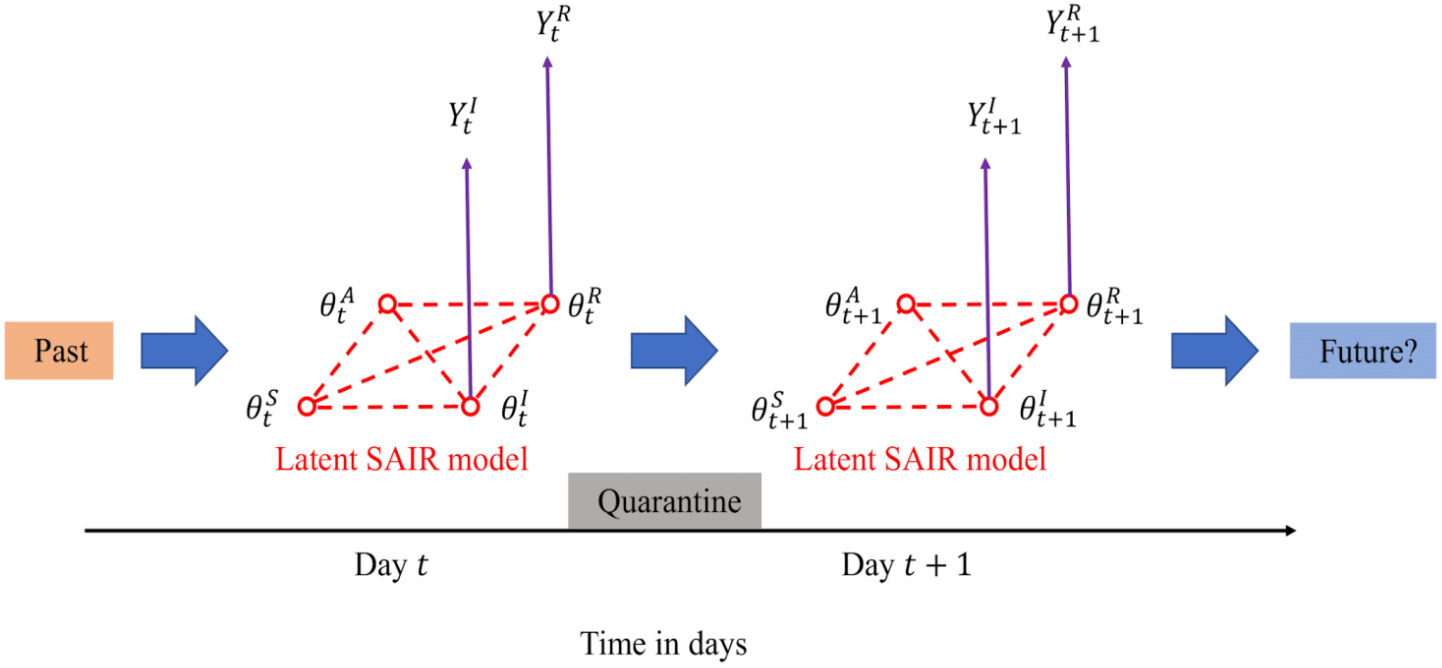
Aims

- **Aim 1:** Build statistical models to learn the pandemic dynamics **accounting for both self-immunization and social distancing**.

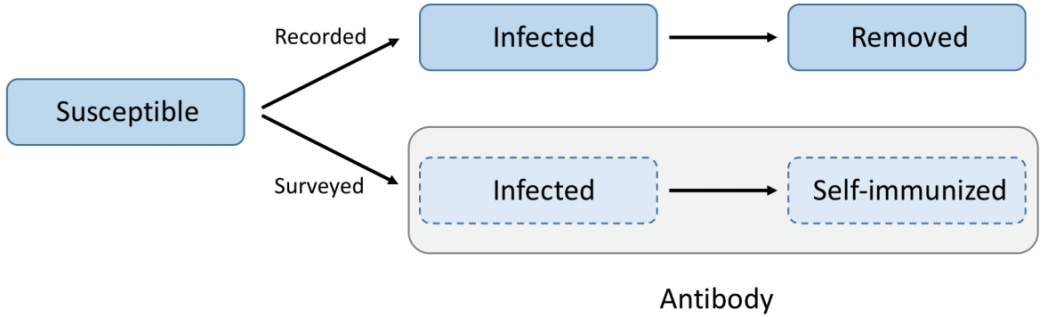
- **Aim 2:** Build a **community-level risk information system** that informs people in the US about personal risk score of infection.

- **Aim 3:** Deliver a **travel risk score information** via Google map assisted app accessible by mobile device users.

Aim 1: Statistical Model: **Generating Data from a Mechanistic Infection Dynamics Model**



Mechanistic Model: SAIR



$$\frac{d\theta_t^A}{dt} = \alpha(t)\theta_t^S,\ \frac{d\theta_t^S}{dt} = -\alpha(t)\theta_t^S - \beta\pi(t)\theta_t^S\theta_t^I,\ \frac{d\theta_t^I}{dt} = \beta\pi(t)\theta_t^S\theta_t^I - \gamma\theta_t^I,\ \text{and}\ \frac{d\theta_t^R}{dt} = \gamma\theta_t^I,$$

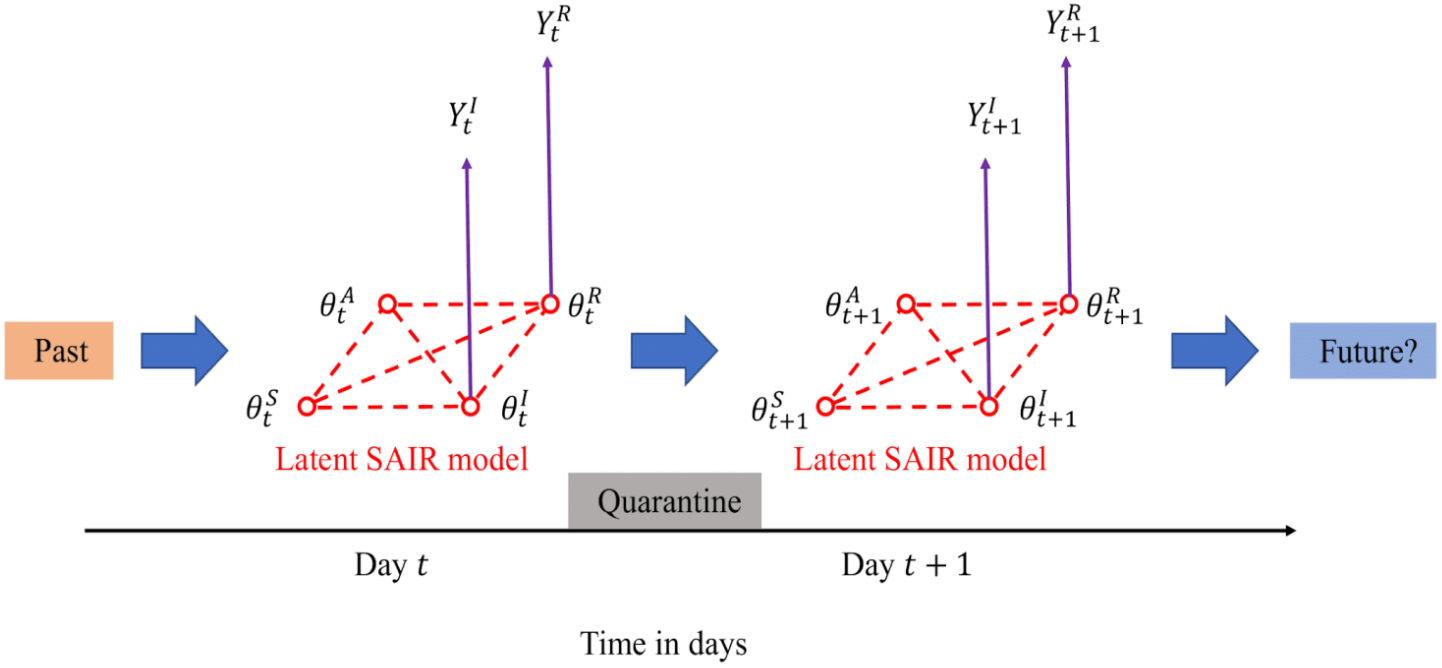
- **Assumption:** Asymptomatic subjects got infected and then recovered and developed antibody to the coronavirus.
- This is a way to overcome the underreporting issue mainly caused by the subpopulation of self-immunized cases.

Parameters in the SAIR Model

$$\frac{d\theta_t^A}{dt} = \alpha(t)\theta_t^S,\ \frac{d\theta_t^S}{dt} = -\alpha(t)\theta_t^S - \beta\pi(t)\theta_t^S\theta_t^I,\ \frac{d\theta_t^I}{dt} = \beta\pi(t)\theta_t^S\theta_t^I - \gamma\theta_t^I,\ \text{and}\ \frac{d\theta_t^R}{dt} = \gamma\theta_t^I,$$

- $\beta > 0$ is the disease transmission rate, $\gamma > 0$ is the removal rate
- $\pi(t)$ is the transmission modifier that characterizes the effect of human interventions
- $R_0 = \beta/\gamma$ is the basic reproduction number that indicates the expected number of cases generated by one infected case without intervention
- $\alpha(t)$ is the self-immunization rate that characterizes the proportion of people moving from the susceptible compartment to the antibody compartment

Statistical Learning of the SAIR Model:
Markov Chain Monte Carlo (MCMC) for State Space Models



Statistical Learning of SAIR Model via State-level Data

$$\frac{d\theta_t^A}{dt} = \alpha(t)\theta_t^S,\ \frac{d\theta_t^S}{dt} = -\alpha(t)\theta_t^S - \beta\pi(t)\theta_t^S\theta_t^I,\ \frac{d\theta_t^I}{dt} = \beta\pi(t)\theta_t^S\theta_t^I - \gamma\theta_t^I,\ \text{and}\ \frac{d\theta_t^R}{dt} = \gamma\theta_t^I,$$

- Why state-level data? Sample size/Policies/Covid-19 diagnostic tests
- A macro model to robustly learn a state-level infection dynamics.
- Approximate the dynamic system of four ODEs by the 4th order Runge-Kutta Method
- Derive posteriors and Implement the MCMC in RJAGS in R software to draw MCMC samples from the posteriors
- Summarize the results from 200,000 MCMC draws after burn-in

Refer to our R package eSIR available on the Github:
https://github.com/lilywang1988/eSIR/tree/master/install_pkg)

Aim 2: A Community-level Risk Information System

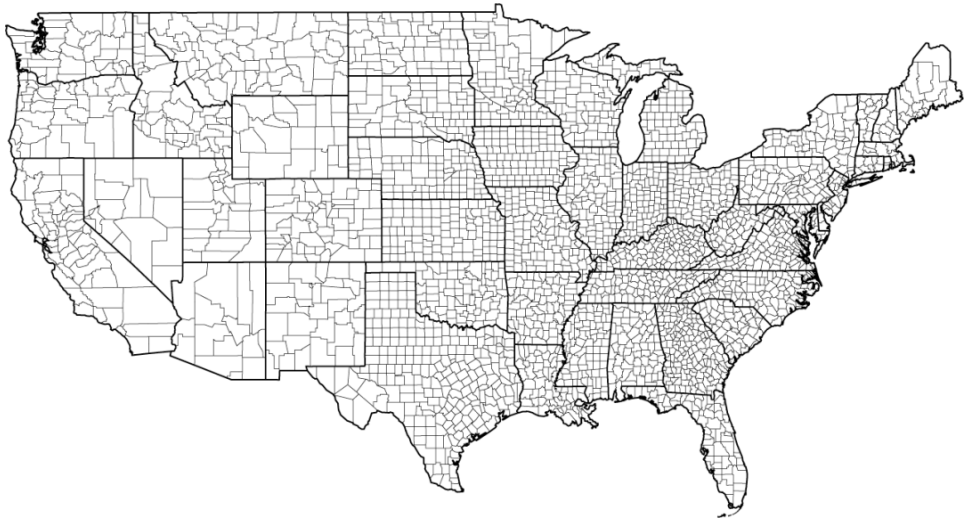
- As the COVID-19 pandemic continues worsening in the US, it is of critical importance to develop **a health information system** that provides **timely risk evaluation** and **prediction of the COVID-19 infection** in **communities**.

- Such community-level risk information is valuable for local governments and residents to
  - Assess the preparedness of medical resources (PPEs and ICU beds).
  - Determine the allocation of the COVID-19 test kits.
  - Adjust various intervention policies.
  - Enforce the conduct of social distancing.

A Community-level Risk Information System

This risk information system should

- Provide **timely** community-level information of the COVID-19 infection risk both now and over a future time period for residents in **each county** of the 3019 continental counties in the US.
- Account for temporally varying **control measures** and **spatial variations** in the COVID-19 infection dynamics.



A Micro Risk Information System

Challenges:

- Substantial **heterogeneity** in the urbanization, ethnic distribution, political views, and economic composition across regions.
- A **micro** model with a **finer-resolution** is needed.
- A location-specific forecast requires combining a **temporal** epidemiological model for time-course infection dynamics and a **spatial** model for the changes of the infection risk over 3109 counties of the continental US.

CA-eSAIR micro prediction model

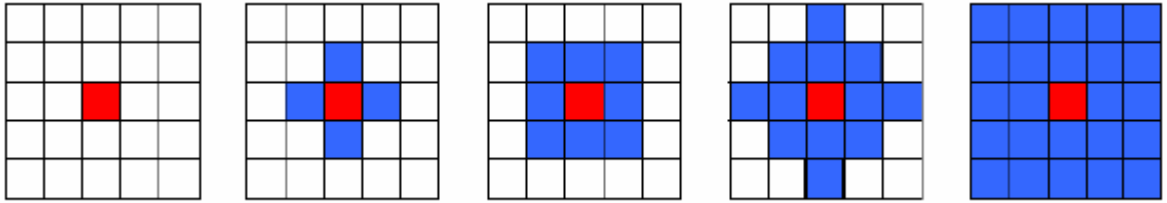
The combination of

- temporal epidemiological model:
  - extended Susceptible-Antibody-Infected-Removed (eSAIR) model
- spatial model
  - cellular automata (CA)

Cellular Automata (CA) model

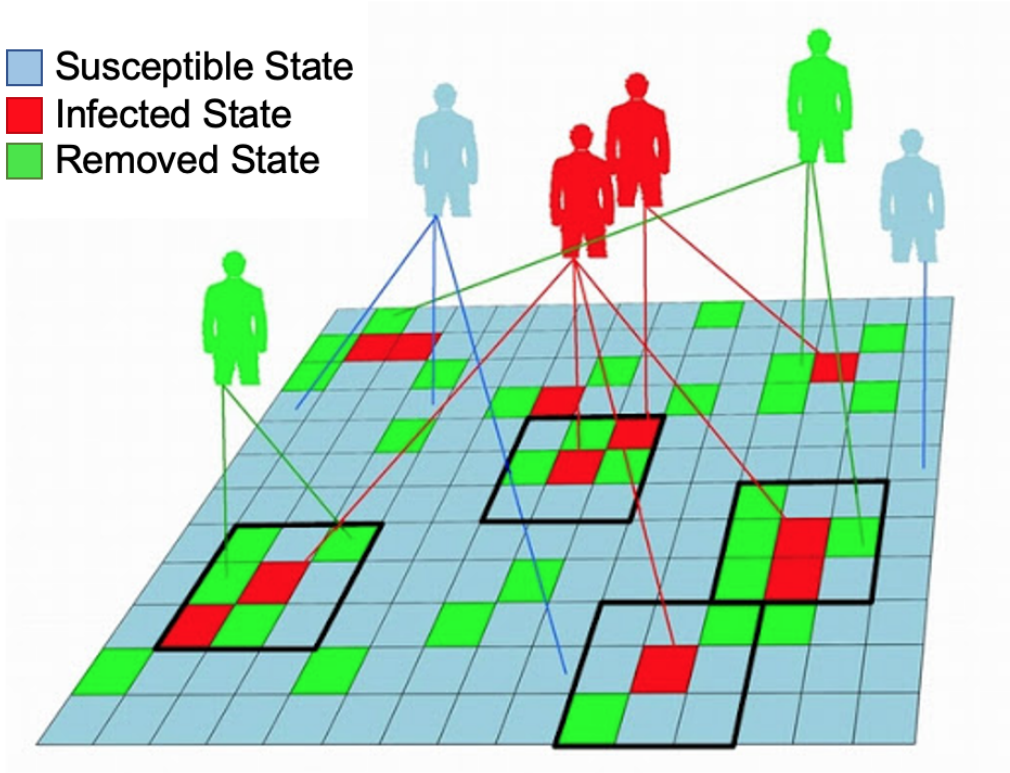
Cellular Automata (CA)

- A collection of "colored" cells on a grid of specified shape.
- Evolves through a number of discrete time steps according to a set of rules based on the states of neighboring cells.
- The rules are applied iteratively for as many time steps as desired.



CA neighborhoods: (a) Empty, (b) Von Neumann, (c) Moore, (d) MvonN, (e) Extended.

Visualizing a cellular automata model



CA-eSAIR model to predict county-level risk

- Let $\mathcal{C}$ be the collection of 3109 counties. For $c \in \mathcal{C}$, N_c denotes the county population size, and $\mathcal{C}_{-c}$ denotes the set of all the other counties except c.
- The county-specific prevalence vector is denoted as $\boldsymbol{\theta}_c(t) = (\theta_c^S(t), \theta_c^A(t), \theta_c^I(t), \theta_c^R(t))^T$.

The local transition function of the CA-based model is the following:

$$\theta_c^A(t) = \theta_c^A(t-1) + \alpha_c(t)\theta_c^S(t-1),$$

$$\theta_c^S(t) = (1-\alpha_c(t))\theta_c^S(t-1) - \beta\pi_c(t)\theta_c^S(t-1)\theta_c^I(t-1) - \beta\pi_c(t)\theta_c^S(t-1)\sum_{c'\in\mathcal{C}_{-c}} \omega_{cc'}(t)\{N_{c'}\theta_{c'}^I(t-1)/N_c\},$$

$$\theta_c^I(t) = (1-\gamma)\theta_c^I(t-1) + \beta\pi_c(t)\theta_c^S(t-1)\theta_c^I(t-1) + \beta\pi_c(t)\theta_c^S(t-1)\sum_{c'\in\mathcal{C}_{-c}} \omega_{cc'}(t)\{N_{c'}\theta_{c'}^I(t-1)/N_c\},$$

$$\theta_c^R(t) = \theta_c^R(t-1) + \gamma\theta_c^I(t-1),$$

CA-eSAIR model to predict county-level risk

The specification of parameters

- $\alpha_c(t)$ is specified as the self-immunization rate proportional to the reported value in New York and the state-specific estimated R_0.
- $\pi_c(t)$ is set as $1 -$ social distancing index/100 if the stay-at-home policy does not expire at the prediction day; otherwise $\pi_c(t)$ is set as 1.00.

- On April 29, NY Governor Andres Cuomo released results of statewide antibody tests surveys; about 15-20% of NY residents had antibody tested positive. Thus, this self-immunization rate function is specified with a jump with size 0.20, say, on April 29.
- The self-immunization function in another state is calibrated by a ratio R_{0s}/R_{0NY} under the assumption that **the higher reproduction number the higher proportion of asymptomatic infections and thus the higher rate of self-immunization.**
- Social distancing scores are estimated by mobile devices data done by the Transportation Institute of the University of Maryland. It changes when a state-level control measure is relaxed.

State	α	$\pi(5.2)$	$\pi(5.3)$	$\pi(5.4)$	$\pi(5.5)$	$\pi(5.6)$	$\pi(5.7)$	$\pi(5.8)$	$\pi(5.9)$
AL	0.153	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AK	0.117	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AZ	0.131	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
AR	0.121	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
CA	0.146	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
CO	0.146	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CT	0.186	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53
DE	0.172	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
DC	0.173	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36
FL	0.153	0.62	0.62	1.00	1.00	1.00	1.00	1.00	1.00
GA	0.156	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
HI	0.116	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44
ID	0.153	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
IL	0.183	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65
IN	0.151	0.76	0.76	1.00	1.00	1.00	1.00	1.00	1.00
IA	0.127	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
KS	0.125	0.75	0.75	1.00	1.00	1.00	1.00	1.00	1.00
KY	0.117	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76
LA	0.153	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
ME	0.115	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MD	0.171	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53
MA	0.233	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46
MI	0.145	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63

Daily self-immunization rate in population

- On April 29, let the estimated accumulated proportion of self-immunized cases be *alpha.*
- Consider a 7 day incubation period for the first confirmed covid-19 case reported on March 1 in New York, under a geometric distribution, the cell probability (the daily self-immunization rate) *p* satisfies *pgeom(67, p) = 0.20.*
- The estimated daily self-immunization rate in New York is 3/1,000, which will used in the prediction for the period of May 3-9.

CA-eSAIR model to predict county-level risk

The inter-county connectivity coefficient is specified as $\omega_{cc'}(t) = \mu_{cc'}\exp\{-\eta r(c, c')\}$

- $\mu_{c,c'}$ is the inter-county mobility factor characterizing the decrease of human encounters in terms of their potential movements between counties.
- $r(c, c')$ is a certain travel distance between two counties in terms of both geodesic distance and "air distance" based on the accessibility to nearby airports.
- η is a tuning parameter that enables us to adjust the scale of the travel distance by minimizing the sum-of-squared prediction error (SSPE) for the one-step ahead risk prediction of the infection rate.

For more details, refer to Zhou et al. (2020). A Spatiotemporal Epidemiological Prediction Model to Inform County-level COVID-19 Risk in the USA. *Harvard Data Science Review*, to appear.

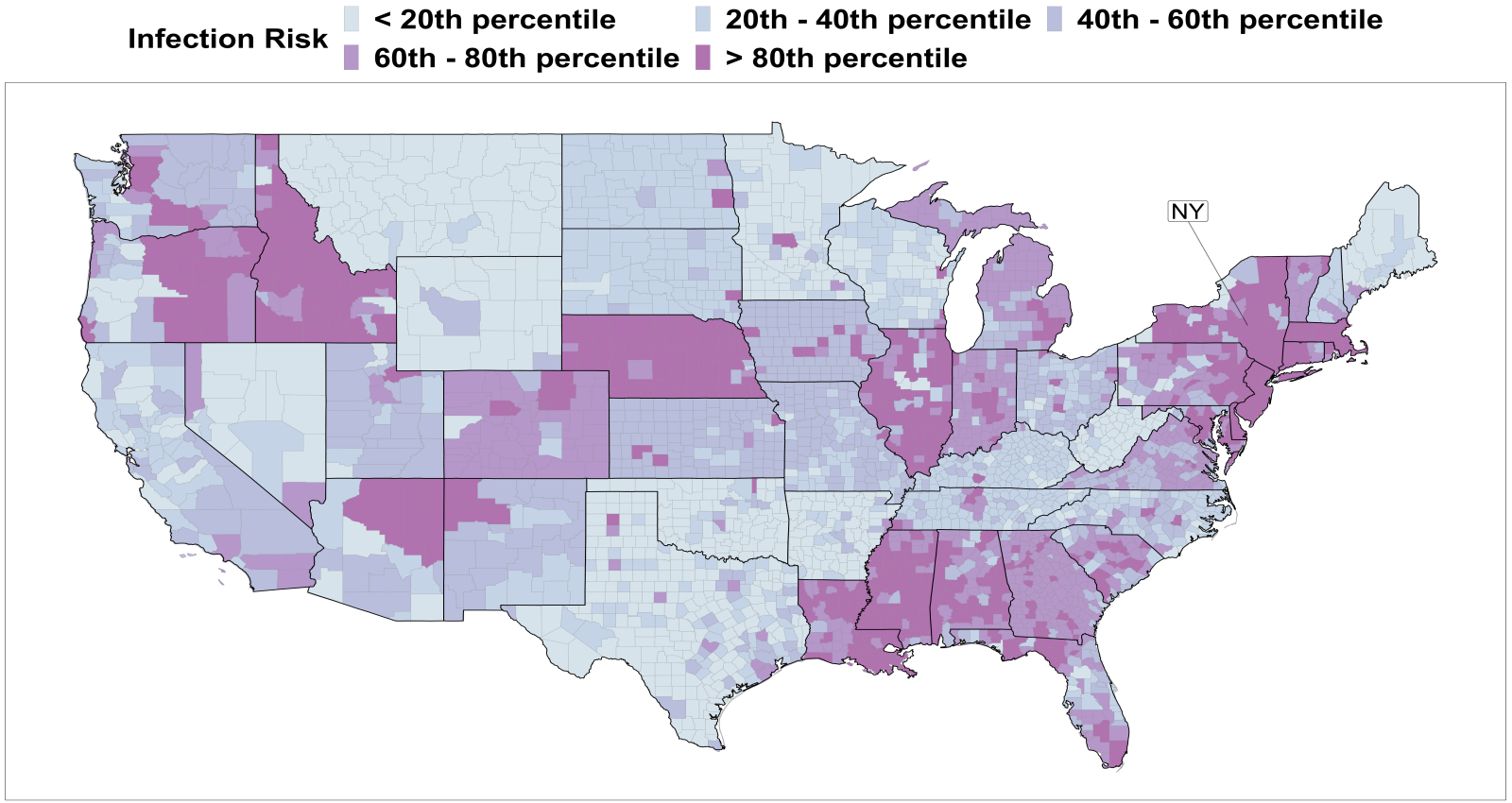
County-level Risk Forecast

One-day ahead risk prediction. Applying the equation for $\theta_c^I(t)$, we obtain a one-day ahead county-level risk prediction $\theta_c^I(t_0+1)$ for each of 3109 continental US counties.

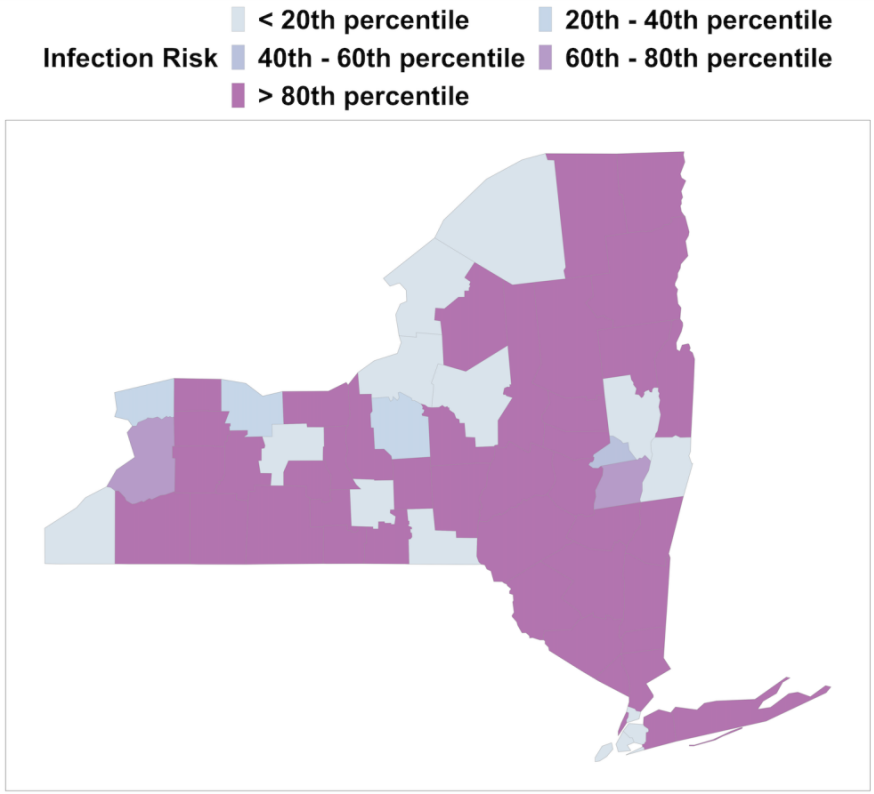
t-day ahead risk prediction. To predict the county-level risk scores over a period of t future days from t_0, the calculation is based on the entire CA-eSAIR model. The risk scores is given by:

$$RS_c(t|t_0) = \theta_c^I(t_0+1) + \{1-\theta_c^I(t_0+1)\}\theta_c^I(t_0+2) + \{1-\theta_c^I(t_0+1)\}\{1-\theta_c^I(t_0+2)\}\theta_c^I(t_0+3)$$
$$+ \cdots + \{1-\theta_c^I(t_0+1)\}\ldots\{1-\theta_c^I(t_0+t-1)\}\theta_c^I(t_0+t).$$

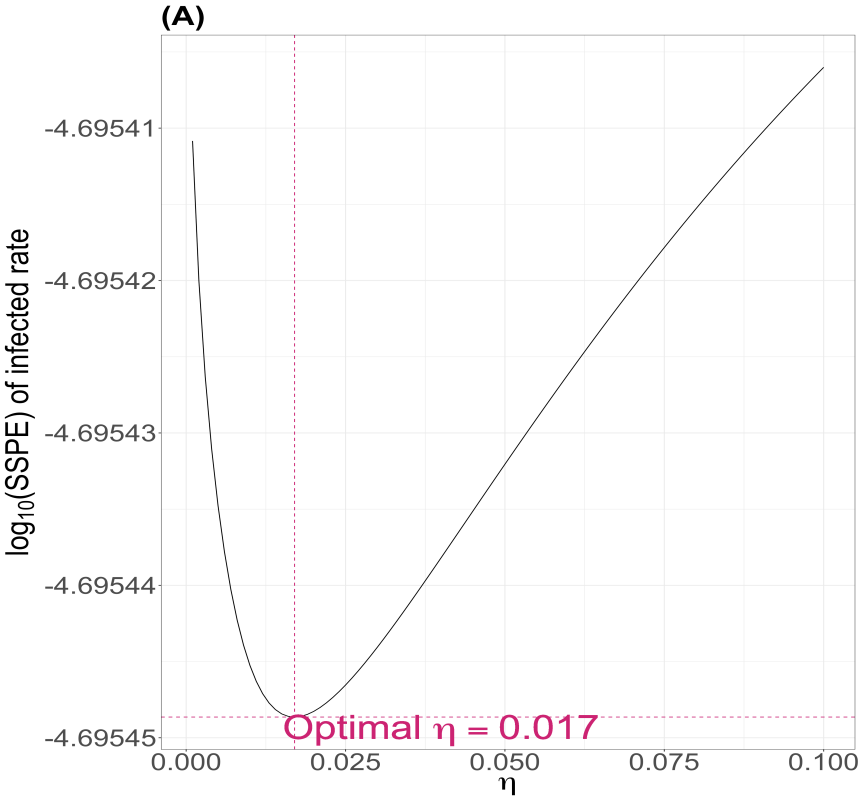
County-level Risk Forecast (7-day ahead from May 2nd)

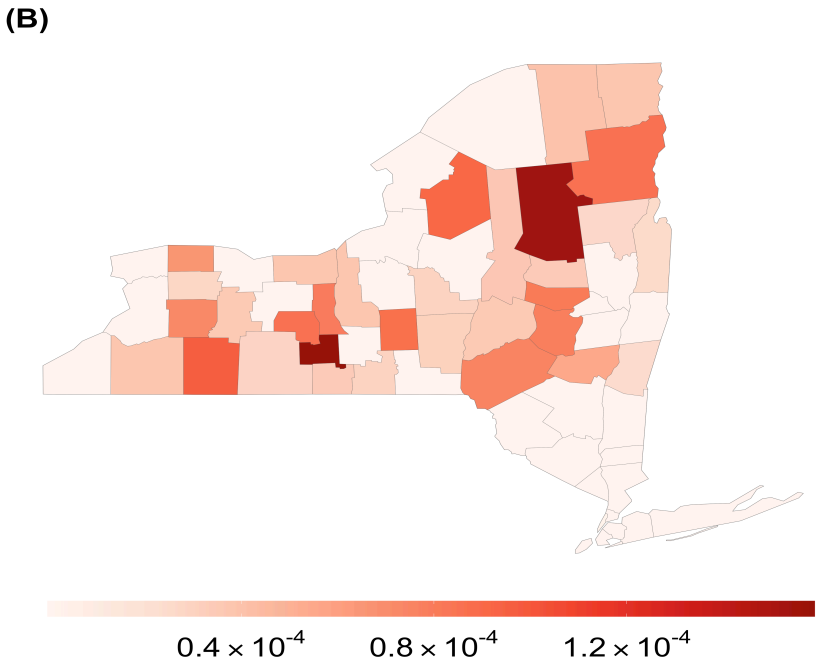


County-level Risk Forecast for NY (7-day ahead from May 2nd)



Optimal tuning based on minimal one-day prediction error



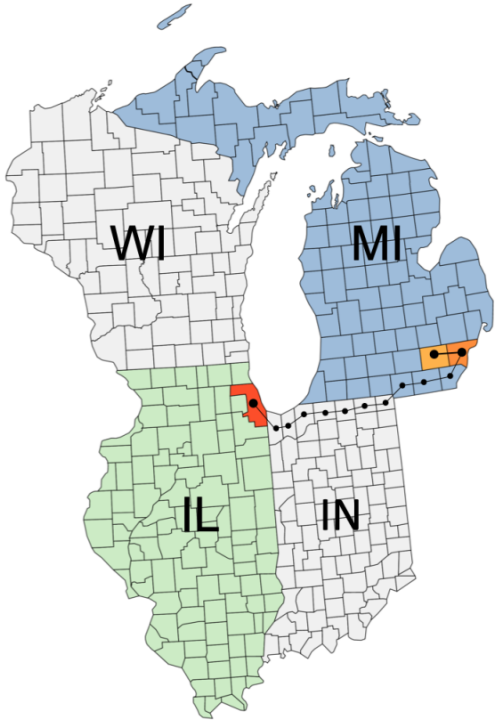


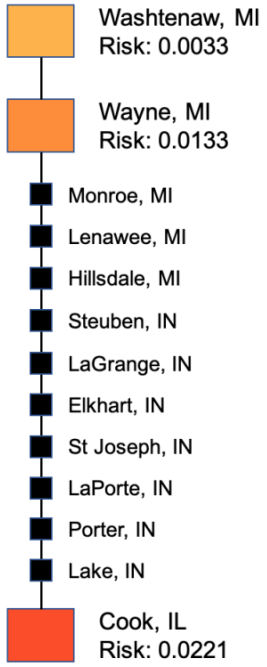
Travel Risk Forecast

Risk prediction of travel. Let C be a set of counties that a traveler plans to stop by over next t days. For simplicity, suppose the traveler stops at one county per day, denoted as $C = \{c_1, \ldots, c_t\}$ with c_j being the county visited on day $t_0 + j, j = 1, \ldots, t$. The risk score associated with the trip is given by:

$$RS(C, t|t_0) = \theta^I_{c_1}(t_0+1) + \{1-\theta^I_{c_1}(t_0+1)\}\theta^I_{c_2}(t_0+2) + \{1-\theta^I_{c_1}(t_0+1)\}\{1-\theta^I_{c_2}(t_0+2)\}\theta^I_{c_3}(t_0+3)$$
$$+ \cdots + \{1-\theta^I_{c_1}(t_0+1)\}\ldots\{1-\theta^I_{c_{t-1}}(t_0+t-1)\}\theta^I_{c_t}(t_0+t).$$

Risk Prediction of Travel





The risk score for this travel route is calculated to be 3.84×10^{-4}.

Considering this, the traveler may revise their travel plan if this projected risk is too high.

One situation may arise where a traveler takes a longer route (in terms of time/distance) in order to lower the travel risk.

Future Directions

- Extend the proposed CA-eSAIR model to predict the risk of COVID-19 in other areas across the world.
- Predict the risk at locations of household, factory and business sites.
- Make more data available in the future:
  - More local data of extensive viral diagnostic and antibody tests available across the country.
  - More individual health and behavior data from better surveillance systems such as mobile device apps for the tracking of personal movements.
- Develop data-driven strategies for precision public health intervention and strategies based on subgroups (age, race, existing medical conditons) for the COVID-19 containment.

Acknowledgements

Thanks to China Data Institute and Shanghai University of Finance and Economics, especially Drs. Bao and Zhang for organizing this series of open seminar.

Appreciative of valuable discussions fby Drs. Gong and Deng from Tsinghua University.

Thanks to Harvard China Data Lab for continuous supplies of high-quality data used in our project.

Research is partially supported by NSF DMS 1811734.

References

1. 1point3acres covid-19 us data. https://coronavirus.1point3acres.com/zh, 2020.
2. China data lab 2020-02-202020 draft version covid-19, us, map harvard dataverse. https://doi.org/10.7910/DVN/HIDLTKdoi/10.7910/DVN/HIDLTK, 2020.
3. List of airports in the united states. https://en.wikipedia.org/wiki/List_of_airports_in_the_United_States, 2016.
4. Social distancing scoreboard. https://www.unacast.com/covid19/social-distancing-scoreboard?view=county&fips=08097, 2020.
5. University of maryland covid-19 impact analysis platform. https://data.covid.umd.edu/, 2020.
6. Peter Xue-Kun Song. Monte Carlo Kalman filter and smoothing for multivariate discrete state space models. Canadian Journal of Statistics, 28(3):641{652, 2000.
7. John Von Neumann, Arthur W Burks, et al. Theory of self-reproducing automata. IEEE Transactions on Neural Networks, 5(1):3{14, 1966.
8. Lili Wang, Yiwang Zhou, Jie He, Bin Zhu, Fei Wang, Lu Tang, Michael Kleinsasser, Daniel Barker, Marisa Eisenberg, and Peter X Song. An epidemiological forecast model and software assessing interventions on covid-19 epidemic in china (with discussion). Journal of Data Sciences, 2020.