Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Community venue exposure risk estimator for the COVID-19 pandemic

Ziheng Sun\textsuperscript{a,b,*}, Liping Di\textsuperscript{a,b}, William Sprigg\textsuperscript{c}, Daniel Tong\textsuperscript{a,d}, Mariana Casal\textsuperscript{e}

\textsuperscript{a} Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA, USA
\textsuperscript{b} Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, USA
\textsuperscript{c} Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, AZ, USA
\textsuperscript{d} Department of Atmospheric, Oceanic and Earth Sciences, George Mason University, Fairfax, VA, USA
\textsuperscript{e} Infectious Disease and Epidemiology Section, Pinal County Health Department, Florence, AZ, USA

\begin{abstract}
Complexities of virus genotypes and the stochastic contacts in human society create a big challenge for estimating the potential risks of exposure to a widely spreading virus such as COVID-19. To increase public awareness of exposure risks in daily activities, we propose a birthday-paradox-based probability model to implement in a web-based system, named COSRE (community social risk estimator) and make in-time community exposure risk estimation during the ongoing COVID-19 pandemic. We define exposure risk to mean the probability of people meeting potential cases in public places such as grocery stores, gyms, libraries, restaurants, coffee shops, offices, etc. Our model has three inputs: the real-time number of active and asymptomatic cases, the population in local communities, and the customer counts in the room. With COSRE, possible impacts of the pandemic can be explored through spatiotemporal analysis, e.g., a variable number of people may be projected into public places through time to assess changes of risk as the pandemic unfolds. The system has potential to advance understanding of the true exposure risks in various communities. It introduces an objective element to plan, prepare and respond during a pandemic. Spatial analysis tools are used to draw county-level exposure risks of the United States from April 1 to July 15, 2020. The correlation experiment with the new cases in the next two weeks shows that the risk estimation model offers promise in assisting people to be more precise about their personal safety and control of daily routine and social interaction. It can inform business and municipal COVID-19 policy to accelerate recovery.
\end{abstract}

1. Introduction

1.1. Motivation

History repeats from pandemic to pandemic (Ghendon, 1994). Reducing the impact of pandemics relies on early detection and mitigation strategies that slow the spread to allow more preparation and reduce impacts on patient care. However, as the COVID-19 pandemic continues to spread, the burden weighs heavy on all of us to control further spread and consequences through social distancing. Lifestyles leading up to this pandemic are capable of spreading disease further and faster than ever before. New and more intense factors amplify disease transmission (Dashraath et al., 2020). The risks apply reasonably well to nearly all densely populated areas. Environmental changes such as disruptive climate, may also contribute (Ogen, 2020). As coronavirus is monitored across the U.S., some communities see a diminishing number of cases, while others have yet to see the peak. Reporting errors and delays also play a role in changes in disease incidence. An objective, publicly-accessible means to estimate risk of infection is needed in order to cope safely with the consequences of social distancing and sequestration. Although various self-quarantine and self-isolation policies are in place in most of the U.S., the virus is not universally contained. People need to visit grocery stores, pharmacies and hospitals. They wish to attend sport, recreation and entertainment events, and community activities and family gatherings. Considering the peak in coronavirus cases might have passed, some states have started to reopen business to restore the economy. Risk assessment tools are essential in order to assist reopening with key information, as to where and when it may be safest to return to places of need or interest.

Reopening is a very daunting challenge for many businesses in the
United States this year (WORKPLACES, 2020). The duration of the pandemic is uncertain and its influence could last for a long time. The current guidelines for reopening are very general and not very customizable to accommodate the real scenarios of life in different communities. Communities differ in population, store types, average visit time, patients, air quality, etc. There is no one-size-for-all solution to cope. A safer plan should be able to flexibly fit each circumstance by considering specific factors like the locally confirmed cases, potential asymptomatic patients, store space, layout, people density, local population composition, commodities, age distribution, capacity limit, social distancing, temperature check, violation enforcement, etc. However, precise recommendations are rarely available for business managers to use. Very few quantitative models can estimate the impact of varying these factors. Thus, few individual businesses can tailor their reopening policies to make smarter decisions in balancing life, safety, and economy.

1.2. Significance

To find a balancing point between a full lockdown in fear of the virus and a plan of reopening businesses out of socioeconomic concerns, a mechanism to help assist that decision is urgent. The decision-makers to be supported include municipal policymakers, business managers and customers. The mechanism, or system, should provide a customized, real-time risk estimation for store managers, who may then adapt their operations to the unpredictable pandemic. An estimation of risk will also give potential customers the information necessary to assess their own risks of visiting businesses or other venues in an area, with increased awareness and confidence about their decision.

To achieve that goal, this project proposes a straightforward social probability model based on the birthday paradox theory (if n people are selected at random, what is the probability that at least r people will have the same birthday?) (McKinney, 1966; Wagner, 2002) to estimate the venue risk of people meeting at least one COVID case, either asymptomatic, pre-symptomatic or actively symptomatic, in various public venues such as shopping centers, grocery stores, gyms, recreational areas, restaurants, DMVs (Departments of Motor Vehicles), and federal agency public field offices. A prototype web-based system named COSRE (community social risk estimator) (Sun, 2020a) is implemented to take zip codes as inputs and output a percentage value indicating the probability of acquiring the virus in the community. Risk is calculated based on real-time data collected from public resources. The purpose of the tool is to give people a reasonable quantitative estimate of their risks of exposure to the COVID-19 virus in their communities.

2. Background

Pandemics are large scale outbreaks of infectious diseases that can increase mortality significantly over great geographic area and cause significant economic, social and political disruption (MadHAV et al., 2017). Extensive global travels and long-distance contacts in human society raise the likelihood of pandemics. One recent example, the 2003 severe acute respiratory syndrome (SARS) pandemic, was a critical threat to public health (Ksiazek et al., 2003). The World Health Organization compels member states to meet specific standards to detect, report on, and respond to infectious disease outbreaks. This international framework contributed to a more coordinated global response during the 2009 influenza pandemic than in the previous 2003 event (Katz, 2009). Global public health departments look to improve pandemic preparedness through refined standards and responding plans to flatten the time-incidence curves and reduce the deaths. Management of community risks in a pandemic needs to apply more restrictive emergency management strategies than other hazards. The objectives of risk management are to strengthen our responding capacities to contain the diseases, enable and promote linkage and integration across the governments and societies. Accurate estimation of risks is the first and fundamental step in making an effective risk management plan. Common responding practices are to limit the point of entry to reduce the possibility of virus traveling. People, communities, states, and countries should continue to communicate with information and advice. Precise information provided early and often will enable the community to correctly understand the health risks they face, and will make it easier to engage in actions to protect themselves (WHO, 2018).

The SARS-COV-2 virus was first isolated from a patient with a pneumonia of unknown origin in Wuhan, China. Genetic analysis revealed that it is closely related to SARS-Cov-1 and genetically clusters within the genus Betacoronavirus, subgenus Sarbecovirus (Cont, 2020). Although pathology studies have been done for the COVID-19 disease, as an ongoing pandemic few tools are available to evaluate the real-time social exposure risks. An accepted classification groups risk assessment into several levels: little, lower, medium, high, very high, and severe. Nicholas et al. proposed a conceptual full-risk-spectrum, Comprehensive Pandemic Risk Management System (CPRMS) to prevent, prepare for, respond to, and mitigate the multisector impacts of severe pandemics (Studzinski, 2020). It contains six institutional building blocks in the global domain: governance and leadership, sustainable financing, information and knowledge systems, human capital resources, essential commodities and logistics, and operational service delivery. But the detailed framework still needs refinement in order to implement in each geographical area (Studzinski, 2020). In the COVID-19 pandemic, there is still a long way to go to implement such a framework.

The US Centers for Disease Control and Prevention (CDC) has developed an online tool called Influenza Risk Assessment Tool (IRAT) to assess the potential pandemic risk posed by influenza A viruses (Cox et al., 2014). The IRAT uses 10 scientific criteria to measure the potential pandemic risk associated with each of the life scenarios mentioned previously. These 10 criteria can be grouped into three overarching categories: “properties of the virus”, “attributes of the population”, and “ecology & epidemiology of the virus”. A composite score for each virus is calculated based on the given scenario. The score gives the means to rank and compare influenza viruses to each other in terms of their potential pandemic risks. It is an evaluation tool, not a predictive tool (CDC, 2020). The tool should include actual knowledge about the virus. However, for a novel virus, like SARS-COV-2, the transmissible properties for the virus are ambiguous; this makes the tool’s reliability for COVID-19 risk evaluation uncertain.

To take timely, effective and appropriate actions to slow an ongoing disease outbreak, it is vital to identify risk factors such as age, gender, occupation, and health conditions. With these, an estimate can be made of death risk in real-time. For infectious risk management, Jung et al. proposed an approach for real-time estimation of the risk of death from COVID-19 infection by inference using the reported cases (Jung et al., 2020). They used the exponential growth rate of the incidence to estimate the confirmed case fatality risk (cCFR) in mainland China. Tanaka et al. studied the reproduction number of COVID-19 and tried to get more accurate R0 (pronounced “R naught”), the average number of people who will contract a contagious disease from one person with that disease) by using time-dependent contact and diagnosis rates to retrain their dynamics transmission model (Tanaka et al., 2020). Castro et al. presented a quantitative framework for real-time ZIKV (short for Zika virus, a mosquito-borne flavivirus) risk assessment to capture uncertainty in case reporting, importations, and vector-human transmission dynamics (Castro et al., 2017). Their approach is divided into three sections: (1) co-factors that influence ZIKV importation and relative transmission rates, (2) country-specific ZIKV outbreak simulations, and (3) ZIKV risk analysis. They fit a probabilistic model (maximum entropy) to build a predictive model for the virus import risk.

Similar kinds of geographical risk models have been developed for other communicable diseases. Cromley discussed how to use GIS technology and related technologies to analyze the geography of disease, the relationships between pathological factors and their environments (Cromley, 2003). To develop a predictive risk mapping of leptospirosis,
an environmentally-driven infectious disease using spatial Bayesian network was proposed (Mayfield et al., 2018). In a risk model developed for Rift Valley Fever, exposure was measured as the proportion of total outbreak years for each district in prior epizootics, whereas the district’s risk of outcome was assessed as severity of observed disease in humans and animals (Munyua et al., 2016). Berke et al. proposed a new approach for spatial relative risk mapping as a tool for geographical risk assessment based on data of pseudorabies virus infections at the farm level in a region of high animal density in Germany (Berke and Grosse Beilage, 2003). Waller and colleagues extended hierarchical spatial models to create spatio-temporal maps of disease rates (Waller et al., 1997). The methods of assessing and mapping the potential economic and public health risks associated with avian influenza outbreaks have been discussed (Dudley, 2008). Olaf Berke proposed an exploratory approach to map spatial relative risk using the background risk in the unexposed population. The results can reveal the importance and geographical distribution of unknown spatial risk factors (Berke, 2005).

Another paper offered a critical review of the methods of disease mapping and spatial regression based on male lip cancer incidence data in Scotland (Wakefield, 2007). However, these existing models require input data that are unavailable or require contact tracing, or the targeted population is assumed unable to infect others and are not counted in potential exposed cases; (Flajolet et al., 1992) was reused here. The probability distribution is similar to the Binomial distribution (Altham, 1978). The basic formula is:

\[ Pr(p, k) = \begin{cases} \frac{\binom{k}{n} p^n (1-p)^{k-n}}{\binom{n}{a_i} p^{a_i} (1-p)^{n-a_i}} & \text{if } p \neq 0 \text{ and } n \neq 0 \text{ and } a_i \neq 0 \\ 0 & \text{if } p = 0 \text{ or } n = 0 \text{ or } a_i = 0 \end{cases} \]

where \( Pr \) is the probability function of meeting someone with active disease; \( p \) is the total community population (town, village, city, county, region, country); \( a_i \) and \( n \) is the total number of potential COVID-19 cases in the area on day \( i \) and day \( i-d \) (defined as all accumulated COVID hosts in the community at the given time, to include asymptomatic, symptomatic, hospitalized and self-quarantined; \( d \) is the average number of days that the disease takes for people to recover or die (~30 days for COVID-19); people who have recovered from the virus are assumed unable to infect others and are not counted in potential cases); \( n \) is the number of the people in the business venue, e.g. grocery stores, shopping centers, gyms, restaurants, workplaces and recreational areas (Sun, 2020b).

First, we calculate the odds of NOT meeting any infected person and subtract those odds from 1 to get the probability of meeting at least one infected person in that specific group of people.

This model assumes that the 100 people in the population \( p \) visit a store with equal chance. If the chance is not equal for everyone, the values of \( n \) and \( p \) should be changed to represent their varying

3. Data and method

3.1. Data

This study uses data from reliable providers. The COVID data comes from John Hopkins University CSSE (Center for Systems and Science Engineering) (Dong et al., 2020), the population data for zip codes and the contiguous United States map are from the United States Census (2010). The COVID data updates daily. The population data of most counties were collected in 2010. The people in the room is assumed to be 100. Estimations are made on the county level.

3.2. Risk estimator core algorithm

The idea of calculating birthday paradox probability (Flajolet et al., 1992) is reused here. The probability distribution is similar to the Binomial distribution (Altham, 1978). The basic formula is:

\[ Pr(p, k) = \begin{cases} \frac{\binom{k}{n} p^n (1-p)^{k-n}}{\binom{n}{a_i} p^{a_i} (1-p)^{n-a_i}} & \text{if } p \neq 0 \text{ and } n \neq 0 \text{ and } a_i \neq 0 \\ 0 & \text{if } p = 0 \text{ or } n = 0 \text{ or } a_i = 0 \end{cases} \]

where \( Pr \) is the probability function of meeting someone with active disease; \( p \) is the total community population (town, village, city, county, region, country); \( a_i \) and \( n \) is the total number of potential COVID-19 cases in the area on day \( i \) and day \( i-d \) (defined as all accumulated COVID hosts in the community at the given time, to include asymptomatic, symptomatic, hospitalized and self-quarantined; \( d \) is the average number of days that the disease takes for people to recover or die (~30 days for COVID-19); people who have recovered from the virus are assumed unable to infect others and are not counted in potential cases); \( n \) is the number of the people in the business venue, e.g. grocery stores, shopping centers, gyms, restaurants, workplaces and recreational areas (Sun, 2020b).

First, we calculate the odds of NOT meeting any infected person and subtract those odds from 1 to get the probability of meeting at least one infected person in that specific group of people.

This model assumes that the 100 people in the population \( p \) visit a store with equal chance. If the chance is not equal for everyone, the values of \( n \) and \( p \) should be changed to represent their varying

3.3. Result interpretation

This risk estimation model can help people to evaluate risk and react accordingly instead of panicking or just being indifferent towards the virus. The formula takes three parameters as inputs and the output is the probability of meeting at least one infected person in the store. The result ranges from 0 to 1. The risk can be generally interpreted as the probability of exposure to another COVID-infected person. Besides, for policy making, it may be useful to classify the probabilities into several levels. For example: the risk less than 25% may be considered relatively low, 25–50% means the probability is medium, 50–75% reflect the place has a high probability for meeting a SARS-COV-2 carrier, >75% means the risk is very high and people might think twice before taking actions. This is just an example and the classification should be validated with real exposure data collected in businesses and public areas. Based on the correlation analysis results, policymakers and community members will have a better understanding of the risk when going outside in the community and may take corresponding actions to avoid being exposed in their specific county.

3.4. Evaluation

To evaluate the proposed risk score, we use several correlation methods to analyze its relationships with new case data for the next two weeks. For non-spatial correlation analysis, Spearman’s correlation coefficient (Spearman’s \( p \)) (Spearman, 1961), a statistical measure of the strength of a monotonic relationship between two variables, is utilized. Its interpretation is similar to Pearson’s \( R \) (Pearson, 1895) and can be used for analyzing non-linear relationships. In addition, spatial autocorrelation analysis is used to study the relationship between risk estimates in one county and those in neighboring counties. If there is a significant correlation, a cluster will be generated to identify the hot spot regions (Oster, 2020). Moran’s I is the most common indicator for spatial correlation estimation. Moran, 1950. Moran’s I ranges from –1 to 1. Large and positive Moran’s I values mean that the nearby areas tend to have similar risk values. In a random spatial distribution, Moran’s I will be close to zero. Univariate Moran’s I, Bivariate Moran’s I, and Differential Moran’s I are used to analyze the correlation between COSRE risk estimates and the new cases two weeks later and the spatiotemporal trends of COSRE risks. The tools utilized include Python matplotlib (Hunter, 2007), GeoPandas (Jordahl, 2014), PyShp (2018) and GeoDa (Anselin et al., 2010).

4. Results

4.1. Risk mapping

Using the proposed algorithm, we calculated the risks for all counties of the contiguous United States and generated a series of maps since Apr 1, 2020. Fig. 1 shows six of the maps on the dates of April 1, April 15, May 1, May 15, July 1, July 15, 2020, from which the trends of social
Fig. 1. Community Exposure Risk of the United States on April 1, April 15, May 1, May 15, July 1, and July 15, 2020. All the maps are in EPSG:4269 projection.
Fig. 1. (continued).
The map of April 1 (Fig. 1a) shows that New York City, Albany GA, New Orleans, Denver, Salt Lake City, and Sun Valley in Idaho became the centers with highest exposure risks. On the contrary, the two states, California and Washington, where the very first COVID-19 patients were found, have relatively low risk. Overall, the epidemic centers had emerged by April 1 and began to spread among neighboring counties. At that time, from abundant news media accounts, most states had instituted social distancing and stay-at-home orders. These orders required people to stay at least 6 feet apart and avoid gathering in groups. Most mass gatherings were canceled following CDC guidelines. However, some virus-affected people started to show symptoms and confirmed cases were reported from coast to coast. As Fig. 1a displays, almost every county showed social risk, but risks were low compared to succeeding months, with the exception of several states that exhibited outbreaks. On April 1st, risks were highest in the metropolitan areas of New York City, New Orleans, and Denver, and in Blaine County, Idaho, and Summit County, Utah. A potential Outbreak could be foreseen from these areas as the time-incidence curves began to rise.

The April 15 map (Fig. 1b) shows that serious areas on April 1 grew worse after two weeks. More counties became red, and red counties became redder. The outbreak centers on April 1st changed little, while areas surrounding them became red. The virus spread fast from New York, Albany and New Orleans, less so than from Denver, Salt Lake City and Sun Valley. New York City was hardest hit as neighboring counties in New Jersey and Connecticut grew deep red. So, too, were counties in Louisiana and Mississippi. In the case of Louisiana and Mississippi, most counties’ confirmed cases were high compared to their populations.

It is important to note at this point that state-to-state and region-to-region comparisons of critical parameters, such as confirmed cases or population-patient ratios, depend on consistency and calibration of methods, tools, training and, in some cases, infrastructure. With such inter-comparability, the information content of the raw data and model output increases in important ways, e.g., nation-wide priority allocation of medical resources and emergency assistance.

Another example may be found in the contiguous counties in northern Arizona, northeastern New Mexico. Their change to deeper pink in the April 15 map from the April 1 map signals a center of high risk. Navajo county in Arizona only has a population of 110,924 (survey in 2019), but 390 confirmed COVID-19 cases make a very high case to population ratio. Unlike metropolitan areas, Navajo county has fewer stores and public places, which would make it even risker if businesses reopen and people gather. Also, in mid-April, the Detroit and Chicago metropolitan areas show signs of spreading incidence. Other regional differences appear. For example, risks of contagion in Washington and California remain relatively low compared to outbreak centers in the U.S. south and east. Florida shows a pattern of relatively high risk around Miami and less risk elsewhere.

The map of May 1 (Fig. 1c) shows that New York, Pennsylvania, New Jersey, and Massachusetts are at high exposure risk in their coastal counties. New York City and its associated counties are the areas of highest risk. For these counties, the risk map highlights the importance of following health guidelines to wear masks and maintain social distancing in public places.

Other east coast states, including Maryland, Virginia, Washington D.C., North Carolina, South Carolina all turn darker red. Inland regions, such as western Pennsylvania, West Virginia, western Virginia, western North Carolina, eastern Tennessee, show relatively lower exposure risks. On May 1 (Fig. 1c), the populous counties in the midwestern states like Iowa, Nebraska, Ohio, Indiana, Kentucky, Minnesota, have turned dark red. But there is less outward spread than in highly urbanized areas like New York. The risks are restricted in certain counties and show no obvious signs of a state-wide outbreak. Those red counties are scattered and isolated. One explanation is that the other less populous counties intentionally reduced their contacts with the populous counties. The connections among different counties in these regions are not as close as those in the east coast counties. The risk in Washington, California, Nevada, Texas, and Florida, is rising slowly and remains lower than in the other outbreak states.

According to the map of May 15 (Fig. 1d), the riskiest areas are still along the east coast, however with a sign that risks may be on the decline in some suburban counties in New York, Pennsylvania and, further south. In Georgia, counties surrounding Columbus County saw less risk, suggesting the new case load may have peaked. New York appears to show a downward trend in new confirmed COVID-19 cases. High-risk areas on July 1 (Fig. 1e) and July 15 (Fig. 1f) include, high-populated counties in the Mississippi basin and the corn belt states, which have become darker, and western counties in Oklahoma and Kansas show that coronavirus has spread further (Fig. 1e). Also, risk across Georgia, Florida, Texas, and Louisiana has turned more serious, as has risk in southern California which has become higher after staying low in the first two months. Other concentrations include counties in northern Arizona and in northwestern New Mexico with large native American populations. In addition, the Miami metropolis turns red and seems headed for dark red on July 1. Conversely, New York and the New England states have remained low and changed little in July.

Strategic as well as tactical operations’ officers find evidence in these maps to focus attention on the corn belt states, including Illinois, Michigan, Ohio, Kentucky, Tennessee, Wisconsin, Iowa, Indiana and Missouri. None show downward trend of COVID-19 case incidences. However, in many of these states, stay-at-home rules were slackened in order to “reopen” and spur the economy. Health experts warn that without data and without proper assessment, decisions to reopen will likely result in a spike of coronavirus cases. Yet, the success of any reopening will depend on decisions made by each and every citizen. Our research helps inform everyone’s individual, smart choices during reopening.

4.2. Spearman correlation with the upcoming new cases

To evaluate the reliability of the COSRE risk score, we validate its accuracy by comparing the risk score for a county with the number of new COVID-19 cases confirmed in the following two weeks. A two-week interval was used because asymptomatic patients are normally detected within two weeks of infection. Fig. 2 shows the Spearman correlation analysis results. In Fig. 2, the x axis is the COSRE risk score at time t, the y axis is new cases per thousand people in the next two weeks after t. Each spot in the plot represents a county. Spearman’s correlation coefficient is consistently above 0.6, which falls in the “strong” category.1 That means a large COSRE score is strongly correlated to more new cases in two weeks, and a small COSRE score is strongly correlated to fewer new cases in two weeks. Chances of coincidence are diminished because all the time steps have the similar correlation. The results show that the COSRE risk scores well represent future exposure risks. Specifically, the April results show that risks for most counties are below 30% while the new cases per thousand in two weeks are less than 5. On April 1 the confirmed numbers surpass exposure risks in many counties (many dots are above the diagonal line). This is probably caused by insufficient testing in the early stage and the risk is underestimated. In May, the increase of risk scores doesn’t correspond to obvious increase in new cases. The new cases per thousand for most counties remain under 5. One reason is that people became concerned and most states have stay-at-home orders in place. So even though the confirmed cases are climbing, the number of people with confirmed cases in public places are actually much lower than expected, resulting in low numbers of infections. Things changed in July. Stay-at-home orders in most states expired, and people gradually returned to normal life routines. The Spearman’s R values are higher in July than in May, meaning that the

1 http://www.statsfutor.ac.uk/resources/uploaded/spearmans.pdf.
increased new cases are more positively correlated to COSRE risk score. In summary, the Spearman correlation results indicate that COSRE risk has a strong positive correlation with the new cases in the next two weeks, which suggests that the risk scores reflect real exposure risk when the quarantine orders relax or expire and people start to go back to normal life.

4.3. Spatial correlation analysis

4.3.1. Spatial autocorrelation and Moran’s I

We calculated the Global Univariate Moran’s I of all the semi-month COSRE risk maps from April to July. As shown in Fig. 3, the global univariate Moran’s I of COSRE risk stays above 0.4 for all time steps. It reaches a peak (~0.6) on April 15, falls back to around 0.4 on June 1, and bounces back to reach another peak on July 15. The positive Moran’s I values prove that most neighbor counties do share similar COSRE risk scores. The significance has a spike on April 15 and weakens during May and June, which can be explained by the pause of venue activities among counties in those two months. The expiration of stay-at-home orders and the increase of family and community gatherings might be considered as two reasons the intense in-person venue activities across counties resulted in the apparent rising of global Moran’s I in July.

To analyze the temporal correlation of the risk score, we used the differential Moran’s I to calculate the correlation between one time-step and its previous time-step. The higher Moran’s I value means the counties have similar risk scores to the last time steps. The near-zero values mean the risk scores change randomly between the two time steps. The value is zero on April 1 because it is the start date with no

![Fig. 2. Correlation between COSRE risk score and the new confirmed cases in the following two weeks (14 days); Spearman’s correlation coefficients are added in the tiles, the x axis is COSRE risk score, the y axis is the new case per thousand people in the following two weeks.](image-url)
previous time step. Except April 1, all the values are positive. Basically, the temporal trends share the same pattern to the global univariate Moran’s I. April and July have higher correlation with the last time-step, and the correlation decreases in May and June. The variance of the differential Moran’s I values means that for each county the restriction on in-person activities in public places in late April, May and June works to reduce risk. However, after entering July the correlation began to show significance between two continuous time steps; the county with high confirmed cases will very likely have a similar amount of confirmed cases at the next checkpoint. Overall, correlation of the risk scores of two continuous time steps is positive and will have higher significance when venue activities increase.

Cluster maps based on local univariate Moran’s I between one county’s COSRE score and its neighbor counties are also generated to extract those counties that share the same pattern of case spikes or lower cases (as shown in Fig. 4). A p-value less than 0.05 (typically \( \leq 0.05 \)) is
normally considered to be statistically significant. We use 0.05 as p-value threshold to extract the significantly correlated counties; they are plotted in Fig. 4. All the colored counties have a p-value smaller than 0.05, which means significant correlation between the COSRE score of the county and its spatial lags. The high-high hotspot regions agree well with our observations in Section 4.1. The New York, New Orleans, Detroit, Denver and Seattle metropolitan areas and neighboring counties are the major hotspot clusters in April. In May, the Chicago metropolitan area joined the list while Seattle and Detroit drop off, along with many major cities in the Midwest states. The Navajo Nation became a hotspot and persists since then. In July, the New York and the other northeastern states turn blue after consistent low risks are observed. Most southern states on the south including southern California, Arizona, Texas, Louisiana, Mississippi, Alabama, Georgia, and Florida, undergo serious exposure risk due to a dramatic increase in confirmed cases. They became the new hotspots. On the other side, the hotspots in April and May in the Midwest states to the north mostly turn blue or non-color and are removed on the July map. These clusters show counties with concurrent high/low risks. The red regions should be considered as very dangerous for venue activities, and the blue areas are relatively safe for small gatherings and use of face protectives (e.g., grocery shopping).

4.3.2. Bivariate correlation between COSRE and new cases in two weeks

Bivariate Moran’s I between COSRE risk and 2-week new cases in neighboring areas is calculated and plotted in Fig. 3. All time steps have positive values. Each value reflects the association between the COSRE risk score of one county and the new confirmed cases in the next two weeks. The trends are different in April from the spatial and temporal trends of the univariate Moran’s I. Instead of peaking in April, the bivariate Moran’s I falls into a trough on April 15. This may be the result of responses to the sudden increase in confirmed cases and people reduced their venue activities, either voluntarily or policy enforced. In May, June and July, the bivariate Moran’s I changes to follow a similar pattern with the global univariate Moran’s I and the differential Moran’s I. It rockets high in July 15 to reflect a strong correlation of high COSRE risk with the new cases of neighbor counties in the next 2 weeks at that time.

Cluster maps are derived based on the calculated bivariate Moran’s I. Only those counties with p-values less than 0.05 are considered as significantly correlated regions, and colored in Fig. 5. The results show that COSRE risk scores can actually reveal those hotspot counties on April 1, like those in New York, Louisiana, and Detroit; case increases in the next two weeks agree very well with model estimates in those regions. In all cluster maps, those hotspot regions are very consistent with the high-high regions identified in Fig. 4, which means that the correlation between COSRE risks and the new cases in the following weeks is significantly positive, especially in those hotspot regions and when little venue activity restrictions are in place. Overall, our COSRE risk shows reliable performance, with stable risk scores that indicate true exposure risks, from the early stage to the reopening stage.

5. Discussion

5.1. Community exposure risk and transmission risk

It should be clarified that exposure risk is not equal to the actual transmission risk. The community risk of exposure is an important contributing factor to the risk of transmission. The relationship is delineated using the following generalized equation:

\[ R_{\text{transmission}} = R_{\text{exposure}} \times R_{\text{contract}} \times I_{\text{immunity}} \]

where \( R_{\text{exposure}} \) is the exposure risk in community daily activities, \( R_{\text{contract}} \) is the infection risk. The two risks come from two steps. The former risk estimates the possibility of people coming in contact with virus hosts and sources. For example, people who stay home are at lower risk of exposure than people who go to the grocery store. The latter risk estimates the possibility of being infected by the virus after contact with a source. If people wear mouth and nose masks, safety glasses and disposable gloves and footwear, their contracting risk will be much lower than for people who wear no protections. \( I_{\text{immunity}} \) represents the ability of the person or the community to be immune to the virus. People with antibodies could be immune to certain types of viruses and their immunity is higher. The transmission risk is inversely proportional to
the immunity of the person or the entire community.

5.2. Exposure risk evaluation

The work reported herein gives a customized quantitative estimation on the social exposure risks of people who show up in public places. The results are tailored for every community by considering local at-risk population, store numbers, potential guest flow. The purpose is to help people increase awareness of their risks and help U.S. businesses adjust their store policies. Following this guidance should accelerate the reopening of business while maintaining a low risk of virus contact for the wellbeing of both employees and customers. Successful application of tools from this project will give both customer and business a clearer understanding of the risk of opening up and going to a place of business or recreation venue. Decisions can be made on the fly. The correlation evaluation has demonstrated that the COSRE risk score has a strong positive correlation with new cases for the following two-week period. The rate of new cases is basically determined by regional transmission risks in which the venue exposure risk is a preconditional factor. The positive correlation can be used as evidence to validate the accuracy and effectiveness of the COSRE risk score. We would expect the score will serve as a simple, stable and reliable indicator to measure current risk and estimating new cases for the next two-week period.

5.3. Model limitation and uncertainty

A reliable model designed for real-world application must consider many factors. More refinement is needed to improve accuracy of the three input parameters. For example, the model assumes that the confirmed patients comprise all the existing patients and all are free to move about the community; in reality, this is not entirely true. As an improvement, it should be used to calculate new potential COVID patient numbers, as close to the real situation as possible. As official virus tests become more available, the ability to monitor and predict community-level risks amid a pandemic will improve. For the population, the model assumes that everyone in the population has the same chance of showing up in one store, which is also not true. Individual preferences exist for which grocery store, shopping center, coffee shop or library to frequent. To make the model more realistic, one can obtain store visiting data from SafeGraph (2020), which allows analysis of the age, gender, and other characteristics of a store’s customers. From this, a more accurate customer pool is possible based on the national population census. One may also remove people from the customer pool who will not show up from the population used for calculation, as when data to populate the April and May maps of Fig. 1 were adjusted to accommodate finding of physicians. Further improvement will come by adjusting the real-time store visitation based on popular hours, business density, county income level and real-time customer counts.

Also, real-world exposure data are scarce. Apple and Google are among the exceptions with more accurate and more extensive data, where smartphone Bluetooth has been used to trace potential coronavirus patients. Based on patient tracing data such as this, the real probability of a venue with COVID patients present could be compared with the estimated risks derived from our formula. Since the pandemic is still ongoing, such a dataset is relatively sensitive and hard to retrieve at present. Model evaluation with real-world data is our next step of work.

6. Conclusion

Public awareness of virus exposure risks is important. Individual decisions will be made during a pandemic as people decide to leave home sanctuaries and reengage in social activities. To inform and assist, we propose a birthday-paradox-based probability model, coupled with a publicly-accessible web-based system to calculate community exposure risks in public gatherings. Model derived risks are generated based on the real-time potential COVID-19 cases, the population in local communities, and the number of people number in a given venue. With this web-accessed system, people may explore effects of the pandemic through a geographical spatiotemporal view, moving through time, and testing different venues and the expected numbers of people in them in order to assess changes in risk as the pandemic unfolds. The system integrates the risk estimation model, computational tools, and the analysis of evolutionary pathways, together with refinements to virus surveillance and to research-based, new understanding of this novel virus.

The model and system proposed is an improvement in assessing risks posed by the SARS-Cov-2 virus and other virus outbreaks, epidemics or pandemics. We are subsequently better equipped to prepare and respond to the ongoing pandemic and to all future vector borne diseases. Yet, caveat and caution are advised. The application scenarios of the system are made to be set in the middle of an outbreak or pandemic, after testing and tracking of patients are in place. Today, without extensive and comparable national testing, the estimated risks might be far from reality. An important objective of our study is to show what can be accomplished with extensive testing and tracking, where methods, materials and objectives are comparable—where apples are compared with apples, without regard to state or region. Valuable information can be drawn from the current web-based system. Complex interactions among factors and accuracy of any or all of the input parameters will distort model from reality. We will continue to work to limit these distortions, improve accuracy and reliability through peer and crowd-sourced review.

Declaration of competing interest

The authors declare they have no actual or potential competing financial interests.

An early draft of this manuscript was deposited in arXiv.org preprint.

Acknowledgment

The authors would like to thank the anonymous reviewers, Professor Dieter Pfoser, Professor Andreas Zufle, and Professor Pat Gillevet from George Mason University, NASA project (17-HAQ17-0044) team members, the MIT Datathon team for their valuable advice. Thanks for the COVID-19 dataset provided by John Hopkins University, the population data from U.S. Census, and the foot traffic data kindly provided by SafeGraph Inc. Thanks to all the developers of the Python libraries and tools used in this work.

References

Altham, P.M., 1978. Two generalizations of the binomial distribution. J. Roy. Stat. Soc.: Series C (Applied Statistics) 27, 162–167.
Anselin, L., Syabri, I., Kho, Y., 2010. GeoDa: an Introduction to Spatial Data Analysis, Handbook of Applied Spatial Analysis. Springer, pp. 73–89.
Berke, O., 2005. Exploratory spatial relative risk mapping. Prev. Vet. Med. 71, 173–182.
Berke, O., Grosse Beilage, E., 2003. Spatial relative risk mapping of pseudorabies-seropositive pig herds in an animal-dense region. J. Vet. Med. Ser. B 50, 322–325.
Castro, L.A., Fox, S.J., Chen, X., Liu, K., Bellan, S.E., Dimitrov, N.B., Galvani, A.P., Meyers, L.A., 2017. Assessing real-time Zika risk in the United States. BMC Infect. Dis. 17, 284.
CDC, 2020. Influenza Risk Assessment Tool (IRAT): Questions & Answers. Census, U.S., 2010. County Population Totals: 2010-2019.
Cent, C.E.F.D.P., 2020. Outbreak of Acute Respiratory Syndrome Associated with a Novel Coronavirus, Wuhan, China; First Update – 22 January 2020. ECDC, Stockholm.
Cox, N.J., Trock, S.C., Burke, S.A., 2014. Pandemic Preparedness and the Influenza Risk Assessment Tool (IRAT), Influenza Pathogenesis and Control-Volume I. Springer, pp. 119–136.
Cromley, E.K., 2003. GIS and disease. Annu. Rev. Publ. Health 24, 7–24.
Dashrath, P., Jeslyn, W.J.L., Karen, L.M.X., Min, L.L., Sarah, L., Biwas, A., Choolani, M. A., Matar, C., Lin, S.L., 2020. Coronavirus disease 2019 (COVID-19) pandemic and pregnancy. Am. J. Obstet. Gynecol.
Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis.
Dudley, J.P., 2008. Public health and epidemiological considerations for avian influenza risk mapping and risk assessment. Ecol. Soc. 13,
Flajolet, P., Gardy, D., Thominier, L., 1992. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. Discrete Appl. Math. 39, 207–229.

Glendon, Y., 1994. Introduction to pandemic influenza through history. Eur. J. Epidemiol. 10, 451–453.

Hunter, J.D., 2007. Matplotlib: a 2D graphics environment. Comput. Sci. Eng. 9, 90–95.

Jordahl, K., 2014. GeoPandas: Python tools for geographic data. https://github.com/geopandas/geopandas.

Jung, S.-m., Akhmetzhanov, A.R., Hayashi, K., Linton, N.M., Yang, Y., Yuan, B., Kobayashi, T., Kinoshita, R., Nishiura, H., 2020. Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: inference using exported cases. J. Clin. Med. 9, 523.

Katz, R., 2009. Use of revised international health regulations during influenza A (H1N1) epidemic. Emerg. Infect. Dis. 15, 1165, 2009.

Ksiazek, T.G., Erdman, D., Goldsmith, C.S., Zaki, S.R., Peret, T., Emery, S., Tong, S., Urbani, C., Cofer, J.A., Lim, W., 2003. A novel coronavirus associated with severe acute respiratory syndrome. N. Engl. J. Med. 348, 1953–1966.

Mayfield, H.J., Smith, C.S., Lowry, J.H., Watson, C.H., Baker, M.G., Kama, M., Nilles, E.J., Lau, C.L., 2018. Predictive risk mapping of an environmentally-driven infectious disease using spatial Bayesian networks: a case study of leptospirosis in Fiji. PloS Neglected Trop. Dis. 12, e0006857.

McKinney, E.H., 1966. Generalized birthday problem. Am. Math. Mon. 73, 385–387.

Moran, P.A., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.

Open, Y., 2020. Assessing nitrogen dioxide (NO2) levels as a contributing factor to the coronavirus (COVID-19) fatality rate. Sci. Total Environ. 138605.

Oster, A.M., 2020. Trends in number and distribution of COVID-19 hotspot counties—United States, March 8–July 15, 2020. MMWR. Morbidity and Mortality Weekly Report 69.

Pearson, K., 1895. Notes on regression and inheritance in the case of two parents. Proc. Roy. Soc. Lond. 58, 246–242.

Z. Sun et al. Health and Place 66 (2020) 102450

WHO, 2018. A Checklist for Pandemic Influenza Risk and Impact Management: Building Capacity for Pandemic Response. WHO.