**Harvard Data Science Review • Special Issue 1: COVID-19:
Unprecedented Challenges and Chances**

A Spatiotemporal Epidemiological Prediction Model to Inform County- Level COVID-19 Risk in the United States

**Yiwang Zhou^{1,2} Lili Wang^{1,3} Leyao Zhang¹ Lan Shi^{4,5} Kangping Yang⁶
Jie He⁷ Zhao Bangyao⁷ William Overton⁷ Soumik Purkayastha¹
Peter Song¹**

¹Song Lab, Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America,

²Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America,

³Howard Hughes Medical Institute, Department of Molecular Cellular and Developmental Biology, University of Michigan, Ann Arbor, Michigan, United States of America,

⁴School of Nursing, University of Michigan, Ann Arbor, Michigan, United States of America,

⁵Department of Biostatistics, Vanderbilt University Medical Center, Vanderbilt University, Nashville, Tennessee, United States of America,

⁶University of Michigan, Ann Arbor, Michigan, United States of America,

⁷Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America

Published on: Aug 06, 2020

DOI: <https://doi.org/10.1162/99608f92.79e1f45e>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](#)

ABSTRACT

As the COVID-19 pandemic continues worsening in the United States, 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. We propose a spatiotemporal epidemiological forecast model that combines a spatial cellular automata (CA) with a temporal extended susceptible-antibody-infectious-removed (eSAIR) model under time-varying state-specific control measures. This new toolbox enables the projection of the county-level COVID-19 prevalence over 3109 counties in the continental United States, including t -day-ahead risk forecast and the risk related to a travel route. In comparison to the existing temporal risk prediction models, the proposed CA-eSAIR model informs the projected county-level risk to governments and residents of the local coronavirus spread patterns and the associated personal risks at specific geolocations. Such high-resolution risk projection is useful for decision-making on business reopening and resource allocation for COVID-19 tests.

Keywords: cellular automata, coronavirus infectious disease, risk prediction, SAIR model

This article includes a select anonymous review document. *Anonymous review is a vital process for high quality publications in HDSR. With permissions of the authors and reviewers, we selectively post anonymously review reports, sometimes with authors' responses, that we believe can further enrich the intellectual journey generated by the corresponding publication.*

[Zhou-et-al_Review-Report.pdf](#)

429 KB

Media Summary

As of May 22, 2020, of 3109 counties in the continental United States, 2895 (93.1%) have confirmed COVID-19 cases and 1666 (53.6%) have reported case fatalities. The COVID-19 pandemic continues worsening in the United States nationwide and, according to CDC data, at least 97.1% of the U.S. population are living with their contagious neighbors. We seek to develop a health information system that provides communities with COVID-19 infection risk predictions in a similar way to the weather forecast. In this article, we establish a COVID-19 forecast paradigm based on an automated operation of evolving intercounty disease-spread patterns with various types of state-level control measures. We estimate the effectiveness of social distancing for each state using mobile device data, and derive intercounty connectivity from county-level residents' mobility variables. In addition to traditional geodesic distance, we use air-distance through nearby airports to quantify the effective distance between any two counties in the United States. This prediction model is tuned by the minimal prediction error of one-day-ahead infection projection. Using this well-tuned forecast system, we can predict county-level COVID-19 prevalence over 3109 counties in the continental United States (e.g., the risk of

infection over the next week in Washtenaw County, Michigan). We can also approximate the infection risk imposed by intercounty travel on a specific date. In comparison to existing temporal risk prediction models, our model informs projected county-level risk and potential spread patterns to governments and residents, and the associated personal risks at specific geolocations. Such high-resolution risk projection is useful for tailored decision-making on business reopenings and resource allocation for both COVID-19 tests and medical equipment.

1. Introduction

Coronavirus disease 2019 (COVID-19), an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) ([World Health Organization, 2020](#)), became a global pandemic that spread out swiftly across the world since its original outbreak in Hubei, China, in December 2020. Since mid-March, the number of confirmed infections in the United States experienced a rapid growth. Many states are facing great challenges in mitigating the spread of the virus, including New York, New Jersey, Massachusetts, Illinois, and Michigan. As of May 8, 2020, this pandemic has caused a total of 1,318,553 confirmed cases and 78,303 deaths in the United States. The United States is now leading the daily contributions to new infections in the current phase of the global pandemic. Being a lethal communicable infectious disease, COVID-19 is expected to continue spreading in the U.S. population, causing an even higher number of infections and deaths in the next few months. With no effective medical treatments nor vaccines available at this moment, public health interventions such as social distancing have been implemented in most of the states at different degrees of stringency in order to mitigate the spread of COVID-19. To address this continuing pandemic, it is of great urgency to develop a risk prediction model that provides 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 United States. The regional risk evaluation and prediction can help account for temporally varying control measures and spatial variations in the COVID-19 infection dynamics. Such community-level risk information is valuable for local governments and residents to assess the preparedness of medical resources (personal protective equipment and ICU beds), to determine the allocation of the COVID-19 test kits, to adjust various intervention policies, and to enforce the conduct of social distancing.

Predicting the county-level COVID-19 infection risk is especially challenging because of the substantial heterogeneity in the urbanization, ethnic distribution, political views, and economic composition across regions. Existing state-level epidemiological prediction models focus fully on time-domain analysis and are not adequate to address such geographic, racial, and economic discrepancies in the United States. A prediction model with a finer resolution is needed to assess the local risk of COVID-19. Technically, a local risk prediction needs to be performed collectively with all counties, since Americans are highly mobile and connected by highways and airways. Therefore, a county-level risk is more informative and appealing to the public and governments. Making such a location-specific forecast requires combining a temporal epidemiological model for the time-course infection dynamics and a spatial model for changes of the infection

risk over 3109 continental counties of the country. We propose a new forecast paradigm by incorporating the extended susceptible-antibody-infected-removed (eSAIR) model and a cellular automata (CA) ([Von Neumann & Burks, 1966](#)). Being an important extension of the extended susceptible-infected-removed (eSIR) model proposed in Wang et al. ([Wang et al., 2020](#)), the eSAIR model enables us to capture the temporal evolution of the disease, in which a time-varying transmission rate modifier is used to account for state-specific control measures. The proposed extension pertains to adding a new compartment of antibody (A) to accommodate the ongoing self-immunization in the U.S. population, so that the resulting model helps address the underreporting issue concerning data available in the public databases. The CA model is based on a spatial process of disease spread over 3109 counties. This new epidemiological forecast model, termed as CA-eSAIR model, can inform residents in different geolocations with timely and areal-evolving risk of COVID-19 infection. It can also help travelers make a plan for a trip in the United States with relevant risk information to avoid counties with high risk. The novelty of the community-level risk projection is twofold: the proposed spatiotemporal prediction model allows people and policymakers to envision how the pandemic may evolve across the United States to, for example, identify rising hotspots of contagion; more importantly, the projection model informs the public of future spread patterns and associated personal risk scores in each county of the United States. Moreover, due to the use of the Markov Chain Monte Carlo (MCMC) estimation in the proposed state-space model, the quantification for the subsequent prediction uncertainty can be easily carried out in the proposed statistical framework.

This article is organized as follows. Section [2](#) begins with an introduction of the eSAIR model with the inclusion of an antibody compartment, and then presents the proposed CA-eSAIR model for county-level risk prediction. Section 3 concerns an empirical study on community-level risk prediction over 3109 counties in the continental United States. Section 4 gives some concluding remarks.

2. Spatiotemporal County-Level Risk Projection

We propose a statistical model, termed as the CA-eSAIR model in this article, to map time-varying nationwide projected infection risk of COVID-19 for each county in the continental United States. This forecast framework consists of a temporal statistical model for infection dynamics and a spatial model for spatial spread patterns. The temporal statistical model is based on an eSIR model ([Wang et al., 2020](#)) that accounts for the time-varying state-level control measures that modify the transmission rate of the virus due to state-specific interventions (e.g., social distancing) as well as the proportion of self-immunized individuals that have developed antibodies to COVID-19. Being an important phenomenon of the COVID-19 contagion, the existence of self-immunization in the U.S. population has been largely ignored in most of the existing models, which will be addressed via our eSAIR model in this article. The spatial model is built upon a CA ([Ahmed & Agiza, 1998](#); [Beauchemin et al., 2005](#); [Chopard & Droz, 1998](#); [Fuentes & Kuperman, 1999](#); [Fuks & Lawniczak, 2001](#); [Schneckenreither et al., 2008](#); [Sirakoulis et al., 2000](#); ([Von Neumann & Burks, 1966](#))[White et al., 2007](#); [Wilcox et al., 2003](#)) that links local disease spread conditions over 3109 counties via a certain

spatial connectivity function characterizing the intercounty mobility, geodistance, and air-distance via accessibility to nearby airports.

2.1. Extended Susceptible-Antibody-Infected-Removed (eSAIR) Model

Figure 1. The compartment composition of the extended susceptible-antibody-infected-removed (eSAIR) model. Three compartments on the top thread form the classical SIR model, including susceptible, infected and removed. The eSAIR model adds an antibody compartment (the bottom thread) to account for the proportion of people who are infected and self-immunized without being tested and recorded.

Motivated by COVID-19 data in Hubei province, China, (Wang et al., 2020) developed an eSIR model under the framework of state-space models (Jørgensen & Song, 2007; Song, 2000). This epidemiological SIR model is driven by a latent Markov SIR model (Kermack & McKendrick, 1927) with the probability of being susceptible (or at risk), θ_t^S ; the probability of being infected, θ_t^I ; as well as the probability of being removed, θ_t^R at a given time t . A useful contribution in the eSIR model is the introduction of a viral transmission rate modifier, namely, the term $\pi(t)$ in Equation (1) that accounts for state-specific preventive policies such as lockdown, social distancing, and use of face masks. Such changing regimes of disease infection are also incorporated into the eSAIR model in this article. To address the underreporting issue associated with available public databases and to build self-immunization into the infection dynamics, we then further extend the previous eSIR model to an eSAIR model by adding an antibody (A) compartment. Shown in the bottom thread of Figure 1, this A compartment accounts for the probability of being self-immunized with antibodies to COVID-19, denoted by θ_t^A ; see Equation (1), where $\alpha(t)$ is a function describing the proportion of people moving from the susceptible compartment to the antibody compartment over time. Compartment A helps circumvent limitations of data collection, especially embracing individuals who were infected but self-cured at home with no confirmation by viral real-time polymerase chain reaction (RT-PCR) diagnostic tests. This new eSAIR model characterizes the underlying population-level dynamics of the pandemic. The following system of ordinary differential equations defines collectively the continuous-time dynamics for the eSAIR model, which governs the law of movements among four compartments of susceptible, self-immunized, infected and removed:

$$\begin{aligned}
 (1) \quad & \frac{d\theta_t^A}{dt} = \alpha(t)\theta_t^S, \\
 & \frac{d\theta_t^S}{dt} = -\alpha(t)\theta_t^S - \beta\pi(t)\theta_t^S\theta_t^I, \\
 & \frac{d\theta_t^I}{dt} = \beta\pi(t)\theta_t^S\theta_t^I - \gamma\theta_t^I, \\
 & \frac{d\theta_t^R}{dt} = \gamma\theta_t^I,
 \end{aligned}$$

where $\alpha(t)$ is the self-immunization rate, β is the basic disease transmission rate, $\pi(t)$ is a time-varying transmission rate modifier, and γ is the rate of being removed from the system (either dead or recovered). The basic reproduction number is $R_0 = \beta/\gamma$, which represents the number of individuals who can contract the virus from an infectious case under no interventions in the system. The product term $\beta\pi(t)$ describes a modified transmission rate due to control measures. Thus, from the eSAIR model the parameter β is estimated by adjusting preventive measures. So the resulting estimate of R_0 can more adequately reflect the underlying viral transmissibility than the unadjusted estimates published in the current literature. The transmission modifier $\pi(t)$ can be specified by some social distancing score measured based on mobile data describing the mobility reduction of individuals during the epidemic. The self-immunization rate $\alpha(t)$ can be specified by an estimate from some recent surveys of antibody prevalence conducted in several states in the United States, including New York, Massachusetts, and California. Due to schedules of the survey sampling, $\alpha(t)$ may be specified with some values at discrete times when survey results become available.

Figure 2. A conceptual framework of the extended susceptible-antibody-infected-removed (eSAIR) model. The latent probabilities of susceptible (S), self-immunized (A), infected (I) and removed (R) at a certain time point t are indicated by $\theta_t = (\theta_t^S, \theta_t^A, \theta_t^I, \theta_t^R)^T$.

Time series of Y_t^I and Y_t^R are, respectively, the daily observed numbers of infections and removed cases (a total of recovered cases and deaths). The dynamics of the four compartments evolve over time according to a Markov system of ordinary differential equations.

The proposed eSAIR model is a kind of state-space model that allows sampling uncertainty in the collection of observed counts of infection, death, and recovery; see Figure 2. One defining feature of this statistical model is that proportions or rates, not counts, are considered in which the size of the underlying population is adjusted. Let $\theta_t^j, j \in \{S, A, I, R\}$ be a probability (or a population-level proportion/fraction) of being susceptible (S), self-immunized (A), infected (I), or removed (R) at a given day t such that $\theta_t^S + \theta_t^A + \theta_t^I + \theta_t^R = 1$. Let τ be a generic term denoting the set of all model parameters to be estimated. This dynamic system of $\theta_t = (\theta_t^S, \theta_t^A, \theta_t^I, \theta_t^R)^T$ is latent and follows a Markov process,

$$(2) \quad \theta_t | \theta_{t-1}, \tau \sim \text{Dirichlet}(\kappa f(\theta_{t-1}, \beta, \gamma)),$$

where parameter κ scales the variance of the Dirichlet distribution, while $f(\cdot)$ is a four-variate vector that determines the mean of the Dirichlet distribution. The function f is the engine of the infection dynamics driven by the extended eSAIR model (Equation 1). A fourth-order Runge–Kutta (RK4) approximation is obtained for $f(\theta_{t-1}, \beta, \gamma)$.¹ In the framework of the state-space models, the observed time series of daily proportions Y_t^I (the proportion of infected cases) and Y_t^R (the proportion of removed cases) are emitted from the latent Markov dynamic system θ_t according to the beta distributions:

$$(3) \quad \begin{aligned} Y_t^I | \theta_t, \tau &\sim \text{Beta}(\lambda^I \theta_t^I, \lambda^I (1 - \theta_t^I)), \\ Y_t^R | \theta_t, \tau &\sim \text{Beta}(\lambda^R \theta_t^R, \lambda^R (1 - \theta_t^R)), \end{aligned}$$

where λ^I and λ^R are the respective variances of the observed proportions. It is easy to see that $E(Y_t^I | \theta_t) = \theta_t^I$ and $E(Y_t^R | \theta_t) = \theta_t^R$. Using the standard MCMC method, we can estimate both the model parameters τ and latent probabilities θ_t over the observational period.² Next, these estimates at the last date of the observed data will be used as initial values in the CA-eSAIR model for the county-level risk prediction. Our experience from empirical studies suggests that the above state-space eSAIR model should be used with data from a relatively large population, preferably at a state level, in order to yield reliable numerical results.

2.2. CA-eSAIR Model

The eSAIR model discussed is suitable for predicting the time-varying prevalence of infection at the population level, which is particularly useful revisiting the beginning of the COVID-19 outbreak. Along the rapid evolution of this pandemic in the United States with substantial county-level data being available on a daily basis, it becomes possible to predict infection patterns and infection rates at the county-level. Such risk information is of great interest to local governments and residents to understand the current and future conditions of the pandemic. For example, local governments need it to make decisions on whether to reopen local businesses or how to allocate resources of both viral diagnostic and antibody tests in communities. To address such needs, we further extend the population-level temporal eSAIR model into a spatiotemporal epidemiological forecast model that enables the prediction of county-level COVID-19 infection. Using this new model, we can project high-resolution risk scores at a future time point for all 3109 continental counties through certain characterizations of intercounty connectivity. In the literature, Von Neumann's CA model ([Von Neumann & Burks, 1966](#)) has been extended to study spatial disease spread patterns. In this article, we borrow the strength of CA's spatial modeling to be combined with the temporal eSAIR model, resulting in a useful spatiotemporal model to predict the community-level infection risk across the United States.

Technically, we extend the classical CA from spatial lattices to areal locations of counties. Let $\mathcal{C}$ be the collection of 3109 counties. For a county $c \in \mathcal{C}$, N_c denotes the county population size, and $\mathcal{C}_{-c}$ denotes the set of all the other counties except county c . For county c at time t , we denote the county-specific prevalence vector by $\theta_c(t) = (\theta_c^S(t), \theta_c^A(t), \theta_c^I(t), \theta_c^R(t))^T$. For the purpose of daily risk prediction, we express the CA-eSAIR model at discrete times in the following form:

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

where β and γ are the effective population rates of contagion and removal, respectively, which are estimated from the eSAIR model with state-level data. Allowing the spatial intercounty disease transmission is unique in the CA, which may primarily be characterized by two factors. One is the strength of the intercounty connectivity, $\omega_{cc'}(t) \in [0, 1]$, that quantifies both the volume and frequency of the intercounty movements between counties c and c' . The other is the ratio of the number of projected infected cases $N_{c'} \theta_{c'}^I(t-1)$ in county c' over the population size N_c in county c , representing a likelihood of an at-risk person in county c contracting the virus from an infectious person in county c' .

In effect, specifying an objective intercounty connectivity coefficient $\omega_{cc'}(t)$ is challenging, as it involves many variables. As far as the disease contagion concerns, in this article we specify this coefficient as $\omega_{cc'}(t) = \mu_{cc'} \exp\{-\eta r(c, c')\}$. The first parameter $\mu_{cc'}$ is the intercounty mobility factor characterizing the decrease of human encounters in terms of their potential movements between counties (Unacast, 2020). The second factor $r(c, c')$ is a certain travel distance between two counties c and c' in terms of both geodesic distance (Karney, 2013) and “air distance” based on the accessibility to nearby airports. It is specified that

$$r(c, c') = d(c, c')I[d(c, c') \leq 500\text{km}] \\ + b(a, a')\{d(c, a) + d(c', a')\}I[d(c, c') > 500\text{km}],$$

where $d(\cdot, \cdot)$ is the geodesic distance (Karney, 2013) (in km) defined as the shortest distance between two places, either between two counties c and c' or between county c and its closest airport a , on the surface of an ellipsoidal model of the earth. Meanwhile, $b(a, a')$ is a factor that characterizes the transportation capacity of airlines between airports a and a' near counties c and c' , respectively. The travel distance $r(c, c')$ can be interpreted in the following way. If the geo-distance between county c and c' is less than 500 km, we assume that individuals from either county will travel by car, resulting in the travel distance equaling the geodesic distance; otherwise, we assume that residents will choose to fly to the other county via its nearest airport. In the latter case, we ignore the distance between airports since individuals are not exposed to the outside community environment during the flight. In addition, the third factor η is a tuning parameter that enables us to adjust the scale of the travel distance. To regularize the contribution of $r(c, c')$, we propose to tune the η parameter by minimizing the sum of (county-level) weighted absolute prediction error (SWAPE) for the one-step-ahead risk prediction of the infection rate. This tuning procedure will be demonstrated with details in our empirical study.

It is worth commenting that, given the fact that COVID-19 testing policies and strategies are state-specific, we assume that the testing rate within a given state is the same. Under this assumption, we fit the eSAIR model state by state via the MCMC method to estimate state-specific model parameters, including posterior means, medians, and 95% credible intervals. Although there exists some heterogeneity of the test rate across counties within a state, such intercounty differences are deemed much smaller than the inter-state differences. Utilizing the MCMC draws (200,000 draws by default), we can assess prediction uncertainty by summarizing over 200,000 projected risk scores from the CA-eSAIR model. Running this procedure on 3109 counties is indeed computationally expensive. A simpler calculation is to propagate estimation uncertainty into risk prediction in

the way that the limits of the 95% credible intervals are carried over to determine uncertainties for the projected risk. In the empirical study we show this simple solution that manifests the uncertainty in the risk projection. In addition, in our software, we set limits 0 and 1 to confine the projected values of $\theta_c^S(t), \theta_c^A(t), \theta_c^I(t), \theta_c^R(t)$ within $[0, 1]$.

2.3. County-Level Risk Prediction

The CA-eSAIR model above enables us to predict the county-level prevalence of COVID-19 infection $\theta_c^I(t)$, which is a key term describing the spread of the pandemic. The projection for the other three population probabilities of susceptible, self-immunized, and removed can be similarly obtained. At current time (say, today) t_0 , using the estimated parameters τ and $\theta_c(t_0) = (\theta_c^S(t_0), \theta_c^A(t_0), \theta_c^I(t_0), \theta_c^R(t_0))^\top$, $c \in \mathcal{C}$ from the state-space eSAIR model as the initial values, we are able to make several kinds of predictions of the infection prevalence described as follows.

One-Day-Ahead Risk Prediction. Applying the third equation of the CA-eSAIR model in Section 2.2, we obtain a one-day-ahead county-level risk prediction $\theta_c^I(t_0 + 1)$ for each of 3109 continental U.S. counties. The number of infected cases $N_{c'} \theta_c^I(t_0)$ in county c' at time t_0 is specified as the observed number of infected cases. This one-day-ahead predicted prevalence is the risk score for individual counties. Such nationwide risk scores can be updated whenever the eSAIR model is updated with new data, producing an updated nationwide infection map.

t -Day-Ahead Risk Prediction. When wishing to predict county-level risk scores over a period of t future days (e.g., 7-day-ahead), the number of infected cases for each county is calculated from the entire CA-eSAIR model. The risk scores over a period of t days from t_0 are given by:

$$(4) \quad \begin{aligned} \text{RS}_c(t|t_0) = & \theta_c^I(t_0 + 1) + \{1 - \theta_c^I(t_0 + 1)\} \theta_c^I(t_0 + 2) \\ & + \{1 - \theta_c^I(t_0 + 1)\} \{1 - \theta_c^I(t_0 + 2)\} \theta_c^I(t_0 + 3) \\ & + \cdots + \theta_c^I(t_0 + t) \prod_{i=1}^{t-1} \{1 - \theta_c^I(t_0 + i)\}. \end{aligned}$$

The risk of infection above is a cumulative chance during the prediction period through a sequence of conditional events. For example, with $t = 2$,

$$\begin{aligned} P(\text{infected at or before day 2}) = & P(\text{infected at day 1}) \\ & + P(\text{infected at day 2} | \text{not infected at day 1}) P(\text{not infected at day 1}). \end{aligned}$$

Risk Prediction of Travel. The CA-eSAIR model can provide predicted risk for a personal trip in the future. Let $\mathcal{C}$ be a set of counties that a traveler plans to stop by over the next t days. For simplicity, suppose the

traveler stops at one county per day, denoted as $C = \{c_1, \dots, c_t\}$ with c_j being the county visited on day $t_0 + j, j = 1, \dots, t$. Then, by a similar cumulative chance, the risk score associated with the trip is given by:

$$(5) \quad \begin{aligned} RS(C, t|t_0) = & \theta_{c_1}^I(t_0 + 1) + \{1 - \theta_{c_1}^I(t_0 + 1)\}\theta_{c_2}^I(t_0 + 2) \\ & + \{1 - \theta_{c_1}^I(t_0 + 1)\}\{1 - \theta_{c_2}^I(t_0 + 2)\}\theta_{c_3}^I(t_0 + 3) \\ & + \dots + \theta_{c_t}^I(t_0 + t) \prod_{j=1}^{t-1} \{1 - \theta_{c_j}^I(t_0 + j)\}. \end{aligned}$$

3. Empirical Study of Community COVID-19 Spread in the United States

3.1. Nationwide Risk Map

Since the first coronavirus case reported in Washington State in February 2020, the number of confirmed COVID-19 infections has escalated quickly from coast to coast in the United States. New York is the leading state with the largest number of reported COVID-19 infections. As shown in Figure 3A, up to May 2, 2020, the numbers of infected cases for New York, New Jersey, Massachusetts, Illinois, California, Pennsylvania, and Michigan follow similar patterns of the pandemic in different geographic areas. Figure 3B shows the reported cumulative number of deaths caused by COVID-19. New York is the most severely hit state with the largest reported number of deaths (>26,000), while Michigan processes the highest rate of case fatality ($\sim 9.5\%$).

Figure 3. Increasing trends of COVID-19 in heavily hit states of the United States up to May 2, 2020. (A) The cumulative number of reported infections since the number of infections reached 100 per state. (B) The cumulative number of deaths since the number of deaths reached 10 per state.

The daily time series of county-level confirmed infections, deaths and recovered cases are obtained from two data sources: Harvard Dataverse (2020) and 1point3acres (2020). The latter two series are combined to form daily time series of removed cases. We fit the eSAIR model separately for all continental states and Washington, DC. The results of seven severely affected states are shown in Table 1, including both the parameter estimates and their credible intervals for β , γ , and R_0 . The Gelman-Rubin (G-R) statistic proposed by Gelman & Rubin (1992), with the 95% confidence upper bound, is used to monitor MCMC convergence, in addition to other convergence diagnostics (e.g., trace plots and mixing from multiple chains with different initial values). In Table 1, all the G-R statistics for MCMC draws of R_0 in the seven states are close to 1, presenting clearly a piece of evidence for MCMC convergence based on four MCMC chains. Of 48 continental U.S. states, 39 passed the MCMC diagnostic check. The transmission rate, removal rate, and basic reproduction number vary across these states, with Illinois having the highest rate of contagion, $\hat{\beta}_{IL} = 0.067$, 95% CI (0.054, 0.081), and $\hat{R}_{0,IL} = 7.42$, 95% CI (5.46, 9.91). In state-level data analyses based on the temporal eSAIR model, we set the state-specific transmission rate modifier and self-immunization rate. Based on results of the statewide antibody test survey released by New York governor Andrew Cuomo, the cumulative proportion of the population in New York that had antibodies by April 29 is 0.2.³ Due to a lack of similar surveys in the other states, the cumulative proportion of people with antibodies in the other states by April 29 is calibrated proportionally with that of New York with respect to the state-specific basic reproduction number. That is, $P_s = \frac{R_{0,s}}{R_{0,NY}} P_{NY}$, under the assumption that a higher R_0 implies a larger number of

infections, and thus more people having antibodies in a state. Also α_s for the following days is estimated by a state-specific geometric distribution. The resulting cumulative proportion of people with antibodies for each state on April 29 is listed in Table 2. The value of $\pi_s(t)$ is specified by the effectiveness score of state-specific social distancing using cell phone data in the United States from the Transportation Institute at the University of Maryland (2020). Specifically, $\pi_s(t) = 1 - \text{social distancing index}/100$ if the stay-at-home policy does not expire at the prediction day, or as 1.00 if the policy ends at the prediction day. The state-level social distancing scores $\pi_s(t)$ over the period of 7 days for prediction, May 2-9, are listed in Table 2.

Table 1. The estimated transmission rate β , the removal rate γ , and the basic reproduction number R_0 for seven severely hit states in the United States using the eSAIR model.

State	β	γ	R_0	G-R for R_0
California	0.065 (0.043, 0.086)	0.015 (0.010, 0.022)	4.25 (2.91, 5.99)	1.001 (1.003)
Illinois	0.067 (0.054, 0.081)	0.009 (0.007, 0.012)	7.42 (5.46, 9.91)	1.001 (1.003)
Massachusetts	0.065 (0.054, 0.077)	0.011 (0.008, 0.014)	6.21 (4.76, 8.04)	1.002 (1.005)
Michigan	0.058 (0.037, 0.081)	0.020 (0.013, 0.029)	2.88 (1.96, 4.09)	1.000 (1.000)
New Jersey	0.054 (0.040, 0.067)	0.008 (0.006, 0.011)	6.66 (4.75, 8.99)	1.000 (1.000)
New York	0.061 (0.048, 0.075)	0.016 (0.011, 0.021)	3.89 (2.92, 5.15)	1.000 (1.001)
Pennsylvania	0.057 (0.038, 0.075)	0.010 (0.007, 0.014)	5.61 (3.79, 7.86)	1.000 (1.001)

Note. Posterior mean and 95% credible intervals obtained from 200,000 MCMC draws are shown for the three estimated parameters. The Gelman-Rubin (G-R) statistic for the estimation of R_0 is also shown, with the 95% confidence upper bound in parentheses.

Table 2. The cumulative proportion of people with antibodies P_s and the transmission rate modifier $\pi(t)$ for each state.

State	P_s	$\pi(5.2)$	$\pi(5.3)$	$\pi(5.4)$	$\pi(5.5)$	$\pi(5.6)$	$\pi(5.7)$	$\pi(5.8)$	$\pi(5.9)$
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
MN	0.111	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69
MS	0.153	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MO	0.140	0.77	0.77	1.00	1.00	1.00	1.00	1.00	1.00

MT	0.117	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NE	0.153	0.77	0.77	1.00	1.00	1.00	1.00	1.00	1.00
NV	0.140	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
NH	0.136	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
NJ	0.214	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43
NM	0.131	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
NY	0.200	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
NC	0.127	0.73	0.73	0.73	0.73	0.73	0.73	1.00	1.00
ND	0.120	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
OH	0.126	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69
OK	0.122	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
OR	0.153	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66
PA	0.186	0.60	0.60	0.60	0.60	0.60	0.60	1.00	1.00
RI	0.153	0.54	0.54	0.54	0.54	0.54	0.54	1.00	1.00
SC	0.153	0.77	0.77	0.77	1.00	1.00	1.00	1.00	1.00
SD	0.168	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
TN	0.141	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
TX	0.133	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
UT	0.141	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
VT	0.153	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64
VA	0.140	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
WA	0.162	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62

WV	0.120	0.71	0.71	1.00	1.00	1.00	1.00	1.00	1.00
WI	0.123	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74
WY	0.122	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. The value of $\mathbf{p}_s$ is specified as the cumulative proportion of people with antibodies proportional to the reported cumulative proportion of people with antibodies in New York on April 29 and the state-specific estimated basic reproduction number. The value of $\pi(t)$ is set as 1 — social distancing index/100 if the stay-at-home policy does not expire at the prediction day, or as 1.00 if the policy ends at the prediction day ([University of Maryland, 2020](#)). Here $\pi(5.2)$ is the social distancing on May 2, and so on.

From the respective state-level eSAIR models, we can compute the posterior mean probabilities of susceptible, self-immunized, infected and removed, $\hat{\boldsymbol{\theta}}_t = (\hat{\theta}_t^S, \hat{\theta}_t^A, \hat{\theta}_t^I, \hat{\theta}_t^R)^\top$. Using these estimates at time t_0 (say, today), we then apply the spatiotemporal CA-eSAIR model to carry out the county-level risk prediction. To account for potential differences in testing strategies and numbers over counties within a state, we tailor the initial θ_t^I according to the county population size. We propose a shrinkage type estimation, by which the county-specific empirical infection rate $\tilde{\theta}_{t_0}^I$ is shrunk toward the estimated state-level average $\hat{\theta}_{t_0}^I$ by the following rules: (i) If $\tilde{\theta}_{t_0}^I \geq \hat{\theta}_{t_0}^I$, then let the initial infection rate $\bar{\theta}_{t_0}^I = \tilde{\theta}_{t_0}^I$; otherwise, let $\bar{\theta}_{t_0}^I = \varepsilon \tilde{\theta}_{t_0}^I + (1 - \varepsilon) \hat{\theta}_{t_0}^I$, where $\varepsilon = 0, 0.5, 0.75, 1$ if the county population size is $[0, 25000], (25000, 50000], (50000, 100000], (100000, \infty)$, respectively. In the state-level temporal eSAIR analyses, we conduct careful inspections on the convergence of the MCMC runs for each state, where we take a thinning scheme by recording one every 10 random draws to reduce autocorrelation and run four separate chains to determine burn-in and to calculate the Gelman-Rubin statistic. As a result, 39 states passed the MCMC convergence diagnosis. For the other states where the MCMC failed to converge due largely to inadequate data (e.g., very low numbers of deaths), the national average estimates of the model parameters ($\hat{\beta} = 0.077$, $\hat{\gamma} = 0.023$ and $\hat{\boldsymbol{\theta}}_{t_0}$) are used to determine the initial estimates of the infection rate.

To determine the daily rate of self-immunization $\alpha(t)$ in a state, we begin with an assumption that the ratio of the two probabilities or proportions of being asymptomatic and symptomatic is constant. That is, $\alpha/p = c$, where α denotes the probability of a susceptible person being asymptomatic, recovered, and self-immunized with no hospitalization, and p denotes the probability of a susceptible individual contracting the coronavirus and becoming symptomatic. To estimate α and p , we consider an approach based on geometric distributions given as follows. For example, in New York State, the first confirmed COVID-19 case was announced on March 1, and on April 29 the cumulative proportion of self-immunized cases from survey results was reported to be 20%. Adding a 7-day incubation of the first symptomatic case, it is easy to see that

$$\alpha = 1 - (1 - 0.2)^{(1/67)} = 0.00332.$$

Likewise, given that the cumulative infection proportion of confirmed COVID-19 cases on May 6 was 0.017, accounting for a 7-day delay in the reporting of confirmed symptomatic cases from April 29, we obtain

$$p = 1 - (1 - 0.017)^{(1/67)} = 0.000256.$$

Then, $c = \alpha/p = 12.97$, meaning that for one confirmed symptomatic case, there are about 13 asymptomatic cases that are self-immunized. Using this formula $\alpha = cp$ with the estimated constant ratio c , we can determine not only a future value for α with a projected p in New York, which is estimated by θ_t^I from the eSAIR model, but also the daily self-immunization rates $\alpha(t)$ in other states with a suitable state-level proportion of symptomatic cases, p .

Figure 4 shows a nationwide 7-day-ahead risk prediction made on May 2. The values of antibody rate $\alpha_c(t)$ and social distancing score $\pi_c(t)$ are set equal to state-specific values because no detailed county-level data were available. As given in Section 2.2, $\mu_{cc'}$ is the intercounty mobility factor characterizing the decrease of human encounters in terms of their potential movements between counties ([Unacast, 2020](#)); $r(c, c')$ is the travel distance containing a factor $b(a, a')$ that characterizes the transportation capacity of airlines between airports a and a' . Table 3 gives values of $\mu_{cc'}$ obtained from the social distancing scoreboard ([Unacast, 2020](#)). Table 4 presents values of $b(a, a')$ according to year 2016 annual enplanements of U.S. airports. As shown in Figure 4, counties in the region of Massachusetts, Connecticut, New York, and New Jersey appear to have a very high risk of COVID-19 infection due to large numbers of confirmed infections reported in these four states. Other counties with high risks are mainly located in the southeastern United States, including those in Florida, Georgia, Alabama, Mississippi, and South Carolina. Illinois, Texas, and some parts in Arizona are also areas with high risks. Although a large number of COVID-19 infections have been reported in California, counties in the state do not appear to have high risks of infection, due possibly to the large population and an early implementation of preventive measures.

Table 3. The determination of the intercounty mobility $\mu_{cc'}$ parameter in the connectivity coefficient $\omega_{cc'}(t)$.

	A	B	C	D	F
A	0.10	0.25	0.35	0.45	0.55
B	0.25	0.35	0.45	0.55	0.65
C	0.35	0.45	0.55	0.65	0.75
D	0.45	0.55	0.65	0.75	0.85
F	0.55	0.65	0.75	0.85	1.00

Note. Grades for reduction of human encounters: A: >94%; B: 82-94%; C: 74-82%; D: 40-74%; F: <40%. For county c belonging to category A and county c' belonging to category D, the value of $\mu_{cc'} = 0.45$ (Unacast, 2020).

Table 4. The determination of the $b(a, a')$ parameter in the connectivity coefficient $\omega_{cc'}(t)$.

	L	M	S	N
L	0.1	0.3	0.4	0.6
M	0.3	0.4	0.6	0.7
S	0.4	0.6	0.7	0.9
N	0.6	0.7	0.9	1.0

Note. L: Large hub that accounts for at least 1% of total U.S. passenger enplanements (generally 18,500,000 total passengers and above); M: Medium hub that accounts for between 0.25% and 1% of total U.S. passenger enplanements (generally 3,500,000-18,500,000 total passengers); S: Small hub that accounts for between 0.05% and 0.25% of total U.S. passenger enplanements (generally 500,000-3,500,000 total passengers); N: Nonhub that accounts for less than 0.05% of total U.S. passenger enplanements, but more than 10,000 annual enplanements. For airport a belonging to category L and airport a' belonging to category S, the value of $b(a, a') = 0.4$ (Wikipedia, 2020).

It is of interest to compare the risk projection accuracy between the state-level temporal eSAIR model and the county-level spatiotemporal CA-eSAIR model. The former ignores the spatial heterogeneities over counties within a state, and uses the state average infection rate (or set $\varepsilon = 0$) as a constant initial $\bar{\theta}_{t_0}^I$. The latter incorporates the intercounty connectivity and allows county-level variations via the cellular automaton. With regard to the one-day risk prediction of May 3 over 39 states that passed the MCMC convergence diagnosis, we obtain the SWAPE for the eSAIR model equal to 2.85×10^{-3} and that for the CA-eSAIR model equal to 3.54×10^{-4} . The prediction with no use of local information has an 8-fold increase in prediction error than that with the use of county-level data.

Figure 4. Nationwide map of 7-day-ahead projected risk of COVID-19 over 3109 counties in the continental United States from May 2, 2020. Risk is classified into five categories. The bins are defined by the 20th, 40th, 60th, and 80th percentiles of nationwide county-specific risks. The five categories correspond to $[0, 46/10,000)$, $[46/10,000, 80/10,000)$, $[80/10,000, 141/10,000)$, $[141/10,000, 226/10,000)$, and $[226/10,000, 5,637/10,000]$. The state of New York is labeled because the results of antibody surveys were released on April 29, 2020.

As another illustration, we show the 7-day-ahead prediction of the county-level risk in New York State (see Figure 5). We choose to create a risk map for the state because New York released results of state-wide antibody testing surveys on April 29. According to New York governor Cuomo, about 20% of the tested individuals in the state already have the antibodies to COVID-19. Thus, in our analysis, we set the self-immunization rate based on the value of $\alpha = 0.20$ on April 29 and all the other α values are calculated based on this value. This specification may be revised when new antibody survey results become available. In the prediction we use all the continental counties in the United States to calculate the county-level risk scores due to strong ties of the state with other parts of the country; for example, Seattle and New York are close in terms of their air-distance.

Figure 5. A 7-day-ahead risk prediction of COVID-19 for each county in New York State from May 2, 2020. Risk is classified into five categories. The bins are defined by the 20th, 40th, 60th, and 80th percentiles of nationwide county-specific risks. The five categories correspond to $[0, 46/10,000)$, $[46/10,000, 80/10,000)$, $[80/10,000, 141/10,000)$, $[141/10,000, 226/10,000)$, and $[226/10,000, 5,637/10,000]$.

The predicted risk of COVID-19 given in Figure 4 and Figure 5 are calculated using the estimated mean of $(\theta^S, \theta^A, \theta^I, \theta^R)$ from the eSAIR model. To illustrate the size of prediction uncertainty, we also show the prediction range defined as the difference between the upper and lower bounds of the predicted risk scores over the 3109 U.S. continental counties (see Figure 6), as well as the counties in New York state (see Figure 7).

The tuning of the η parameter in the connectivity coefficient $\omega_{cc'}(t)$ is based on the SWAPE, where the weight is the ratio of county population size over the total population size of the 39 states that passed the MCMC convergence diagnosis. The SWAPE optimal value $\eta = 35$ gives the smallest $\text{SWAPE} = 3.54 \times 10^{-4}$. Figure 8A displays the county-level weighted absolute prediction error (WAPE) of the 7-day-ahead projections of infection rates over 3109 counties, along with a scatterplot of county-level weighted prediction errors (WPEs) and $\log_{10}(\text{county population size})$ in Figure 8B. In this figure, the absolute values of county-level weighted prediction errors are used to calculate the SWAPE, with a positive sign corresponding to an overprediction and a negative sign to an underprediction. This figure (8B) shows that the majority of the counties have very small county-level weighted prediction errors, and that counties with large population sizes tend to have larger county-level weighted prediction errors. Such elevated county-level prediction errors occurring in large counties are attributed to more bias in the collected surveillance data. To zoom in the scatterplot, Figure 8B excludes three counties with the largest prediction errors, including Los Angeles county

in California (county-level WPE = -1.030×10^{-5}), Cook county in Illinois (county-level WPE = -1.982×10^{-5}), and New York county in New York (county-level WPE = -4.072×10^{-5}). Figure 9 shows the densities of the county-level WAPes for the 1- to 7-day-ahead infection rate predictions, with the SWAPE from May 3 to May 9 being

$3.54 \times 10^{-4}, 4.01 \times 10^{-4}, 4.60 \times 10^{-4}, 5.16 \times 10^{-4}, 5.83 \times 10^{-4},$
 $6.55 \times 10^{-4}, 7.19 \times 10^{-4},$

respectively. These SWAPes are all at the order of 10^{-4} , namely, one difference for the cumulative number of infections per 10,000 people in a county. For an average county of 100,000 people (in fact 97,118 in our demographic data), with the data up to May 2, the predicted total number of infections on May 3 is about 33 cases more or less than the actual observed number of confirmed infections, and in most of the cases our predicted numbers are larger than the reported. This is due probably to the underreporting issue with the surveillance data. The prediction error increases due to the elevated uncertainty over time. This prediction error should be interpreted with caution, given various potential biases in the collection of confirmed COVID-19 cases in practice ([Angelopoulos et al., 2020](#)). Because the test data may be biased, a small error does not necessarily mean a more accurate prediction.

Figure 6. Nationwide map of the range of 7-day-ahead projected risk of COVID-19 for 3109 counties in the continental United States from May 2, 2020. Range is defined as the difference of the estimated upper and lower confidence bounds of risk. Range is classified into four categories. The bins are defined by the 25th, 50th, and 75th percentiles of nationwide county-specific risk ranges. Larger range value denotes higher uncertainty in risk prediction. The state of New York is labeled since we calculate the self-immunization rate α based on the survey sampling results of New York on April 29.

Figure 7. A map of the range of 7-day-ahead projected risk of COVID-19 for each county in New York from May 2, 2020. Range is defined as the difference of the estimated upper and lower confidence bounds of risk. Range is classified into four categories. The bins are defined by the 25th, 50th, and 75th percentiles of nationwide county specific risk ranges. Larger range value denotes higher uncertainty in risk prediction.

3.2. Projected Risk for a Travel Route

When the nationwide county-level projected risk scores are available over a period of time, such information can be used to assess a potential risk for a planned trip. Let us consider a hypothetical travel path by car from Ann Arbor (Washtenaw County, Michigan) to Detroit (Wayne County, Michigan) and then to Chicago (Cook County, Illinois); see Figure 10. We assume that the traveler would begin the trip on May 2, 2020, and then spend one day in each of these counties for business. Then, the risk of infection during the entire itinerary is calculated as 0.0364 by Equation (5). Using the predicted risk score, the traveler may revise the travel plan if this projected risk is too high.

Figure 8. Distributions of the 7-day-ahead county-level prediction errors for the infection rate. (A) The 7-day-ahead county-level WAPE of infection rate for 3109 counties in the continental United States from May 2, 2020. Seventh-day county-level WAPE (corresponding to predictions made for May 9, 2020) is classified into four categories. The bins are defined by the 25th, 50th, and 75th percentiles of nationwide county-specific 7th-day county-level WAPE values. (B) Scatterplot of county-level weighted prediction errors (WPEs) against $\log_{10}(\text{county population sizes})$. Three counties (Los Angeles, California, Cook, Illinois and New York, New York) are not included in order to amplify the distribution of the points. The WPEs of the 3 counties are -1.030×10^{-5} , -1.982×10^{-5} , and -4.072×10^{-5} , respectively.

Figure 9. Density plots comparing 7-day-ahead $\log_{10}$ (county-level WAPE) for 3109 counties in the continental United States from May 3, 2020 to May 9, 2020. Vertical lines for individual density curves represent the respective median values. The density curves shift to the right from May 3 to May 9, indicating an increase in county-level WAPE resulted from the increasing prediction uncertainty over time.

Figure 10. A hypothetical trip begins on May 2, 2020, with stops in Ann Arbor, Michigan, Detroit, Michigan, and Chicago, Illinois, during May 3-5. Day 1: Washtenaw County (projected risk = 0.0033); Day 2: Wayne County (projected risk = 0.0127); Day 3: Cook County (projected risk = 0.0208).

4. Concluding Remarks

In this article, we propose a spatiotemporal epidemiological forecast CA-eSAIR model that provides timely predictions of community-level COVID-19 infection risk for the 3109 continental U.S. counties. The proposed CA-eSAIR model can make both short-term local risk predictions (e.g., one-day- or 7-day-ahead) and relatively long-term predictions (e.g., one-month-ahead). In addition, the projected county-level risk scores over a time window can be used to calculate a risk for a planned trip in the United States. Such a high-resolution projected risk map is particularly useful for local governments and residents to learn the future spread patterns of the pandemic and their associated personal risks. In this empirical study, we demonstrate that

utilizing local information can greatly improve the prediction accuracy over the classical temporal model with no use of such information.

In addition to the new spatiotemporal prediction model, another major contribution of this article is to incorporate the public survey results of self-immunization as an antibody compartment in the modeling of the infection dynamic system. This new antibody compartment can greatly help circumvent the underreporting issue for data collected in a public database. The current health surveillance system in the United States is incapable of capturing asymptomatic individuals with light symptoms nor is it able to provide sufficient resources for the coronavirus RT-PCR diagnostic test. Recently released surveys of herd immunity in New York, California and Massachusetts have shed light to correct the underreporting. In the future, when more extensive surveys of similar types are to be conducted in the United States, more accurate and more frequent updates of the self-immunization rate can be incorporated in the proposed eSAIR model, even possibly at county level. The resulting CA-eSAIR would make better community-level risk prediction.

In this article, due to various limitations, several functions, for example, $\pi(t)$ and $\omega_{cc'}$, used in the CA-eSAIR model are not fully objective, which need to be improved in the future. The transmission rate modifier $\pi_c(t)$ is included to take time-varying public health interventions (e.g., social distancing) into account. The state-level effectiveness of social distancing is obtained from the published values by the Transportation Institute at the University of Maryland (2020) derived from cell phone mobile data. Our CA-eSAIR model allows county-level value of social distancing, but without access to high-resolution data from the UMD webpage, we must use a state-level value. This may be improved in the future with better data available. To use the proposed CA-eSAIR model, one needs to define a reasonable intercounty connectivity coefficient $\omega_{cc'}(t)$. Since this quantity is related to many variables, it is difficult to determine it with full objectivity. Relaxing this limitation is worthy of a future research project. Relevant to disease contagion, two major factors have been used to form this coefficient. One is an intercounty mobility index derived from mobile cell phone data; the other is a measure of intercounty travel distance. Also, we introduce a tuning parameter in this metric of connectivity to optimize the contribution of travel distance by minimizing the one-day-ahead prediction error. With more personal tracking data available in the future, the quantification of intercounty connectivity can eventually be improved.

One must exercise caution when using the proposed CA-eSAIR model for risk prediction. First, we assume that a person who recovers from the COVID-19 infection is immune to the coronavirus and no longer contagious. This assumption is likely to be true but has not been justified yet. Second, although the underreporting of confirmed infections is addressed by the compartment of antibody, the underreporting of deaths from COVID-19 remains. This gives rise to a potential bias affecting the prediction accuracy. Not only is this a public health surveillance issue but a fundamental research topic in epidemiology. In reality, it is often difficult to determine a defining cause to a death with high certainty. This problem becomes even more

complicated in the case of a death from the coronavirus because the medical diagnosis of “COVID-19 disease” is not clinically well defined yet.

It is noteworthy that the predicted risks at some state borders (e.g., Idaho and Nebraska) in Figure 4 exhibit nonsmooth changes. This may be an outcome of the initial values generated from state-level eSAIR analyses in that we assume both control measures and testing policies and strategies to be state-specific, as well as the percentile-based categorization for the need of color coding in the plot. Such within-state homogeneity is also used in the prediction, resulting in more homogeneous intrastate projected risks than the interstate projected risks. Consequently, some counties at state borders appear to have noticeable discrepancies in their projected risks. Resolving these differences at state borders may not be that simple and is beyond the capacity of our current methodology. It would be interesting future work that develops a method to discern the true interstate differences from artifacts in the risk prediction for counties at state borders, where a certain objective criterion is the key to gauge adequate smoothness.

Further directions of the risk-mapping method include extending the proposed CA-eSAIR model to predict the risk of the COVID-19 in other areas across the world, or to predict the risk at locations of household, factory, and business sites. The latter provides finer-scale information of COVID-19 risk critical for business reopening and allocation of resources for viral RT-PCR diagnostic tests. This in-depth analysis needs more local data of extensive viral diagnostic and antibody tests available across the country and individual health and behavior data from better surveillance systems such as mobile device apps for the tracking of personal movements. The CA-eSAIR model provides a useful prediction paradigm to develop data-driven strategies for precision public health intervention and strategies for COVID-19 containment.

Acknowledgments

The authors thank the Editor, Associate Editor, Data Visualization Editor, and four anonymous reviewers for their insightful comments and constructive suggestions that helped us improve the manuscript greatly. The authors also appreciate the Health Data Science Concentration offered by the Department of Biostatistics at the University of Michigan, through which five first-year MS students (Zhang, Shi, Yang, Zhao, and Overton) had the opportunity to become involved in the COVID-19 data collection and analysis. The authors are grateful to Drs. Bhramar Mukherjee and Jian Kang for their valuable suggestions on this work.

Disclosure Statement

This research is partially supported by the NSF grant DMS1811734.

References

1point3acres. (2020). *Global COVID-19 tracker and interactive charts*. <https://coronavirus.1point3acres.com/>

- Ahmed, E., & Agiza, H. (1998). On modeling epidemics including latency, incubation and variable susceptibility. *Physica A: Statistical Mechanics and Its Applications*, 253(1–4), 347–352. [https://doi.org/10.1016/S0378-4371\(97\)00665-1](https://doi.org/10.1016/S0378-4371(97)00665-1)
- Angelopoulos, A. N., Pathak, R., Varma, R., & Jordan, M. I. (2020). On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate. *Harvard Data Science Review*, (Special Issue 1). <https://doi.org/10.1162/99608f92.f01ee285>
- Beauchemin, C., Samuel, J., & Tuszynski, J. (2005). A simple cellular automaton model for influenza A viral infections. *Journal of Theoretical Biology*, 232(2), 223–234. <https://doi.org/10.1016/j.jtbi.2004.08.001>
- Chopard, B., & Droz, M. (1998). *Cellular automata* (Vol. 1). Springer.
- Fuentes, M., & Kuperman, M. (1999). Cellular automata and epidemiological models with spatial dependence. *Physica A: Statistical Mechanics and Its Applications*, 267(3–4), 471–486. [https://doi.org/10.1016/S0378-4371\(99\)00027-8](https://doi.org/10.1016/S0378-4371(99)00027-8)
- Fuks, H., & Lawniczak, A. T. (2001). Individual-based lattice model for spatial spread of epidemics. *Discrete Dynamics in Nature and Society*, 6(3), 191–200. <https://doi.org/10.1155/S1026022601000206>
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. <https://doi.org/10.1214/ss/1177011136>
- Jørgensen, B., & Song, P. X.-K. (2007). Stationary state space models for longitudinal data. *Canadian Journal of Statistics*, 35(4), 461–483. <http://doi.org/10.1002/cjs.5550350401>
- Karney, C. F. (2013). Algorithms for geodesics. *Journal of Geodesy*, 87(1), 43–55. <https://doi.org/10.1007/s00190-012-0578-z>
- Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A*, 115(772), 700–721. <https://doi.org/10.1098/rspa.1927.0118>
- Lab, C. D. (2020). US COVID-19 daily cases with basemap. *Harvard Dataverse*. <https://doi.org/10.7910/DVN/HIDLTK>
- Office of the Governor, N. Y. (2020). *Amid ongoing COVID-19 pandemic, Governor Cuomo announces results of completed antibody testing study of 15,000 people showing 12.3 percent of population has COVID-19 antibodies*. <https://www.governor.ny.gov/news/amid-ongoing-covid-19-pandemic-governor-cuomo-announces-results-completed-antibody-testing>
- Schneckenreither, G., Popper, N., Zauner, G., & Breitenecker, F. (2008). Modelling SIR-type epidemics by ODEs, PDEs, difference equations and cellular automata—A comparative study. *Simulation Modelling Practice*

and Theory, 16(8), 1014–1023. <http://doi.org/10.1016/j.simpat.2008.05.015>

Sirakoulis, G. C., Karafyllidis, I., & Thanailakis, A. (2000). A cellular automaton model for the effects of population movement and vaccination on epidemic propagation. *Ecological Modelling*, 133(3), 209–223. [https://doi.org/10.1016/S0304-3800\(00\)00294-5](https://doi.org/10.1016/S0304-3800(00)00294-5)

Song, P. X.-K. (2000). Monte Carlo Kalman filter and smoothing for multivariate discrete state space models. *Canadian Journal of Statistics*, 28(3), 641–652. <https://doi.org/10.2307/3315971>

Unacast. (2020). *Social distancing scoreboard*. <https://www.unacast.com/covid19/social-distancing-scoreboard>

University of Maryland. (2020). University of Maryland COVID-19 Impact Analysis Platform, Maryland Transportation Institute.

Von Neumann, J., & Burks, A. (1966). *Theory of self reproducing automata*. University of Illinois press, Urbana.

Wang, L., Zhou, Y., He, J., Zhu, B., Wang, F., Tang, L., Kleinsasser, Michael, Barker, D., Eisenberg, M., & Song, P. X. K. (2020). An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China (with discussion). *Journal of Data Science*, 18(3), 409–432. [https://doi.org/10.6339/JDS.202007_18\(3\).0003](https://doi.org/10.6339/JDS.202007_18(3).0003)

White, S. H., Del Rey, A. M., & Sánchez, G. R. (2007). Modeling epidemics using cellular automata. *Applied Mathematics and Computation*, 186(1), 193–202. <https://doi.org/10.1016/j.amc.2006.06.126>

Wikipedia. (2020). *List of airports in the United States*. https://en.wikipedia.org/wiki/List_of_airports_in_the_United_States

Wilcox, R., Grammaticos, B., Carstea, A., & Ramani, A. (2003). Epidemic dynamics: Discrete-time and cellular automaton models. *Physica A: Statistical Mechanics and Its Applications*, 328(1–2), 13–22. [http://doi.org/10.1016/S0378-4371\(03\)00552-1](http://doi.org/10.1016/S0378-4371(03)00552-1)

World Health Organization. (2020). *Naming the coronavirus disease (COVID-19) and the virus that causes it*. [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it)

©2020 Yiwang Zhou, Lili Wang, Leyao Zhang, Lan Shi, Kangping Yang, Jie He, Bangyao Zhao, William Overton, Soumik Purkayastha, and Peter Song. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.

Footnotes

1. See the details in Appendix A of Wang et al. (2020). [↵](#)
2. See more details in Wang et al. (2020) and an R package at <https://github.com/lilywang1988/eSIR>. [↵](#)
3. Refer to Cuomo’s press conference (Office of the Governor, 2020) for the details of the antibody testing survey. [↵](#)

References

- 1point3acres. (2020). *Global COVID-19 tracker and interactive charts*. [↵](#)
- Ahmed, E., & Agiza, H. (1998). On modeling epidemics including latency, incubation and variable susceptibility. *Physica A: Statistical Mechanics and Its Applications*, 253(1–4), 347–352. [↵](#)
- Angelopoulos, A. N., Pathak, R., Varma, R., & Jordan, M. I. (2020). On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate. *Harvard Data Science Review*. [↵](#)
- Beauchemin, C., Samuel, J., & Tuszynski, J. (2005). A simple cellular automaton model for influenza A viral infections. *Journal of Theoretical Biology*, 232(2), 223–234. [↵](#)
- Chopard, B., & Droz, M. (1998). *Cellular automata* (Vol. 1). Springer. [↵](#)
- contributors, W. (2020). *List of airports in the United States* — *Wikipedia, The Free Encyclopedia*. [↵](#)
- Fuentes, M., & Kuperman, M. (1999). Cellular automata and epidemiological models with spatial dependence. *Physica A: Statistical Mechanics and Its Applications*, 267(3–4), 471–486. [↵](#)
- Fuks, H., & Lawniczak, A. T. (2001). Individual-based lattice model for spatial spread of epidemics. *Discrete Dynamics in Nature and Society*, 6(3), 191–200. [↵](#)
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. [↵](#)
- Jørgensen, B., & Song, P. X.-K. (2007). Stationary state space models for longitudinal data. *Canadian Journal of Statistics*, 35(4), 461–483. [↵](#)
- Karney, C. F. (2013). Algorithms for geodesics. *Journal of Geodesy*, 87(1), 43–55. [↵](#)
- Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772), 700–721. [↵](#)
- Lab, C. D. (2020). *US COVID-19 daily cases with basemap*. Harvard Dataverse. <https://doi.org/10.7910/DVN/HIDLTK> [↵](#)
- of Maryland, U. (2020). *University of Maryland COVID-19 Impact Analysis Platform, Maryland Transportation Institute*. [↵](#)
- of Maryland, U. (2020). *University of Maryland COVID-19 Impact Analysis Platform, Maryland Transportation Institute*. [↵](#)
- Organization, W. H. (2020). *Naming the coronavirus disease (COVID-19) and the virus that causes it*. [↵](#)

- Schneckenreither, G., Popper, N., Zauner, G., & Breitenecker, F. (2008). Modelling SIR-type epidemics by ODEs, PDEs, difference equations and cellular automata—A comparative study. *Simulation Modelling Practice and Theory*, 16(8), 1014–1023. [↵](#)
- Sirakoulis, G. C., Karafyllidis, I., & Thanailakis, A. (2000). A cellular automaton model for the effects of population movement and vaccination on epidemic propagation. *Ecological Modelling*, 133(3), 209–223. [↵](#)
- Song, P. X.-K. (2000). Monte Carlo Kalman filter and smoothing for multivariate discrete state space models. *Canadian Journal of Statistics*, 28(3), 641–652. [↵](#)
- Unacast. (2020). *Social distancing scoreboard*. [↵](#)
- Von Neumann, J., & Burks, A. (1966). *Theory of self reproducing automata*. University of Illinois press, Urbana. [↵](#)
- Wang, L., Zhou, Y., He, J., Zhu, B., Wang, F., Tang, L., Kleinsasser, Michael, Barker, D., Eisenberg, M., & Song, P. X. K. (2020). An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China (with discussion). *Journal of Data Science*, to appear. [↵](#)
- White, S. H., Del Rey, A. M., & Sánchez, G. R. (2007). Modeling epidemics using cellular automata. *Applied Mathematics and Computation*, 186(1), 193–202. [↵](#)
- Willox, R., Grammaticos, B., Carstea, A., & Ramani, A. (2003). Epidemic dynamics: Discrete-time and cellular automaton models. *Physica A: Statistical Mechanics and Its Applications*, 328(1–2), 13–22. [↵](#)