SERRA manuscript No.
(will be inserted by the editor)

Contextual Contact Tracing based on Stochastic Compartment Modeling & Spatial Risk Assessment

Mateen Mahmood · Jorge Mateu · Enrique Hernández-Orallo

Received: date / Accepted: date

Abstract The current situation of COVID-19 highlights the paramount importance of infectious disease surveillance, which necessitates early monitoring for effective response. Policymakers are interested in data insights identifying high-risk areas as well as individuals to be quarantined, especially as the public gets back to their normal routine. This paper investigates both requirements by the implementation of disease outbreak modeling and exploring its induced dynamic spatial risk in the form of risk assessment, along with its real-time integration back into the disease model. The paper implements a contact tracing-based stochastic compartment model as a baseline, to further modify the existing setup to include the spatial risk. This modification of each individual-level contact's intensity to be dependent on its spatial location has been termed as *Contextual Contact Tracing*. The results highlight that the inclusion of spatial context tends to send more individuals into quarantine which reduces the overall spread of infection. With a simulated example of an induced spatial high-risk, it is highlighted that the new spatio-SIR model can act as a tool to empower the analyst with a capability to explore disease dynamics from a spatial perspective. The paper concludes that the proposed spatio-SIR tool can be of great help for policymakers to know the consequences of their decision prior to their implementation.

Keywords Compartment Modeling · Contact Tracing · Digital Epidemiology · Human Mobility · Self Organizing Maps · Trajectories

M. Mahmood
Universitat Jaume I, Castellón, Spain
E-mail: al394323@uji.es

J. Mateu
Universitat Jaume I, Castellón, Spain
E-mail: mateu@uji.es

E. Hernández-Orallo
Universitat Politécnica de Valencia, Valencia, Spain
E-mail: ehernandez@disca.upv.es

1 Introduction

Detection and control of COVID-19 in particular, and infectious diseases in general, have irrupted as a major societal challenge. As of 31^{st} January 2021, the COVID-19 pandemic has over 101 million confirmed cases with above 2.1 million deaths worldwide (WHO, 2021). This explosive dissemination is not only a universal threat to public health organizations, but it also jeopardizes social functioning, industry, economy and international relations (Zhou et al, 2020). Countries such as Israel and South Korea which took prompt actions towards testing and identification of previous contacts in case of an identified individual were able to restrict the disease spread. However, countries that did not proceed with the initial massive testing and contact tracing had to go for extreme measures of lockdown, quarantining and contact precautions (social-distancing, facemasks, etc.) (Hernández-Orallo et al, 2020).

Detecting all infected individuals among the population requires massive testing on a regional scale. Though authorities have followed ingenious medical methods to rapidly detect the infected individuals, it has a considerable economical burden and implementation barriers. In a situation like this, detection of an infectious disease requires non-pharmaceutical interventions (NPI) and is to be supported by methods outside of the medical system, which sets the basis of the term *Digital Epidemiology* (DE) (Salathé, 2018).

One such DE based method is *Digital Contact Tracing* (DCT), which can provide prior contacts of a detected individual. This rapid identification of exposed individuals (who need to be tested or quarantined), can support the health system by restricting the uncontrolled asymptomatic propagation of infection. In DCT, the key to track the infectious transmission is to keep an eye on the physical *interaction* (contacts) of individuals, and understanding these interactions are as important as understanding the contagion process.

These interactions are much more than just recording of a *contact*, and when studied from a spatio-temporal perspective provide a comprehensive understanding of disease dynamics. While the temporal domain deals with the duration and instance of contacts, the spatial aspect refers the influence of a geographical location on the outcome of a contact, with a notion that some areas are inducing disease transmission more than others due to their urban function (Wang et al, 2017), environment and overall infectious activities.

On the other hand, these interactions based on individuals movement is subject to tracking of human mobility, where detection of an infected individual means that infectious trajectories can be tracked. Such tracking is critical to understand how an infection propagates in population and in space, as it not only identifies future infectious contacts but also highlights the places these infectious trajectories have visited (Benreguia et al, 2020). Identification of such high-risk areas is critical for policymakers in decisions related to smart lockdown, areal curfew, etc.

This scenario makes contact tracing, mobility tracking and spatial risk interconnected processes. It is a recursive sequence as illustrated in Figure 1, where the probability of transmission of a contact is proportional to the

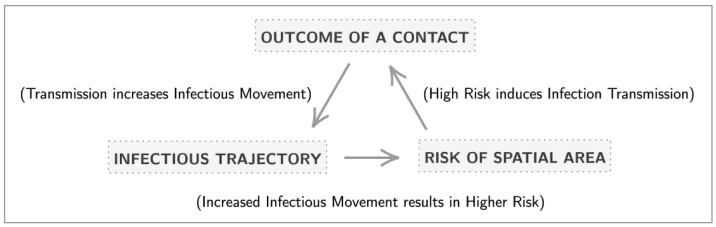

Fig. 1 Relation of temporally varying spatial risk affecting epidemic model and vice versa

risk intensity of its spatial location. This spatial location evolves based on infectious movement, which itself is an outcome of an infectious contact. Hence, there is a requirement of an approach to thoroughly fuse the effect of space into a disease model while dealing with infectious trajectories. In this paper, we focus on the inclusion of the spatial aspect of these physical interactions termed as *Contextual Contact Tracing*. The idea is that contacts taking place in contextually distinct geographical locations are to be treated differently based on the vulnerability they pose to the susceptible individual.

In human infectious diseases, where the pathogen is another human being, there is a requirement to track human movement. Tracking the known infected individuals and their interactions are already demanding, but the existence of asymptomatic individuals makes this monitoring even more challenging (Müller et al, 2020). These undetected individuals are transmitting the infection to a larger set of individuals, who themselves are infecting the community in an uncontrollable domino effect. Early detection of asymptomatic individuals followed by isolation or treatment is the key to restrict pandemic growth, where state of the art highlights the accepted practise of digital methods in such detection studies (Anglemyer et al, 2020). Ongoing research (Van Doremalen et al, 2020; Simmerman et al, 2010) highlights the aerosol and surface stability of infectious diseases, where COVID-19, SARS-CoV1 and Influenza A/H1N1, all have indicated up to days surface transmission. Both these aspects, tracking of individuals and risk assessment of space, sets the basis of infectious disease surveillance in this digital era.

Tracking human movement relies on human mobility data, which is of prime importance in individual-level research on infectious disease dynamics (Brockmann et al, 2009). Recent advancement in location-aware technologies and computing procedures have resulted in a massive influx of mobility data, which is capable of representing the movement of an individual to a very small scale up to less than a meter (Zheng, 2015). This high-level detail makes these datasets an ideal candidate for high precision tasks such as contact tracing.

Despite that, an important consideration is that continuous recording of an individual movement is highly invasive (Reichert et al, 2020), which is why there is no infectious disease related individual-level trajectory dataset publicly available so far. To minimize this concern, the use of bluetooth has been proposed (Martinez-Martin et al, 2020), though it only collects the con-

tact information as and when it happens. At the same time, Benreguia et al. (Benreguia et al, 2020) suggest that in preparation for an extremely critical scenario where entire humanity is at stake and the requirement of saving lives is of highest priority, the use of continuous recordings of individual's movement is justified given it is implemented by government and with a guarantee of privacy protection. Similarly, for spatial risk assessment, individual-level work is only executed on a sparse scale. Souza et al. (Souza et al, 2019) detected spatial clusters using spatial scan statistics, based on Twitter feed data. Another spatial clustering application on aggregated data is available in (Desjardins et al, 2020) where a countywide space-time clustering is executed.

The long-standing COVID-19 has amplified research in this domain with several studies involving individual-level-mobility for investigation of disease dynamics. Many of these studies described the spatio-temporal trends inclusive of stochastic aspects, proposing statistical foundations to fit models to data. However, the spatial aspects focused more on spatial separation rather than spatial location. Even if the spatial location was considered, it was in the aggregated form of spatially varying demographic factor (Mahsin et al, 2020).

In epidemic modeling, compartment models distribute each individual in the population based on their disease states. Generally, they are of Susceptible, Infected and Recovered (SIR) type, however many versions such as SEIR, SEIAHCRD (Berger et al, 2020; Bardina et al, 2020) exist which depend on the type of disease and applied methodologies. Though the temporal aspect is well addressed in these SIR models, the *spatial context* is generally new.

In spatio-epidemic modeling, the idea of a space-dependent SIR model has been presented in (Takács and Hadjimichael, 2019) in form of a numerical experiment. They considered a generalized SIR model where population size differed over space. Another spatial-SIR model is explained in (Bisin and Moro, 2020) to understand spatial diffusion of disease based on quantitative effects of geographical context in determining that diffusion. Modifying epidemic parameters based on the spatial location have also been proposed. A space-time dependent *basic reproductive ratio* is implemented in (Martinez-Beneito et al, 2020), while Lang et al. (Lang et al, 2018) discuss a framework of a SIR model on spatial networks where the probability of transmission is based on spatial distances along the edges. A *bayesian maximum entropy* based extension is also available for metapopulation-level epidemic modeling (Angulo et al, 2013).

All these models propose population-level frameworks for the inclusion of space in SIR modeling. Complete integration of a spatial context in an individual-level study of contact tracing is still missing, which can consider the influence of *space* (location) for each specific contact.

This paper proposes a new spatio enhanced setup of SIR modeling, where contacts are associated with an intensity of its risk score based on its spatial location. This association of risk with a contact is executed by reforming the quantitative value of a contact, where enhancement is in a manner that a riskier contact has a higher probability of disease transmission than the one which is of relatively lower risk. For temporally varying spatial risk, we re-evaluate spatial risk scores based on infectious activities of the recent past.

We analyze here real-life mobility data of NCCU Trace (Tsai and Chan, 2015) which provides movement of 115 students recorded for 15 days. In the implementation, we first execute contact tracing to construct temporal network graphs. These contact graphs are further used to implement an epidemic model with self-induced infection, which was later enhanced to a spatially-enhanced epidemic model inclusive of spatial risk. In parallel, we track infectious trajectories and the location of contacts as elements for spatial risk assessment. The results highlight that the inclusion of spatial context tends to send more individuals into quarantine which reduces the overall spread of infection.

The reason behind pursuing this study in the absence of real information about infection is because a methodology is also missing, which can consider spatial risk in a contact tracing process and track mobility trajectories. Therefore, the feasibility of this idea is developed in the form of a *spatio-epidemic tool*, which is an established proposal for future works, not only to work with a real dataset as they become available but also in the domain of spatial risk.

A recent publication from February 2021 presents movement data of infected (COVID-19) individuals from South Korea (Park et al, 2021). However, data is not in the form of continuous trajectories, but are recordings of individual's interactions with others through a contact tracing application. This availability is a motivating fact as more real-world datasets related to infection information as well as mobility trajectories will be publicly available offering a definite way forward for this work.

The remainder of this paper is as follows: Section 2 describes the methodology of both, SIR model (baseline setup) and its enhancement to a spatio-SIR model. Section 3 introduces the selected dataset with discussion on the experimental design. This section further presents results of both models supported by varying simulations to effectively understand the new spatio-SIR setup. Section 4 concludes the paper and presents limitations and future work.

2 Methodology

2.1 Baseline-SIR modeling

The baseline-SIR model for this study is motivated by (Hernández-Orallo et al, 2020), in which contact tracing technologies are evaluated along with a comparison of stochastic versus deterministic approaches. In this paper, we reproduce their stochastic setup (hereby referred to as *base-SIR*) as our baseline-SIR model, with a rationale that a stochastic model is more realistic than a deterministic one due to its probabilistic nature. Similarly, an event-based method as followed in base-SIR is preferred due to the incorporation of event-driven chance element. An overview of methodology for the implementation of baseline-SIR, and its modification to a spatio-SIR model (see Section 2.2) is available in Figure 2.

Base-SIR brings forth a novel addition of *Quarantine Susceptible* and *Quarantine Infected* related compartments which add a new perspective in the mod-

eling of a real-world scenario related to contact tracing based compartment modeling. Base-SIR implements *Gillespie's First Reaction Method* (GFRM) (Keeling and Rohani, 2011), which handles efficiently a contact tracing problem, especially on a trajectory-based dataset.

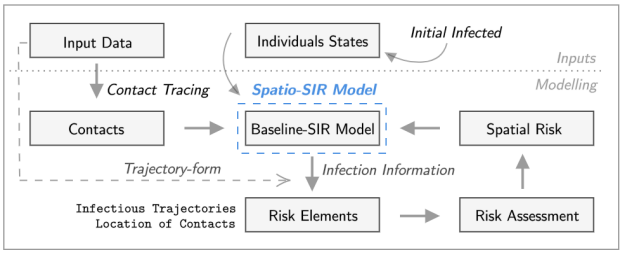

Fig. 2 Summarized workflow of overall methodology

2.1.1 Contact Tracing

Contact Tracing is the identification of *colocation* of two or more individuals. However, this colocation is not restricted to a single point or a single instance of time, but a range of an area and duration which are based on epidemiological aspects. Contact with a possibility of transmission is the one within two meters of an infected individual with an exposure of at least one minute (Hernández-Orallo et al, 2020). Therefore, we define d_c as the *distance threshold* and t_c as the *duration threshold* for considering a contact as risky.

Identification of infectious contacts needs accurate information about the possible transmissible pathways from an infected person to each individual in the population (Eames and Keeling, 2003). A network graph is a computationally efficient representation of such interactions where in individual-level studies, *nodes* refers to individuals and *edges* represents their contacts (Enright and Kao, 2018). A temporal network graph can be denoted as $G(t)$, with ν (nodes) and ε (edges), where t represents the instance of time. In epidemic modeling, it is common to have a temporal frequency of a *day* (Keeling and Rohani, 2011), hence $\varepsilon_{ij}(t)$ will exist between individual i and j if there exists a contact between the two on day t.

An adjacency matrix is commonly used to store graph information. It is a *graph matrix*, where rows and columns represent nodes (individuals), with a third dimension corresponding to the day of contact. A contact is represented with a value of either 0 or 1, where 1 depicts the existence of an edge (contact) between the two. Figure 3 presents a toy example of a network graph and the associated graph matrix.

Directions of edges as in directed/undirected graph is ignored as contacts are independent of direction. This highlights the assumption that infection can be transmitted in both directions depending on the disease state of the

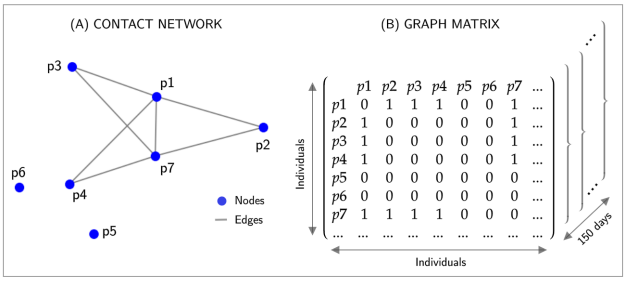

Fig. 3 A toy example: contacts of a single day in the form of contact network and graph matrix

individual and not the structure of the network. For a pair (i, j) of individuals, this symmetry can be viewed as $(G_{ij}(t) = G_{ji}(t))$.

In contact network, *degree* shows the count of connections of a node with the other nodes in the network. The *temporal degree* $K_i(t)$ is the count of contacts of a person i with other individuals in the network $G(t)$ on day t. Hence, an *average degree* κ for a time period T can be computed as (1)

$$\kappa = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{T} \sum_{j=1}^{T} K_i(t_j) \right] \tag{1}$$

As the rate of infection is influenced by the count of infected individuals, hence it is useful to have a degree only involving contacts with infected individuals. Such a degree of diffusion can be represented as (2), where $I_j(t)$ denotes that individual j can infect others (return 1) or not (returns 0).

$$\mathcal{K}_i(t) = \sum_{j=1}^{N} G_{ij}(t) \cdot I_j(t) \tag{2}$$

Identifying prior contacts is the overall essence of contact tracing in order to restrict next generation of cases. This requires a backward time window $\triangle$ depending on the type of disease (infectious period, incubation time, etc.), and can be used in the form of (3) to extract all prior contacts $\mathcal{C}_i(t, \triangle)$ of an individual i at t with window $\triangle$

$$\mathcal{C}_i(t, \triangle) = \sum_{j=1}^{N} \left(\max_{\tau \in [t-\Delta, t]} G_{ij}(\tau) \right) D_j(t) \tag{3}$$

Here, $D_j(t)$ is 1 if at time t person j is infected and traced. Algorithm (1) explains the process flow of contact tracing. Once the contacts are identified, a baseline setup can be formulated to simulate SIR events. The model also evaluates the efficiency of the contact tracing methods. Contact tracing can be manual (that is, based on interviewing the detected and infected individuals) or smartphone based (using contact tracing apps). We define a value q as

the fraction of traced individuals being quarantined. For example, this value can reflect the number of individuals that use the mobile contact tracing app. In the case where the tracing time is greater than 1, the q value must be normalised by the average tracing time $(1/\tau_T)$, as $q^{'} = q/(1/\tau_T) = q \cdot \tau_T$ in order to distribute the tracing quarantine over the days. The idea is that if the tracing time is long (for example, by using interviews), it is precisely because it takes time to trace back the prior contacts, so the whole number of traced individuals during this tracing time is equally distributed over these days. Finally, apart from contacts, baseline-SIR model relies on (a) *infection states* and (b) *epidemic parameters*.

Algorithm 1 *Pseudocode of contact tracing process*

input: trace_dataset, n ▷ *(n← count of individuals)*
output: G ▷ *($G \leftarrow$ contact graph)*
// initialize $G(n, n, d) \leftarrow 0$ ▷ *($d \leftarrow$ number of days)*
for *($t \leftarrow$ 1 to $d \times 86400$)* **do** ▷ *($t \leftarrow$ time in seconds)*
  for *($p1 \leftarrow$ 1 to n)* **do** ▷ *($p1 \leftarrow$ first person)*
    for *($p2 \leftarrow$ 2 to n)* **do** ▷ *($p2 \leftarrow$ second person)*
      if (p1 != p2) **then**
        dist = distance between p1 and p2
        if (dist $< d_c$) **then** ▷ *($d_c \leftarrow$ distance threshold)*
          record duration
          if (duration $> t_c$) **then** ▷ *($t_c \leftarrow$ duration threshold)*
            $G(p1, p2, day) \leftarrow 1$ ▷ *($day \leftarrow t/86400$)*
return: G

2.1.2 Infection states (compartments)

Infection states refer to the compartments an individual can be during an epidemic. As in base-SIR, a total of five compartments are considered which represent the states of Susceptible (S), Infected (I), Recovered (R), Quarantine Susceptible (Q_S) and Quarantine Infected (Q_I). With these five compartments, there are seven possible SIR events that will imply the transition of an individual from a compartment to another. Figure 4 presents the possible events (transfers among compartments), which are:
$(S \rightarrow I,\ \ S \rightarrow Q_S,\ \ S \rightarrow Q_I,\ \ I \rightarrow Q_I,\ \ Q_S \rightarrow S,\ \ I \rightarrow R,\ \ Q_I \rightarrow R)$

As information about the latency states of individuals is not available, a self-induced infection approach is followed. This means that out of the total population, a certain count of individuals in the population are initiated as infected being in compartment (I), to have a sense of disease propagation based on their future contacts, as epidemic progresses.

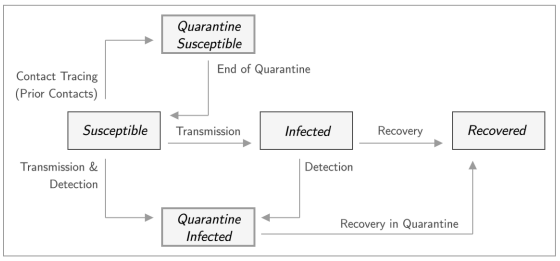

Fig. 4 Overview of compartments and possible transfers between them

2.1.3 Epidemic parameters

Epidemic parameters, as introduced in Table 1, refer to the disease-specific elements in the form of coefficients that contribute to computing the rates of each event associated with individuals.

Table 1 Summary of infectious disease related modeling parameters

parameter	description
κ	*Average degree (daily contacts per individual)*
$\mathcal{K}_i$	*Contacts of individual i with infected individuals*
$\mathcal{R}_0$	*Basic reproductive ratio*
δ	*Rate of detection*
γ	*Recovery rate* ($1/\gamma$ = *disease specific days for recovery*)
b	*Probability of transmission of infection* ($b = \mathcal{R}_0 \cdot \gamma/\kappa$)
β	*Transmission rate* ($\beta = \kappa \cdot b$)
τ_Q	*Time in quarantine*
$q^{'}$	*Tracing efficiency* ($q^{'} = q \cdot \tau_T$)
$\mathcal{C}_i(\triangle)$	*Backward contact tracing of individual i with detected infected individuals*

The core of the model is to answer the question of how individuals move from one compartment to another. In a closed environment where births, deaths and migration are ignored, transition ($\boldsymbol{S} \rightarrow \boldsymbol{I}$) is subject to disease transmission and is a function of three aspects: (i) the presence of infected individuals, (ii) contacts between susceptible and infected ($S{\leftrightarrow}I$) and (iii) the probability of transmission. Considering κ as the *degree* of ($S{\leftrightarrow}I$) contacts and b representing the *probability of transmission* of infection, the *transmission rate* β can be deduced as $\beta = \kappa{\cdot}b$. The transition from ($\boldsymbol{I} \rightarrow \boldsymbol{R}$) is simpler as it can be considered a constant around a mean value based on clinical data of *infectious period*. The probability of an infected individual to be recovered relies on how long they have been infected, which can be denoted as *recovery rate* γ, a constant value representing the inverse of infectious period. The ratio β/γ is called as *basic reproductive ratio* $\mathcal{R}_0$. It represents the expected count

Table 2 Rate equations related to each SIR event

event	description of rate	rate equation
$\boldsymbol{S} \rightarrow \boldsymbol{I}$	*Transmission of infection*	$(1 - \mathcal{C}_i(t, \triangle)) \cdot b \cdot \mathcal{K}_i(t)$
$\boldsymbol{S} \rightarrow Q_S$	*Susceptible person being quarantined*	$q^{'} \cdot \mathcal{C}_i(t, \triangle) \cdot (1 - b \cdot \mathcal{K}_i(t))$
$\boldsymbol{S} \rightarrow Q_I$	*Susceptible person being infected & detected*	$q^{'} \cdot \mathcal{C}_i(t, \triangle) \cdot b \cdot \mathcal{K}_i(t))$
$\boldsymbol{I} \rightarrow Q_I$	*Infected person being detected*	δ
$Q_S \rightarrow \boldsymbol{S}$	*End of quarantine after quarantine period*	$\mathcal{T}_Q$
$\boldsymbol{I} \rightarrow \boldsymbol{R}$	*Recovery after infectious period*	γ
$Q_I \rightarrow \boldsymbol{R}$	*Recovery after infectious period in quarantine*	$\mathcal{T}_Q$

of cases directly affected by a single case and is considered as one of the more representative parameters of disease in epidemiology.

Similarly, based on the epidemic parameters enlisted in Table 1, rates of each event can be computed using equations provided in Table 2. In this paper, we propose a generalized framework for spatio-SIR modeling through the use of values corresponding to COVID-19, as given in Table 3; however, any disease-specific model can be developed by adjusting these parameters.

Table 3 Estimated values for COVID-19 (extracted from (Ferretti et al, 2020; Hellewell et al, 2020; Li et al, 2020))

parameter	estimated value
$\mathcal{R}_0$	*3*
δ	*0.1*
γ	*1/15*
$\mathcal{T}_Q$	*1/14*

2.1.4 Event-based modeling

In an event-driven model, each possibility is considered as an event and then a random element will decide which event may happen, based on the cumulative rates of all events and converting those rates into probabilities. This highlights that even if the probability of an event is similar, an individual may experience a varied event based on the chance element. There are numerous methods to implement event-driven approaches, one of which is *Gillespie's Method* (Gillespie, 1977), common in SIR modeling (Keeling and Rohani, 2011).

Gillespie's algorithm, initially intended for the study of chemical reactions, is also applicable in scenarios such as SIR modeling where an outcome of the contact is like a biochemical reaction of a cell with fluctuating possibilities of events. It is a variant of a Monte Carlo method, with a computationally feasible solution. Gillespie's First Reaction Method (GFRM) is a simplified version of the original Gillespie's Direct Method with a scalable approach.

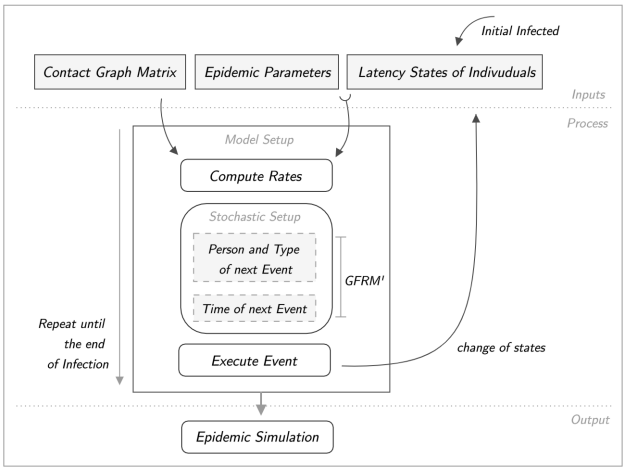

Fig. 5 Process flow of an event-driven stochastic SIR model

In GFRM, there are two stochastic elements. First is the *type* of event which includes the person to which event will happen and the kind of event (out of defined seven events) that will happen. The other stochastic element is of the *time* of next event which refers to the duration since the latest event after which this event will take place. The former, as per GFRM, is determined by computing the rates of each event and then stochastically drawing the next event. For the latter, in our approach, it is completely stochastic based on a random element instead of computing time for each event. This modification is due to the reason that there is no inherent time of an event in a contact tracing process. Based on this modified GFRM, event-based stochastic SIR model can be implemented on identified contacts using *infection states* and *epidemic parameters*. Figure 5 shows the workflow of GFRM-based SIR model.

2.2 Spatio-SIR modeling

This section focuses on two aspects: *temporally varying spatial risk* and *spatio-SIR model*. Here, spatial risk refers to the transmission vulnerability a spatial location poses to a susceptible individual involved in an infectious contact. This spatial risk is for a certain period and is continuously evolving based on previous infectious activities. Our spatio-SIR model extends the baseline setup taking into account the spatial risk in the future tracing of contacts.

As the goal is to associate a risk score to each contact based on its spatial location, it is important to address the definition of *location*. A simple and computationally efficient approach is to consider a regular lattice (grid) structure segmenting the study area into smaller cells, each one having a risk score. From this, location of a contact can be defined as the corresponding cell of the grid in which the contact is taking place.

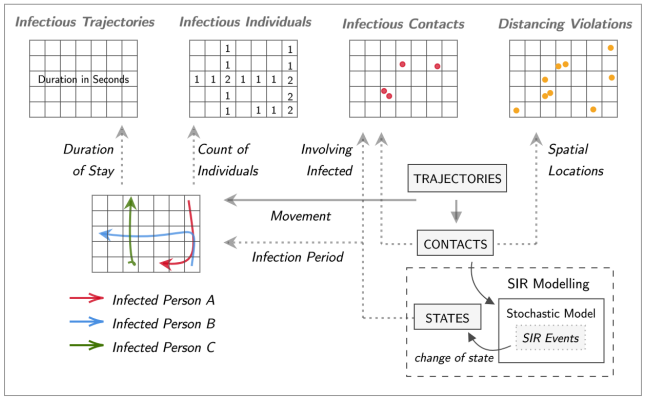

Fig. 6 Computing risk basis from trajectories and contacts

2.2.1 Temporally varying spatial risk

In this study, spatial risk relies only on monitoring of SIR events, to track infectious trajectories and location of contacts (Benreguia et al, 2020). With such monitoring, we computed risk scores based on four risk basis, as follows:

(a) *Infectious trajectories* refer to the amount of time an infectious trajectory has spent in each cell, where infected person is only tracked from the time of infection until recovery or quarantine.

(b) *Infected individuals* refer to the count of infected individuals in each cell. This is distinct from (a) in a sense that shorter cumulative duration of many individuals is riskier than a longer duration of a single individual.

(c) *Infectious contacts* refer to the locations of all contacts involving an infected individual. They are more than the count of times infection is actually transmitted as this involves all ($S \leftrightarrow I$) contacts; and on the other end, transmission is dependent on the randomly chosen event. From this, a relation *transmissive contacts* $\in$ *all* ($S \leftrightarrow I$) *contacts*, can be deduced.

(d) *Social distancing violations* refer to all ($S \leftrightarrow S$) contacts. This property reflects population density and also captures the notion that a place (cell) with higher precautionary violations must be of higher risk than a place following the public health regulations (Rezaei and Azarmi, 2020).

As contact graphs are developed per day, the same frequency can be followed in order to develop these four risk types. This means that risk scores of each cell is based on the cumulative effect of activities from the previous day and are to be updated every next day. Figure 6 depicts the process of computing risk basis, by tracking infectious trajectories for their duration and count, alongside the monitoring of contacts for their spatial locations.

With four risk basis, there comes a need for integrating these risk attributes into a single representation. This requires normalizing all grids to a common range and to further combine them into a single grid. This results in a *risk*

map, based on activities from the previous day to provide an evolved risk for the next day.

2.2.2 Unsupervised classification using SOM and K-means

For combining multiple aspects into a single map form, implementation of a multi-criteria analysis approach is not applicable as there is no prior information of which criterion is significant over others. For classification, a supervised method requires information about the characteristics of the target class and pre-existing labels for the method to cluster data and label them accordingly. However, lack of validation data restricts the application of supervised classification. A possible solution is to implement an unsupervised learning method, as it does not rely on pre-existing labels for reinforcement. Such methods only require input patterns to highlight relationships and can assist in the exploration of the available covariates to develop a single classified risk map.

One such unsupervised clustering technique is *Self Organizing Map* (SOM), which can serve to the purpose of combining information of multiple grids into a single one. SOM is basically a dimensionality reduction technique but as SOM preserves the topographic relationships in feature space to ensure nearby objects are clustered together, it has been extensively used for the clustering of geospatial data (Henriques et al, 2012; Gopal, 2016).

Considering the dimensionality reducing capability, SOM is similar to the statistical equivalent of Principal Component Analysis, whereas Baccao et al. (Bação et al, 2005) suggest SOM as a possible substitute for K-means clustering when the neighbourhood is not considered. Besides, in comparison to statistical techniques, SOM offers three main advantages due to its non-parametric nature: (i) it works independent of variable's distributions, (ii) it is computationally efficient to non-linear problems and (iii) it caters for noise or missing data more effectively (Asan and Ercan, 2012).

As highlighted in (Vesanto and Alhoniemi, 2000), the best approach to implement SOM is a two-step process. First, input data is to be transformed into a two-dimensional neurons network; secondly, SOM neurons are to be clustered using a hierarchical or partitive approach. The major benefits of this two-step approach is; (i) the computational efficiency even with a smaller dataset and (ii) noise reduction in case of imperfect data as input for clustering.

An important consideration here is to choose the size of SOM neurons network which is dependent on the size of input dataset. In our implementation, we use a regular lattice of $12 \times 15 = 180$ cells, hence an optimal size of SOM neurons can be acquired as $5 \cdot \sqrt{180} = 67.08$. For a two-dimensional structure of neurons network, we considered the total of 64 neurons instead of 67, which could be arranged in symmetric shape of 8x8 neurons network.

After the establishment of SOM network as a representation of input data (multiple grids), a hierarchical clustering process can cluster the neurons into the desired number of groups. In this paper, we follow the partitioning approach of K-means clustering as they do not rely on previously found clusters as the hierarchical approach does (Vesanto and Alhoniemi, 2000).

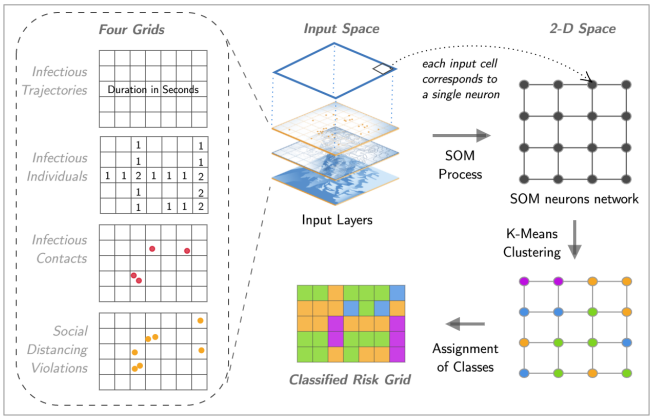

Fig. 7 Unsupervised classification workflow using SOM and K-means

As risk values of grid cells vary over time, and a flexible value of k (number of clusters) in K-means can result in a different count of classes for different days, we fixed the count of classes to 5 so as so to have an equal number of groups every time a new spatial risk is computed. However, in case data does not allow to have five classes, then an optimized number of classes is chosen for an appropriate representation. This results in a classified grid-based risk where each cell corresponds to a class of risk. As the output after K-means is an un-ordered classification, which is the same as segmenting the cells in different groups but not knowing which group is of higher risk, classes are assigned with appropriate labels by comparing the cumulative average of the risk score in all cells associated with each class and further assignment of ordered labels in descending order with greater average as the highest risk class. Figure 7 illustrates the complete process of combining multiple grids through the use of SOM followed by K-means and further labelling.

2.2.3 Contextual contact tracing

In order to include the temporally varying spatial risk for each specific contact, we modify the previous contact graph G to obtain a new contact graph G'. For this modification, we use the daily-acquired risk-based grids (spatial risk) and based on the location of the contacts we obtain G' which considers the risk score of each contact's location. In G', each contact value is a varying intensity depending on the spatial risk instead of the previous constant value of 1 (which represented a contact). Here, we introduce a new range representation for a contact between the value of 0.5 *(lowest risk)* to 1.5 *(highest risk)*. The rationale behind this range is to be able to compare with baseline-SIR setup (see Figure 13) where previous value 1 is the mean of the new range representation. Once real data about spatial risk as well as infection information

is available to fit the model to the data, different values for this range can be configured to identify the best fit.

Using this new matrix G' in (2), we obtain new *degree of diffusion* $\mathcal{K}'_i(t)$, which is used on the rate equations defined on Table 2. As rates of events in SIR model are based on the cumulative infectious contacts represented by $\mathcal{K}_i$, a varying contact value (between 0.5 to 1.5) will result in a varied influence to the transmission process for each specific contact, meaning a direct effect of spatial risk on the disease transmission. As in the baseline-SIR model, this spatio-SIR model can be solved using the GFRM's rate equations as stated in Table 2. Besides, the consideration of varying $\mathcal{K}'_i$ in these equations only influence events related to *Susceptible* population ($\boldsymbol{S} \rightarrow \boldsymbol{I}$, $\boldsymbol{S} \rightarrow Q_S$ and $\boldsymbol{S} \rightarrow Q_I$). However, there is no influence of spatial risk on events related to *Infected* individuals and those in *Quarantine*. This process of dynamically computing risk scores based on daily movement and reflecting its effect by modifying contact graph is termed as *Contextual Contact Tracing*.

3 Data analysis and simulations

3.1 Dataset and experimental design

NCCU Trace (Tsai and Chan, 2015) refers to an android application to trace movements of 115 students in a campus environment of National Chengchi University (Taiwan), for a period of 15 days with measurement interval up to 10 minutes and spatial position rounded to meters. The application was designed to capture information regarding GPS, WiFi, and Bluetooth devices in proximity, resulting in their movement traces. The Appendix contains details of the NCCU dataset with an overview of the study area and sample of recordings. For an epidemic, a period of 15 days is very short to assess the spread of infection. A possible solution is to extend the period of the dataset by concatenating the same dataset multiple times, as the pattern of human mobility shows a regularity over the same weekdays. Such joining can produce a data for 150 days, an appropriate duration for epidemic modeling.

Both baseline-SIR and the spatio-SIR models are evaluated over NCCU data. The experiments assume 10 individuals as *initial infected* ($I_0 = 10$) on the first day of the epidemic with no recovered individual ($R = 0$). Sum of individuals in all compartments is 115 at all times. We used the values of COVID-19 parameters as discussed in Table 3 and *tracing efficiency* q' of 0.1. For the stochasticity, 10 realizations of the same initial conditions but the random allocation of initial infection are executed. This means that in each realization, infected individuals are different. Averaging the results over 10 realizations, average curves are obtained, where a curve represents the count of individuals in each compartment. Due to stochasticity, duration of the epidemic in these realizations varies, hence we extrapolate trends of other realizations to the epidemic with the longest duration to obtain an average representation.

In each run within a single realization of a model, only one epidemic event is executed. The *time* of next event is a stochastic duration as a part of day, hence there are multiple events per day, with at least one event in a day, and overall hundreds of events even for a short epidemic of few weeks. Hence, a single realization of model moves forward executing events after each *time* step. As a whole, these executed events simulates a disease outbreak scenario.

3.2 Data analysis

3.2.1 Results - Baseline-SIR model

The ability of an individual-level compartment model to monitor the latent state of each individual at all times highlights its importance in the infectious diseases realm. To understand this capability, Figure 8 illustrates individual-level latency of a subset of the population (33 out of 115). At the start of the epidemic (day 1), four individuals (2, 21, 26 & 32) are infected as the initial outbreak, whereas the remaining all are *Susceptible*. The first stochastic event (second column from left) is of infection for individual (8). In every iteration, there is only one event, where the time of the next event (a part of the day) is also random, hence there can be multiple events in a single day. Individual (8) remains infected for a week and gets detected around the 11^{th} day. Individual (2) gains recovery only after few days. Individual (21) remains infected and undetected for the whole shown period. Similarly, the state of each individual can be observed based on the time-series review of their associated compartment.

Exploring the modification of a spatial context needs the setup of a baseline model to experiment over. Figure 9 presents the output of such a baseline setup in form of an outbreak scenario using parameters from Table 1. At the beginning of the epidemic, everyone except the *Infected* is in the *Susceptible* compartment, which means there is no *Recovered* individual. Initially, the count of *Infected* individuals increases from 10 to 14 in the first few days as *Susceptible* population interacts (contacts) with already infected (initial outbreak). However, not only their count decreases afterwards as they are sent into *Quarantine Infected*, but the *Susceptible* count also diminishes from initial count of 105 to 40 in a fortnight. Due to backward tracing $\mathcal{C}_i$, higher number of individuals are identified as *exposed* and sent into *Quarantine Susceptible* as a precautionary measure. These plummeted trends of the count of *Susceptible* and *Infected* forces less population on the streets, which not only restricts the future infectious contacts but ultimately the overall disease outbreak.

The peak of individuals in *Quarantine Susceptible* is around 19^{th} day with 40 plus individuals, where afterwards the sum of individuals remains nearly constant which depicts an equal frequency of individuals moving between ($S \leftrightarrow \mathrm{Q}_S$) compartments. *Quarantine Infected* compartment reaches its highest count of 5 twice on the 13^{th} and 22^{nd} days. Once a person is *Recovered*, that individual remains in that compartment, which is evident from the continuous

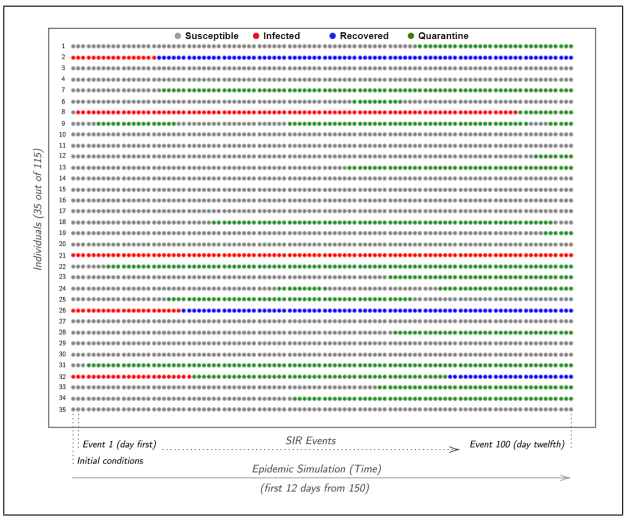

Fig. 8 Individual-level change in latency of 35 out of 115 total individuals is shown based on the SIR events as model runs forward. Each row belongs to a single individual, where the compartment they belong to at an instance of time is represented column-wise chronologically from left to right. There is only one event per column with multiple events per day, where figure illustrates first 100 events from the initial 12 days of epidemic.

increase in its count from 0 at the start to 38 at its end. Even after there is no *Infected* person on the street after the 45^{th} day, the model continues in anticipation of risk due to the presence of individuals in *Quarantine Infected*; and ultimately ends the epidemic with their recovery around 113^{th} day.

3.2.2 Results - Enhanced spatio-SIR model

Spatio-SIR enhancement is achieved by computing the spatial risk out of events in the baseline setup, which requires monitoring of SIR events for infectious activities. Figure 10 presents a 1-day sample of such infectious activities. Figure 10-A illustrates the movement of infectious individuals shown over the study area. In this sample, there are 7 infected individuals with mobility concentrated inside the NCCU campus (center-top). Out of these infectious trajectories, two sorts of attributes are extracted. First is the collective duration of time spent by these individuals in each area, and secondly how many individuals were located in each area. The other two basis are of *Infectious Contacts* and *All Contacts*, where the latter is shown in Figure 10-B. It identifies locations of all contacts termed as *social distancing violations* in order to highlight the notion that a place with a higher number of contacts means it is of higher risk than a place with a lower number of contacts. This concept has been also implemented by (Rezaei and Azarmi, 2020) for infection risk assessment.

Based on the risk basis shown in Figure 10, grid-based risks are developed as presented in Figure 11. Here, the trajectories and contacts are transformed into a grid structure with an intensity of associated attributes normalized to

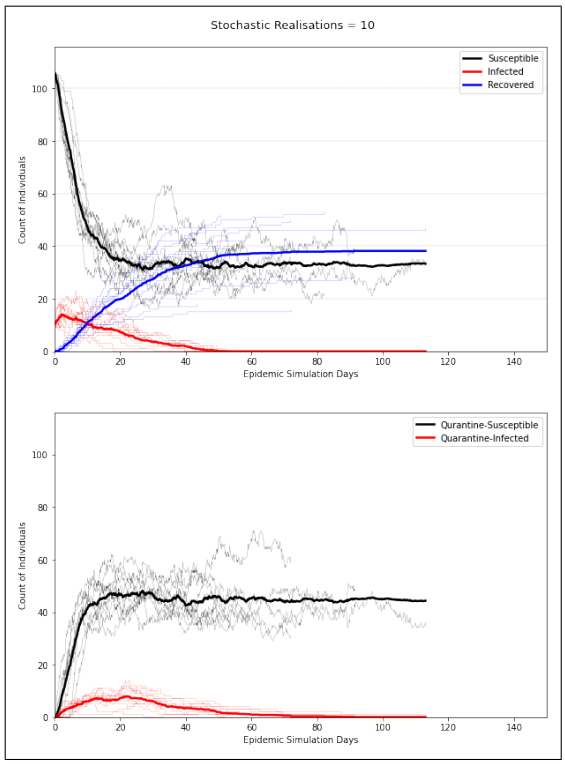

Fig. 9 Output of a disease outbreak scenario by baseline-SIR model. *(Top)* presents trends related to counts of *Susceptible, Infected & Recovered*, whereas *(Bottom)* illustrates count of individuals in *Quarantine* related compartments. Each of the 10 realizations of stochastic model is shown *(light in color)*, with their average curves represented with *(dark bold) lines*. Count of total population is 115 which are represented over the *Y-axis*.

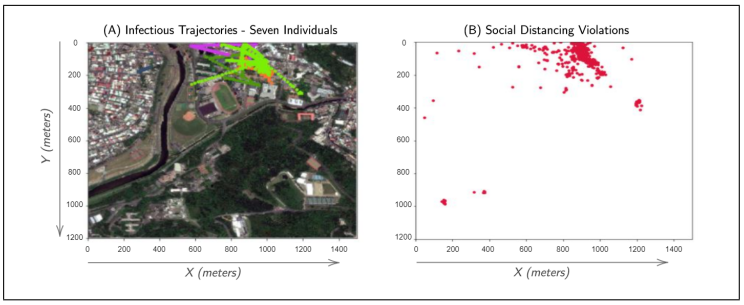

Fig. 10 Risk basis of *infectious trajectories* and *location of contacts* from one complete day during an epidemic. In the sample shown in Sub-Figure (A), there are total of 7 infected individuals from that day. Sub-Figure (B) shows locations of all contacts from the same day.

[0,1]. Figure 11-A and 11-B capture information of infectious trajectories in terms of *duration* and *count* respectively. Similarly, the location of different

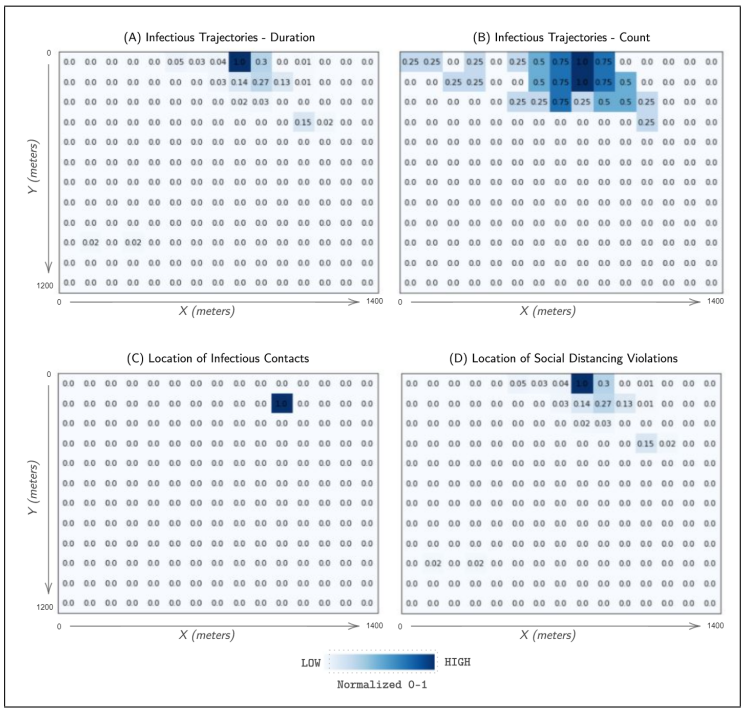

Fig. 11 Grid-based risk computed out of risk basis. Sub-Figure (A) and (B) refer to grid-based representation of *duration* and *count* of trajectories shown in Figure 10-(A). Sub-Figure (C) depicts location of an infectious contact occurred on the same day, whereas Sub-Figure (D) presents translation of Figure 10-(B) into grid form. All values are normalized to the range of 0 to 1, where 1 refers to highest risk.

nature of contacts is captured in Figure 11-C and 11-D. Based on the previous day, these attributes serve as the basis of risk for the next day.

To identify spatial risk for the future contextual tracing of contacts, multiple grids from Figure 11 are integrated into a single grid as shown in Figure 12. In order to classify the output to segment areas of higher or lower risk, risk scores are grouped into 5 classes with their labels corresponding to their intensity of risk. The classes of risk are (0.50, 0.75, 1.00, 1.25, 1.50) with 1.50 referring to the highest risk. A review of this result shows that based on activities from the previous day (Figure 10), the highest risk area is at the centre-top cells, whereas the surrounding areas are also of higher risk. While there is no spatial risk in the remaining study area on this particular day, however, due to the temporally varying nature, the spatial risk may evolve in future instants.

Results of the spatio-SIR model are compared with the results of baseline-SIR in Figure 13. As the inclusion of spatial risk tends to affect the rates of events related to *Susceptible* individuals and getting infected is subject to an infectious contact, hence in the spatio-SIR model, there are more events of the population moving into *Quarantine Susceptible*. Though the trends of

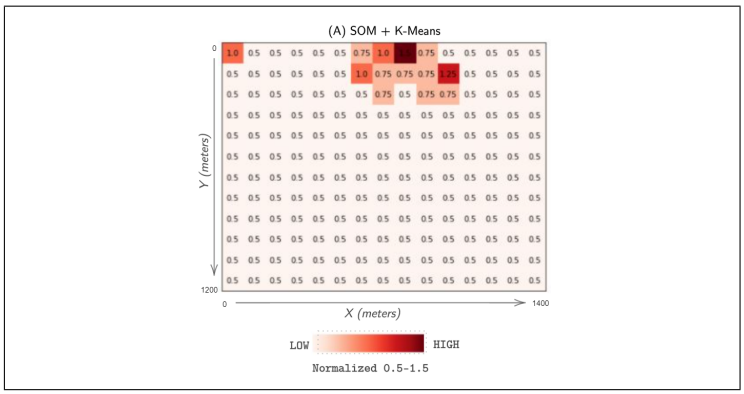

Fig. 12 Combining risk from multiple grids shown in Figure 11 into a single grid output using SOM and K-means. Due to lack of validation data, integration is executed in unsupervised manner through the implementation of SOM followed by K-means. Risk scores are computed in the range of [0.5,1.5], where 1 refers to the previous normal (existing SIR model with a constant spatial risk and all contacts being of equal nature)

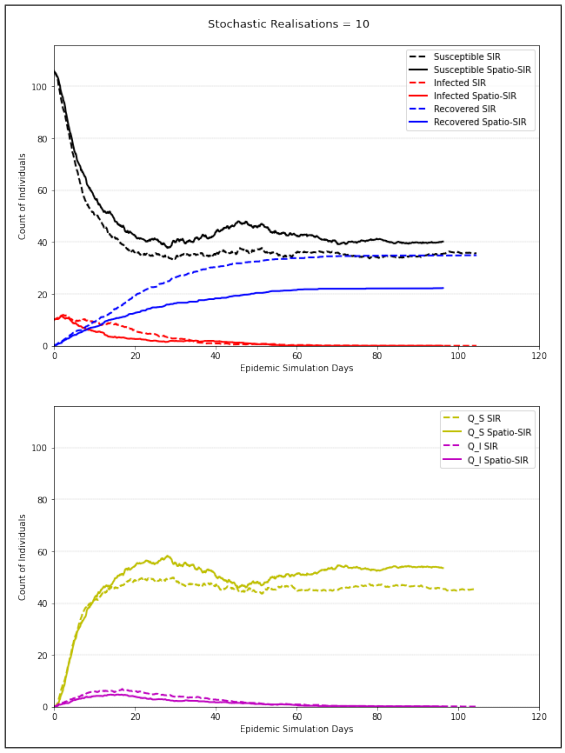

Fig. 13 Comparing *average* of 10 stochastic realizations of a disease outbreak scenario from baseline-SIR *(dashed)* and spatio-SIR *(solid)*. *(Top)* presents trends related to count of *Susceptible, Infected & Recovered*, whereas *(Bottom)* illustrates counts of individuals in *Quarantine* related compartments. Count of total population is 115 which are represented over *Y-axis*.

Quarantine Susceptible in both models are similar till day 15^{th}, however, the mentioned phenomenon is evident afterwards where peak of individuals in *Quarantine Susceptible (spatio-SIR)* is 59 on 29^{th} day, whereas there are less than 50 individuals in *Quarantine Susceptible (baseline-SIR)* by the same day.

Early events of quarantining reduces the counts of *Infected*, *Recovered* and *Quarantine Infected*. Comparing the trends of *Susceptible* population, it can be observed that in the first week both are more or less similar, however, the first week onward the susceptible population in baseline-SIR decreases to 40 by 17^{th} day, whereas it takes an extra week (23^{rd} day) for the same decline up to 40 in spatio-SIR. This highlights that due to the additional aspect of spatial risk, a greater fraction of the population remains susceptible. Similarly, an increase in the count of *Susceptible* around 45^{th} day depicts the return of quarantined population after a period of two weeks, whereas such a return is not visible in baseline-SIR as there is no consideration of spatial risk. With a higher count of total individuals in *Quarantine Susceptible*, overall infection is controlled which can be confirmed from the trend of *Infected* and *Recovered*. As in spatio-SIR model, total recovered are 21 compared to the count of 35 in baseline-SIR model. The same can be observed in the trends of *Quarantine Infected*, as with less *Infected* on the streets, the spread of infection is controlled; hence, a lower count of *Quarantine Infected* in spatio-SIR compared to baseline-SIR, apart from the start and end of an epidemic which is nearly similar.

3.2.3 Simulated scenarios over the NCCU data

This section reinforces the need for a spatio-epidemic tool. As model simulates a scenario based on the initial values configured for it, changing the initial setup can help assess impact of the change on the overall disease outbreak simulation. Here, baseline-SIR model executes one such variation as presented in Figure 14 with different intensities of the *Initial Infected* I_0.

In general, the higher quantum of initial outbreak results in a longer epidemic which is evident in all subplots. In Figure 14-A, *Susceptible* population is compared, where higher count of initial outbreak reflects in early departure of individuals from the susceptible compartment; either getting *Infected* (due to greater frequency of infectious contacts) or *Quarantined* (because of prior contact tracing of *Infected* individuals). Higher infected count ($I_0 = 10$ and $I_0 = 15$) results in decrease of *Susceptible* count from 105/100 to approximately 40 within 2 weeks, whereas when $I_0 = 5$ reaches the count of 40 after six weeks. Figure 14-B illustrates the effect of varying initial outbreak on the total counts of *Infected*, where a directly proportional relationship is evident in the initial spread of infection up to 19^{th} day. However, once a majority of *Infected* are sent into *Quarantine Infected* and higher count of individuals are already in *Quarantine Susceptible*, all scenarios tends to have a similar pattern afterwards. Similarly, Figure 14-C depicts a likewise trend of initial difference, where two setups of ($I_0 = 05$) and ($I_0 = 10$) later (after 70^{th} day) coincide to have a similar pattern (around 30 *Recovered* individuals). However, ($I_0 = 15$) results in a massive outbreak with almost 50 *Infected* individuals by the 70^{th}

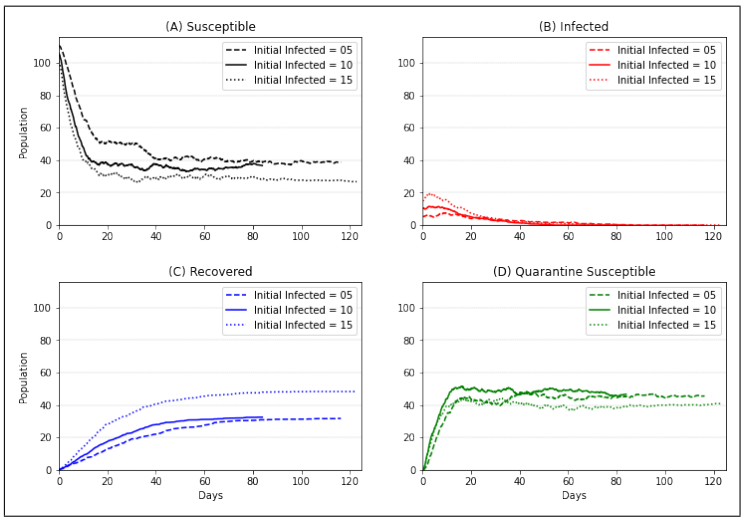

Fig. 14 Comparing the average of 10 stochastic realizations (of baseline-SIR model) with varying count of *Initial Infected* I_0, to observe their effect on the overall disease outbreak. Three scenarios of $I_0 = 5/10/15$ are shown with a population size of 115. Subplots *(A,B,C and D)* shows comparison of *Susceptible, Infected, Recovered & Quarantine Susceptible*, respectively. Due to stochasticity, final duration of epidemic varies depending on the overall spread of infection.

day. Figure 14-D highlights that higher count of initial infected will either send more contacts into *Quarantine Infected* or *Quarantine Susceptible*, which is dependent on (i) the *transmission rate* ($\beta = \kappa \cdot b$) and (ii) the chance element of event-based setup. Hence, the relation of initial infected with *Quarantine* related compartments is not straightforward. However, trend of ($I_0 = 15$) specially after the 40^{th} day depicts that due to greater initial outbreak, more individuals were *Infected*, thus more people are in *Recovered* and *Quarantine Infected*, because of which overall count of *Quarantine Susceptible* is low.

Another possible variation on the analyzed scenarios can be the *Tracing Efficiency* which is available in Figure 15. Tracing efficiency refers to the fraction of identified prior contacts based on backward tracing. As 100% tracing is not plausible, only a proportion is evaluated as an estimate of tracing. In the case of no backward tracing (zero efficiency) shown in Figure 15-A, there are no individuals in *Quarantine Susceptible*. Only *Infected* who gets detected are sent into quarantine, which results in a massive disease outbreak with count of *Recovered* more than 80 individuals. In Figure 15-B, 62 individuals are in *Quarantine Susceptible* by the 10^{th} day, whereas with efficiency of 0.50 (Figure 15-C) and 0.75 (Figure 15-D), there are 77 and 88 individuals in *Quarantine Susceptible* by the same period of 10 days. It can be deduced that for every 25% increase in the *Tracing Efficiency*, 10% more population is sent into quarantine. In general, it can be said that with greater tracing efficiency, the greater amount of population is early forced for quarantine, which ultimately reduces the overall spread of infection (less *Infected* and less *Recovered*). The

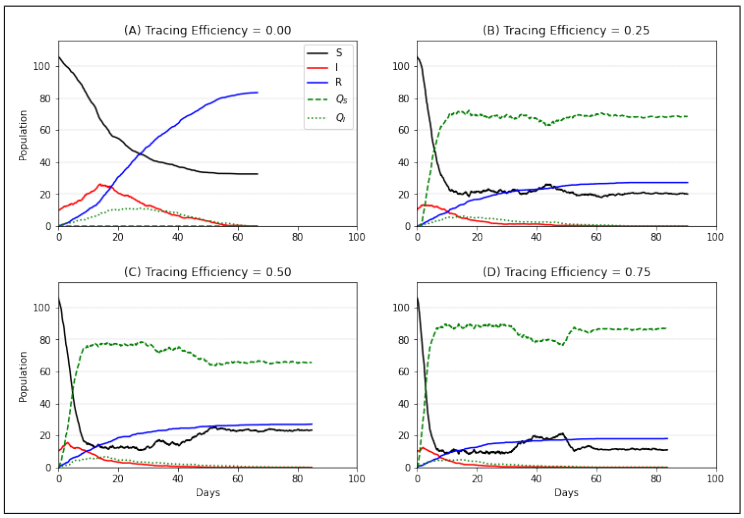

Fig. 15 Comparing the average of 10 stochastic realizations (of baseline-SIR model) with varying *Tracing Efficiency* $q^{'}$, to observe their effect on the overall disease outbreak (population size = 115). Subplots *(A,B,C and D)* present four cases of $q^{'}$=0, 0.25, 0.50 and 0.75, respectively, where 1 means 100% backward tracing.

population being forced to quarantine means they are leaving the *Susceptible* compartment, which is evident by the degree of slope in the downward trend of *Susceptible* count proportional to tracing efficiency. Due to high tracing efficiency in the Figure 15-D, a huge subset of the population is sent into *Quarantine* immediately as the infection breaks out. This large amount of individuals when collectively comes out of quarantine (after a period of 14 days), results in a sudden drop of Q_S count around the 40^{th} day. An opposite can be observed in the count of *Susceptible*.

Other than varying the initial configuration, another capability of the spatio-epidemic tool is the ability to simulate real-world scenarios such as relaxation in social-distancing, spatio-temporal curfew/lockdown or a holiday season with more population on the streets. This capability of the tool can assist policymakers to simulate scenarios, visualizing the consequence of their decisions prior to their actual implementation. One such real-world scenario of a holiday season with higher count of public on the streets is presented in Figure 16. In this experiment, we introduce a holiday season as an *Intervention* in a specific period from day 11^{th} to 20^{th}. Quantitatively, this intervention is in the form of spatial high-risk of value 1.5 at all areas (cells).

A major difference is in the overall period of epidemic, where the *Intervention* setup executes an epidemic of more than 100 days considering the added spatial risk from day 11 to 20, whereas in spatio-SIR modeling the epidemic is finished in less than 60 days. Observing the trend of *Recovered* individuals, a continuous increase after day 10 is evident in *Intervention* setup, compared to spatio-SIR output. This escalation ends up with a total of 37 recovered in the

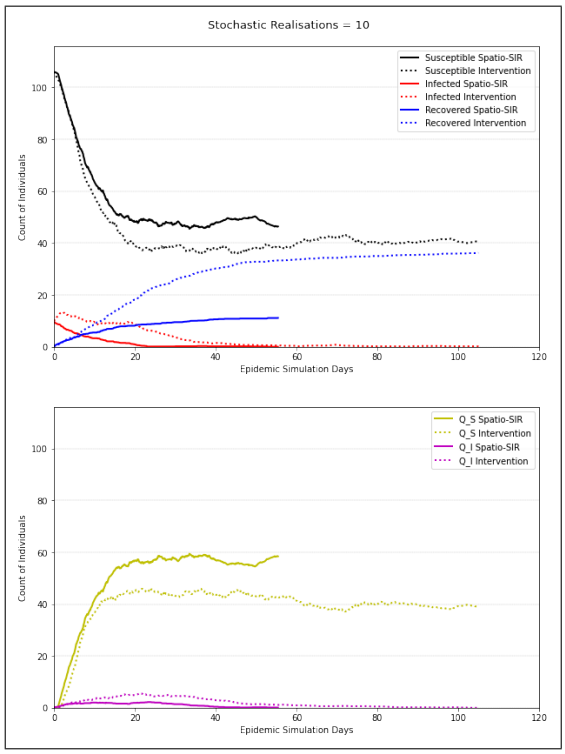

Fig. 16 Comparing average of 10 stochastic realizations of a disease outbreak scenario from spatio-SIR *(solid)* and a case of Intervention - spatio-SIR model with spatial risk of 1.5 from day 11 to 20 *(dotted)*. *(Top)* presents trends related to counts of *Susceptible, Infected & Recovered*, whereas *(Bottom)* illustrates counts of individuals in *Quarantine* related compartments. Count of total population is 115 which are represented over *Y-axis*.

former, while total recovered individuals in the latter are 11. A similar pattern is identifiable while observing the trends of *Infected* population, where since day 10^{th}, the rate of infection is more or less constant (a horizontal line) until the 20^{th} day. This is different from the infected trend in spatio-SIR model where the rate of infection is decreasing after the initial increase in the first few days of the epidemic. Observing the trend of *Quarantine Susceptible*, a spike is noticeable after day 11 in the *Intervention* setup. Counts of susceptible in quarantine in *Intervention* setup is 56 on day 20^{th}, whereas under the spatio-SIR model there are only 43 susceptible individuals in quarantine by the same day, confirming the capability of new setup to capture spatial high-risk.

3.3 Mobility simulation in a new space

As individual-level mobility datasets are scarcely available, a possible solution is to self-simulate movement trajectories for the study area (new space). This

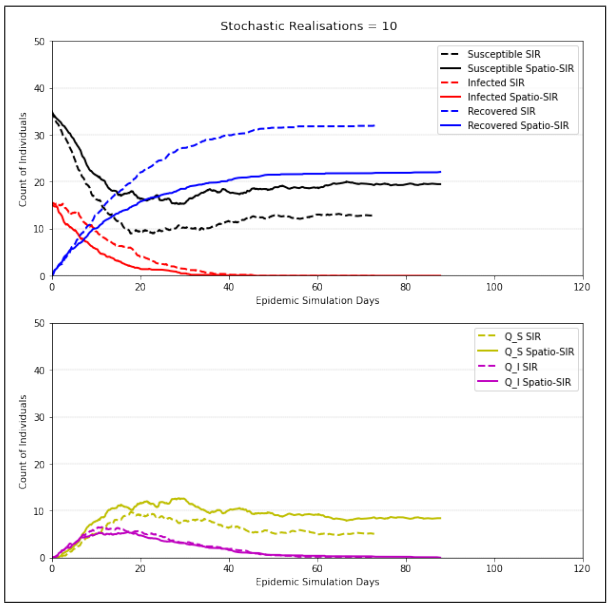

Fig. 17 Comparing average of 10 stochastic realizations of a disease outbreak scenario from baseline-SIR *(dashed)* and spatio-SIR *(solid)*. *(Top)* presents trends related to counts of *Susceptible, Infected and Recovered*, whereas *(Bottom)* illustrates counts of individuals in *Quarantine* related compartments. Count of total population is 50 which are represented over *Y-axis*.

can also help in the application of methodology on multiple datasets to assess its performance in different spaces.

In this study, we have also generated a synthetic dataset using spatial movements from Geolife Data (Zheng et al, 2011). Geolife, a project by Microsoft, provides trajectory movements of 178 users for a period of four years with temporal resolution of 1 to 5 seconds and spatial resolution of 5 to 10 meters. The dataset in total contains 17621 trajectories, total distance of 1251654 kilometers holding information of 48203 hours. As for contact tracing and spatial risk assessment, we require mobility to be highly concentrated on a small study area. Unfortunately, this is not the case in the original Geolife dataset. However, we used only its spatial movement but modified the temporal and user-related attribute to reflect the daily movement of 15 days for 50 users, for a study area of 20×16 square kilometres. Figure 22 in the Appendix presents a visualization of this modified construction, whereas Figure 17 illustrates the comparison of the SIR model and spatio-SIR model over this new dataset. Both models, SIR and spatio-SIR, depict a similar trend on synthetic trajectories as over NCCU trace. The consideration of spatial risk tends to send more people in *Quarantine Susceptible* which initially protects them from the infection, but the population remain susceptible in general as the quarantined population comes back to *Susceptible* stage after quarantining period.

Similarly, the overall infection propagation is reduced due to lower count of infectious contacts which results in lower counts of *Recovered*.

4 Conclusions and discussion

We conclude that the inclusion of spatial risk in epidemic modeling can greatly support the public health system by identification of infectious contacts and highlighting places carrying the high risk. It is a bi-fold domino effect that relies on both, persons and places, and breaking the chain is necessary not only in terms of individuals but also high-risk areas. For a critical time such as COVID-19, an integrated approach as the one introduced here can be developed into a comprehensive system of *infectious disease surveillance*. In terms of modeling, consideration of spatial risk as in spatio-SIR model increases the tracing efficiency, where a greater number of individuals are highlighted as exposed depending on the location of contacts; as in this study contacts are mostly concentrated in a small region that is at high-risk at all times. These vulnerable individuals who are currently *Susceptible* will either be infected or sent into quarantine, depending on the chance element of event-based setup. This consideration of exposure based on spatial risk tends to perform more *meaningful events* [1] concerned with the *Susceptible* population rather than events to the *Infected* or *Quarantined*. Furthermore, it is shown that this framework can act as a tool for policymakers to execute scenarios, visualizing the consequence of their decisions prior to their actual implementation.

We have proposed a generalized framework for spatio-SIR modeling, however, a disease-specific model can be developed by adjusting the parameters available in Table 3. With regards to contact tracing, the study highlights that for contact tracing to be effective, the maximum fraction of the population needs to be digitally activated, using the contact tracing app or other implemented mode of tracking (Hernández-Orallo et al, 2020).

The major limitation of the study is the non-availability of actual information about infected individuals. With that, the proposed methodology could have been configured to fit a model to data. In this paper, this limitation was handled through a self-induced initial outbreak. Another aspect is that the selected dataset is not of an epidemic scenario. A dataset from an era of an epidemic situation can assist in the analysis of such patterns and further explore its spatial risk. Similarly, 15 days recording of movements is an inadequate period for a long-standing scenario of an epidemic. In this paper, this limitation was handled by concatenating the same dataset multiple times for 150 days. However, a better option would be to have a mobility dataset of a longer duration. A limitation of the followed approach is that the contacts were identified per day. This approach helped in establishing a setup to understand disease dynamics in a spatial context, however, a finer frequency such as hourly contacts graphs or a real-time application of tracing in terms of recording a contact as they happen can be followed for higher accuracy.

[1] Infection or Quarantining, compared to Recovery while in Quarantine

This modification of an existing SIR model into a spatio-SIR model through the inclusion of spatial risk serves only as a foundation of an idea. This leads to many way forwards opening new avenues for the integration of spatial component into digital epidemiology. Spatial risk is a complete domain in itself that includes the identification of factors stimulating the vulnerability of being infected at a certain place and time. Hence, it is recommended to incorporate the spatial context from additional perspectives other than just infectious trajectories. A suggested idea is to integrate spatial information such as points of interests (restaurants, parks, etc.), public transits, urban functions, demographic details and environmental factors, for the overall spatial risk assessment. Such a study will explore the spatial effect of covariates in disease transmission by understanding their intrinsic underlying relationships presenting a higher or lower score of risk, and additionally, how these covariates amalgam as a whole.

Implementation of this study was based on an event-based SIR model where rates of events were computed to randomly draw the next event, as well as the time of the event and the person to which event will occur. This complete stochasticity can be adjusted in a sense to develop a semi-stochastic setup where the person to which event will occur is not completely random but a factor based on their exposure. Such a factor can be associated with each individual based on their movement in infectious places and the frequency of their contacts. Though semi-stochastic in nature, a specific model like this can also provide with the exposure profile for each individual.

Given that this domain of infectious diseases generally lacks data availability related to infection and/or movement, a practical way forward is to transform this spatio-enhanced model into a comprehensive tool for simulations. Such a tool can allow users to feed in movement data and then based on infectious movements, the user can execute spatio-SIR modeling while configuring the initial setup. Furthermore, the tool can have the capabilities to implement real-world scenarios like spatial curfew, commercial lockdown, relaxation in social distancing etc. The overall situation of COVID-19 signifies the importance of such a tool that can support public health policymakers as and when required.

Overall, this paper concludes that tracking of individual-level infectious trajectories is critical not only for person-to-person contact tracing but also to identify spatial risk which is transmitting (surface/aerosol transmission) as well as propagating (inducing riskier contact) in nature. The study also highlights that accurate modeling of this sort is restricted due to the data unavailability, and there is a critical requirement of datasets to ensure a practical application of the proposed approach.

The authors conclude this study with the remarks, that even if this domain[2] is generally hindered due to the lack of data availability, the investigation process related to it should keep on exploring methods to effectively understand disease dynamics. This is beneficial not only for literature but also critical for the overall well being of humanity.

[2] individual-level trajectory-based infectious diseases SIR modeling

References

Anglemyer A, Moore TH, Parker L, Chambers T, Grady A, Chiu K, Parry M, Wilczynska M, Flemyng E, Bero L (2020) Digital contact tracing technologies in epidemics: a rapid review. Cochrane Database of Systematic Reviews (8) 3

Angulo J, Yu HL, Langousis A, Kolovos A, Wang J, Madrid AE, Christakos G (2013) Spatiotemporal infectious disease modeling: a bme-sir approach. PloS one 8(9):e72,168 4

Asan U, Ercan S (2012) An introduction to self-organizing maps. In: Computational Intelligence Systems in Industrial Engineering, Springer, pp 295–315 13

Bação F, Lobo V, Painho M (2005) Self-organizing maps as substitutes for k-means clustering. In: International Conference on Computational Science, Springer, pp 476–483 13

Bardina X, Ferrante M, Rovira C (2020) A stochastic epidemic model of covid-19 disease. arXiv preprint arXiv:200502859 4

Benreguia B, Moumen H, Merzoug MA (2020) Tracking covid-19 by tracking infectious trajectories. arXiv preprint arXiv:200505523 2, 4, 12

Berger DW, Herkenhoff KF, Mongey S (2020) An seir infectious disease model with testing and conditional quarantine. Tech. rep., National Bureau of Economic Research 4

Bisin A, Moro A (2020) Learning epidemiology by doing: The empirical implications of a spatial-sir model with behavioral responses. Tech. rep., National Bureau of Economic Research 4

Brockmann D, David V, Gallardo AM (2009) Human mobility and spatial disease dynamics. Reviews of nonlinear dynamics and complexity 2:1–24 3

Desjardins M, Hohl A, Delmelle E (2020) Rapid surveillance of covid-19 in the united states using a prospective space-time scan statistic: Detecting and evaluating emerging clusters. Applied Geography p 102202 4

Eames KT, Keeling MJ (2003) Contact tracing and disease control. Proceedings of the Royal Society of London Series B: Biological Sciences 270(1533):2565–2571 6

Enright J, Kao RR (2018) Epidemics on dynamic networks. Epidemics 24:88–97 6

Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dorner L, Parker M, Bonsall D, Fraser C (2020) Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. Science 10

Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. The journal of physical chemistry 81(25):2340–2361 10

Gopal S (2016) Artificial neural networks in geospatial analysis. International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology pp 1–7 13

Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, Munday JD, Kucharski AJ, Edmunds WJ, Sun F, Flasche S, Quilty BJ, Davies N, Liu Y, Clifford S, Klepac P, Jit M, Diamond C, Gibbs H, van Zandvoort

K, Funk S, Eggo RM (2020) Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts. The Lancet Global Health 8(4):e488–e496 10

Henriques R, Lobo V, Bação F (2012) Spatial clustering using hierarchical som. Applications of Self-Organizing Maps pp 231–250 13

Hernández-Orallo E, Manzoni P, Calafate CT, Cano JC (2020) Evaluating how smartphone contact tracing technology can reduce the spread of infectious diseases: the case of covid-19. IEEE Access 2, 5, 6, 26

Keeling MJ, Rohani P (2011) Modeling infectious diseases in humans and animals. Princeton University Press 6, 10

Lang JC, De Sterck H, Kaiser JL, Miller JC (2018) Analytic models for sir disease spread on random spatial networks. Journal of Complex Networks 6(6):948–970 4

Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J (2020) Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2). Science 10

Mahsin M, Deardon R, Brown P (2020) Geographically dependent individual-level models for infectious diseases transmission. Biostatistics 4

Martinez-Beneito MA, Mateu J, Botella-Rocamora P (2020) Spatio-temporal small area surveillance of the covid-19 pandemics. arXiv preprint arXiv:201103938 4

Martinez-Martin N, Wieten S, Magnus D, Cho MK (2020) Digital contact tracing, privacy, and public health. Hastings Center Report 50(3):43–46 3

Müller M, Derlet PM, Mudry C, Aeppli G (2020) Testing of asymptomatic individuals for fast feedback-control of covid-19 pandemic. Physical biology 17(6):065,007 3

Park J, Chang W, Choi B (2021) An interaction neyman-scott point process model for coronavirus disease-19. arXiv preprint arXiv:210202999 5

Reichert L, Brack S, Scheuermann B (2020) Privacy-preserving contact tracing of covid-19 patients. IACR Cryptol ePrint Arch 2020:375 3

Rezaei M, Azarmi M (2020) Deepsocial: Social distancing monitoring and infection risk assessment in covid-19 pandemic. Applied Sciences 10(21):7514 12, 17

Salathé M (2018) Digital epidemiology: what is it, and where is it going? Life sciences, society and policy 14(1):1 2

Simmerman JM, Suntarattiwong P, Levy J, Gibbons RV, Cruz C, Shaman J, Jarman RG, Chotpitayasunondh T (2010) Influenza a Virus Contamination of Common Household Surfaces during the 2009 Influenza A (H1N1) Pandemic in Bangkok, Thailand: Implications for Contact Transmission. Clinical Infectious Diseases 51(9):1053–1061, DOI 10.1086/656581, URL https://doi.org/10.1086/656581, https://academic.oup.com/cid/article-pdf/51/9/1053/827213/51-9-1053.pdf 3

Souza RC, Neill DB, Assunção RM, Meira W (2019) Identifying high-risk areas for dengue infection using mobility patterns on twitter. Online Journal of Public Health Informatics 11(1) 4

860 Takács B, Hadjimichael Y (2019) High order discretization methods for spatial
861 dependent sir models. arXiv preprint arXiv:190901330 4
862 Tsai TC, Chan HH (2015) Nccu trace: Social-network-aware mobility trace.
863 IEEE Communications Magazine 53(10):144–149 5, 15
864 Van Doremalen N, Bushmaker T, Morris DH, Holbrook MG, Gamble A,
865 Williamson BN, Tamin A, Harcourt JL, Thornburg NJ, Gerber SI, et al
866 (2020) Aerosol and surface stability of sars-cov-2 as compared with sars-
867 cov-1. New England Journal of Medicine 382(16):1564–1567 3
868 Vesanto J, Alhoniemi E (2000) Clustering of the self-organizing map. IEEE
869 Transactions on neural networks 11(3):586–600 13
870 Wang P, Fu Y, Liu G, Hu W, Aggarwal C (2017) Human mobility synchro-
871 nization and trip purpose detection with mixture of hawkes processes. In:
872 Proceedings of the 23rd ACM SIGKDD international conference on knowl-
873 edge discovery and data mining, pp 495–503 2
874 WHO (2021) WHO Coronavirus Disease (COVID-19) Dashboard — WHO
875 Coronavirus Disease (COVID-19) Dashboard. URL https://covid19.who.
876 int/ 2
877 Zheng Y (2015) Trajectory data mining: an overview. ACM Transactions on
878 Intelligent Systems and Technology (TIST) 6(3):1–41 3
879 Zheng Y, Fu H, Xie X, Ma W, Li Q (2011) Geolife gps trajectory dataset-user
880 guide. microsoft research 25
881 Zhou C, Yuan W, Wang J, Xu H, Jiang Y, Wang X, Wen QH, Zhang P (2020)
882 Detecting suspected epidemic cases using trajectory big data. arXiv preprint
883 arXiv:200400908 2

5 Appendix: NCCU data

This section provides details related to the study area and the selected dataset. Figure 18 depicts coordinates of the study area, whereas Figure 19 illustrates the complete dataset of all 115 individuals for the period of 15 days where each user is shown with a different colour.

Fig. 18 Coordinates of the study area

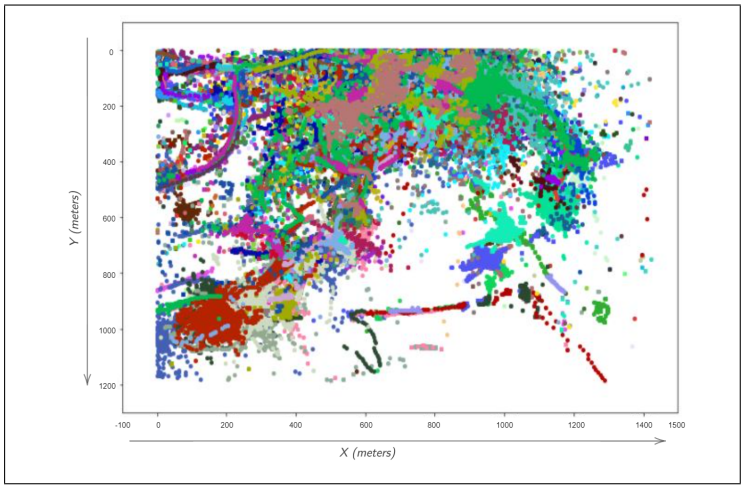

Fig. 19 Extent of recorded dataset

Figure 20 illustrates mobility trajectories of a single user for a period of one day, and Figure 21 shows 1-day movement for 5 users.

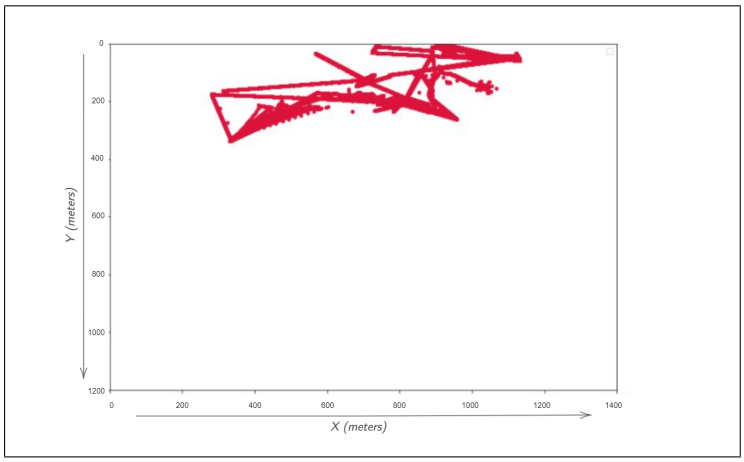

Fig. 20 Mobility trajectory of a single user for a 1-day period

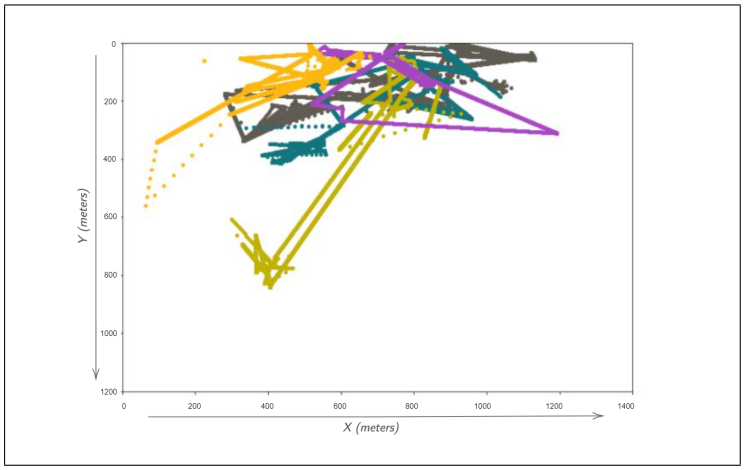

Fig. 21 Mobility trajectory of five users for a 1-day period

Figure 22 shows self-generated mobility trajectories of all 50 users for a period of 15 days.

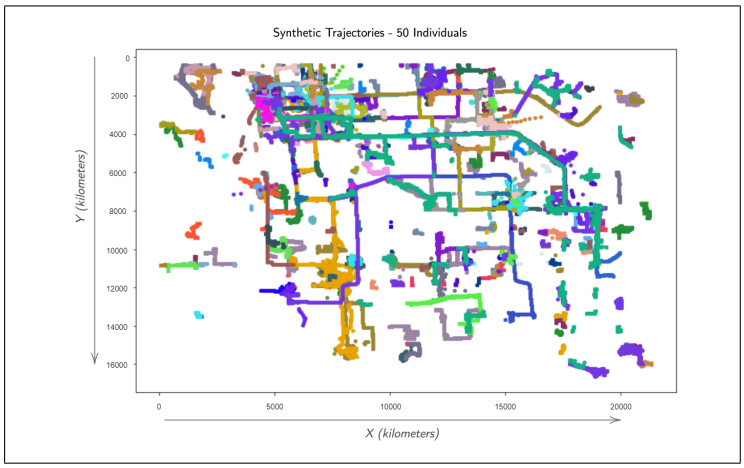

Fig. 22 Geolife data based self simulated mobility trajectories of 50 users for complete data period of 15 days