Contextual Contact Tracing based Spatio Enhanced Compartment Modelling & Spatial Risk Assessment

Dissertation submitted in partial fulfillment of the requirements for the Degree of Master of Science in Geospatial Technologies

February 21, 2021

MUHAMMAD MATEEN MAHMOOD

Supervised by:

Prof. Dr. JORGE MATEU
Department of Mathematics, Universitat Jaume I

Co-Supervised by:

Jun. Prof. Dr. JUDITH VERSTEGEN
Institute for Geoinformatics, University of Münster

&

Dr. ANA CRISTINA COSTA
NOVA Information Management School, Universidade Nova de Lisboa

ifgi
Institute for Geoinformatics
University of Münster

UNIVERSITAT JAUME I

NOVA IMS
Information Management School

Declaration of Academic Integrity

I Muhammad Mateen Mahmood hereby confirm that this thesis on ***Contextual Contact Tracing based Spatio Enhanced Compartment Modelling & Spatial Risk Assessment*** is solely my own work and that I have used no sources or aids other than the ones stated. All passages in my thesis for which other sources, including electronic media, have been used, be it direct quotes or content references, have been acknowledged as such and the sources cited.

February 21, 2021 ______________________________

I agree to have my thesis checked in order to rule out potential similarities with other works and to have my thesis stored in a database for this purpose

February 21, 2021 ______________________________

Acknowledgements

First of all, I would like to pay my highest gratitude to ***Prof. Jorge Mateu*** who has been an inspiration throughout my thesis journey. His encouragement, motivation and guidance not only helped me in the completion of this thesis but also elevated my enthusiasm for spatial statistics. I am eternally grateful to him for this experience as I find myself privileged to have been given a chance to work under the supervision of Prof. Jorge and to have an association with him through this thesis.

In addition, I am equally thankful to ***Jun.Prof. Judith Verstegen*** and ***Dr. Ana Cristina Costa*** for their valuable remarks. In hindsight, their confidence in me has helped in more ways than I can count; from believing in myself to truly achieving what I had intended to accomplish. I could not have imagined a better advisory team than them.

I would like to thank for the guidance services of ***Prof. Enrique Hernández*** *(Universidad Politécnica de Valencia, Spain)*, for his support regarding the understanding of his recent work which ended up being the motivation of this study. I would like to thank all members of the *Department of Mathematics*, especially ***Somnath Chaudhuri***, ***David Payares*** & ***Poshan Niraula*** for their time, support and motivation throughout the thesis process.

Among the class fellows, I would like to mention specially the names of ***Ganesh Prasad***, ***Hamid Ali***, ***Saadoon Hammad*** and ***Anu Shrestha*** for sharing their experiences and giving me their precious time. I thank all my fellow mates of *M.Sc. Geospatial Technologies 2019-2021* family, who helped me in whatever form was possible for them.

My acknowledgement would not be complete without paying my highest gratitude to my family back home in *Pakistan and Canada*, for their ever-enduring support. It is all due to their prayers, especially my mother's, that I have come this far in life. They are the real strength behind every success that came my way throughout my pursuits.

Last but not least, I would like to pay my respect, love and gratitude to my wife, ***Urwa Ali***, who patiently stood by me during the course of my study in Europe and has been a constant motivation in these terrible times of COVID-19. Thank you for your appreciated sacrifices, for your patience and tolerance, and for keeping me sane over the past few months. Thank you for being my advisor and proofreader. I owe you everything.

M. Mateen. M

Abstract

The current situation of COVID-19 appears as a paradigm shift that seems to have far-reaching impacts on the way humans will now continue with their daily routine. The overall scenario highlights the paramount importance of infectious disease surveillance, which necessitates immediate monitoring for effective preparedness and efficient 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 thesis research investigates both requirements in a hybrid approach by the implementation of disease outbreak modelling 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 study implements human mobility based contact tracing in the form of an event-based stochastic SIR model as a baseline and further modifies the existing setup to be inclusive of 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 suggest that the Spatio-SIR model tends to perform more meaningful events concerned with the Susceptible population rather than events to the Infected or Quarantined. With an example of a real-world scenario of induced spatial high-risk, it is highlighted that the new Spatio-SIR model can empower the analyst with a capability to explore disease dynamics from an additional perspective. The study concludes that even if this domain is hindered due to lack of data availability, the investigation process related to it should keep on exploring methods to effectively understand the disease dynamics.

Keywords:

Epidemiology, Contact Tracing, Trajectories, Compartment Modelling, Self Organizing Maps

Contents

Declaration of Academic Integrity **i**

Acknowledgements **ii**

Abstract **iii**

1 Introduction **1**

1.1 Context Development 1

1.2 Motivation & Problem Statement 3

1.3 Research Aims and Objectives 4

1.4 General Methodology 5

1.5 Contributions 6

1.6 Thesis Outline 6

2 Literature Review **7**

2.1 Infectious Disease Surveillance 7

2.1.1 Contact Tracing 8

2.1.2 Digital Contact Tracing 8

2.1.3 Digital Epidemiology 8

2.1.4 Human Mobility Trajectories 9

2.1.5 Contextual Contact Tracing 10

2.2 Mathematical Modelling of Infectious Diseases 10

2.2.1 Deterministic & Stochastic Models 11

2.2.2 Compartment & Agent Based Models 11

2.2.3 Hybrid Models 11

2.2.4 Individual Level Compartment Modelling *(ILCM)* 11

2.2.5 Point Processes & Compartment Models 12

2.2.6 Graph Networks 13

2.2.7 Spatio-Epidemic Modelling 13

2.3 Spatial Risk Assessment 14

2.3.1 Spatial Multi Criteria Analysis 14
2.3.2 Unsupervised Learning 15
2.3.3 Self Organizing Maps 15
2.4 Selection of Baseline 16

3 Theoretical Background **17**
3.1 Contact Network Graph 17
3.1.1 Degree of Network 18
3.1.2 Influence Set, Branching Tree & Superspreaders 18
3.1.3 Prior Contacts 19
3.1.4 Graph Matrix 20
3.2 SIR Modelling 20
3.2.1 Stochastic SIR Modelling 21
3.2.2 Event-Driven Modelling 22
3.2.3 Gillespie's First Reaction Method 22
3.3 Self Organizing Map 23

4 Methodology **25**
4.1 Implementation Framework 25
4.1.1 Methodological Baseline 25
4.1.2 Implementation Workflow 26
4.2 Implementation Process 27
4.2.1 Stage-1: SIR Modelling 27
1. Pre-Processing 28
2. Contact Tracing 29
3. SIR Model 30
4. Comparison 32
4.2.2 Stage-2: Spatial Risk 34
1. Risk Elements 34
2. Risk Assessment 36
4.2.3 Self Organizing Maps (SOM) 36
3. Spatial Risk 37
4.2.4 Stage-3: Hybrid Spatio-SIR Model 37

5 Data and Experimental Design **39**
NCCU Trace Data 39
Experimental Design 39

6 Results & Discussion 41

6.1 SIR Modelling 41

6.1.1 Baseline-SIR Model 41

6.1.2 Individual Level Latency 43

6.1.3 Varying Initial Configurations 44

6.1.4 Comparison with base-SIR 47

Contact Tracing 47

SIR Model 47

6.2 Spatial Risk 47

6.2.1 Risk Basis 48

6.2.2 Risk Grids 48

6.2.3 High Risk Areas 50

6.3 Hybrid Spatio-SIR Model 50

6.3.1 Dynamic Spatial Risk 51

6.3.2 Spatio-SIR Model 52

6.3.3 Real World Scenario 54

6.4 Limitations 54

6.5 Future Work 57

7 Conclusion 58

7.1 In Case of Data Availability 60

8 Appendix 61

8.1 Overview of Datasets 61

8.2 NCCU Trace Dataset 63

8.2.1 Study Area 63

8.2.2 Extent of Recorded Dataset 64

8.2.3 Initial Location of Individuals 65

8.2.4 Mobility Trajectories 66

References 68

List of Figures

1.1 Relation of temporally varying spatial risk affecting epidemic model and vice versa 4

1.2 Summarized workflow of overall methodology 5

3.1 Sample Trajectories and their corresponding Network form 17

3.2 Branching Process in the form of a Transmission Tree 19

3.3 Process of *(A)* Forward Tracing and *(B)* Backward Tracing 19

3.4 Graph Matrix corresponding to trajectories of Figure 3.1 20

3.5 Process flow of a typical SIR Model 21

3.6 Process flow of an Event Driven Stochastic SIR Model 22

3.7 Process Flow of Self Organizing Map 24

4.1 Work modules of Stage-1 *(Baseline SIR Modelling)* 26

4.2 Work modules of Stage-2 *(Spatial Risk Assessment)* 26

4.3 Work modules of Stage-3 *(Hybrid Spatio-SIR Model)* 27

4.4 Implementation workflow of overall methodology 28

4.5 Single day contacts in the form of (A) Network & (B) Matrix 30

4.6 Implementation Process of Stage-1 *(SIR Modelling)* 33

4.7 Computing Risk Basis from Trajectories & Contacts 35

4.8 Unsupervised classification workflow using SOM & K-Means 37

4.9 Assignment of ordered labels to un-ordered classes 38

4.10 Assignment of spatial risk scores to contacts 38

5.1 Simulating 150 days data from 15 days NCCU Trace data 40

6.1 Output of a disease outbreak scenario by Baseline-SIR model. *(Top)* presents trends related to count of *Susceptible, Infected & Recovered*, whereas *(Bottom)* illustrates count of individuals in *Quarantine* related compartments. Each of the 10 realisations 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 *Y-axis*. 42

6.2 Individual-level change in latency of *35* out of total *115* individuals is shown based on the SIR events as model simulates. 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. 43

6.3 Comparing the *Average* of 10 stochastic realisations 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 & 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. 45

6.4 Comparing the *Average* of 10 stochastic realisations with varying *Tracing Efficiency* $q^{'}$, to observe their effect on the overall disease outbreak (population size = *115*). Subplots *(A,B,C&D)* presents four cases of $q^{'}$=0, 0.25, 0.50 & 0.75, respectively, where 1 means 100% backward tracing. 46

6.5 Risk basis of *Infectious Trajectories* (Sub-Figure-A) & *Location of Contacts* (Sub-Figure-B) from one complete day during an epidemic. In the sample shown in (A), there are total of *Seven* infected individuals from that day. (Sub-Figure-B) shows locations of all contacts from the same day. . . 48

6.6 Risk Grids computed out of Risk Basis (discussed in Section 6.2.1). Subfigure (A) & (B) refers to grid based representation of *Duration* & *Count* of trajectories shown in Figure 6.5-(A). Sub-Figure-(C) depicts location of an infectious contact occurred on the same day, whereas Sub-Figure-(D) presents translation of Figure 6.5-(B) into grid form. All values are normalized to the range of *0 to 1*, where 1 refers to Highest Risk. 49

6.7 Combining risk from multiple grids shown in Figure 6.6 into a single grid output. 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] as discussed in section 4.2.4, where '1' refers to the previous normal (existing SIR model with a constant spatial risk and all contacts are of equal nature) 50

6.8 Temporally varying Spatial Risk for the first two weeks of epidemic is shown for a sample of six cells labelled as (*A, B, C, D, E* & *F*). Risk values vary between *0.5-1.5*, while there is a new value for every day. 51

6.9 Comparison of $\mathcal{K}_i$ values of all individuals between Baseline-SIR model and the new Spatio-SIR model. *(Dashed red lines)* shows previous discrete numbers of infectious contacts from baseline-SIR model, whereas *(Green dotted lines)* presents the enhanced continuous values from the spatio-SIR model. Grey boxes focuses on the notion that for some individuals like (i=20), the SIR model considered a higher count of infectious contacts ($\mathcal{K}_{20}$=1) whereas Spatio-SIR consider less ($\mathcal{K}'_{20}$=0.75) as the contacts were of lower risk. 52

6.10 Comparing *Average* of 10 stochastic realisations of a disease outbreak scenario from Baseline-SIR *(dashed)* & Spatio-SIR *(solid)*. *(Top)* presents trends related to count of *Susceptible, Infected & Recovered*, whereas *(Bottom)* illustrates count of individuals in *Quarantine* related compartments. Count of total population is *115* which are represented over *Y-axis*. 53

6.11 Comparing *Average* of 10 stochastic realisations 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 count of *Susceptible, Infected & Recovered*, whereas *(Bottom)* illustrates count of individuals in *Quarantine* related compartments. Count of total population is *115* which are represented over *Y-axis*. 55

8.1 Coordinates of Study Area 63

8.2 Extent of Recorded Dataset 64

8.3 (Above) - Initial Locations of Individuals in Dataset (Bottom) - NCCU (Tsai and Chan, 2015) provided map 65

8.4 Mobility trajectory of a single user for 1 day period 66

8.5 Mobility trajectory of five users for 1 day period 67

List of Tables

2.1 Latest technologies for Digital Contact Tracing 9
2.2 Use of Human Mobility data in Infectious Disease Dynamics 10
2.3 List of Variants of SIR Compartment Models 12

4.1 Summary of infectious disease parameters 32
4.2 Rate equations related to each SIR event 32
4.3 Estimated values related to COVID-19 33

5.1 Overview of experimental design for different scenarios 40

6.1 Comparison of Contact Tracing of SIR Model with base-SIR 47

List of Abbreviations

ABM	Agent based Model
AHP	Analytic Hierarchy Process
AIDS	Acquired Immunodeficiency Syndrome
API	Application Programming Interface
BME	Bayesian Maximum Entropy
BMU	Best Matching Init
CCTV	Closed Circuit Television
CM	Compartment Model
CRD	Call Record Data
DCT	Digital Contact Tracing
DE	Digital Epidemiology
GFRM	Gillispie First Reaction Method
GPS	Global Positioning System
HIV	Human Immunodeficiency Virus
ILM	Individual Level Model
MERS	Middle East Respiratory Syndrome
NBM	Network Based Model
NCCU	National Chengchi University (Taiwan)
NPI	Non Pharmaceutical Interventions
OWA	Ordered Weighted Average
PCA	Principal Component Analysis
SARS	Severe Acute Respiratory Syndrome
SIR	Susceptible Infected Recovered
SMCA	Spatial Multi Criteria Analysis
SOFM	Self Organizing Feature Map
SSS	Spatial Scan Statistics
UL	Unsupervised Learning
WHO	World Health Organization

“Analyse Problems, not Data”

*Peter Diggle **

***'Analyse Problems, not Data: One World, One Health'** - *Journal of Spatial Statistics (2018)*

Dedicated to COVID-19,
which changed our lives forever

1 Introduction

1.1 Context Development

The emergence and re-emergence of infectious diseases and their explosive dissemination have been a major concern for mankind. Humanity has experienced devastating epidemics like the Zika virus, MERS, SARS, and Spanish Flu, but the global burden of the SARS-CoV2 (COVID-19) pandemic highlighted a universal threat to public health departments. According to the World Health Organization[1] (WHO, 2021), as of 31st January 2021, the COVID-19 pandemic has over 101 million confirmed cases with above 2.1 million deaths worldwide. Apart from the health-related aspects, this pandemic scenario jeopardizes social functioning, economies and international relations (Zhou et al., 2020).

Detection and control of COVID-19 in particular and infectious diseases in general have irrupted as a major societal challenge (Oliver et al., 2020). Countries like Israel and South Korea, which took prompt actions towards testing and identifying previous contacts in case of an identified individual, were able to restrict the spread of the disease. Whereas the countries which did not proceed with the massive testing and manual contact tracing had to go for draconian measures of lockdown, quarantining and contact precautions (social-distancing, facemasks etc.) (Hernández-Orallo et al., 2020; Park et al., 2020).

To overcome an infectious disease, the goal is to detect all infected individuals among the population, which necessitates massive testing on a regional scale. This requirement has major limitations, such as the availability of medical resources and enforceability. Though authorities have followed ingenious medical methods (nasal swabs, X-rays etc.) to rapidly detect the infected individuals, considerable economical burden and implementation barriers still exist. In a situation like this, detection of an infectious disease requires non-pharmaceutical interventions (NPI), which sets the basis of the term *Digital Epidemiology* (Salathé, 2018). This refers to the use of data and methods from outside of the public health system, not generated for the epidemiological purpose at the first place.

[1] WHO Coronavirus Disease (COVID-19) Dashboard

One of those methods is called *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 infectious transmission is to keep an eye on the physical ***interaction*** (contacts) of individuals. Additionally, understanding these interactions is just as important as understanding the contagion process itself.

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

On the other hand, these interactions based on individuals movement are subject to tracking human mobility. As continuous tracking of a human being is not possible, their associated digital devices can be tracked, which are a considerable proxy of their movement (Oliver et al., 2020). With DCT based detection of infectious individuals, infectious trajectories can be tracked. Such tracking not only identifies future infectious contacts but also highlights the places these infectious trajectories have visited. Identification of such risk areas is critical for policymakers for decisions related to smart lockdown or areal curfew.

Despite that, an important consideration is that continuous recording of an individual movement for a considerably long duration is highly invasive in nature. The mere idea of tracking raises privacy concerns as these datasets record an individual's routine, behaviour and personal matters (Reichert et al., 2020). That is the reason why there is no individual-level trajectory dataset publicly available related to infectious diseases. In order to minimize this concern, the use of Bluetooth was proposed (Martinez-Martin et al., 2020), as being short-range in nature, it only collects the contact information as and when they happen. 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 trajectory form is justified given *(i)* it is implemented by government and *(ii)* with a guarantee of privacy protection.

In summary, digital methods are of great value to understand disease dynamics, especially in the early identification of prior contacts and predicting future propagation. They can guide policymakers by simulating scenarios of different control measures and can assist the public health system with data insights that are critical for their sustainability.

1.2 Motivation & Problem Statement

State of the art highlights the accepted practise of digital methods in infectious disease studies. Continuous research is on the monitoring of aggregated-level-mobility which is indisputably available in comparison to individual-level monitoring which is seldom in practice due to its invasive nature. However, an alarming scenario as of COVID-19 calls attention to working with individually-specific trajectory datasets, which adds to the motivation of this study.

The long-standing COVID-19 has amplified research in this domain with several studies involving individual-level-mobility for investigation of disease dynamics (Benreguia et al., 2020; Park et al., 2021; Souza et al., 2019b). 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 the spatial location (Mahsin et al., 2020). Even if the spatial location was considered, it was in the aggregated form of spatially varying demographic factor. Despite all the efforts, complete integration of a spatial component still seems to be missing, which can consider the effect of ’space’ (location) for each specific individual-level contact.

Similarly, the lack of individual-level-mobility data, especially high-frequency recordings of movements, restricts the analysis of the trajectory-form. Knowledge about *Infectious Trajectories* is critical to understand how an infection propagates in population and in space. A scenario that makes contact tracing, mobility tracking and temporally varying spatial risk an interconnected process. It is a recursive sequence as illustrated in Figure 1.1, where the probability of transmission of contact is proportional to the risk intensity (of its spatial location), which evolves based on infectious movement, which itself is an outcome of an infectious contact. Hence, there is a requirement of a hybrid spatio-epidemic model to thoroughly fuse the effect of space into a disease model while dealing with infectious trajectories.

In epidemic modelling, *Compartment Models* (Gallagher and Baltimore, 2017) are common which distribute each individual in the population-based on their disease states. Generally, they are of Susceptible, Infected and Recovered (SIR) type, however many versions like SEIR, SIS, SEIAHCRD (see Table 2.3) exists which depends on the type of disease and improvised methodologies. Though the temporal aspect is well addressed in these SIR models, as identified earlier, the ‘spatial context’ is yet to be fully explored, especially in individual-level studies, which serves as the basis of this thesis research.

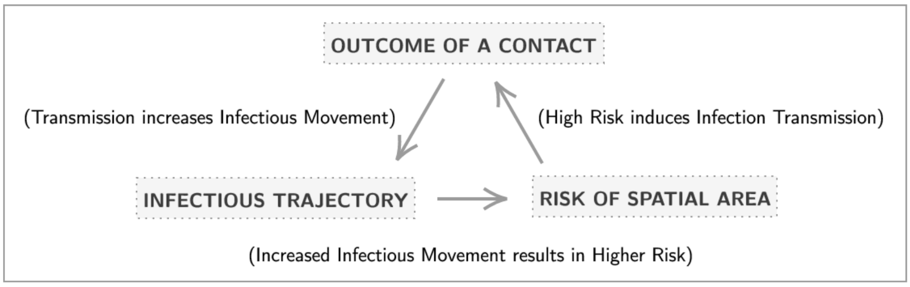

OUTCOME OF A CONTACT
(Transmission increases Infectious Movement)
(High Risk induces Infection Transmission)
INFECTIOUS TRAJECTORY
RISK OF SPATIAL AREA
(Increased Infectious Movement results in Higher Risk)

FIGURE 1.1: Relation of temporally varying spatial risk affecting epidemic model and vice versa

1.3 Research Aims and Objectives

This research focuses on the inclusion of spatial aspect (temporally varying spatial risk) of these physical interactions termed as *Contextual Contact Tracing*. The idea being that contacts taking place in contextually distinct geographical locations must be treated differently based on the vulnerability they pose to the susceptible individual.

The research intends to work with infectious trajectories in a twofold aim study. One is to explore the development of a hybrid spatio-epidemic model where spatial context is integrated into the epidemic model for each specific contact, and the second is to identify how temporally varying spatial risk evolves based on the onset of an infectious disease scenario. Following research questions are defined to formulate the problem:

- *How does the new spatio-epidemic model inclusive of spatial context implements a disease outbreak scenario in comparison to the existing non-spatial approach?*
- *What is the effect of explicitly induced spatial high risk on the epidemic propagation?*

To answer these research questions, the following mentioned objectives are defined:

- *Identify element in an epidemic model that can include the spatial context of contact tracing*
- *Monitor latency states of each individual to track infectious trajectories*
- *Identify the basis of spatial risk from only available information of trajectories*
- *Calculate a new risk scoring system for the spatial risk representation*
- *Evaluate the results of spatio-epidemic model to compare with a simple epidemic model*

1.4 General Methodology

In order to implement this two-aimed study, the structure of the methodology can be organized in three broad stages; *(i)* implementation of a baseline-SIR model; *(ii)* tracking infectious trajectories and contacts for *Spatial Risk Assessment* and *(iii)* enhancement of baseline-SIR setup for the development of a hybrid spatio-SIR model. A brief overview of this methodological organization is shown in Figure 1.2.

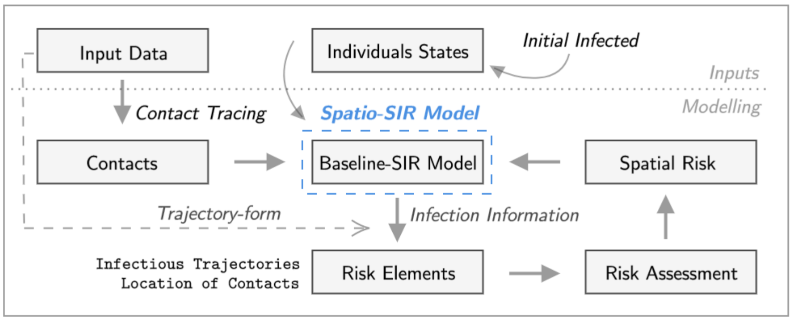


FIGURE 1.2: Summarized workflow of overall methodology

At first, an existing contact tracing based stochastic SIR Model motivated from (Hernández-Orallo et al., 2020) is implemented which offers additional compartments of *Quarantine Susceptible & Quarantine Infected* as well. For contact tracing, mobility trajectories are tracked based on two thresholds of distance and duration to identify daily contacts, where a *Network Graph* (West et al., 1996) based implementation is followed. Established SIR model is of an event-based approach following the principles of *Gillespie's Method* (Gillespie, 1977). Additionally, the SIR model is compared with Hernandez's work (Hernández-Orallo et al., 2020) to ensure baseline setup is well established.

Secondly, to compute *temporally varying spatial risk*; infectious trajectories, infectious contacts and social distancing violations are explored as elements of risk. Here, a regular lattice (grid) based structure is implemented for computational efficiency. Due to the lack of validation data related to the spatial risk of the study area, an unsupervised learning technique is investigated to recognize high-risk clustering based on data characteristics.

Finally, using SIR model as the baseline and computed spatial risk from the second stage, a complete spatio enhanced SIR model is implemented. The modification brings forth the spatial risk for every day based on the activities of the previous day. This procedure takes into account the spatial risk for each specific contact based on its spatial location, where the probability of next SIR event is affected by the overall spatial risk of the area.

1.5 Contributions

The research proposes a new spatio enhanced setup of SIR modelling, where contacts are associated with an intensity that is representative of its risk score based on its spatial location. This association of risk to a contact is executed by reforming the network graph, where enhancement is in a manner that a riskier contact has a higher possibility (probability) of disease transmission than the one which is of relatively lower risk. For temporally varying spatial risk, risk scores are reevaluated based on infectious activities of the recent past. As per the author's knowledge, this enhancement of the existing SIR model to be inclusive of spatial risk is a novel contribution to the literature.

As data about infection as well as spatial risk is not available due to its invasive nature, the feasibility of this idea is developed in the form of a framework that can serve as a *Spatio-Epidemic Tool* for such spatio enhanced epidemiological analysis. Even with these data limitations, implementation persists due to the absence of a methodology for the consideration of spatial risk in contact tracing and working with individual-level mobility trajectories, a knowledge gap exploited by the scenario of COVID-19. The implemented methodology of a spatio-SIR model 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 assessment.

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

1.6 Thesis Outline

The remainder of the dissertation is as follows: *Chapter 2* presents a comprehensive review of the literature. *Chapter 3* details the theoretical background of the concepts and methods chosen for the implementation. *Chapter 4* explains the methodological baseline and implementation process in detail. *Chapter 5* introduces the selected dataset along with Experimental Design. Results of the developed framework are presented, interpreted and discussed in *Chapter 6*, along with limitations and future works. *Chapter 7* ends this document with conclusions and a scenario in case desired data was available.

2 Literature Review

This chapter details a comprehensive review of the concepts supporting the thesis from existing state-of-the-art. Initially, infectious disease surveillance has been discussed from the lens of contact tracing, digital epidemiology and human mobility. Secondly, the mathematical modelling of infectious diseases has been presented, introducing work so far in the integration of *'spatial'* context in this domain. And finally, spatial risk assessment has been viewed in the scenario of unsupervised learning.

2.1 Infectious Disease Surveillance

In infectious diseases where the pathogen (the infectious agent) is another human being, and the infection is transmitted through person-to-person interaction (Yang et al., 2020); the key to tracking the infectious transmission is to keep an eye on the physical interaction of individuals. Tracking the known infected individuals and their interactions is already demanding, but the existence of asymptomatic individuals makes this monitoring even more challenging (Müller et al., 2020). These undocumented individuals are transmitting the infection to a larger set of individuals, who themselves are infecting community in an uncontrollable domino effect. Hernández (Hernández-Orallo et al., 2020) highlights that early detection of infected population including asymptomatic individuals followed by their isolation or treatment is the key to restrict pandemic growth.

Similarly, the contaminated area is another critical perspective. Ongoing research (Simmerman et al., 2010; Van Doremalen et al., 2020) highlights the aerosol and surface stability of infectious diseases. COVID-19, SARS and Influenza, all have indicated up to days surface transmission which highlights the critical need for spatial risk monitoring.

Both these aspects, tracking of individuals and assessment of space, sets the basis of infectious disease surveillance. Such a vigil mechanism can assist policy-makers with the provision of both, high-risk areas as well as individuals to be quarantined. In order to implement such a surveillance system, *Contact Tracing* is highly beneficial especially when the disease is in its early stages of an outbreak (Hernández-Orallo et al., 2020).

2.1.1 Contact Tracing

The primary focus of surveillance is to reduce the next generation of cases, which implies that for an infected individual we cease future contacts and backtrack the latest ones. This tracking process of identification is called Contact Tracing (Eames and Keeling, 2003).

Contact tracing is mostly common in sexually transmitted diseases like HIV/AIDS, where it is easy to remember previous contacts (Udeagu et al., 2013). For other infectious diseases, traditional methods are common which are based on an interview of the infected individual recalling their prior contacts generally dating back to only a few days before the onset of symptoms. Afterwards, the contacts are notified of their exposure to an infected individual and further tested. These manual methods are slow in nature, require a higher number of resources and most importantly dependent on a persons memory. The severity of COVID-19 highlights that such practises are no longer beneficial and there is a requirement for an efficient and reliable mode of contact tracing (Anglemyer et al., 2020).

2.1.2 Digital Contact Tracing

Contact Tracing is the tracing of individuals and identifying their contacts. But it is not practically possible to track an individual all the time as there is no active (physical tracking) or remote (CCTV cameras) method available for such continuous monitoring. However, nowadays, a digital device (mobile phone, smartwatch, wearable sensor, GPS etc.) is accompanying an individual most of the time and can serve as a suitable proxy of individuals movement. Tracing those devices instead of humans, and assessing their proximity distances as the basis for potential contacts is the concept of *Digital Contact Tracing* (Martinez-Martin et al., 2020). Hereby, contact tracing refer to digital contact tracing.

2.1.3 Digital Epidemiology

As explained by Salathe in (Salathé, 2018), apart from relying on digital data, Digital Epidemiology (DE) in principle is the same as Epidemiology. Both deal with the understanding of disease dynamics and then to use that knowledge to mitigate disease. However, the prime difference being that DE relies on data outside the public health system which was not generated for medical purpose in the first place. Few examples of such datasets include search engine analysis like Google Flu Trend (Lazer et al., 2014), web-based participatory surveillances like InfluenzaNet (Paolotti et al., 2014), social media profiling and mobile phones. In the application of contact tracing, DE opens new avenues for the use of digital technologies for tracing individuals movement and identifying their contacts. Table 2.1 describes examples of such latest technologies alongside their examples.

TABLE 2.1: Latest technologies for Digital Contact Tracing

Technology	Examples
Internet/Wireless	WiFi, Bluetooth, Social Media (Twitter, Facebook, Google etc.)
Telecommunication	Cellular/Mobile/Call Record Data (CRD)
Remote Sensing	Cameras, Image Processing, Artificial Intelligence
GPS Tracking	Digital Devices, Wearable Sensors (Smartwatch)

2.1.4 Human Mobility Trajectories

Contact Tracing is subject to knowledge of human mobility; which is of prime importance in individual-level research on infectious diseases (Brockmann et al., 2009), and understanding that mobility is as important as understanding the contagion process (Soriano-Paños et al., 2020). Human mobility is much more than just recordings of geographical locations, and when studied from spatio-temporal perspective provide a comprehensive understanding of the human interactions (Basole, 2004). Researchers have worked with individual-level human mobility with multiple sources as previously discussed in Table 2.1, whereas Table 2.2 entails domain of their work with such data sources.

Recent advancement in location-aware technologies and computing procedures have resulted in a massive influx of movement data. Such a continuous recording results in a trajectory-form, which is capable of representing the movement of an individual to a small scale up to less than of a meter (Zheng, 2015). This high-level detail makes these datasets an ideal candidate for high precision tasks like contact tracing. Gonzalez (Gonzalez et al., 2008) highlights that mobility trajectories generally have a spatial regularity, and modelling mobility pattern can identify highly visited locations for each individual. Examples of continuous data recordings are GPS trajectories and Mobile phone, whereas other sources like Twitter API and CCTV Camera offers sparsely recorded collection or aggregated data. Overview of available trajectories datasets is available in Appendix 8.1.

Trajectory, also known as Path, Track or Segment, in the context of human mobility is the evolution of the individual's spatial component over time. It results in a sequence of movements and stops in a chronological order where movements are to be analyzed for contact tracing purpose. Such a detailed collection of data comes at a cost of two major constraints, *(i) Privacy Concern* (previously discussed in section 1.1) and *(ii) Data Size*.

The large volume is a distinctive feature of a trajectory dataset, which mainly depends on the temporal resolution of sample recordings (Zhou et al., 2020), however with the availability of computing power and big data techniques, this restriction can be addressed.

TABLE 2.2: Use of Human Mobility data in Infectious Disease Dynamics

Data Source	**Application**	**References**
Twitter API	Spatial Risk Modelling	Souza et al., 2018
	Spatial Clustering	Souza et al., 2019a; Souza et al., 2019b
Mobile Data	Epidemic Modelling	Lima et al., 2015; Tizzoni et al., 2014
	Disease Mitigation Strategy	Rubrichi et al., 2017
CCTV Cameras	Spatial Risk Assessment	Rezaei and Azarmi, 2020
GPS Trajectories	Compartment Modelling	Hernández-Orallo et al., 2020
Wearable Sensor	Epidemic Modelling	Mastrandrea and Barrat, 2016
Google Mobility	Epidemic Modelling	Chiang et al., 2020

2.1.5 Contextual Contact Tracing

Siła-Nowicka et al. highlights in (Siła-Nowicka et al., 2016) that the unprecedented improvements in the quality and quantity of data collection regarding mobility are not complemented with methods to extract useful patterns out of them. Recent developments in the domain of trajectory data mining put forth the availability of methods for understanding patterns of the underlying trajectory, however, there is still a need to investigate movement data beyond only their own but in a contextual aspect (Purves et al., 2014). In contact tracing, this promotes the idea that a ’contact’ taking place in contextually distinct locations must be treated differently, which serves as a cornerstone of this research.

2.2 Mathematical Modelling of Infectious Diseases

Mathematical models have been a constant application in the epidemic analysis, where the main objective is the quantitative measurement of disease elements (population, host, pathogen, periods etc.). These models tend to mimic reality in order to predict future behaviour (Hau and Kranz, 1990). Earlier mathematical models focused more on the epidemiological parameters such as transmission rates and infection period and quantified results in a simple form of overall pandemic size (Hethcote, 2009). Nowadays, with increasing computational power, high-resolution data and enhanced awareness about the heterogeneous nature of infectious disease dynamics; mathematical models are able to capture the complexity of disease phenomenon (Enright and Kao, 2018).

2.2.1 Deterministic & Stochastic Models

Epidemic mathematical models are of two types. ***Deterministic*** and ***Stochastic***. Deterministic models give a constant result for the same initial parameters whereas Stochastic possess inherent randomness and results in a different output to even similar initial conditions. Stochastic models mimic the probabilistic nature of disease phenomenon and are better suited for real-world applications whereas Deterministic models are well suited when dealing with a larger size of population (Hernández-Orallo et al., 2020).

2.2.2 Compartment & Agent Based Models

In terms of the implementation framework, epidemic models are of two kinds, equation-based *Compartment Model (CM)* and simulation oriented *Agent based Model (ABM)*. ABM is also called as *Individual Level Model (ILM)*. CM deal with aggregated data in population-level dynamics, implementing overtime transition of entities between discrete compartments. A typical example is a Susceptible-Infectious (SI) framework where the population shift from *'S'* to *'I'* as they are infected. On a finer scale, ABM simulates individuals (agents) and their interaction. Here, the infectious event is subject to the activities of the agent and are otherwise non-random in a simulated environment (Gallagher and Baltimore, 2017). ABM are more realistic in their application however comes with a drawback of high input data requirements and being computationally expensive.

2.2.3 Hybrid Models

Many hybrid models exist to seek advantages of both, simplicity of CM and thorough control of ABM. Bobashev et al. (Bobashev et al., 2007) implemented a hybrid model where after a certain number of infected individual, ABM shifted to be a CM. The idea being that after a certain size of the infected population, CM is stable enough to model the problem. Similarly, Banos et al. (Banos et al., 2015) presented an inter-city model with agents travelling between them. These hybrid implementations highlight the role of individual-level compartment modelling in infectious disease epidemiology.

2.2.4 Individual Level Compartment Modelling *(ILCM)*

Nowadays, many improvised methodologies are available derived from the original SIR Compartment model presented in (Kermack and McKendrick, 1927). In that model, Kermack and McKendrick considered three compartments of **S**, **I** and **R** where S referred to Susceptible, I meant Infected and R denoted those that have Recovered forever. Transitions between the compartment were based on a constant rate of changes and outside

TABLE 2.3: List of Variants of SIR Compartment Models

Variant	Description of Addition	References
SIS	**S**usceptible	El-Doma, 1999; Nåsell, 1996
SEIR	**E**xposed	Gu et al., 2020; He et al., 2020
mSEIR	metapopulation SEIR	Chen et al., 2020
SEIHR	**H**ospitalized	Ferrante et al., 2016; Niu et al., 2020
SEIHRD	**D**eceased	Leitao and Vázquez, 2020
SEIAR	**A**symptomatic	Pribylova and Hajnova, 2020
SAYRD	s**Y**mptomatic	Bisin and Moro, 2020
SEIAHCRD	**C**ritically Hospitalized	Bardina et al., 2020

factors like births and deaths were ignored. In this dissertation, ILCM is also referred to as *SIR Models* whereas their theoretical background is explained in Section 3.2.

Since the advent of the SIR concept back in 1927, many extended models have been proposed. An extension in the form of 'SEIR' is available (Yamana et al., 2020), where compartment **E** refers to *Exposed* individuals who have contracted the virus but are not infectious yet. Similarly, the majority of the changes were in a number of compartments based on disease specifications and implementation methodology. Overview of such variations is available in Table 2.3. Apart from compartment wise changes, other changes include transformation into partial differential equations, implementation variant like discrete and continuous-time models, integration of Bayesian inference, mobility networks and machine learning techniques (Bardina et al., 2020; Kresin et al., 2020).

2.2.5 Point Processes & Compartment Models

Point Processes (González et al., 2016), especially *Hawkes Process* (Chiang et al., 2020) have also gained recognition in epidemiological modelling. Hawkes integrated CM like Hawkes-SIR (Rizoiu et al., 2018) is also in use. Hawkes is typically common in clustering patterns or spatio-temporal point patterns, with the involvement of multiple spatial covariates. In comparison, Hawkes allows for non-parametric estimation whereas CM require estimation using the expert opinion of epidemiologists. Though CM put forward a more plausible implementation in form of a natural mathematical representation, they are also vulnerable to parametric bias (Kresin et al., 2020).

2.2.6 Graph Networks

In infectious diseases of a directly transmissible nature, graph networks are of fundamental importance. Network-based models (NBM) based on graph theory (West et al., 1996) allow structuring the physical interaction of population in a more realistic scenario, as in a real-world, an individual has a finite set of contacts instead of population-wide mixing (Keeling and Eames, 2005). ILCM combined with NBM are an ideal choice for epidemic modelling when sufficient information regarding contact patterns is available (Renardy and Kirschner, 2020). Detailed description on NBM is explained in Section 3.1.

2.2.7 Spatio-Epidemic Modelling

Spatial aspects have long been of interest in public health surveillance. However, due to the non-availability of individual-level data, work mostly focused on aggregated data on a population-level (Dlamini et al., 2020; Gomes et al., 2020; Pourghasemi et al., 2020). Modelling spatial heterogeneity at the level of each individual is often found *Veterinary Epidemiology* (Barlow, 1991; Parham and Ferguson, 2006). In human-related studies, individual-level work is either executed on a sparse scale or is only a proposal of a methodology. One such implementation has been proposed in (Benreguia et al., 2020) related to tracking of individual infectious trajectories and serves as a basis of thesis.

Space has been a consistent consideration as an illustration tool to visually understand the distributions, risk maps and spatial clustering. However, in SIR modelling, space as an input to a disease model is ***only*** found in aggregate-level modelling.

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* $\mathcal{R}_0$ is implemented in (Martinez-Beneito et al., 2020), while Lang et al. (Lang et al., 2018) discusses a theoretical framework of an SIR model on spatial networks where *Probability of Transmission* β is based on spatial distances along the edges. A Bayesian Maximum Entropy (BME) based extension of the SIR model is also available for metapopulation level stochastic modelling of infectious diseases (Angulo et al., 2013). All these models propose aggregate-level modelling, however, *Space* as an input in an individual-level SIR model is still missing, a knowledge gap highlighted in this research.

2.3 Spatial Risk Assessment

As presented in the Introduction, the aim of this thesis research is twofold; disease modelling with spatial context as input and secondly, how spatial risk evolves based on the output of the disease model. This section discusses the latter in terms of risk assessment.

Risk, a scenario of vulnerability to danger (Merriam-Webster, 2021), is a common topic of interest in disaster and crime-related studies. In epidemiology, *Risk* is the possibility of disease exposure, which varies in space and time. In this context, *Spatial Risk Assessment* is the process of observing disease-related covariates in space-time dimensions in order to identify the risk of each unit in space (Pfeiffer et al., 2008). A geostatistical process (Matheron, 1963) will result in a continuous risk surface, whereas a lattice-based (Saveliev et al., 2007) approach will result in an aggregated form of spatially arranged units.

Several techniques are useful for executing such an assessment depending on the structure of data and the aim of the study. *Spatial* or *Spatio-Temporal Point Processes* are one of them which have been previously discussed in Section 2.2.4. Similar methods include one-dimensional point processes as *Scan Statistics* or their extension in multi-dimensions as *Spatial Scan Statistics (SSS)* (Kulldorff, 1997). Souza et al. (Souza et al., 2019a; Souza et al., 2019b) conducted Spatial Risk Assessment using SSS, where they proposed two new spatial scan methods in the form of conditional and unconditional logistical models and detected spatial clusters using sparsely sampled Twitter feed data. Another spatial clustering application on aggregated data is available in (Desjardins et al., 2020) where a county-wide space-time clustering is executed.

As previously highlighted in Section 2.2.7, the spatial aspect is scarce in individual-level SIR studies. For the same, spatial risk assessment based on individual-level infectious trajectories is also missing which serves as the secondary basis for this bifold aimed study.

2.3.1 Spatial Multi Criteria Analysis

Techniques like Scan Statistics and Spatial Clustering are useful for *Point Variance* [1] inputs like human mobility based on the Twitter feed, or residence address of an infected individual. These techniques are designed to capture and highlight the underlying trend of data points. In a scenario of *Geostatistical* or a *Lattice* [2] inputs, such as mobility trajectories or population density, there is a need for a mechanism which can combine the contribution of multiple covariates as an overall representation. One such process is *Multi-Criteria*

[1] Point Variance: Finite collection of locations in a particular space, region or a window

[2] Lattice: Aggregated data in form of a finite collection of spatial regions

Analysis (MCA) (Nijkamp and van Delft, 1977) which can be extended into the spatial domain in form a *Spatial Multi Criteria Analysis (SCMA)* (Chakhar and Mousseau, 2008).

Implementation of MCA or SCMA relies on *Ordered Weights*, which defines the relative significance of a criterion over others. Methods like Analytical Hierarchical Process (AHP) (Saaty, 2014) and Ordered Weighted Averaging (OWA) (Yager and Kacprzyk, 2012) are based on such weighted average concepts, However, this requires knowledge for the allocation of weights; or in case of model fitting, requires *True Value* (validation data) so weights of *Predicted Value* can be adjusted accordingly.

In a scenario where such *priori* is not available for allocation of weights, an unsupervised learning method is required which can form a representation of multi-dimensional input data based on the characteristics of data (Hinton et al., 1999).

2.3.2 Unsupervised Learning

Unsupervised Learning (UL), also known as Self-Organization refers to a Machine Learning technique that highlights the undetected patterns of input data layers especially its statistical structure. UL does not rely on pre-existing labels or explicit target outputs, as in *Supervised Learning* which involves human supervision (Hinton et al., 1999).

Other than Principal Component Analysis (PCA), another major application of unsupervised learning is Cluster Analysis. As unsupervised methods only require input patterns, it is highly useful in scenarios where the grouping of objects or segmentation can highlight relationships (Asan and Ercan, 2012). This can assist in tasks like Spatial Risk Assessment, to explore the covariates that affect space to be of a higher or lower risk for disease transmission, and additionally, how these covariates amalgam as a whole.

2.3.3 Self Organizing Maps

One such unsupervised learning-based clustering technique is Self Organizing Maps (SOM) which is also known as Self Organizing Feature Map (SOFM). SOM was initially presented by Teuvo Kohonen in (Kohonen, 1982). SOM is a dimensionality reduction technique to convert high-dimensional data into 1,2 or 3 dimensions. Due to its topology-preserving nature, SOM has been extensively used for the visualization (Ultsch, 1993) and clustering of geospatial data (Gopal, 2016; Henriques et al., 2012), as well as actively used in epidemiology (Basara and Yuan, 2008; Pearce et al., 2015; Zhang et al., 2009).

Similarly, SOM has been used as an unsupervised classification technique in multiple spatial domains like land cover classification (Gonçalves et al., 2011), maritime environment (Lobo, 2009), hydrologic modelling (Hsu et al., 2002), solar wind classification (Amaya et al., 2020) and water resource modelling and analysis (Kalteh et al., 2008).

Considering the dimensionality reducing capability, SOM is similar to the statistical equivalent of PCA or multidimensional scaling (Krasznai et al., 2016). Baccao et al. (Bação et al., 2005) suggest SOM as a possible substitute for K-Mean clustering in case the neighbourhood is ignored. Considering that, SOM offers the following advantages in comparison to statistical techniques, due to its non-parametric nature. *(i)* SOM works independent of variable's distributions, *(ii)* SOM is computationally efficient to non-linear problems and *(iii)* SOM caters for noise or missing data more effectively (Asan and Ercan, 2012).

Extended SOMs in the form of *Hierarchical (H-SOM)* (Henriques et al., 2012), *Growing (G-SOM)* (Villmann and Bauer, 1998) and *Growing Hierarchical (GH-SOM)* (Pampalk et al., 2004) are also available with additional benefits. H-SOM considers the output of SOM as an input for hierarchical clustering algorithms. G-SOM innovates with an iterative application of SOM where the size of SOM increases with every iteration, this is especially useful when output map size is not known.

The result of SOM is a network of neurons that are the representatives of input features. SOM Network can be visualized in the form of U-Matrix (Westerlund, 2005) or an advance version as U*Matrix (Ultsch, 2003).

2.4 Selection of Baseline

In a nutshell, the pre-existing techniques of contact tracing focus exclusively on the spatial separation of individuals, when, in fact, the spatial location of contact is equally vital to understanding the overall disease dynamics. While extensive research on the monitoring of aggregated-level-mobility is indisputably available, individual-level monitoring is still missing due to its invasive nature. Besides, research on trajectory-based mobility has not yet been thoroughly explored. However, the alarming situation of the COVID-19 outbreak impels us to work with individually-specific trajectory datasets.

Individual-level trajectory-based mobility comes with computational complexity, however, network-based implementation can support their efficient handling. SIR modelling can be explored for a modification inclusive of a spatial component. Besides, the space part is to be explored in an unsupervised manner due to a lack of validation data.

3 Theoretical Background

This chapter serves as the theoretical background explaining the basics of concepts and methods used in this thesis. The first section elaborates on Contact Network Graph and its associated terminologies, as well as the construction of a Graph Matrix. The second section presents SIR modelling from deterministic and stochastic perspectives. In the end, the theory of Self Organizing Maps is explained in detail.

3.1 Contact Network Graph

Contact Tracing necessitates accurate information about the possible transmissible pathways for each individual in the population (Eames and Keeling, 2003). A graph network in the form of nodes and edges is an intuitive and computationally efficient representation of such an interaction where in individual-level studies, *Nodes* refers to *Individuals* and *Edges* represent 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 modelling, it is common to have temporal frequency of a 'day', hence $\varepsilon_{ij}(t)$ will exist between individual i and j if there exists a contact between the two on day 't' (Hernández-Orallo et al., 2020). Figure 3.1 shows sample trajectories and their corresponding network form with 'contacts' represented in the form of edges.

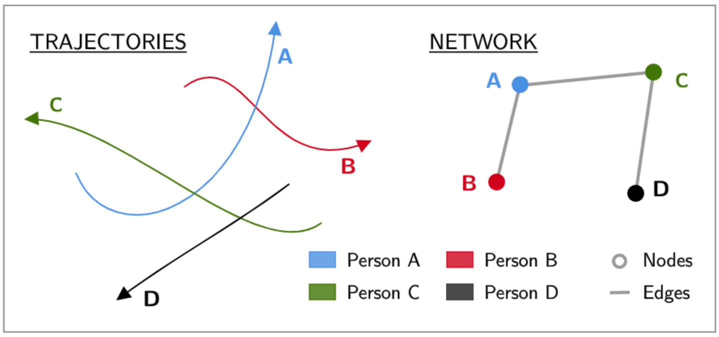


FIGURE 3.1: Sample Trajectories and their corresponding Network form

In terms of structural configuration, Network Graph is of two types. *Static Networks* and *Dynamic Networks*. Static Networks represent contacts that are constant and permanent and are mostly applicable in numerical experiments whereas Dynamic Networks capture the heterogeneity of a real-world scenario with a dynamicity of continuously changing contact structure. In contact network graph, directions of edges as in directed or undirected graph is ignored as contact is independent of direction. This highlights the assumption that infection can be transmitted in both directions depending on the disease state of individual and not the structure of the network (Hernández-Orallo et al., 2020). For a pair (i, j) of individuals, this symmetry can be viewed as $(G_{ij}(t) = G_{ji}(t))$.

3.1.1 Degree of Network

In contact networks, *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 individual i with other individuals in the network graph $G(t)$ on day $'t'$. From this, an *Average Degree* κ for a time period T can be computed as (3.1)

$$\kappa = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\int_{0}^{T} K_i(t)dt\right) \tag{3.1}$$

As explained in discussion of Figure 3.5, the rate of infection is influenced by the count of infected, hence it is useful to have a degree only involving contacts of infected individuals. Such a degree of diffusion can be represented as (3.2), where $I_j(t)$ denotes $\varepsilon_{ij}(t)$ in case a contact exists between pair (i, j) on day t.

$$\mathcal{K}_i(t) = \sum_{j=1}^{N} G_{ij}(t) \cdot I_j(t) \tag{3.2}$$

3.1.2 Influence Set, Branching Tree & Superspreaders

Based on the degree of contacts of each individual, their *Reachable Set*, also called as *Influence Set* can be extracted, which are a list of nodes that can get infected in case this node contracts the infection. Hence, ν_j is reachable from ν_i if there exists ε_{ij} during T (Enright and Kao, 2018). In dynamic networks, influence sets are continuously changing based on daily movements as contact tracing based edges $\varepsilon_{ij}(t)$ are evolving over t.

Transmission of infection within the Influence Set of individuals and the overall transmission chain can be represented in the form of *Branching Process*. Most common formulation

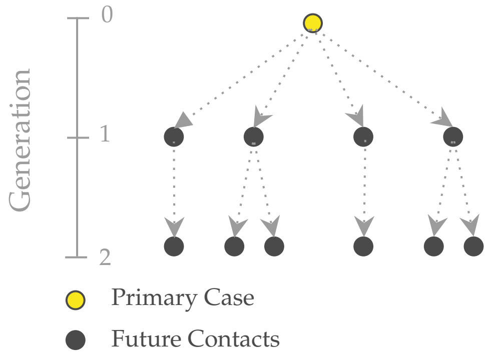


FIGURE 3.2: Branching Process in the form of a Transmission Tree

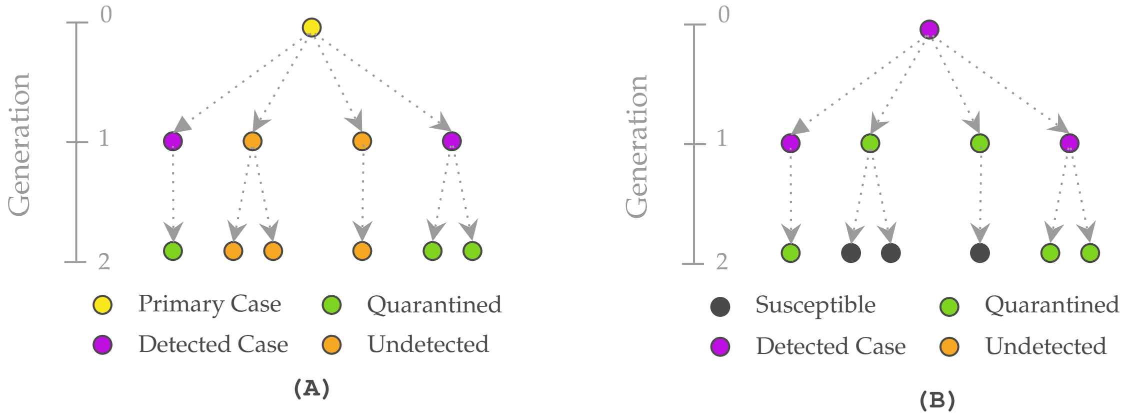


FIGURE 3.3: Process of *(A)* Forward Tracing and *(B)* Backward Tracing

of branching process is of Galton–Watson process which is widely used in epidemiology (Jacob, 2010). Such a chain of transmission or a *Transmission Tree* can assist in *Forward* and *Backward* tracing of infectious contacts. Backward tracing is highly beneficial to identify *Prior Contacts* who have been exposed to an infected person and are required to be tested or quarantined. These transmission trees are also helpful in identification of *Superspreaders*, individuals who are responsible for transmitting the disease to a large number of contacts (Bradshaw et al., 2020; Chang et al., 2020). Figure 3.2 illustrates a Branching Process in form of a Transmission Tree, whereas Figure 3.3 exemplify Forward and Backward Tracing for the same Tree.

3.1.3 Prior Contacts

Identifying prior contacts is the overall essence of contact tracing in order to restrict the next generation of cases. Endo et al. (Endo et al., 2020) highlights the benefits of Backward contact tracing in case of an overdispersed transmission. This requires a backward time window $\triangle$ depending on the type of disease (like infectious period, incubation time etc.), and can be used in the form of (3.3) to extract all prior contacts $\mathcal{C}_i(t, \triangle)$ of an individual i at t with window $\triangle$. Here, $D_j(t)$ is 1 if at time t person j is infected and traced.

$$\mathcal{C}_i(t, \triangle) = \sum_{j=1}^{N} \left(\max_{\tau \in [t-\Delta, t]} \mathbf{G}_{ij}(\tau) \right) D_j(t) \tag{3.3}$$

3.1.4 Graph Matrix

In graph theory, *Adjacency matrix* also known as *Connection Matrix* is commonly used to store graph information, where rows and columns represent *Nodes* (individuals) and the respective values either 0 or 1 depicts existence of an *Edge* (contact) between the two. In contact network graph, diagonal of adjacency matrix is always zero as it is a self-contact. Adjacency matrix of trajectories and network of Figure 3.1 is shown in Figure 3.4.

$$\begin{pmatrix} & A & B & C & D \\ A & 0 & 1 & 1 & 0 \\ B & 1 & 0 & 0 & 0 \\ C & 1 & 0 & 0 & 1 \\ D & 0 & 0 & 1 & 0 \end{pmatrix}$$

FIGURE 3.4: Graph Matrix corresponding to trajectories of Figure 3.1

3.2 SIR Modelling

As introduced in Section 2.2.4, SIR model distributes the population into discrete compartments of predefined disease states like ***S, I or R*** [1]. The core of the model is to answer the question that how individuals move from one compartment to another. In a closed environment where births, deaths and migration are ignored, transition is only in one direction as ***S*** → ***I*** and ***I*** → ***R***. The latter is simple as it can be considered a constant around a mean value based on clinical data of *infectious period*. The probability of an infected to be recovered relies on how long they have been infectious, which can be denoted as *Recovery Rate* γ, a constant value representing inverse of infectious period. The former is subject to disease transmission and is a function of three aspects; *(i)* the presence of infected individuals, *(ii)* contacts between susceptible & infected and *(iii)* the probability of transmission. Considering κ as the degree of $(S \leftrightarrow I)$ contacts and b representing probability of transmission of infection, the *Transmission Rate* β can be deduced as $\beta = \kappa \cdot b$.

[1] **S**: Susceptible, **I**: Infected, **R**: Recovered

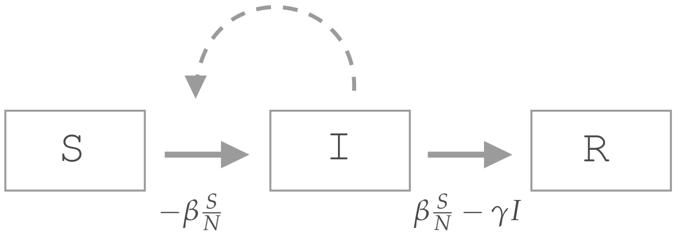


FIGURE 3.5: Process flow of a typical SIR Model

Hence, rate of changes can be represented in the form of non-linear ordinary differential equations (ODE) as in (3.4) (*for details, see Keeling and Rohani, 2011*)

$$\frac{dS}{dt} = -\beta I \frac{S}{N} \quad (a) \qquad \frac{dI}{dt} = \beta I \frac{S}{N} - \gamma I \quad (b) \qquad \frac{dR}{dt} = \gamma I \quad (c) \tag{3.4}$$

with initial conditions: S(t_0)>0, I(t_0)>0, R(t_0)=0 & S + I + R = N (N = population)

The ratio β/γ is called as *Basic Reproductive Ratio* $\mathcal{R}_0$. It represents the expected count of cases directly affected by a single case and is considered as the representative parameters of a disease in epidemiology.

A simplified visual overview of SIR modelling is available as Figure 3.5 where dotted line highlights that the quantum of infectious individual influences the rate of infection. A detailed implementation of SIR on Contact Networks is explained in Section 4.2.1.

3.2.1 Stochastic SIR Modelling

The above-explained SIR model remains deterministic as for given values of β and γ, disease dynamics are constant. As detailed in Section 2.2, Stochastic Models are inclusive of the probabilistic element related to disease transmission, and to include that randomness, probability distributions are to be used to model the transfer between compartments.

Deriving *Per Susceptible Rate* from (3.4-a) as $\beta \cdot I/N$, based on which the probability of a susceptible moving to infection compartment can be represented as (3.5).

$$P(S \to I) = 1 - e^{-\beta\frac{I}{N}} \tag{3.5}$$

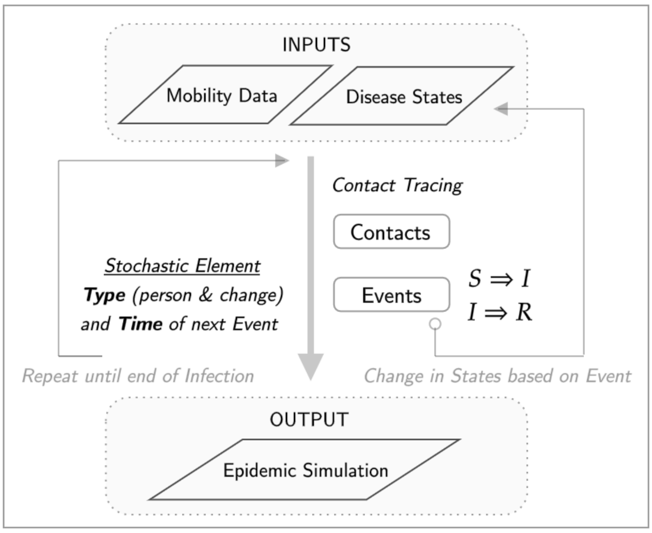


FIGURE 3.6: Process flow of an Event Driven Stochastic SIR Model

3.2.2 Event-Driven Modelling

In SIR models, the stochastic element can be incorporated into the demographics of each compartment and can be referred to as *Demographic Stochasticity*, which refers to the variations in the demographic process that are affected by a random event at the level of each individual. 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. Figure 3.6 presents the workflow of an event-driven stochastic SIR model.

This highlights that even if the probability of an event is constant, an individual may experience a varied fate based on the chance element. There are several methods to implement such an event-driven approach, one of them is *Gillespie's Method* (Gillespie, 1977) which is common in SIR modelling (Keeling and Rohani, 2011).

3.2.3 Gillespie's First Reaction Method

Gillespie's algorithm initially intended for the study of chemical reactions is also applicable in scenarios like SIR modelling where an outcome of the contact is like a biochemical process of a cell with fluctuating possibilities of events. It is a variant of 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 in which the simulation time scales with the size of the population, a problem addressed in GFRM method. The methodology of GFRM is explained in Algorithm (1).

Algorithm 1 Pseudocode of Gillespie's First Reaction Method

```
Label all events: ($E_1, E_2, ..., E_n$)
Determine respective Rates ($R_1, R_2, ..., R_n$) of all events
Calculate the occurrence (next) time of each event $(m) \Rightarrow \delta t_m = \frac{-1}{R_m} \log(RAND_m)$
Event with smallest $(\delta t)$ will be the next event $(p)$
Update Time $t \rightarrow t + \delta t_p$, and perform $(p)$
Repeat from step 2
```

3.3 Self Organizing Map

As introduced in Section 2.3.3, SOM is basically a dimensionality reduction method by transforming multi-dimensional input layers into discrete low dimensional (mostly two-dimensional) space, by preserving the topological relationships of features and not their actual distances (Asan and Ercan, 2012). From a neural network perspective, it is of *feed-forward* nature which allows information flow in one direction, that is from the input features to the output SOM neurons. Here, every input feature is connected to each output node for the computations of weights. Hence, it is a completely connected network (Larose and Larose, 2014), as illustrated in Figure 3.7. SOM can be defined based on three characteristics of (1) Competition, (2) Adaption and (3) Cooperation, which are:

- ***Competition*** refers to the nature of neurons to compete for the representation of an input sample. Winner (known as *Best Matching Unit - BMU*) is decided through a discriminant function computing distances between input samples and weight vectors of each node and selected based on similarity (least distant). Generally, the distance measure is of Euclidean type, but others are also used in SOM applications.

- ***Adaption*** reflects the learning phenomenon of a neural network where BMU is adjusted in favour of the input sample. Such a learning process impels output nodes to become similar to the input sample, finally capturing the complete information.

- ***Cooperation*** reflects the topography preserving property of SOM and highlights that the nearby locations in output SOM neurons represent similar input properties. This is achieved through a neighbourhood function, where neighbour neurons of BMU are also adjusted to learn from the same input.

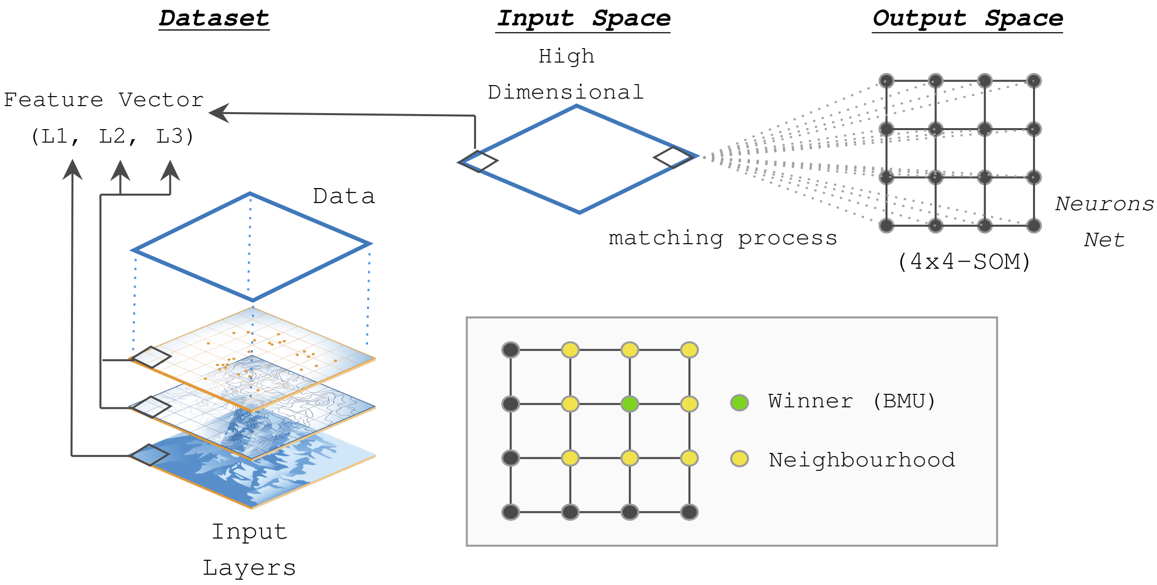


FIGURE 3.7: Process Flow of Self Organizing Map

Implementation of SOM is a two-step process. First is the *Training Phase* which involves weights adjustments to match input dataset, and second is the *Testing Phase* in which weights are fixed to test the performance of SOM.

In Training Phase, formulation of SOM is achieved as explained in Algorithm (2). (*for details, see Asan and Ercan, 2012*) .

Algorithm 2 Pseudocode of SOM Training Process

```
1: Assign random weights to all neurons
   (elements of each weight vector is equal in count to number of elements in each input vector)
2: For input vector $x_i$, compute distance from weight vector $W_k \Rightarrow d_k = \sum(x_i - W_{ik})^2$
   (repeat step 2 for all input vectors and compute distance of each input with each neuron)
3: Compute a winner $\Rightarrow d_{min} = min(d_0, d_1)$
4: Update weights for winner; $W_{ik}(t+1) = W_{ik}(t) + \alpha(x_i - W_{ik}(t))$ [$\alpha$ = learning rate]
   (change in weight $\Delta W_{ij}$ is proportional to $(x_{ij} - W_{ij})$ and $\alpha$)
5: Repeat from step 2
```

In Testing Phase, weights obtained from Training phase are fixed for each neuron. These weights are now used to test each input vector which is expected to have least distance to its assigned node compared to others.

4 Methodology

This chapter focuses on the methodology adopted to answer the research questions. First, an implementation framework is discussed to elaborate baseline model and overall workflow. This is followed by an explanation of implementation process including SIR model setup, computing risk out of SIR events [1] and introducing that risk back into the model.

4.1 Implementation Framework

4.1.1 Methodological Baseline

The methodology focuses on the establishment of a baseline setup by reproducing an available SIR model, and further exploring the baseline model for modification to include the spatial context. Spatial risk assessment is executed to supplement new spatio-SIR setup with *temporally varying spatial risk* for the future tracing of contacts.

Baseline SIR model for this study is motivated from (Hernández-Orallo et al., 2020) *(hereby referred as **base-SIR**)*, with following reasons as the basis of this selection:

- ***Additional Compartments:*** This model put forward a novel addition of Quarantine related compartment which are beneficial when dealing with individual-level contact tracing based compartment modelling. Additional compartments introduced in *base-SIR* are of *Quarantine Susceptible & Quarantine Infected*.
- ***Event based Stochasticity:*** As event-driven approach possesses inherent randomness, and stochasticity tends to mimic the probabilistic nature of disease phenomenon; both attributes makes this combination better suited for a real-world application.
- ***Computational Efficiency:*** This model implements an improvisation of Gillespie's First Reaction Method which supports in efficient handling of the computational complexity of a real-world contact tracing, especially on a trajectory-based dataset.

[1] Events of the susceptible person being Infected, or Recovery of Infected

4.1.2 Implementation Workflow

As introduced in Section 1.4, implementation of this two-aimed study is broadly categorized into three stages, which can be sub-divided into modules as follows. A comprehensive overview of the overall implementation workflow is available in Figure 4.4.

1. ***Baseline-SIR Modelling:*** This stage includes the implementation of the baseline model with all its components, as illustrated in Figure 4.1. First, trajectory dataset is processed for data interpolation. Then Contact Network and Graph Matrix are extracted based on *Contact Tracing*. Lastly, the stochastic SIR model is implemented based on identified contacts and is compared with *base-SIR* model.

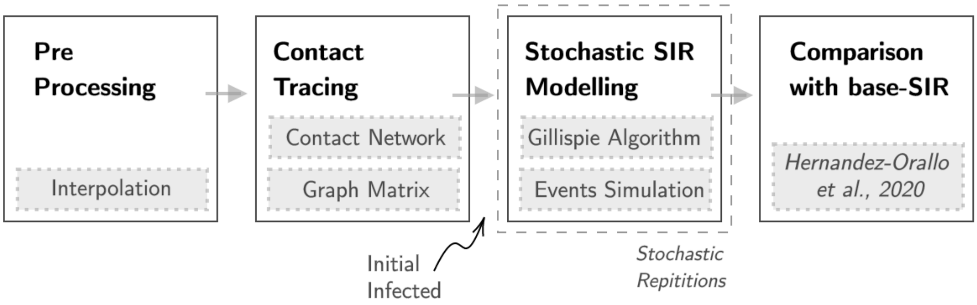


FIGURE 4.1: Work modules of Stage-1 *(Baseline SIR Modelling)*

2. ***Spatial Risk:*** This stage is a continuous process for each iteration of Baseline-SIR model to compute new spatial risk based on SIR events from the previous iteration. First, the basis of risk assessment is identified and computed based on movement data and SIR events of previous iterations. Secondly, these attributes are combined to have an overall risk representation. This result in a value of risk for each future contact which is considered as the 'spatial context' of the location of the contact. Figure 4.2 explains this stage in a stepwise work modules.

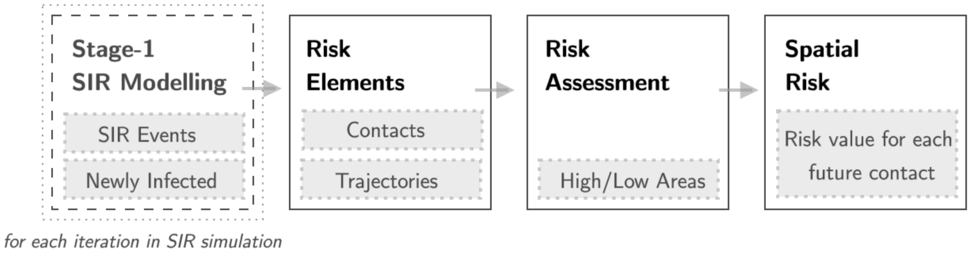


FIGURE 4.2: Work modules of Stage-2 *(Spatial Risk Assessment)*

3. ***Hybrid Spatio-SIR Model:*** The last stage is the arrangement of the *Stage-1* & *Stage-2* in a cyclic process to have an overall hybridization of Baseline-SIR model and Spatial Risk. In this stage, the core step is of including the spatial context in the baseline-SIR model as highlighted in Figure 4.3, so as to have a risk intensity associated with each contact, and further assess how these values influence the future SIR events.

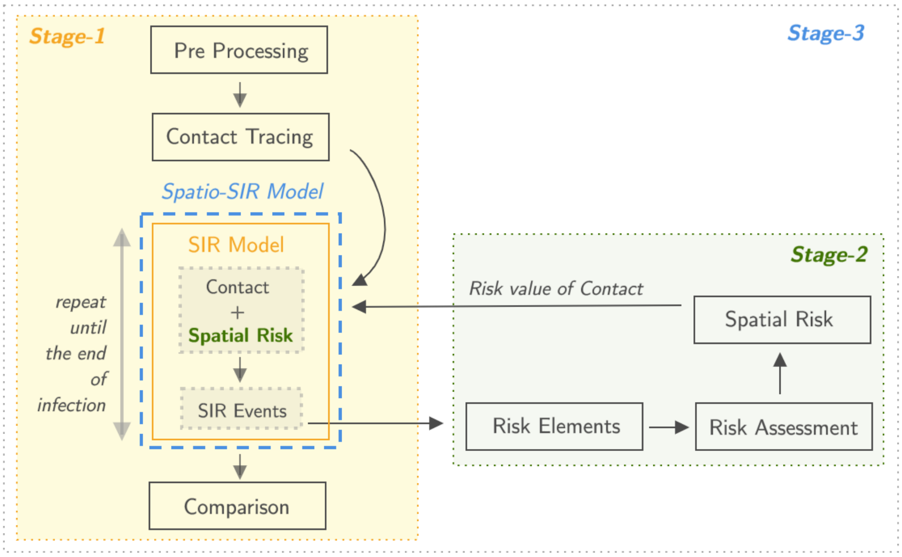


FIGURE 4.3: Work modules of Stage-3 *(Hybrid Spatio-SIR Model)*

4.2 Implementation Process

4.2.1 Stage-1: SIR Modelling

Stage-1 of SIR Modelling consists of *four* work modules:

1. ***Pre-Processing:*** Interpolating data to have locations of all individuals at all times.
2. ***Contact Tracing:*** Identification of contacts based on distance & duration thresholds.
3. ***SIR Model:*** Executing a disease outbreak scenario based on contacts and SIR states.
4. ***Comparison:*** Comparison of SIR model and epidemic details with *base-SIR* model.

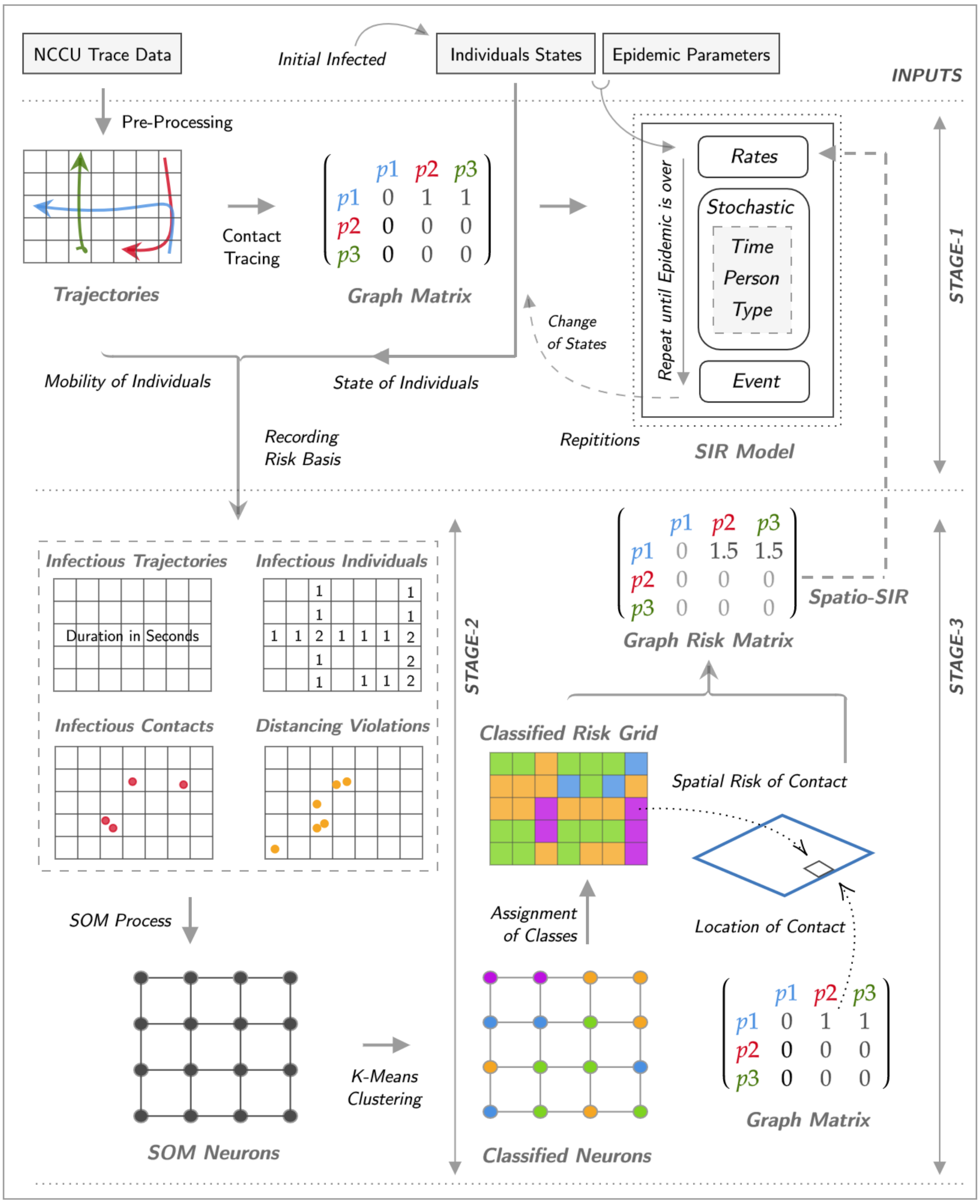


FIGURE 4.4: Implementation workflow of overall methodology

1. Pre-Processing

Pre-Processing involves only the step of *Interpolation of Trajectories*. NCCU trace dataset provides continuous recordings of individuals mobility however data also contains off-set of varying temporal frequency or break in recordings. In order to ensure there is data for each timestamp, trace dataset is interpolated so as to have spatial location of all individuals at all timestamps with a temporal frequency of seconds.

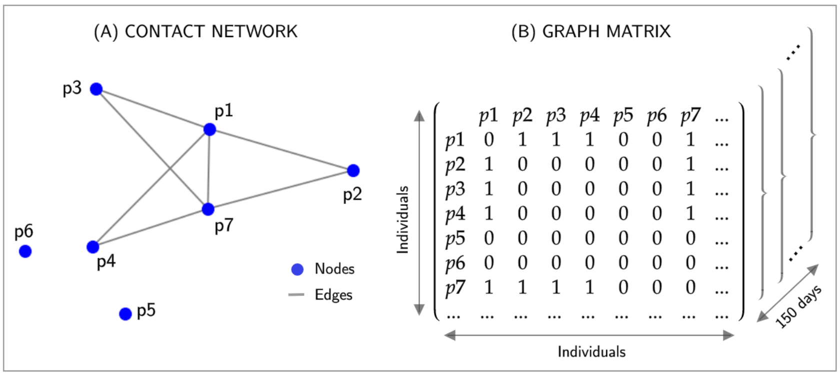
2. Contact Tracing

Contact Tracing is the identification of *'collocation'* of two or more individuals. However, this collocation 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. Hernandez et al. (Hernández-Orallo et al., 2020) highlight that contact with a possibility of transmission is the one within 2 meters of an infected individual with at least a duration of 1 minute.

Information of these identified contacts can be stored in Contact Graph Matrix (a three-dimensional adjacency matrix, as explained in section 3.1, with third dimension corresponding to the day of contact). This can support in the transfer of only useful information (contacts) from computationally complex trajectory dataset to a simpler matrix form. However, it is required to have a temporal frequency of graph matrix, as it is computationally expensive to construct such matrices for each minute or a second. In epidemic modelling, it is common to have a temporal frequency of a 'day', hence a daily contact graph matrix can be developed capturing information about contacts happening in that day. This means that even a single contact in a day between two individuals will be represented as a contact among the two on that day. Algorithm (3) presents the method for extracting daily contact graph matrix out of interpolated trajectories, whereas Figure 4.5 illustrates an example of a resulting adjacency matrix and network graph.

Algorithm 3 Contact Tracing

```
input: trace, n                              ▷ (trace←dataset) & (n←# of individuals)
output: $G$
// initialize $G(n, n, 150) \leftarrow 0$                 ▷ ($G$ ← contact matrix)
for ($t$) do                                        ▷ ($t$ ← time in seconds)
    for ($p1$ ← 1 to $n$) do                          ▷ ($p1$ ← first person)
        for ($p2$ ← 2 to $n$) do                      ▷ ($p2$ ← second person)
            dist = distance ( p1↔p2 )
            if (dist < $d_c$) then                  ▷ ($d_c$ ← distance threshold)
                record duration
                if (duration > $t_c$) then          ▷ ($t_c$ ← duration threshold)
                    $G(p1, p2, day) \leftarrow 1$               ▷ (day←this-day)
return: $G$
```



	p1	*p2*	*p3*	*p4*	*p5*	*p6*	*p7*	...
p1	0	1	1	1	0	0	1	...
p2	1	0	0	0	0	0	1	...
p3	1	0	0	0	0	0	1	...
p4	1	0	0	0	0	0	1	...
p5	0	0	0	0	0	0	0	...
p6	0	0	0	0	0	0	0	...
p7	1	1	1	1	0	0	0	...
...	...	...	...	...	...	...	...	...

FIGURE 4.5: Single day contacts in the form of (A) Network & (B) Matrix

3. SIR Model

Once the contacts are identified and stored in a contact matrix, next is to formulate a setup to execute SIR events based on contacts and latency state of individuals. In order to mimic a real epidemic, a stochastic setup is more realistic than a deterministic one as it involves the chance element. Similarly, an event-based approach is better suited as it considers each individual independently. One such event-based stochastic model based on GFRM is introduced in (Keeling and Rohani, 2011) (page 203), which is selected as Baseline-SIR model for this study.

Initial Infected (I_0)

As information about the infection state of individuals is not available, a limitation discussed in section 6.4, 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 the compartment (I), so as to have a sense of disease propagation based on their future contacts, as disease propagates. Section 6.1.3 in the chapter 'Results & Discussion' describes the relation of varying Initial Infected (I_0) with the overall epidemic.

The Stochastic Model

Apart from Graph Matrix (G), model relies on two more components for its execution:

- ***Infection States:*** Tracking infection states of each individual in the population
- ***Epidemic Parameters:*** Infectious disease specific epidemic parameters

Infection States refers to the compartments an individual can be during time period T.

As motivated by *base-SIR*, a total of five compartments are considered which depict following respective states:

Infection Related

$\boldsymbol{S}$: Susceptible (not infected)
$\boldsymbol{I}$: Infected
$\boldsymbol{R}$: Recovered or Removed

Quarantine Related

$\boldsymbol{Q_S}$: Susceptible individual in quarantine
$\boldsymbol{Q_I}$: Infected individual in quarantine

Same as graph matrix for contact tracing, state of individuals can be efficiently stored in state vectors, where 0 or 1 represents if that state is applicable to the individual or not. In an event-based model, an individual will only be in a single state at one point in time. An example of four individuals is presented with their state vectors ($\vec{S}$, $\vec{I}$ & $\vec{R}$) representing their conditions. The example illustrates that person two (p2) is still infected and fourth (p4) have recovered whereas other two (p1 & p3) remain susceptible to the disease as there is an infected person in the population.

$$\vec{S} = \begin{pmatrix} p1 & p2 & p3 & p4 \\ 1 & 0 & 1 & 0 \end{pmatrix} \qquad \vec{I} = \begin{pmatrix} p1 & p2 & p3 & p4 \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad \vec{R} = \begin{pmatrix} p1 & p2 & p3 & p4 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

With five compartments of (S, I, R, Q_S, Q_I) means that are seven possible SIR events that will imply the transition of an individual from one compartment to other. An event means a change of two state vectors, *(i)* from compartment and *(ii)* to compartment. These seven events are:

$$\boldsymbol{S \rightarrow I} \qquad \boldsymbol{S \rightarrow Q_S} \qquad \boldsymbol{S \rightarrow Q_I} \qquad \boldsymbol{I \rightarrow Q_I} \qquad \boldsymbol{Q_S \rightarrow S} \qquad \boldsymbol{I \rightarrow R} \qquad \boldsymbol{Q_I \rightarrow R}$$

Epidemic Parameters refers to the disease-specific elements in the form of coefficients which contribute to computing the rates of each event associated with individuals. Table 4.1 summarizes such parameters used in this study, as explained in section 3.2.2. Based on these parameters, rates of all events can be mathematically obtained as per equations provided in Table 4.2.

Using these parameters, ***Stochastic Model*** is formulated using discussed three components of *(i)* Graph Matrix, *(ii)* Infection States and *(iii)* Epidemic Parameters. In this study, a generalized framework for spatio-SIR modelling through the use of values corresponding to COVID-19 is implemented, as given in Table 4.3, however, any disease-specific model can be developed by adjusting these parameters. Similarly, implementation of GFRM is improvised in this study where *Time* of next event is a stochastic duration based on a random element instead of computing time for each event, as mentioned in line 4

TABLE 4.1: Summary of infectious disease parameters

Parameter	Description
κ	*Average degree (daily contacts per individual)*
b	*Probability of transmission of infection*
β	*Transmission rate* ($\beta = \kappa \cdot b$)
γ	*Recovery rate* ($1/\gamma$ = *disease specific days for recovery)*
$\mathcal{R}_0$	*Basic Reproductive Ratio* ($\mathcal{R}_0 = \beta/\gamma$)
δ	*Rate of detection*
$\mathcal{T}_Q$	*Time in Quarantine*
$\mathcal{K}_i$	*Contacts of individual 'i' with infected individuals*
$q^{'}$	*Tracing efficiency*
$\mathcal{C}_i(\triangle)$	*Backward contact tracing of individual 'i' with detected infected individuals*

TABLE 4.2: Rate equations related to each SIR event

Event	Description of Rate	Equation
$\boldsymbol{S \rightarrow I}$	*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
$\boldsymbol{Q_S \rightarrow S}$	*End of quarantine after quarantine period*	$\mathcal{T}_Q$
$\boldsymbol{I \rightarrow R}$	*Recovery after infectious period*	γ
$\boldsymbol{Q_I \rightarrow R}$	*Recovery after infectious period in quarantine*	$\mathcal{T}_Q$

of Algorithm (1). However, *Type* of the next event is based on GFRM by computing the rates of each event and then stochastically drawing the next event. Figure 4.6 illustrates this formulation and graphically explains the SIR modelling process.

4. Comparison

Results of SIR modelling and its comparison with *base-SIR* is discussed in section 6.1.4.

Stage-1 of SIR Modelling provides the Baseline-SIR model to explore modifications related to the inclusion of 'spatial context'. This contact tracing based model initiates with a self-induced infection to few individuals, and implements disease propagation among the population in an epidemic scenario. Being event-based and individual-level, the

TABLE 4.3: Estimated values related to COVID-19

Parameter	Estimated Value (COVID-19)
γ	*1/15*
$\mathcal{R}_0$	*3*
δ	*0.1*
$\mathcal{T}_Q$	*1/14*

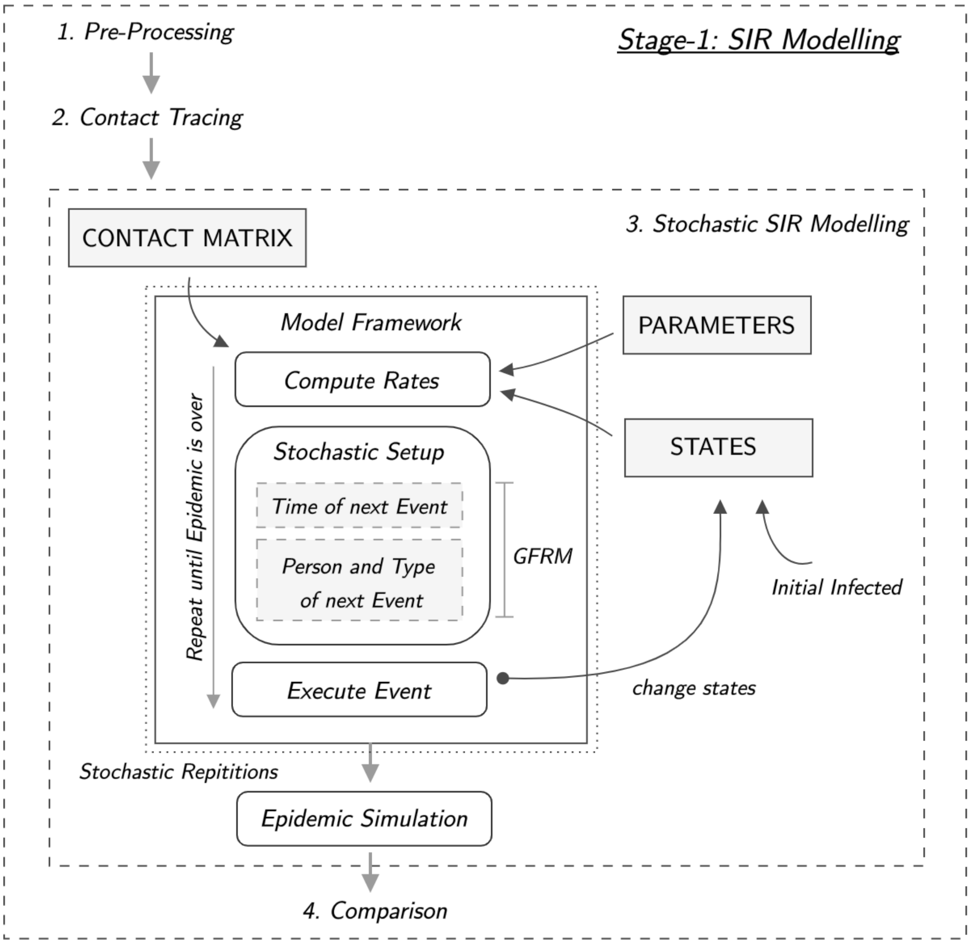


FIGURE 4.6: Implementation Process of Stage-1 *(SIR Modelling)*

model tracks an infected individual from the moment of infection to recovery, which allows to track individual-level infectious trajectories with high temporal precision.

4.2.2 Stage-2: Spatial Risk

Stage-2 of Spatial Risk consists of *three* work modules:

1. ***Risk Elements:*** Recording risk stimulating attributes based on SIR modelling events.
2. ***Risk Assessment:*** Combining multiple attributes to attain classification of risk.
3. ***Spatial Risk:*** Computing a spatial risk score based on identified classes of risk.

Risk Grids

Risk Grids refers to the output structure of Stage-2. As intention is to associate a risk score to each contact based on its spatial location, it is important to address the definition of 'location'. As the only dataset is of mobility trajectories which are discrete recordings of movements per second, a continuous risk surface would require additional steps of data transformations based on spatio temporal point processes. A simpler and computationally efficient approach is to consider a regular lattice (grid) structure segmenting the study area into smaller cells with each cell having a risk score. From this, 'location' of contact can be defined as the corresponding cell in which the contact is taking place.

1. Risk Elements

SIR tracking involves the monitoring of SIR related attributes which can be considered as the basis of high-risk. At first, risk basis are identified based on the available data. Secondly, identified basis are recorded to execute risk assessment and compute risk scores.

Risk Basis

Due to the limitation of data other than only trajectories of movement (discussed in section 6.4), the situation compels to develop risk out of the trajectories dataset. Only considering movements and extracted contacts, four risk basis are identified as:

a. Infectious Trajectories
b. Infected Individuals
c. Infectious Contacts
d. Social Distancing Violations

(a) Infectious Trajectories refers to the amount of time an infectious trajectory has spent in each cell. Such tracking of infectious trajectories is critical in identifying high-risk places (Benreguia et al., 2020). Here, idea is to track an infected individual's movement in cells from the time of infection until recovery or quarantine.

(b) Infected Individuals refers to count of infected individuals in each cell. This is related to (a), but distinct in a sense that shorter cumulative duration of many infected individuals is riskier than a longer duration of a single individual (see Figure 6.6 for an example).

(c) Infectious Contacts refers to the location of those contacts which involves an Infected individual. They are more in number that the total amount of time infection is transmitted as this involves all $(S \leftrightarrow I)$ contacts; while the transmission is dependent on the rates and the randomly chosen event, where [*Transmissive Contacts* ∈ *All (S↔I) Contacts*].

(d) Social Distancing Violations refers to all $(S \leftrightarrow S)$ contacts. This property reflects population density and also capture the notion that a place (cell) with higher precautionary violations must be of higher risk than a place following the public health regulations.

Recording Risk Attributes

This deals with computing the 'risk basis' from the dataset based on SIR events. As a network graph is developed per day (as explained in section 4.2.1), the same can be followed to develop risk grids. This means that the 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 4.7 explains the process of computing risk basis, by tracking *infectious trajectories* for their duration and count; alongside monitoring *contatcs* for their spatial locations.

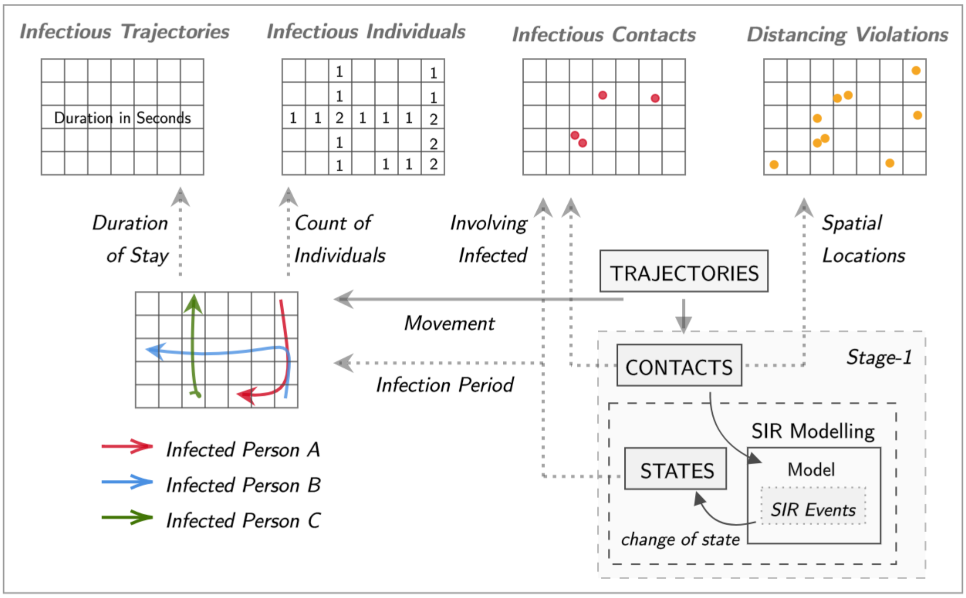


FIGURE 4.7: Computing Risk Basis from Trajectories & Contacts

2. Risk Assessment

Risk assessment deals with the process of integrating multiple risk attributes into a single representation. This requires combining four grids (normalized) into a single grid, a *Risk Grid* based on activities from the previous day to provide classes of risk for the next day.

Here, implementation of a multi-criteria analysis approach like *Ordered Weighted Averaging* is not applicable as it relies on Ordered Weights to define the relative significance of a criterion over others. In this case, there is no prior information of which criterion is significant over others, nor validation data is available to fit the model to data.

Similarly, clustering approach like *Subset Scan Statistics*, or *Spatio-Temporal Point Processes* methods are also not suitable, as in this study the prime entity is of infectious trajectories which are continuous tracks of mobility and not data points over the study area.

For classification, a supervised method requires information about the characteristics of the target class and pre-existing labels (through supervision process), for the method to cluster data and label them accordingly. Same as Ordered Weighted Averaging, lack of validation data restricts the application of supervised classification as well.

A possible solution is to implement an unsupervised learning method, as it does not rely on pre-existing labels or explicit target outputs for reinforcement. Such methods are inclusive of unsupervised clustering and classification techniques, as they only require input patterns to highlight relationships. This approach can assist in the exploration of the covariates available in the form of four grids to develop a single classified risk map.

4.2.3 Self Organizing Maps (SOM)

SOM is an unsupervised clustering technique, which can serve the purpose of combining information of four grids into a single one. SOM preserves the topographic relationships in feature space to ensure nearby objects are clustered together. Applications of SOM are presented in detail in section 2.3.3, whereas section 3.3 explains working of SOM process.

Clustering with SOM is a two step process. Firstly, dimensionality reduction from a four-dimensional (four grids) input space to a two-dimensional SOM net (neurons). Secondly, SOM net is further grouped into desired numbers of clusters using K-Means. Optimal size of SOM neurons can be acquired through $(5 \cdot \sqrt{d_{size}})$, where d_{size} refers to size of dataset. Figure 4.8 illustrates the SOM process of dimensionality reduction followed by K-Means clustering to acquire classes. This results in a classified risk grid where each cell corresponds to a class of risk (e.g High, Medium, Low). The procedure to assign appropriate labels as well as the risk score to each class is detailed below.

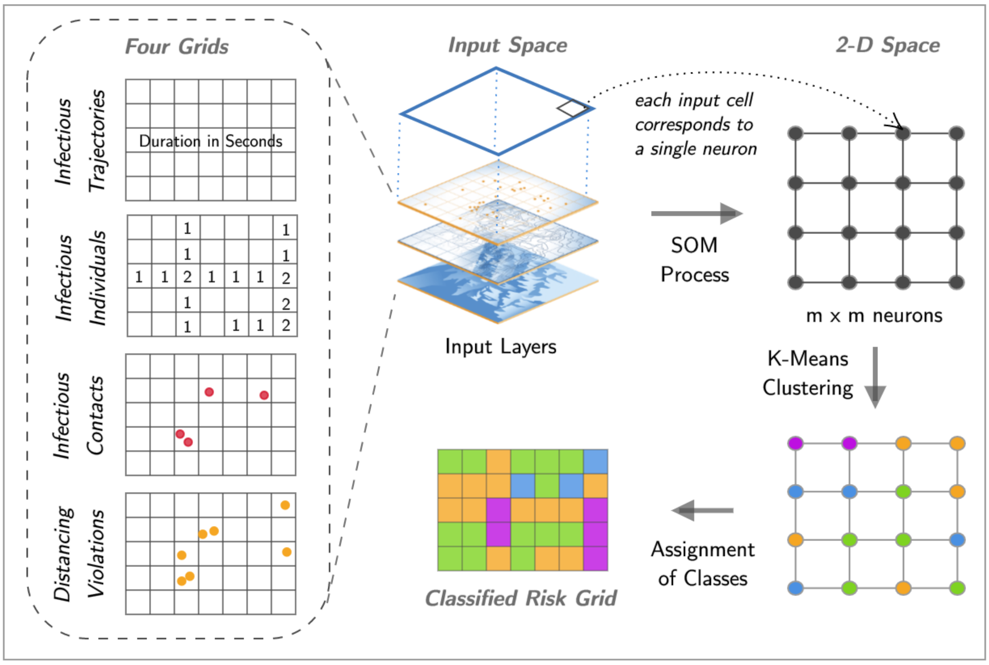


FIGURE 4.8: Unsupervised classification workflow using SOM & K-Means

3. Spatial Risk

This module deals with the assignment of appropriate labels to each class of the classified risk grid. As the output after K-Means is an *un-ordered* classification, which is same as segmenting the cells in different groups but not knowing which group is of higher risk than others. To order the classes, the cumulative average of all cells in the class is compared to assign ordered labels in descending order. Figure 4.9 explains this process of labelling.

4.2.4 Stage-3: Hybrid Spatio-SIR Model

In Stage-3, hybridization means the development of a repetitive setup to fuse *Spatial Risk* in each iteration of Spatio-SIR model until the end of the epidemic. To include the spatial risk for each contact, the graph matrix is modified to have a varying intensity based on risk score instead of the previous constant value of '1' representing a contact. As rates of events in SIR model are based on the cumulative contacts represented by κ, a varying contact value between [0.5,1.5] with mean 1 (existing model) will result in a modified κ resulting in a higher or lower Transmission Rate ($\beta = \kappa \cdot b$). Hence, a contact taking place in higher risk cell will have a higher chances of an event to a susceptible individual involved in the contact (which are of $S \rightarrow I$, $S \rightarrow Q_S$ $S \rightarrow Q_I$). Events related to infected

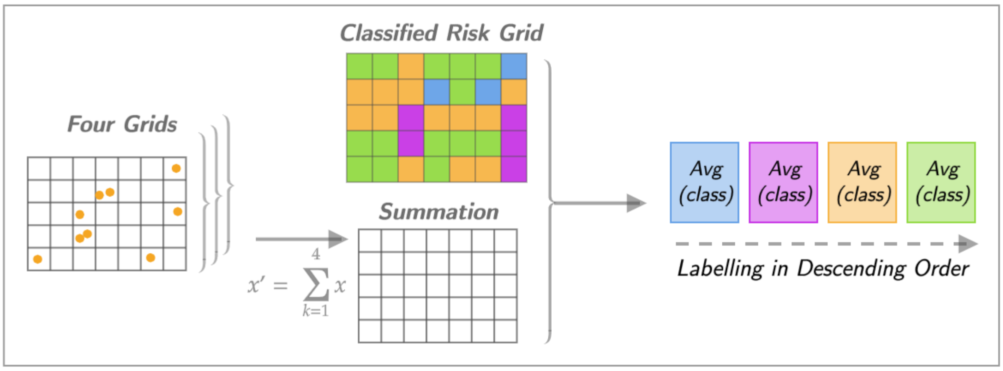


FIGURE 4.9: Assignment of ordered labels to un-ordered classes

individuals and those in quarantine remains as before as there is no effect of spatial risk to them. Figure 4.10 illustrates this modification of graph matrix to reflect spatial risk.

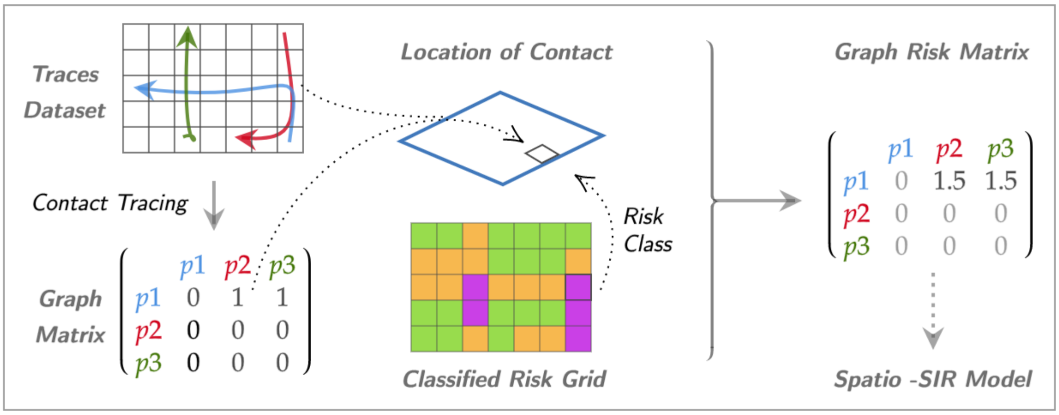


FIGURE 4.10: Assignment of spatial risk scores to contacts

This process of including spatial risk to modify graph matrix is to be executed every day (as graph matrix is on daily basis). But as risk is dependent on the SIR activities from the previous day, they are to be computed dynamically every time as simulation enters the next day. This process of dynamically computing risk score based on daily movement and reflecting its effect by modifying graph matrix is termed as the hybridization of spatial risk & SIR model. A form of a *Spatio-SIR Model* implementing contextual contact tracing. Complete development workflow of Spatio-SIR model is already presented as Figure 4.4.

5 Data and Experimental Design

This chapter introduces the selected dataset and focuses on the experimental design setup for implementation of the explained methodology. Tracking human mobility necessitates continuous traces in the form of trajectories. Appendix 8.1 presents an overview of such datasets used in recent research. However, either the datasets are available in the form of pairwise distances to preserve privacy (a form suffice for contact tracing, but not sufficient when the aim is to track infectious trajectories), or are not publicly of freely available. In this study, mobility data of *NCCU Trace* is used, same as used in *base-SIR* model.

NCCU Trace Data

NCCU Trace (Tsai and Chan, 2015), refers to an android application based mobility model to trace movements of students in a campus environment of National Chengchi University, Taiwan. Model was designed to capture information regarding GPS, WiFi, and Bluetooth devices in proximity, resulting in movement traces in the form of (*user, time, locationX, locationY*). In this study, movements of 115 students were recorded for a period of 15 days, with measurement interval up to 10 minutes and spatial position rounded to meters. Appendix 8.2 contains details of NCCU Dataset with an overview of the study area, sample recording and a map of initial locations. For an epidemic, 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. Figure 5.1 shows this repetition of the dataset to simulate for 150 days, an appropriate duration for epidemic modelling.

Experimental Design

In this study, baseline-SIR model, as well as spatio-SIR model is evaluated over NCCU Trace where the total population is *115* and data duration is of *150* days. The experiments assumes ten individuals as *Initial Infected* ($I_0 = 10$) on first day of the epidemic with no *Recovered Individual* ($R = 0$); with values of COVID-19 parameters as discussed in Table 4.3, where the sum of individuals in all compartments is *115* at all times. For the

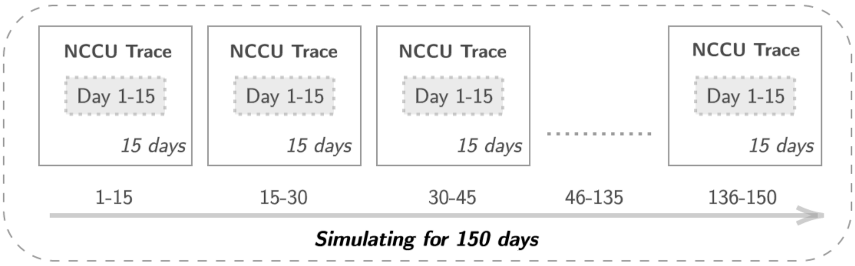


FIGURE 5.1: Simulating 150 days data from 15 days NCCU Trace data

stochasticity, 10 realisations of the same initial conditions but the random allocation of initial infection are executed. This means that in each realisation, infected individuals are different. Averaging the results of 10 realisations, average curves are obtained, where a curve represents the count of individuals in each compartment. Due to stochasticity, duration of the epidemic in these realisations varies, hence we extrapolate trends of other realisations to the epidemic with the longest duration to obtain an average representation.

In each run within a single realisation 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. A single realisation of model proceeds in time based on these stochastic duration *(time steps)* to provide an output of a disease scenario.

The capability of associating spatial component into an infectious disease modelling can act as a tool for policymakers to simulate scenarios, visualizing the consequence of their decisions prior to their actual implementation. One such real-world scenario is presented in Figure 6.11 where a concept of *Intervention* is introduced to assess its impact on the results. Here, intervention in the form of spatial high-risk of (1.5) at all cells is introduced in the system from day 11^{th} to 20^{th}. And the idea is to mimic a real behaviour of holiday season with relaxation in precautionary measures. In such a scenario, there is a lot of movement with an increased count of contacts and asymptomatic individuals; and the spatio-SIR model is expected to capture this sudden increase of spatial risk in its trend. Table 5.1 provides an overview of varying experimental design.

TABLE 5.1: Overview of experimental design for different scenarios

Description	**Baseline-SIR**	**Spatio-SIR**	**Intervention Setup**
Spatial Risk	No Risk	Dynamic	Induced high-risk from day 11-20
Result compared to	Spatio-SIR	Baseline-SIR	Spatio-SIR

6 Results & Discussion

This chapter presents the findings segmented into three stages corresponding to each stage of implementation. First, outputs of the SIR model are illustrated with visualizations of Spatial Risk. Lastly, modification of Spatio-SIR model is presented & discussed.

6.1 SIR Modelling

This section shares output of Baseline-SIR model and a graphical representation of individual-level transitioning of population among compartments. This is followed by a comparison of varied initial setup and comparison of the implemented model with base-SIR.

6.1.1 Baseline-SIR Model

Exploring modification of a spatial context necessitates setup of a baseline model to experiment over. Figure 6.1 presents output of such a baseline setup in form of an outbreak scenario based on experimental design, using parameters from Table 4.1.

At the beginning of epidemic, everyone except the *Infected* is in the *Susceptible* compartment, which means there is no *Recovered* individual ($I_0 = 10$, $S = 105$ & $R = 0$). Initially, count of *Infected* individuals increases from *10 to 14* in 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 fortnight, as due to tracing efficiency (backward tracing) 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.

Peak of individuals in *Quarantine Susceptible* is around 19^{th} day with 40 plus individuals, where afterwards sum of individuals remains more or less constant which depicts an equal frequency of individuals moving between ($S \leftrightarrow Q_S$) compartments. *Quarantine Infected* compartments reaches its highest count of five twice on 13^{th} and 22^{nd} day.

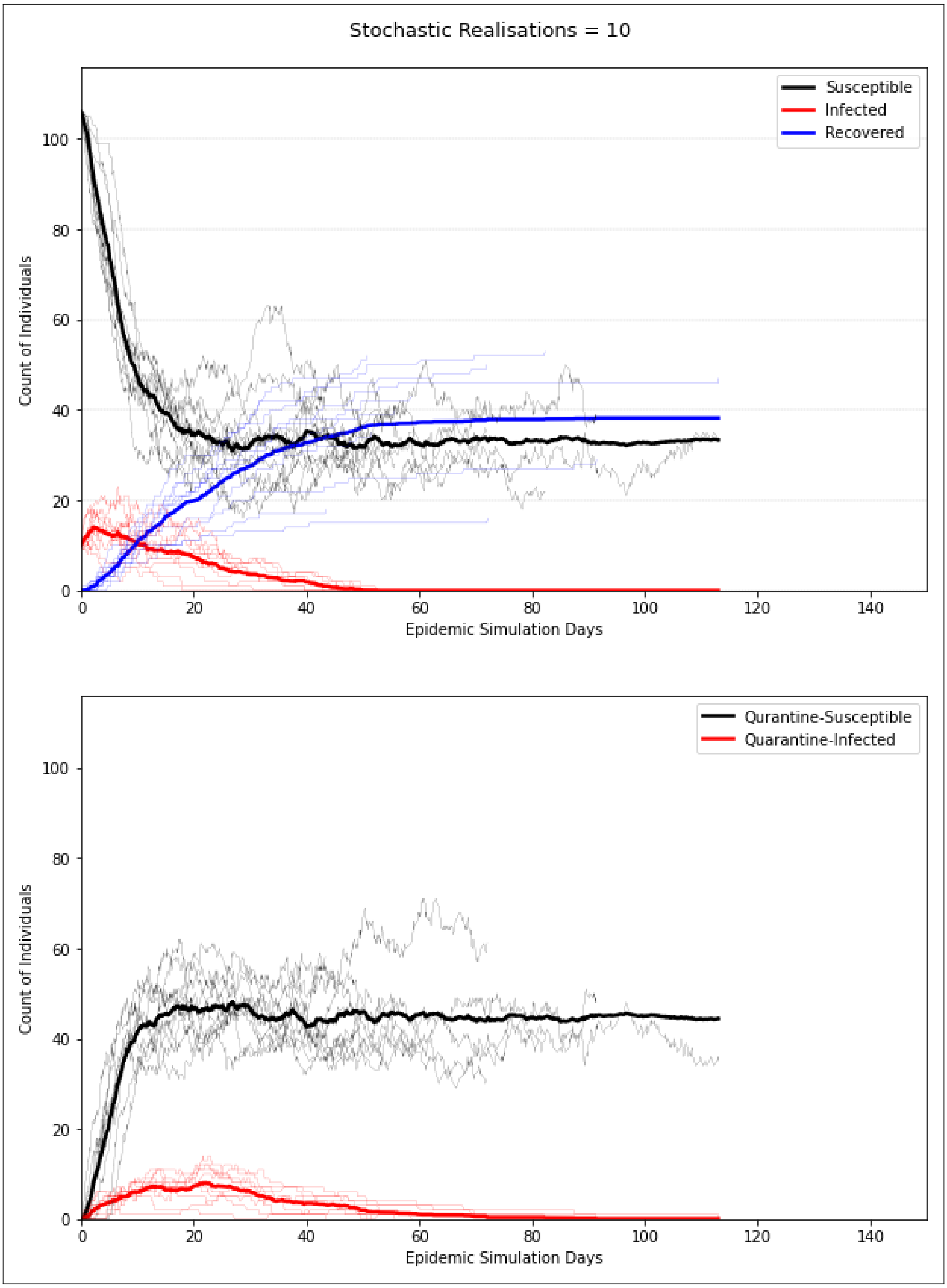


FIGURE 6.1: Output of a disease outbreak scenario by Baseline-SIR model. *(Top)* presents trends related to count of *Susceptible, Infected & Recovered*, whereas *(Bottom)* illustrates count of individuals in *Quarantine* related compartments. Each of the 10 realisations of stochastic model is shown *(light in color)*, with their Average curves represented with *(dark bold) lines*. Count of total population is *115* which are represented over *Y-axis*.

Once a person is *Recovered*, that individual remains in that compartment, which is evident from the continuous increase in its count from zero at the start of the epidemic to *38* at its end. Even after there is no *Infected* person on the street after 45th day, the model continues in anticipation of risk due to the presence of individuals in *Quarantine Infected* compartment; and ultimately ends the epidemic with their recovery around 113th day.

6.1.2 Individual Level Latency

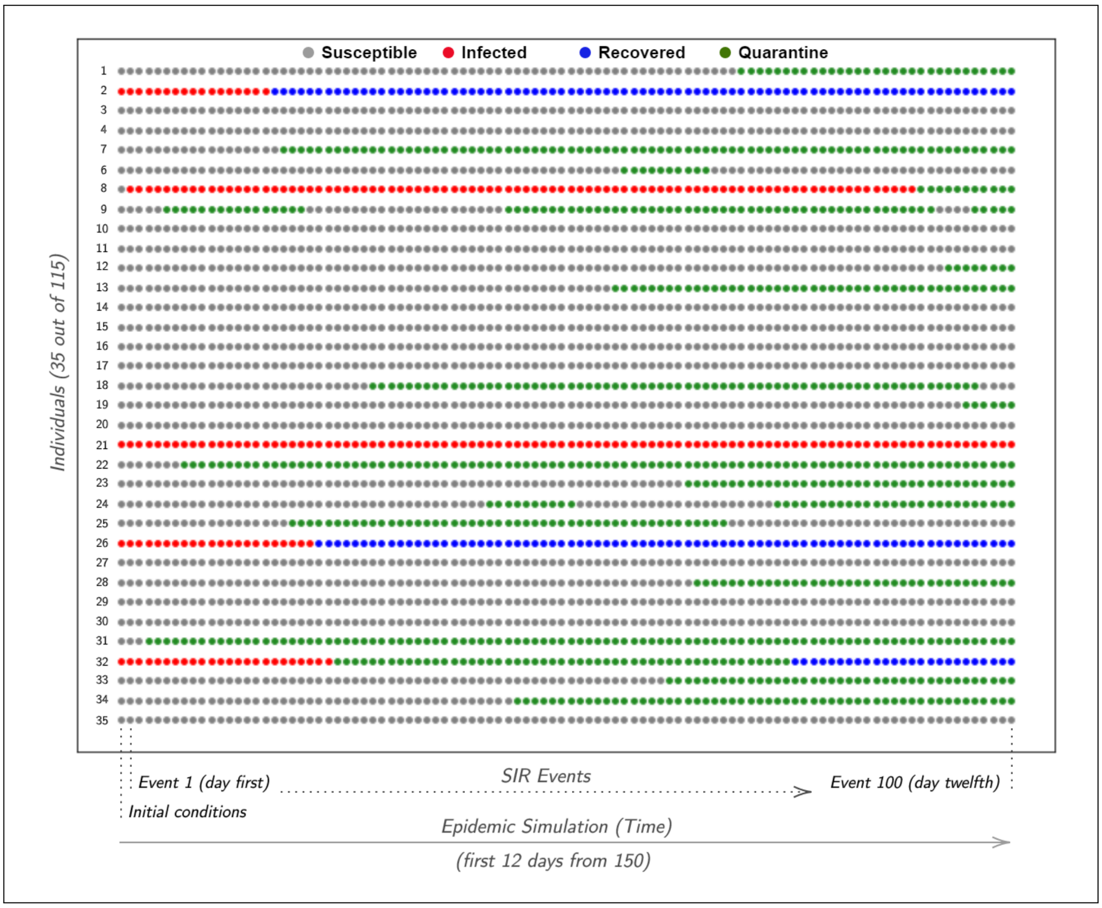


FIGURE 6.2: Individual-level change in latency of *35* out of total *115* individuals is shown based on the SIR events as model simulates. Each row belongs to a single individual, where the compartment they belong to at an instance of time is represented column-wise chronologically from left to right. 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.

The ability of an individual-level compartment model to monitor the latency state of each individual at all times signifies its importance in infectious diseases realm. To understand this capability, Figure 6.2 illustrates *Individual-level latency* of a subset of population.

At the start of the epidemic (day 1), four individuals (2, 21, 26 & 32) are infected as the initial outbreak whereas remaining all are *Susceptible*. The first stochastic event (second column from left) is of infection to the individual (8). In every iteration, there is only one event, where the time of next event (a part of the day based on stochasticity) is also random, hence there can be multiple events in a single day. Individual (8) remain 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.

6.1.3 Varying Initial Configurations

Baseline-SIR model executes a disease outbreak scenario based on the initial conditions of epidemic configured for it. Changing the initial setup can provide a summary of its impact on the overall outbreak and is a helpful tool to analyse varying scenarios. Other than *Contacts* based on contact tracing of mobility trajectories, other critical components are of *Count of Initial Infected* (self introduced in this study), *Tracing Efficiency* (for backward tracing of infectious contacts) and *Basic Reproductive Ratio* (depends on the disease). One such variation is presented in Figure 6.3 with different intensities of initial outbreak.

In general, the higher quantum of initial outbreak results in a longer epidemic which is evident in all subplots. In plot (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 ($I_0 = 10$ & $I_0 = 15$) results in decrease of *Susceptible* count from 105/100 to approximately 40 within 2 weeks, whereas $I_0 = 5$ reaches the count of 40 after *six* weeks.

Plot (B) illustrates the effect of varying initial outbreak on the total count 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, plot (C) depicts a likewise trend of initial difference, where two setups of ($I_0 = 05$) & ($I_0 = 10$) later (after 70^{th} day) coincides 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} day.

Plot (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*

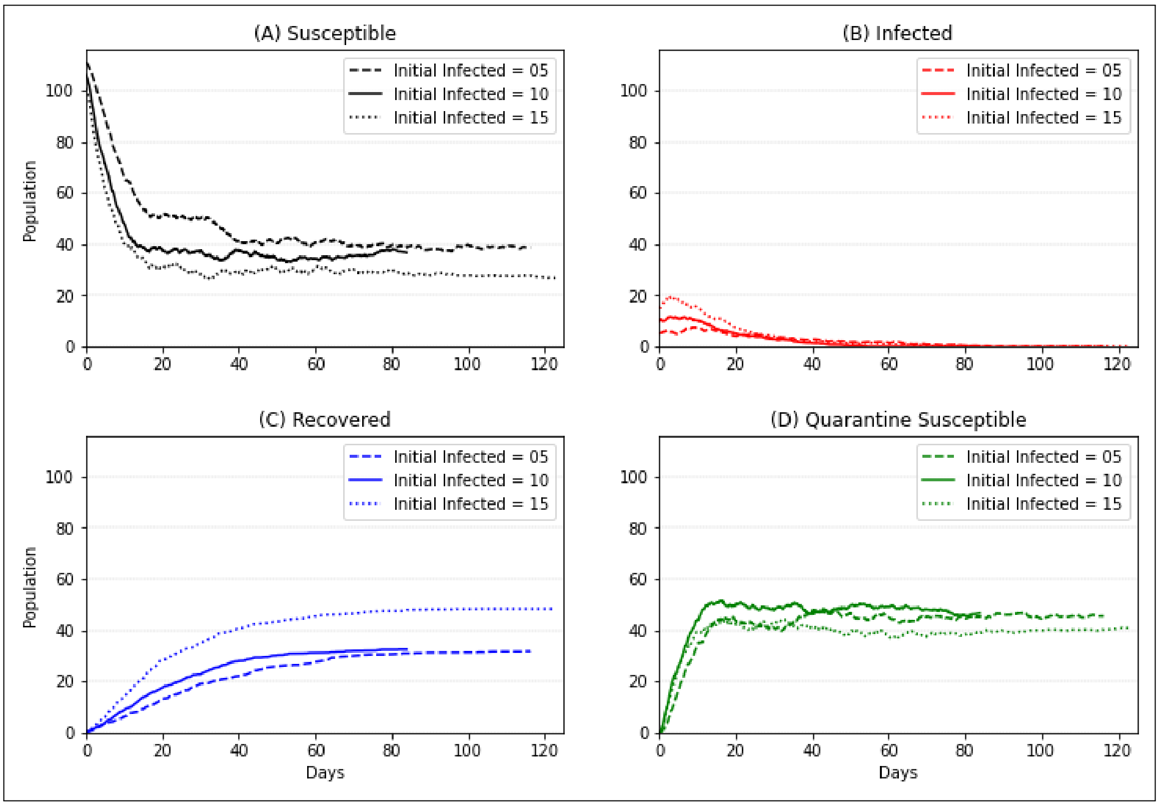


FIGURE 6.3: Comparing the *Average* of 10 stochastic realisations with varying count of *Initial Infected* I_0, to observe their effect on the overall disease outbreak. Three scenarios of $I_0 = 5/10/15$ are shown with a population size of *115*. Subplots *(A,B,C & D)* shows comparison of *Susceptible, Infected, Recovered & Quarantine Susceptible*, respectively. Due to stochasticity, final duration of epidemic varies depending on the overall spread of infection.

rate ($\beta = \kappa \cdot b$) and *(ii)* the chance element of event-based stochastic setup. Hence, the relation of initial infected with *Quarantine* related compartments are 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 can be of *Tracing Efficiency* which is available in Figure 6.4. Tracing efficiency refers to 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 case of no backward tracing (zero efficiency) shown in plot (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.

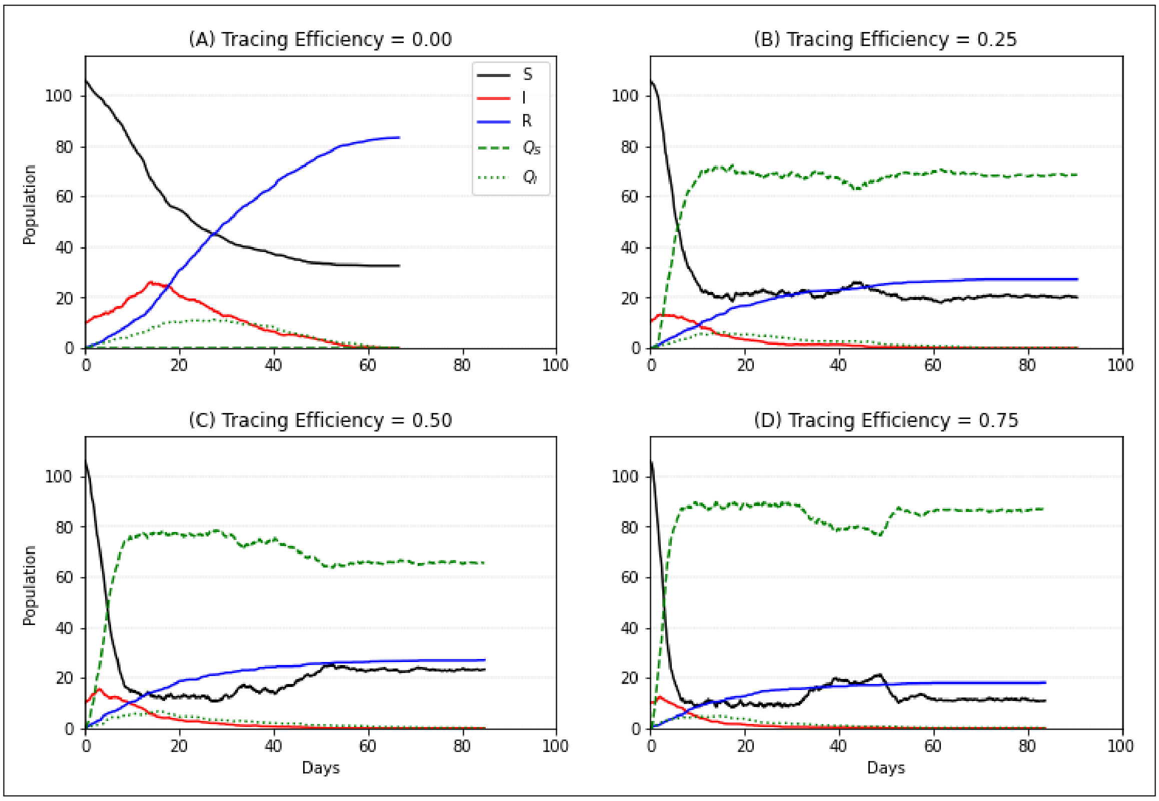


FIGURE 6.4: Comparing the *Average* of 10 stochastic realisations with varying *Tracing Efficiency* q', to observe their effect on the overall disease outbreak (population size = *115*). Subplots *(A,B,C&D)* presents four cases of q'=0, 0.25, 0.50 & 0.75, respectively, where 1 means 100% backward tracing.

In Plot (B), *62* individuals are in *Quarantine Susceptible* by the 10^{th} day, whereas with efficiency of 0.50 (plot (C)) and 0.75 (plot (D)), there are *77* and *88* individuals in *Quarantine Susceptible* by the same period of 10 days. From which 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* & less *Recovered*). The population being forced to quarantine means they 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 plot (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 40^{th} day. An opposite can be observed in the count of *Susceptible*.

6.1.4 Comparison with base-SIR

An important step here is to ensure that the Baseline-SIR model, which serves as the core for enhancement regarding spatial context, is compared with base-SIR. Due to the stochastic nature of modelling, a direct numeric comparison of both is not possible, however the general trend of execution can be compared as a degree of confirmation.

Contact Tracing

As contact tracing process prior to SIR model is executed on the same dataset, Table 6.1 compares the properties of graph matrix of both as authentication of *Contact Tracing* process.

TABLE 6.1: Comparison of Contact Tracing of SIR Model with base-SIR

Graph Property	SIR Model	base-SIR
Rank *(contacts per day per person)*	7.66	7.66
Total Contacts *(two meters, any duration)*	31,769	31,769

SIR Model

General implementation of Baseline-SIR model results in similar behaviour as of base-SIR. The influence of the initial infected population, tracing efficiency and rate of detection tends to capture the same pattern as base-SIR. The computational complexity of both the models is dependent on *(i)* size of population (N) and *(ii) Rank* of contact pattern; which makes it *'Exponential to N'*. The difference lies in the extension of work as base-SIR focuses on the evaluation of tracing technologies and uses different contact networks for different technologies, and further compares stochastic and deterministic approaches. Whereas this study only implements a stochastic setup with one contact network (of two meters distance) as the baseline for modifications related to spatial risk and development of a new Spatio-SIR model.

6.2 Spatial Risk

This section discusses the results of Spatial Risk as of how SIR events produce disease stimulating elements which are the basis of risk for future contacts. This is followed by results of unsupervised learning-based clustering, and identification of high-risk areas.

6.2.1 Risk Basis

As explained in section 4.2.2, elements of risk are extracted after each iteration of Baseline model based on SIR events and infectious activities. Figure 6.5 presents a similar sample.

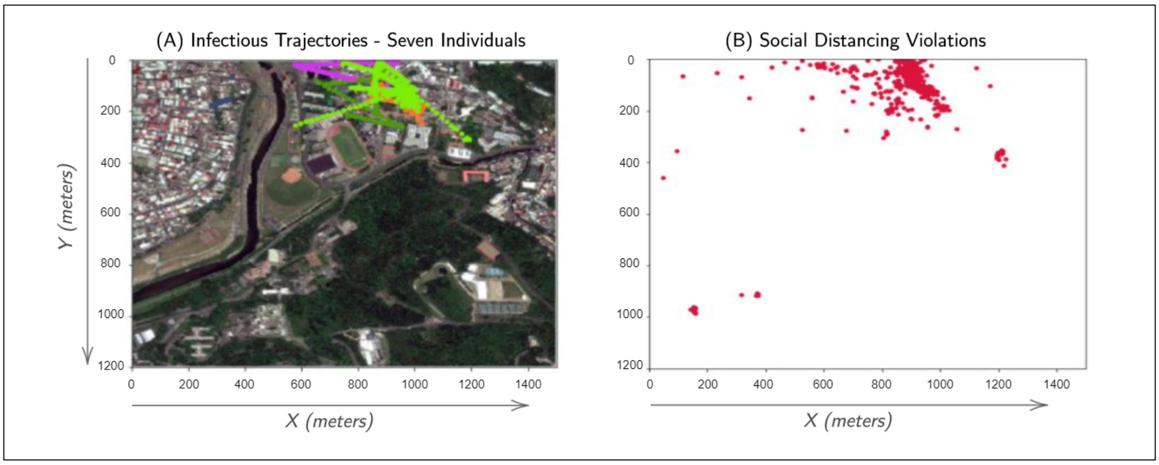

(A) Infectious Trajectories - Seven Individuals
(B) Social Distancing Violations
Y (meters)
X (meters)
X (meters)
0
200
400
600
800
1000
1200
1400

FIGURE 6.5: Risk basis of *Infectious Trajectories* (Sub-Figure-A) & *Location of Contacts* (Sub-Figure-B) from one complete day during an epidemic. In the sample shown in (A), there are total of *Seven* infected individuals from that day. (Sub-Figure-B) shows locations of all contacts from the same day.

Sub-figure (A) illustrates the movement of infectious individuals from a single day (out of the whole epidemic period), which have been shown over the study area. In this sample, there are seven infected individuals with most of the 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 concentrated in each area. Results of which are available in Figure 6.6.

Other two basis related to 'location of contacts' is of *Infectious Contacts* and *All Contacts*, where the latter is shown in Sub-figure (B). It identifies locations of all contacts termed as *Social Distancing Violations* in order to highlight the notion that a place with higher number of contacts means it is of higher risk than a place with lower number of contacts. A concept also implemented by (Rezaei and Azarmi, 2020) for Infection Risk Assessment.

6.2.2 Risk Grids

Based on the risk basis shown in Figure 6.5, risk grids are developed as presented in Figure 6.6. Here the trajectory form of mobility and point form of contacts are transformed to regular lattice (grid) structure with intensity of associated attributes normalized to [0,1].

Sub-figure (A) & (B) captures information of infectious trajectories in two different aspects of *duration* and *count* respectively, and the difference highlights importance of this consideration. Similarly, location of different nature of contacts is captured in sub-figure (C) & (D). Based on previous day, these attributes serve as the basis of risk for the next.

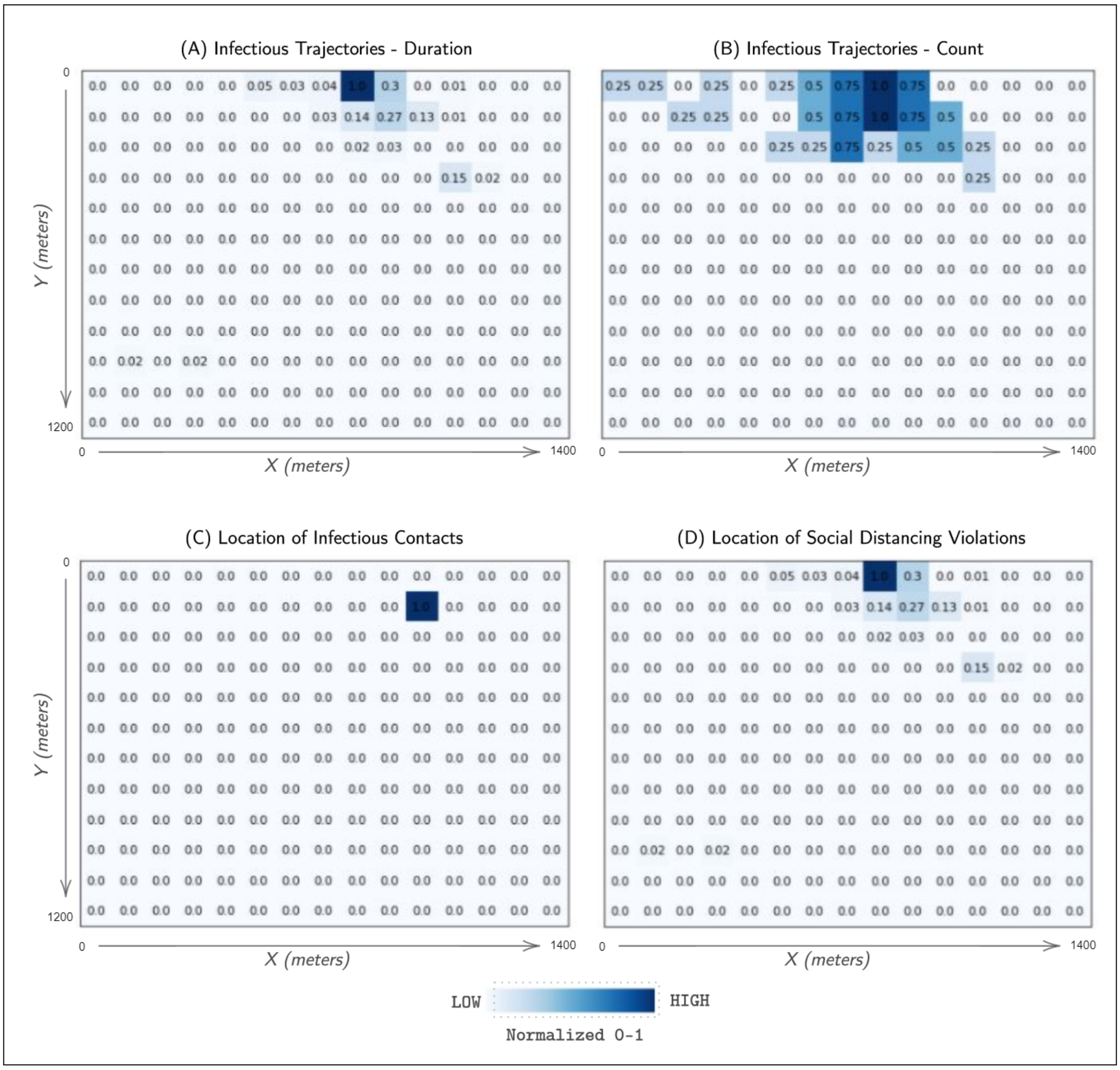
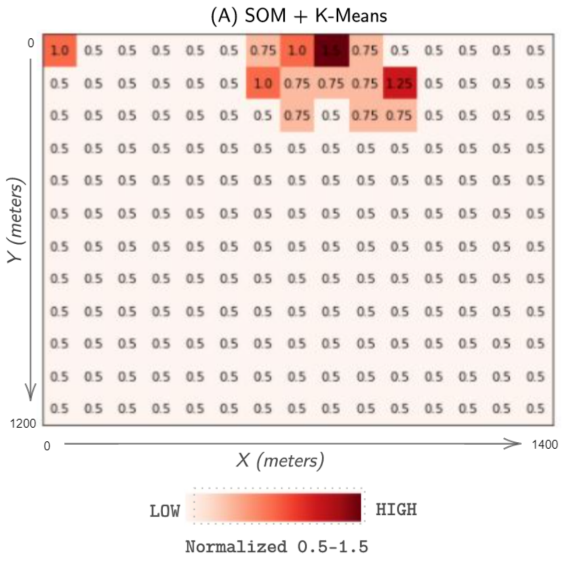


FIGURE 6.6: Risk Grids computed out of Risk Basis (discussed in Section 6.2.1). Sub-figure (A) & (B) refers to grid based representation of *Duration* & *Count* of trajectories shown in Figure 6.5-(A). Sub-Figure-(C) depicts location of an infectious contact occurred on the same day, whereas Sub-Figure-(D) presents translation of Figure 6.5-(B) into grid form. All values are normalized to the range of *0 to 1*, where 1 refers to Highest Risk.

6.2.3 High Risk Areas

To identify temporally varying spatial risk for the inclusion of spatial context in the future tracing of contacts, multiple grids from Figure 6.6 are integrated into a single representation as shown in Figure 6.7. In order to classify the output to segment areas of higher or lower risk, risk scores are grouped into five classes with their labels corresponding to their intensity of risk. Five classes of risk are (0.50, 0.75, 1.00, 1.25, 1.50) with 1.50 referring to the highest value of risk. Review of this result shows that based on activities from the previous day (Figure 6.5), the highest risk area is at the centre top cell, whereas the surrounding areas are also of higher risk. While there is no spatial risk in the remaining part of the study area on this particular day, however, due to the temporally varying nature, the spatial risk may evolve in future.



1.0	0.5	0.5	0.5	0.5	0.5	0.75	1.0	1.5	0.75	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	1.0	0.75	0.75	0.75	1.25	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.75	0.5	0.75	0.75	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

FIGURE 6.7: Combining risk from multiple grids shown in Figure 6.6 into a single grid output. 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] as discussed in section 4.2.4, where '1' refers to the previous normal (existing SIR model with a constant spatial risk and all contacts are of equal nature)

6.3 Hybrid Spatio-SIR Model

This section illustrates the modification results of a Spatio-SIR model. First, the dynamic nature of spatial risk is shown along with its corresponding effect on graph matrix. Later, comparison of the Spatio-SIR model with baseline-SIR model is presented and discussed.

6.3.1 Dynamic Spatial Risk

Continuously evolving risk scores are illustrated in Figure 6.8. As risk scores are computed every day in Spatio-SIR model, each cell brings forth a new risk factor to the contacts associated with it. The values of risk follow the range of [0.5,1.5] where 1 refers to previous normal (SIR Model) and 0.5 means there was no activity in the cell on the previous day. Here, a horizontal line depicts that the cell had a similar activity on this day as it had on the previous day. In this sample set, cell 'B' remains at high-risk the most confirming high infectious activity.

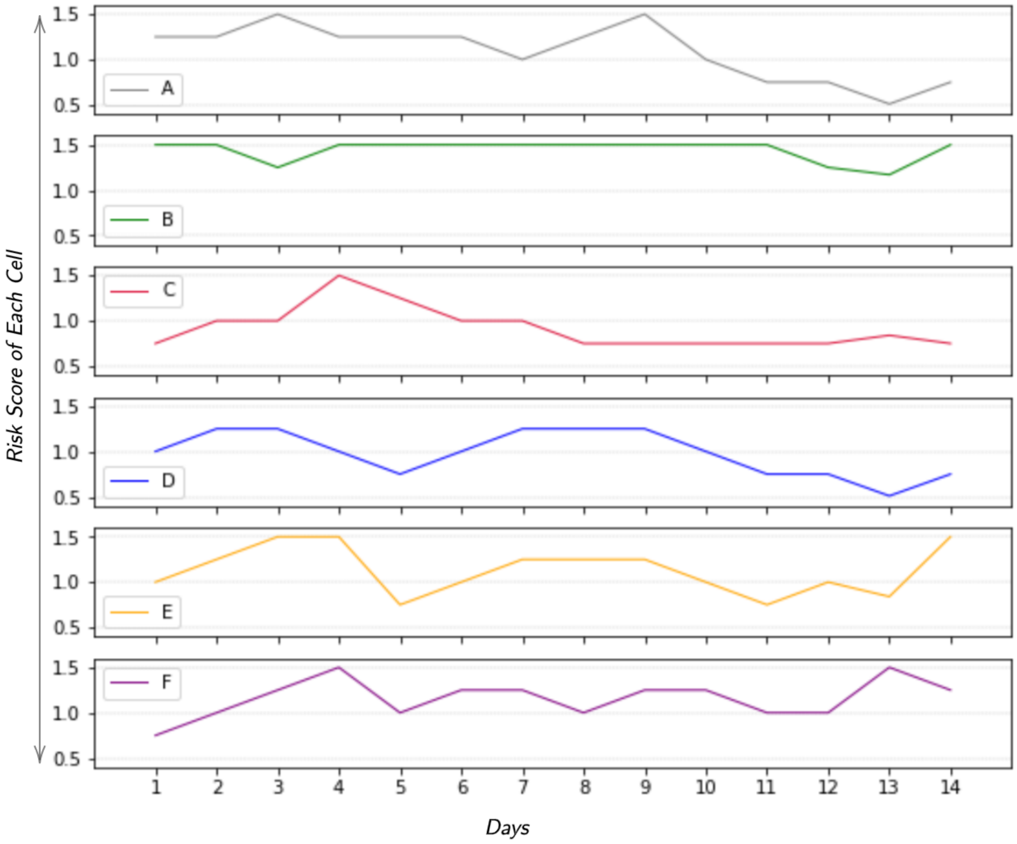


FIGURE 6.8: Temporally varying Spatial Risk for the first two weeks of epidemic is shown for a sample of six cells labelled as (*A, B, C, D, E & F*). Risk values vary between *0.5-1.5*, while there is a new value for every day.

The effect of varying spatial risk of a cell value is on the intensity of the contact which is presented in Figure 6.8. A modified graph (inclusive of spatial risk effect) results in continuous value of $\mathcal{K}_i$ (infectious contacts per individual), which were previously in discrete form of [0,1,2,..n].

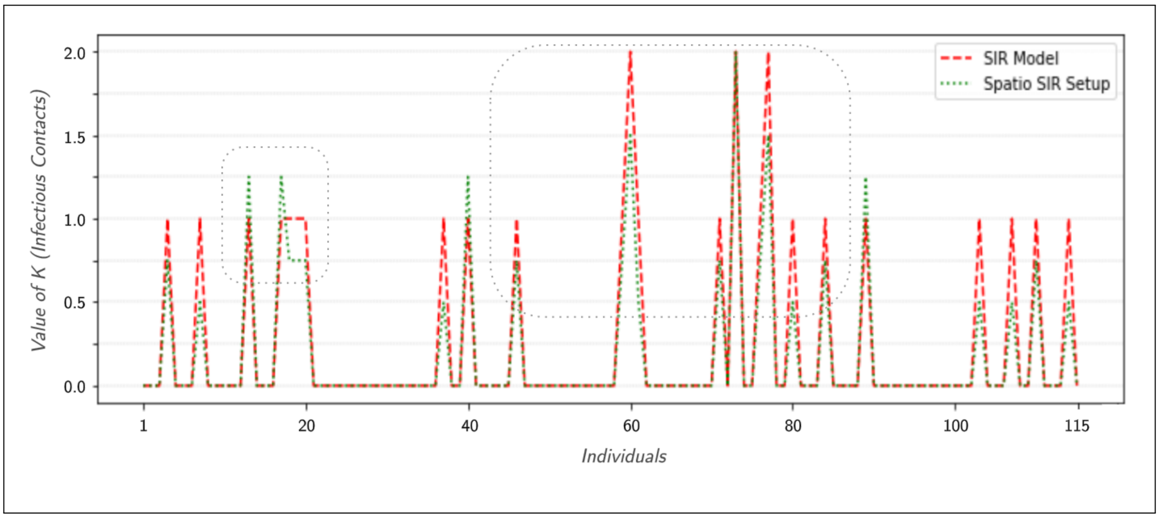


FIGURE 6.9: Comparison of $\mathcal{K}_i$ values of all individuals between Baseline-SIR model and the new Spatio-SIR model. *(Dashed red lines)* shows previous discrete numbers of infectious contacts from baseline-SIR model, whereas *(Green dotted lines)* presents the enhanced continuous values from the spatio-SIR model. Grey boxes focuses on the notion that for some individuals like (i=20), the SIR model considered a higher count of infectious contacts ($\mathcal{K}_{20}$=1) whereas Spatio-SIR consider less ($\mathcal{K}'_{20}$=0.75) as the contacts were of lower risk.

6.3.2 Spatio-SIR Model

Modification of baseline setup as in Spatio-SIR model is compared with the Baseline-SIR model in Figure 6.10. 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 Spatio-SIR model, there are more events of the population moving into *Quarantine Susceptible*. Though the trends of *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 not only reduces *Infected* and in turn *Recovered* and *Quarantine Infected*; but keeps more population susceptible as well. As after quarantine period, population resume their normal routine. 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

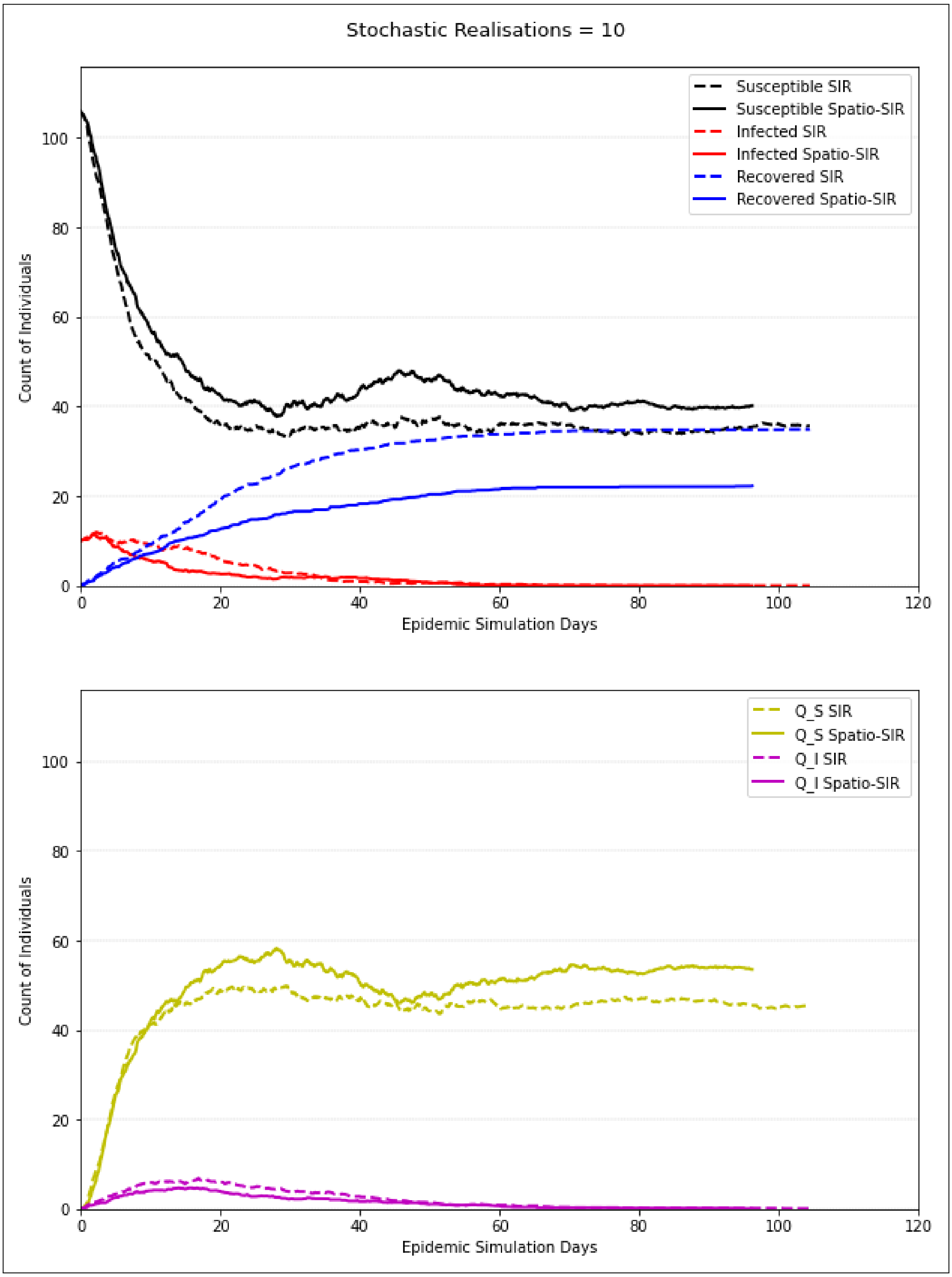


FIGURE 6.10: Comparing *Average* of 10 stochastic realisations of a disease outbreak scenario from Baseline-SIR *(dashed)* & Spatio-SIR *(solid)*. *(Top)* presents trends related to count of *Susceptible, Infected & Recovered*, whereas *(Bottom)* illustrates count of individuals in *Quarantine* related compartments. Count of total population is *115* which are represented over *Y-axis*.

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 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* in both, as with less *Infected* on the streets, the overall spread of infection is controlled, hence a lower count of *Quarantine Infected* individuals in *Spatio-SIR* compared to *Baseline-SIR*, apart from the start and end of an epidemic which is more or less similar.

6.3.3 Real World Scenario

This study also explores a real world scenario as introduced in experimental design (chapter 5), results of which are presented in Figure 6.11. A major difference is in the overall period of epidemic, where the *Intervention* setup executes an epidemic of *100* plus days considering the added spatial risk from day *11 to 20*, whereas in *Spatio-SIR* modelling 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 former, while total recovered individuals in latter are *11*. 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. Count of susceptible in quarantine in *Intervention* setup is *56* on day 20^{th}, whereas in *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.

6.4 Limitations

Even with the presence of many limitations, the persistence behind exploring individual-level contact tracing was the knowledge gap in the individual trajectory-level domain which have been exploited by the scenario of COVID-19. The implemented methodology

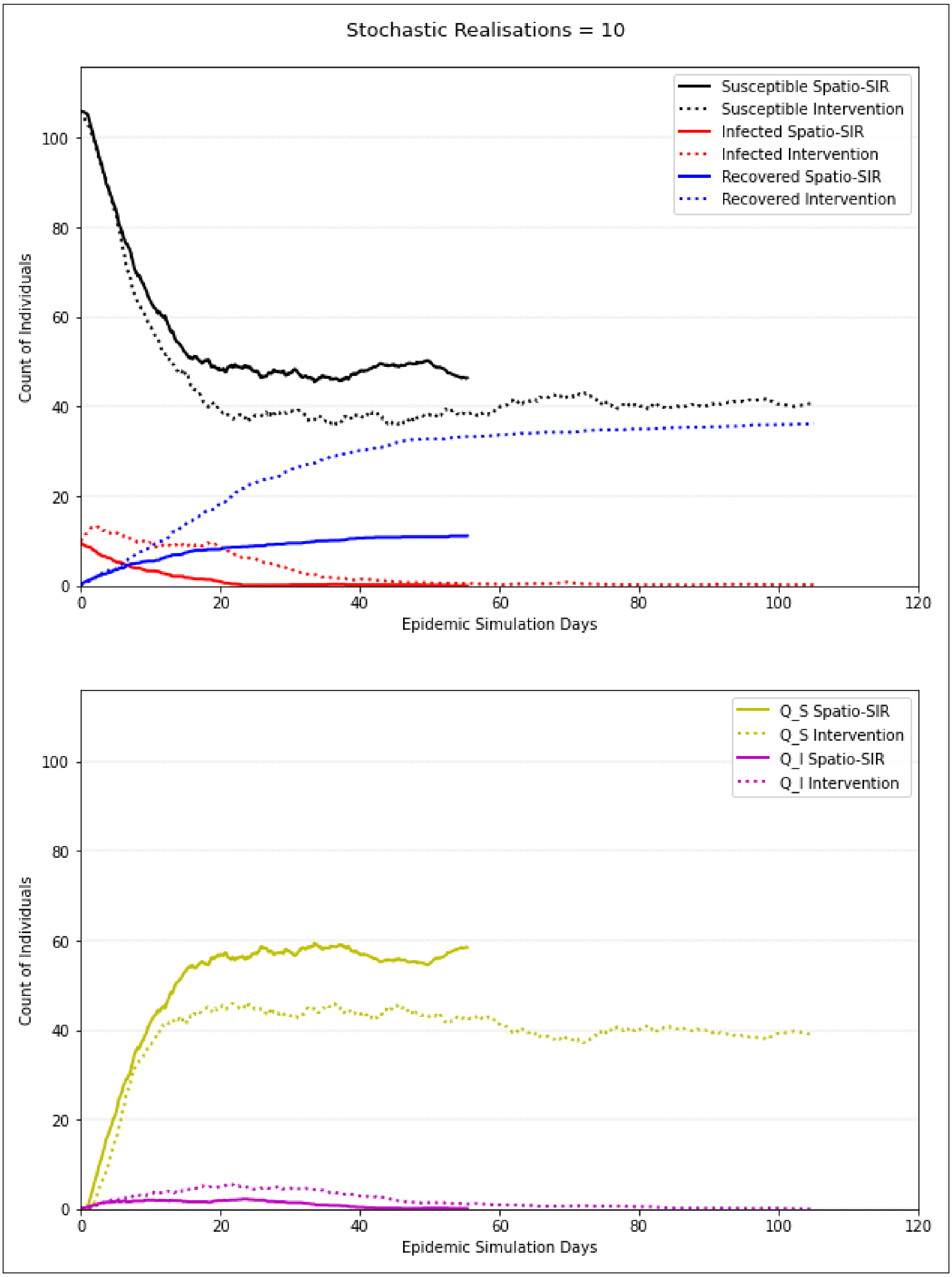


FIGURE 6.11: Comparing *Average* of 10 stochastic realisations of a disease outbreak scenario from Spatio-SIR *(solid)* and a case of Intervention - Spatio-SIR model with spatial risk of *1.5* from day *11 to 20 (dotted)*. *(Top)* presents trends related to count of *Susceptible, Infected & Recovered*, whereas *(Bottom)* illustrates count of individuals in *Quarantine* related compartments. Count of total population is *115* which are represented over *Y-axis*.

is an established proposal for future works with a real dataset and also in the domain of spatial risk. All limitations inclusive of one of the dataset are discussed as follows:

- **Infected Cases**: One limitation of the study is the actual information about infected individuals. With that, the proposed methodology can be configured to fit a model to data. After such fine-tuning, the model will be able to simulate any epidemic scenario, along with visualization of scenarios like varied tracing, outbreak etc. In this study, this limitation was handled through a self-induced initial outbreak.
- **Contact Graphs & Spatial Risk, per Day**: One of the 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 like hourly contacts graphs or a real-time application of tracing in terms of recording a contact as they happen can be followed for a higher accuracy.
- **Extrapolation of Trend lines**: In the representation of results, *Average* curves were obtained by extrapolating output trends of other realisations of shorter duration, an approach followed by (Hernández-Orallo et al., 2020). Even though the focus of this study is not on this aspect of representation, however, such an approach restricts observation of the actual trend of an epidemic with a longer duration.
- **Continuous Traces**: Another limitation is related to interpolation of trajectories, which were in the form of continuous traces of movements and were further interpolated to have the location of an individual at all times. A gap in such a continuous trace, when interpolated is like interpolating between the last point of the previous trip and the starting point of the next trip, which results in unrealistic movement paths. An ideal trajectory dataset will be in form of paths where each trip is identified and can be further self-interpolated to acquire data for all timestamps. In this study, this limitation was not handled for comparison with base-SIR.
- **Data Period**: *15* days recording of movements is an inadequate period for a long-standing scenario like an epidemic. In this study, this limitation was handled by concatenating the same dataset multiple time for 150 days. However, a better option would be to have a mobility dataset of a longer duration.
- **Data Scenario**: Another similar aspect is that the selected dataset, which is not of an epidemic scenario. As a real-world mobility pattern during an epidemic scenario is different from the mobility activities of non-epidemic days. A dataset from an era of an epidemic situation can assist in the analysis of such patterns and further explore its spatial risk.

6.5 Future Work

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. This section puts forth a list of recommended works for the future not only for the improvement of methodology but exploring additional possible scenarios.

- **Spatial Risk**: Spatial Risk is a complete domain in itself that includes 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 (Wang et al., 2017), demographic details and environmental factors, for the overall spatial risk assessment. Such a study will identify and explore the spatial effect of covariates in disease transmission by understanding their intrinsic underlying relationships, and to present a higher or lower score of risk.

- **Vulnerability Scores per Individual**: Implementation of this study was based on event-based stochastic SIR model where rates of each 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 vulnerability. Such a factor can be associated with each individual based on their movement in infectious places, the frequency of their contacts in general and exposure to infectious individuals. Though a semi-stochastic in nature, but a specific model like this will tend to have more critical events than only population moving in & out of quarantine. Nonetheless, such an implementation is subjected to the availability of validation data for its evaluation.

- **Spatio-SIR Tool**: 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 a Spatio-SIR modelling 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 which can support public health policymakers as and when required.

7 Conclusion

This study focused on digital epidemiology to explore existing contact tracing methods for a modification to include spatial context. A particular concentration was on how to associate spatial risk to each individual-level contact, that in an epidemiological model a riskier contact may have a higher possibility of disease transmission than the one which is of relatively less risk. As data about infection as well as the spatial risk was not available due to its invasive nature, the feasibility of this idea was developed in form of a framework that can serve as a tool for such spatio enhanced epidemiological analysis. The implementation included setting up a Baseline-SIR model, based on which spatial risk was identified which was further considered as the spatial context of future contacts. The results suggest that the new Spatio-SIR model tends to perform more meaningful events concerned with the Susceptible population rather than events to the Infected or Quarantined. With an example of a real-world scenario of induced spatial high-risk, it is highlighted that Spatio-SIR model can empower the analyst with a capability to explore disease dynamics from an additional perspective. The conclusions of this study as answers to the research questions are presented as follows:

- Implementing a disease outbreak scenario inclusive of spatial risk in SIR methodology in the form of Spatio-SIR model brings forth two aspects; one is its impact on the execution of disease scenario in terms of different output, and second is the capability of associating spatial component into infectious disease modelling.

  For the former, consideration of spatial risk is like increasing 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 in *Susceptible* compartment will either be infected or sent into quarantine, based on the chance element of stochastic event-based modelling. 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.

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

For the latter, this hybridization framework of spatial risk and disease modelling can act as a tool for policymakers to execute scenarios, visualizing the consequence of their decisions prior to their actual implementation. The potential of such an integration paves way for future research in the domain of 'spatial risk' as discussed in section 6.5.

- The explicitly induced spatial high-risk adversely affect the overall spread of infection. It not only results in higher count of *Infected* and *Recovered*, but also compels a longer epidemic, with greater proportion of population being sent into *Quarantine*. Such an induced spatial intervention application can assist in visualizing the impact of different scenarios and can be further related to socioeconomic factors.

The study proposes a generalized framework for Spatio-SIR modelling, however, a disease-specific model can be developed by adjusting the parameters available in Table 4.2. 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).

Overall, this study 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 (Benreguia et al., 2020). The study also highlights that accurate modelling of this sort is restricted due to the data unavailability (Tizzoni et al., 2014), and there is a critical requirement of datasets to ensure a practical application of the proposed approach. Besides, section 7.1 discusses the scenario in case desired data is available for the implementation of this Spatio-SIR model.

The author concludes 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 modelling

7.1 In Case of Data Availability

This section ends the thesis document with a discussion of a scenario, in case required datasets were accessible. The approach in such an ideal scenario is as follows:

- ***Dataset***: The desired dataset includes *(i)* Mobility trajectories & *(ii)* Infection information with timestamps. As in a real-world scenario, a detected infected person cannot be on the streets, hence the mobility period must be before the initial infection samples to evaluate their presence as infectious trajectories.

- ***Experimental Design***: As in any modelling process, the available dataset will be distributed into *Training Data* and *Test Data*, to train Baseline-SIR Model on training data and evaluate infection propagation simulation using the test data.

- ***Model Training***: Based on the same methodology of event-based stochastic SIR modelling, as already used in the study, the Baseline-SIR model would have been trained. However, due to the availability of actual infection information this time, model parameters as in Table 4.2. especially tracing efficiency $q^{'}$, probability of transmission of infection b and basic reproductive ratio $\mathcal{R}_0$ could have been configured to realistically simulate the actual scenario.

- ***Model Validation***: With the availability of validation data of infected individuals identified at the latter end of the data period, simulation of Baseline-SIR could have been related to it to evaluate how better the model captures the disease dynamics.

- ***Stochasticity***: It is important to highlight that due to the stochastic nature of simulation, the model is not expected to identify the correct infected individuals but to highlight a similar fraction of the infected population in general.

- ***Spatio-SIR Modelling***: In order to incorporate spatial risk into the newly proposed Spatio-SIR model, knowledge about spatial risk is required which is either not available or not offered publicly. If such information about places of high-risk is available even for a short duration, spatial risk assessment can be improved based on validation data which may help in improved results related to SIR Modelling.

As introduced in section 1.5, the recent accessibility of the first publicly available movement trajectories of COVID-19 infected individuals from Seoul, Korea (Park et al., 2021), is a motivating aspect that more real-world datasets will be publicly available offering a definite way forward for this study. Hence, in the meantime, the research must proceed in simulated environments and with available datasets to ensure there is continuous progress in learning infectious disease dynamics.

8 Appendix

8.1 Overview of Datasets

This section presents an overview of existing datasets related to individual-level mobility.

The BBC Four Pandemic

Year:	2018
Study Area:	Town of Haslemere, England
Application:	Pandemic Modelling
Description:	Pairwise distances of 496 volunteers recorded for three days every five minutes with up to one meter of spatial resolution
Reference:	Klepac et al., 2018
Limitation:	Lack of continuous information on mobility

GeoLife (Microsoft)

Year:	2007-2011
Study Area:	China and Europe
Application:	Trajectory Dataset
Description:	Trajectory movements recorded 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. Dataset in total contains 17,621 trajectories, total distance of 1,251,654 kilometers holding information of 48,203 hours
Reference:	Zheng et al., 2011
Limitation:	Trajectory form is available but (1) dataset is old, and (2) not suitable due to scarcity of movement in concentrated regions especially for high scale study of spatial risk

Copenhagen Networks Study

Year:	2019
Study Area:	Technical University of Denmark
Application:	Contact Networks
Description:	Network of physical proximity of 700 volunteers recorded for four weeks with information regarding Bluetooth signal strength, phone calls and social media friendship
Reference:	Sapiezynski et al., 2019
Limitation:	Lack of continuous information on mobility

Cellular Datasets

Dataset:	Orange
Application:	Disease Mitigation Strategies
Reference:	Rubrichi et al., 2018
Dataset:	Telenor
Application:	Human Mobility & Epidemics
Reference:	Wesolowski et al., 2015
Limitation:	Not available publicly

Social Media (Twitter)

Application:	Spatial Risk Modelling of Infectious Diseases
Reference:	Souza et al., 2018
Application:	High Risk Areas for Dengue
Reference:	Souza et al., 2019a
Limitation:	Lack of continuous information on mobility as well as for Contact Tracing

8.2 NCCU Trace Dataset

This section provide details related to study area and selected dataset.

8.2.1 Study Area

University campus of National Chengchi University (Taiwan).

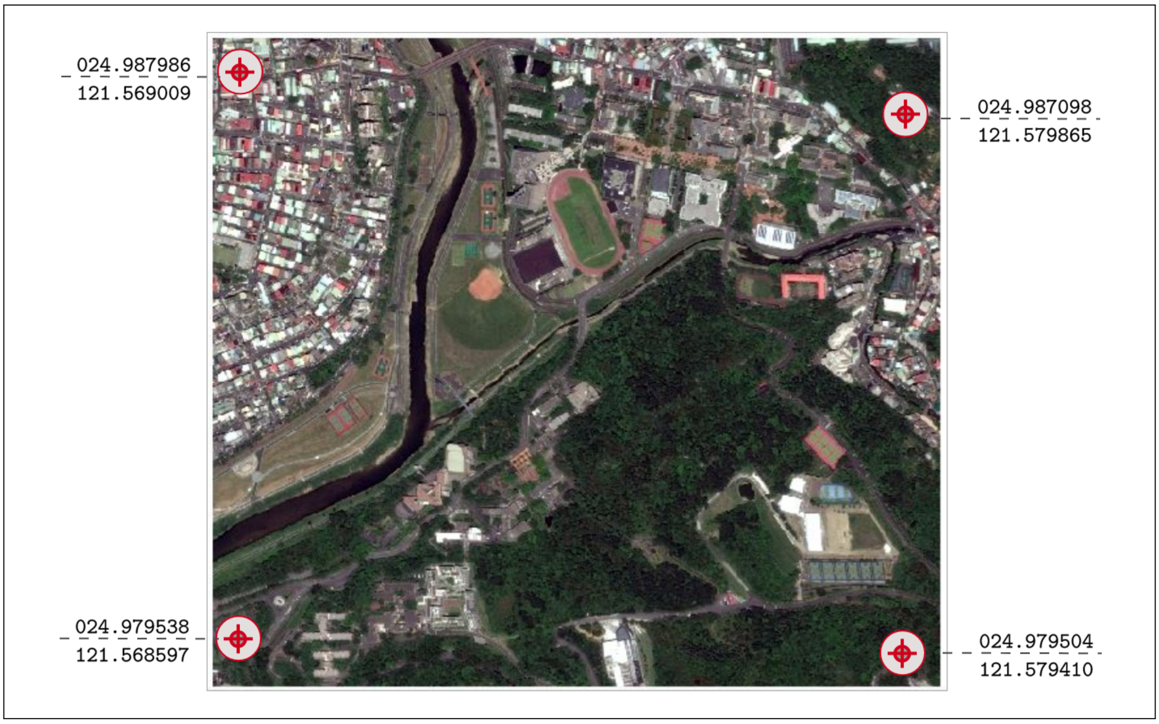


FIGURE 8.1: Coordinates of Study Area

8.2.2 Extent of Recorded Dataset

Figure 8.2 illustrates complete dataset of all 115 individuals for the period of 15 days where each user is shown with a different colour.

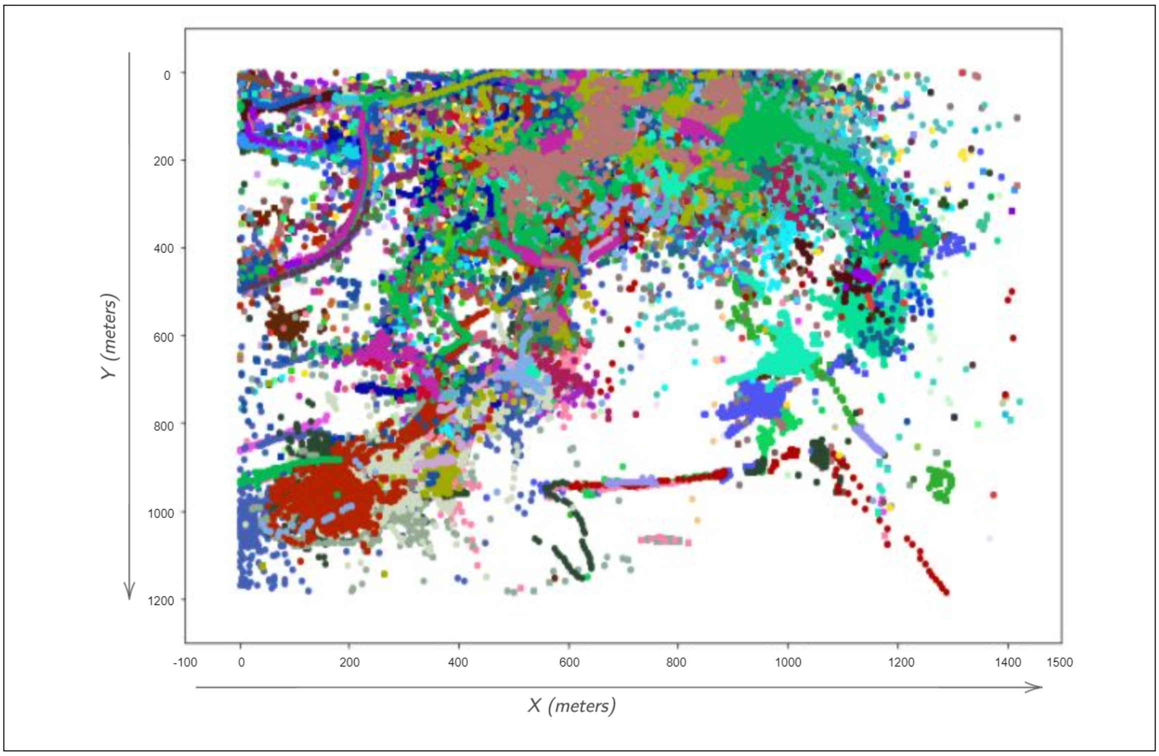


FIGURE 8.2: Extent of Recorded Dataset

8.2.3 Initial Location of Individuals

Figure 8.3 illustrates initial locations of individuals from the dataset to validate with NCCU (Tsai and Chan, 2015) provided snapshot of initial locations shown below.

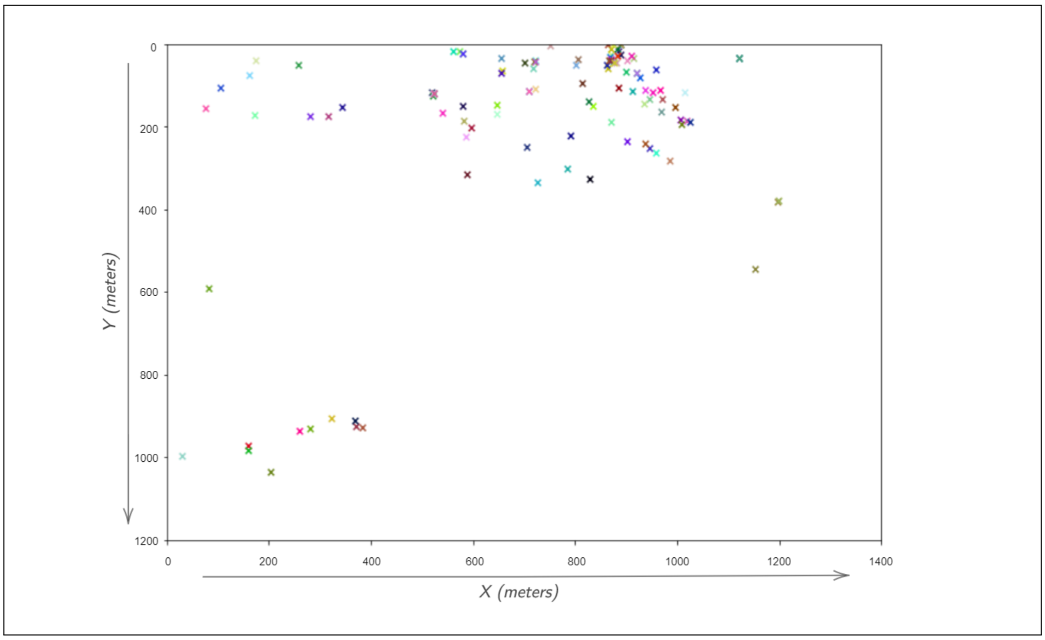
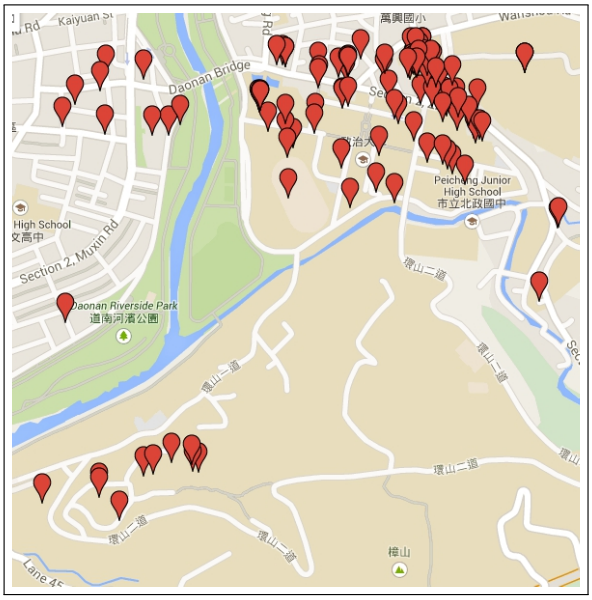


FIGURE 8.3: (Above) - Initial Locations of Individuals in Dataset (Bottom) - NCCU (Tsai and Chan, 2015) provided map



8.2.4 Mobility Trajectories

This section provides sample visualization of trajectory datasets for its understanding.

Single Individual - One Day

Figure 8.4 illustrates mobility trajectories of a single user for a period of one day.

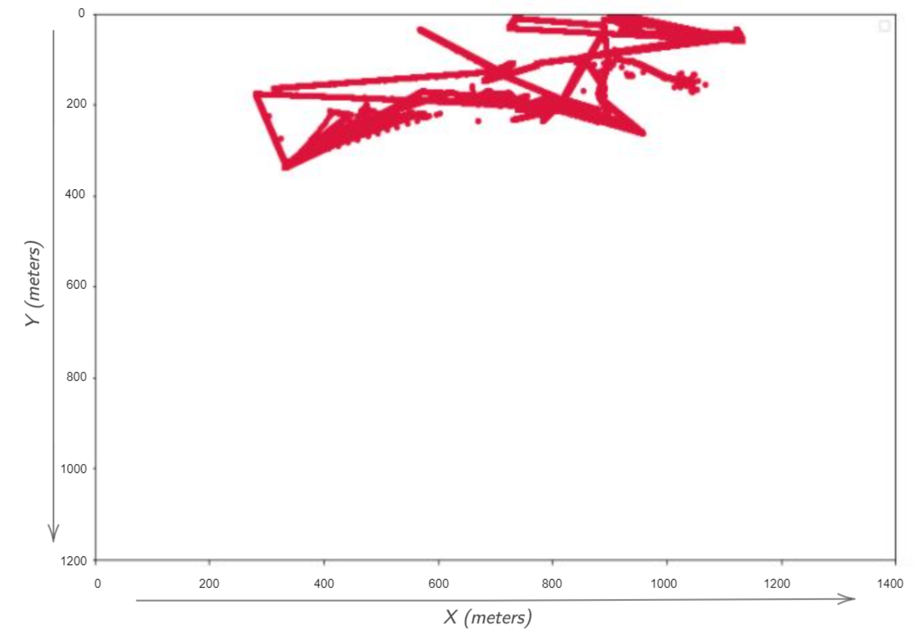


FIGURE 8.4: Mobility trajectory of a single user for 1 day period

Five Individuals - One Day

Figure 8.5 illustrates mobility trajectories of 05 users for first complete day.

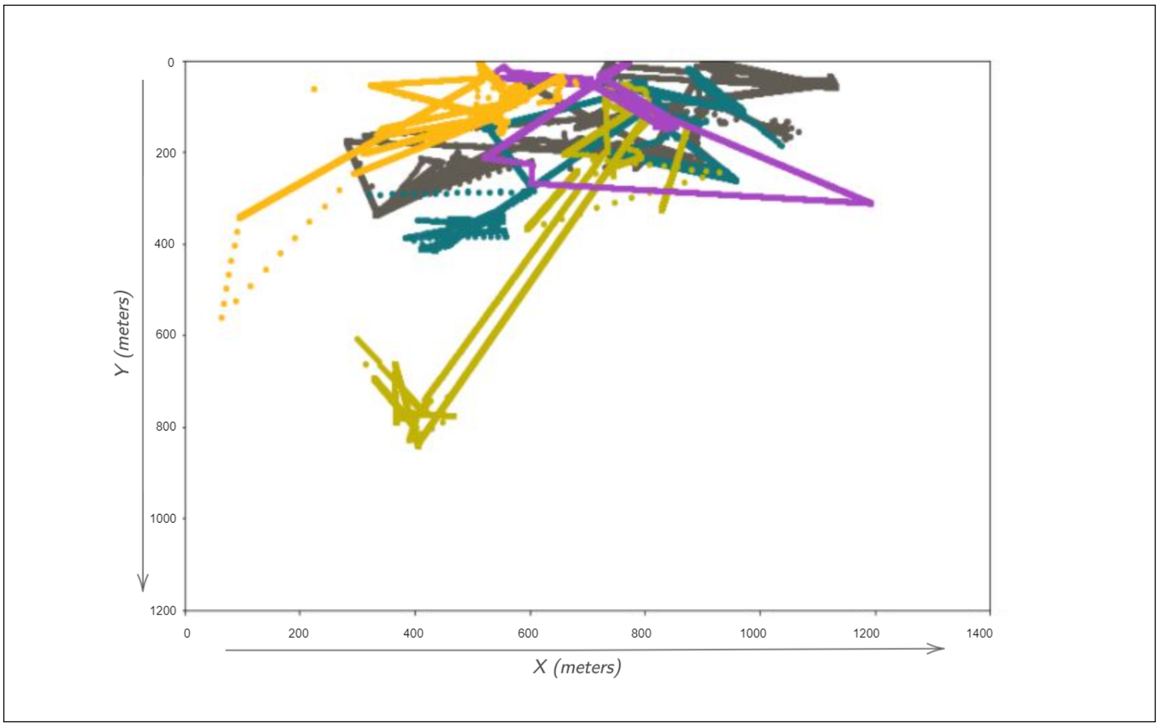


FIGURE 8.5: Mobility trajectory of five users for 1 day period

References

Amaya, J., Dupuis, R., Innocenti, M. E., & Lapenta, G. (2020). Visualizing and interpreting unsupervised solar wind classifications. *arXiv preprint arXiv:2004.13430*.

Anglemyer, A., Moore, T. H., 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).

Angulo, J., Yu, H.-L., Langousis, A., Kolovos, A., Wang, J., Madrid, A. E., & Christakos, G. (2013). Spatiotemporal infectious disease modeling: A bme-sir approach. *PloS one*, *8*(9), e72168.

Asan, U., & Ercan, S. An introduction to self-organizing maps. In: *Computational intelligence systems in industrial engineering*. Springer, 2012, pp. 295–315.

Bação, F., Lobo, V., & Painho, M. Self-organizing maps as substitutes for k-means clustering. In: *International conference on computational science*. Springer. 2005, 476–483.

Banos, A., Corson, N., Gaudou, B., Laperrière, V., & Coyrehourcq, S. R. (2015). The importance of being hybrid for spatial epidemic models: A multi-scale approach. *Systems*, *3*(4), 309–329.

Bardina, X., Ferrante, M., & Rovira, C. (2020). A stochastic epidemic model of covid-19 disease. *arXiv preprint arXiv:2005.02859*.

Barlow, N. (1991). A spatially aggregated disease/host model for bovine tb in new zealand possum populations. *Journal of applied ecology*, 777–793.

Basara, H. G., & Yuan, M. (2008). Community health assessment using self-organizing maps and geographic information systems. *International journal of health geographics*, *7*(1), 1–8.

Basole, R. C. The value and impact of mobile information and communication technologies. In: *Proceedings of the ifac symposium on analysis, modeling & evaluation of human-machine systems*. *9*. 2004, 1–7.

Benreguia, B., Moumen, H., & Merzoug, M. A. (2020). Tracking covid-19 by tracking infectious trajectories. *arXiv preprint arXiv:2005.05523*.

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.

Bobashev, G. V., Goedecke, D. M., Yu, F., & Epstein, J. M. A hybrid epidemic model: Combining the advantages of agent-based and equation-based approaches. In: *2007 winter simulation conference*. IEEE. 2007, 1532–1537.

Bradshaw, W. J., Alley, E. C., Huggins, J. H., Lloyd, A. L., & Esvelt, K. M. (2020). Bidirectional contact tracing dramatically improves covid-19 control. *medRxiv*.

Brockmann, D., David, V., & Gallardo, A. M. (2009). Human mobility and spatial disease dynamics. *Reviews of nonlinear dynamics and complexity*, *2*, 1–24.

Chakhar, S., & Mousseau, V. (2008). Spatial multicriteria decision making. *Encyclopedia of GIS*, *10*, 978–0.

Chang, E., Moselle, K. A., & Richardson, A. (2020). Covidsimvl–transmission trees, superspreaders and contact tracing in agent based models of covid-19. *medRxiv*.

Chen, D., Yang, Y., Zhang, Y., & Yu, W. (2020). Prediction of covid-19 spread by sliding mseir observer. *Science China Information Sciences*, *63*(12), 1–13.

Chiang, W.-H., Liu, X., & Mohler, G. (2020). Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates. *medRxiv*.

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*, 102202.

Dlamini, W. M., Dlamini, S. N., Mabaso, S. D., & Simelane, S. P. (2020). Spatial risk assessment of an emerging pandemic under data scarcity: A case of covid-19 in eswatini. *Applied Geography*, *125*, 102358.

Eames, K. T., & Keeling, M. J. (2003). Contact tracing and disease control. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1533), 2565–2571.

El-Doma, M. (1999). Analysis of an age-dependent sis epidemic model with vertical transmission and proportionate mixing assumption. *Mathematical and computer modelling*, *29*(7), 31–43.

Endo, A. et al. (2020). Implication of backward contact tracing in the presence of overdispersed transmission in covid-19 outbreaks. *Wellcome open research*, *5*.

Enright, J., & Kao, R. R. (2018). Epidemics on dynamic networks. *Epidemics*, *24*, 88–97.

Ferrante, M., Ferraris, E., & Rovira, C. (2016). On a stochastic epidemic seihr model and its diffusion approximation. *Test*, *25*(3), 482–502.

Gallagher, S., & Baltimore, J. Comparing compartment and agent-based models. In: *Joint statistical meeting, baltimore*. 2017.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, *81*(25), 2340–2361.

Gomes, D., Andrade, L., Ribeiro, C., Peixoto, M., Lima, S., Duque, A., Cirilo, T., Góes, M., Lima, A., Santos, M., et al. (2020). Risk clusters of covid-19 transmission in northeastern brazil: Prospective space–time modelling. *Epidemiology & Infection*, *148*.

Gonçalves, M., Costa, J., Netto, M., & Mwasiagi, J. (2011). Land-cover classification using self-organizing maps clustered with spectral and spatial information. *Self-Organizing Maps Applications and Novel Algorithm Design*, *1*, 299–322.

González, J. A., Rodríguez-Cortés, F. J., Cronie, O., & Mateu, J. (2016). Spatio-temporal point process statistics: A review. *Spatial Statistics*, *18*, 505–544.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *nature*, *453*(7196), 779–782.

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*, 1–7.

Gu, J., Yan, H., Huang, Y., Zhu, Y., Sun, H., Zhang, X., Wang, Y., Qiu, Y., & Chen, S. (2020). Better strategies for containing covid-19 epidemics—a study of 25 countries via an extended varying coefficient seir model. *medRxiv*.

Hau, B, & Kranz, J. Mathematics and statistics for analyses in epidemiology. In: *Epidemics of plant diseases*. Springer, 1990, pp. 12–52.

He, S., Peng, Y., & Sun, K. (2020). Seir modeling of the covid-19 and its dynamics. *Nonlinear Dynamics*, *101*(3), 1667–1680.

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

Hernández-Orallo, E., Manzoni, P., Calafate, C. T., & Cano, J.-C. (2020). Evaluating how smartphone contact tracing technology can reduce the spread of infectious diseases: The case of covid-19. *IEEE Access*.

Hethcote, H. W. The basic epidemiology models: Models, expressions for r0, parameter estimation, and applications. In: *Mathematical understanding of infectious disease dynamics*. World Scientific, 2009, pp. 1–61.

Hinton, G. E., Sejnowski, T. J., Poggio, T. A., et al. (1999). *Unsupervised learning: Foundations of neural computation*. MIT press.

Hsu, K.-l., Gupta, H. V., Gao, X., Sorooshian, S., & Imam, B. (2002). Self-organizing linear output map (solo): An artificial neural network suitable for hydrologic modeling and analysis. *Water Resources Research*, *38*(12), 38–1.

Jacob, C. (2010). Branching processes: Their role in epidemiology. *International journal of environmental research and public health*, *7*(3), 1186–1204.

Kalteh, A. M., Hjorth, P., & Berndtsson, R. (2008). Review of the self-organizing map (som) approach in water resources: Analysis, modelling and application. *Environmental Modelling & Software*, *23*(7), 835–845.

Keeling, M. J., & Eames, K. T. (2005). Networks and epidemic models. *Journal of the Royal Society Interface*, *2*(4), 295–307.

Keeling, M. J., & Rohani, P. (2011). *Modeling infectious diseases in humans and animals*. Princeton University Press.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, *115*(772), 700–721.

Klepac, P., Kissler, S., & Gog, J. (2018). Contagion! the bbc four pandemic–the model behind the documentary. *Epidemics*, *24*, 49–59.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, *43*(1), 59–69.

Krasznai, E. Á., Boda, P., Csercsa, A., Ficsór, M., & Várbíró, G. (2016). Use of self-organizing maps in modelling the distribution patterns of gammarids (crustacea: Amphipoda). *Ecological informatics*, *31*, 39–48.

Kresin, C., Schoenberg, F., & Mohler, G. Comparison of the hawkes and seir models for the spread of covid-19. In: 2020.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods*, *26*(6), 1481–1496.

Lang, J. C., De Sterck, H., Kaiser, J. L., & Miller, J. C. (2018). Analytic models for sir disease spread on random spatial networks. *Journal of Complex Networks*, *6*(6), 948–970.

Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: An introduction to data mining* (Vol. 4). John Wiley & Sons.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205.

Leitao, Á., & Vázquez, C. (2020). A stochastic seihrd model: Adding randomness to the covid-19 spread. *arXiv preprint arXiv:2010.15504*.

Lima, A., Pejovic, V., Rossi, L., Musolesi, M., & Gonzalez, M. (2015). Progmosis: Evaluating risky individual behavior during epidemics using mobile network data. *arXiv preprint arXiv:1504.01316*.

Lobo, V. J. Application of self-organizing maps to the maritime environment. In: *Information fusion and geographic information systems*. Springer, 2009, pp. 19–36.

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

Martinez-Beneito, M. A., Mateu, J., & Botella-Rocamora, P. (2020). Spatio-temporal small area surveillance of the covid-19 pandemics. *arXiv preprint arXiv:2011.03938*.

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

Mastrandrea, R., & Barrat, A. (2016). How to estimate epidemic risk from incomplete contact diaries data? *PLoS computational biology*, *12*(6), e1005002.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, *58*(8), 1246–1266.

Merriam-Webster. (2021). Definition of Risk in Merriam-Webster dictionary. Retrieved February 11, 2021, from https://www.merriam-webster.com/dictionary/risk

Müller, M., Derlet, P. M., Mudry, C., & Aeppli, G. (2020). Testing of asymptomatic individuals for fast feedback-control of covid-19 pandemic. *Physical biology*, *17*(6), 065007.

Nåsell, I. (1996). The quasi-stationary distribution of the closed endemic sis model. *Advances in Applied Probability*, 895–932.

Nijkamp, P., & van Delft, A. (1977). *Multi-criteria analysis and regional decision-making* (Vol. 8). Springer Science & Business Media.

Niu, R., Wong, E. W., Chan, Y.-C., Van Wyk, M. A., & Chen, G. (2020). Modeling the covid-19 pandemic using an seihr model with human migration. *IEEE Access*, *8*, 195503–195514.

Oliver, N., Letouzé, E., Sterly, H., Delataille, S., De Nadai, M., Lepri, B., Lambiotte, R., Benjamins, R., Cattuto, C., Colizza, V., et al. (2020). Mobile phone data and covid-19: Missing an opportunity? *arXiv preprint arXiv:2003.12347*.

Pampalk, E., Widmer, G., & Chan, A. (2004). A new approach to hierarchical clustering and structuring of data with self-organizing maps. *Intelligent Data Analysis*, *8*(2), 131–149.

Paolotti, D., Carnahan, A., Colizza, V., Eames, K., Edmunds, J, Gomes, G., Koppeschaar, C, Rehn, M., Smallenburg, R., Turbelin, C., et al. (2014). Web-based participatory surveillance of infectious diseases: The influenzanet participatory surveillance experience. *Clinical Microbiology and Infection*, *20*(1), 17–21.

Parham, P. E., & Ferguson, N. M. (2006). Space and contact networks: Capturing the locality of disease transmission. *Journal of the Royal Society Interface*, *3*(9), 483–493.

Park, J., Chang, W., & Choi, B. (2021). An interaction neyman-scott point process model for coronavirus disease-19. *arXiv preprint arXiv:2102.02999*.

Park, O., Park, Y., Park, S., Kim, Y., Kim, J., Lee, J., Park, E., Kim, D., Jeon, B., Ryu, B., Ko, D., Kim, E., Kim, H., Lee, H., Gwack, J., Jo, J., Lee, J., Hyun, J., Kim, J., . . . Yum,

M. (2020). Contact transmission of covid-19 in south korea: Novel investigation techniques for tracing contacts. *Osong Public Health and Research Perspectives*, (1). https://doi.org/10.24171/j.phrp.2020.11.1.09

Pearce, J. L., Waller, L. A., Mulholland, J. A., Sarnat, S. E., Strickland, M. J., Chang, H. H., & Tolbert, P. E. (2015). Exploring associations between multipollutant day types and asthma morbidity: Epidemiologic applications of self-organizing map ambient air quality classifications. *Environmental Health*, *14*(1), 1–13.

Pfeiffer, D., Robinson, T. P., Stevenson, M., Stevens, K. B., Rogers, D. J., Clements, A. C., et al. (2008). *Spatial analysis in epidemiology* (Vol. 142). Oxford University Press Oxford.

Pourghasemi, H. R., Pouyan, S., Heidari, B., Farajzadeh, Z., Shamsi, S. R. F., Babaei, S., Khosravi, R., Etemadi, M., Ghanbarian, G., Farhadi, A., et al. (2020). Spatial modeling, risk mapping, change detection, and outbreak trend analysis of coronavirus (covid-19) in iran (days between february 19 and june 14, 2020). *International Journal of Infectious Diseases*, *98*, 90–108.

Pribylova, L., & Hajnova, V. (2020). Seiar model with asymptomatic cohort and consequences to efficiency of quarantine government measures in covid-19 epidemic. *arXiv preprint arXiv:2004.02601*.

Purves, R. S., Laube, P., Buchin, M., & Speckmann, B. (2014). Moving beyond the point: An agenda for research in movement analysis with real data. *Computers, Environment and Urban Systems*, *47*, 1–4.

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

Renardy, M., & Kirschner, D. E. (2020). A framework for network-based epidemiological modeling of tuberculosis dynamics using synthetic datasets. *Bulletin of Mathematical Biology*, *82*(6), 1–20.

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

Rizoiu, M.-A., Mishra, S., Kong, Q., Carman, M., & Xie, L. Sir-hawkes: Linking epidemic models and hawkes processes to model diffusions in finite populations. In: *Proceedings of the 2018 world wide web conference*. 2018, 419–428.

Rubrichi, S., Smoreda, Z., & Musolesi, M. (2017). A comparison of spatial-based targeted disease containment strategies using mobile phone data. *arXiv preprint arXiv:1706.00690*.

Rubrichi, S., Smoreda, Z., & Musolesi, M. (2018). A comparison of spatial-based targeted disease mitigation strategies using mobile phone data. *EPJ Data Science*, *7*, 1–15.

Saaty, T. L. (2014). Analytic heirarchy process. *Wiley statsRef: Statistics reference online*.

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

Sapiezynski, P., Stopczynski, A., Lassen, D. D., & Lehmann, S. (2019). Interaction data from the copenhagen networks study. *Scientific Data*, *6*(1), 1–10.

Saveliev, A., Mukharamova, S., & Zuur, A. Analysis and modelling of lattice data. In: *Analysing ecological data*. Springer, 2007, pp. 321–339.

Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from gps trajectories and contextual information. *International Journal of Geographical Information Science*, *30*(5), 881–906.

Simmerman, J. M., Suntarattiwong, P., Levy, J., Gibbons, R. V., Cruz, C., Shaman, J., Jarman, R. G., & 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. https://doi.org/10.1086/656581

Soriano-Paños, D., Arias-Castro, J. H., Reyna-Lara, A., Martínez, H. J., Meloni, S., & Gómez-Gardeñes, J. (2020). Vector-borne epidemics driven by human mobility. *Physical Review Research*, *2*(1), 013312.

Souza, R. C., Neill, D. B., Assunção, R. M., & Meira, W. (2019a). Identifying high-risk areas for dengue infection using mobility patterns on twitter. *Online Journal of Public Health Informatics*, *11*(1).

Souza, R. C., Assunção, R. M., Neill, D. B., & Meira Jr, W. Detecting spatial clusters of disease infection risk using sparsely sampled social media mobility patterns. In: *Proceedings of the 27th acm sigspatial international conference on advances in geographic information systems*. 2019, 359–368.

Souza, R. C., Assunção, R. M., Neill, D. B., Silva, L. G., & Meira Jr, W. (2018). Spatial risk modeling for infectious disease surveillance using population movement data.

Takács, B., & Hadjimichael, Y. (2019). High order discretization methods for spatial dependent sir models. *arXiv preprint arXiv:1909.01330*.

Tizzoni, M., Bajardi, P., Decuyper, A., King, G. K. K., Schneider, C. M., Blondel, V., Smoreda, Z., González, M. C., & Colizza, V. (2014). On the use of human mobility proxies for modeling epidemics. *PLoS Comput Biol*, *10*(7), e1003716.

Tsai, T.-C., & Chan, H.-H. (2015). Nccu trace: Social-network-aware mobility trace. *IEEE Communications Magazine*, *53*(10), 144–149.

Udeagu, C., Bocour, A, Ramos, Y, et al. Bringing sexually-transmitted disease (std) contact tracing into the age of social media and mobile connectivity. In: *Annual conference on youth+ tech+ health, yth live*. 2013.

Ultsch, A. Self-organizing neural networks for visualisation and classification. In: *Information and classification*. Springer, 1993, pp. 307–313.

Ultsch, A. (2003). U*-matrix: A tool to visualize clusters in high dimensional data.

Van Doremalen, N., Bushmaker, T., Morris, D. H., Holbrook, M. G., Gamble, A., Williamson, B. N., Tamin, A., Harcourt, J. L., Thornburg, N. J., Gerber, S. I., et al. (2020). Aerosol and surface stability of sars-cov-2 as compared with sars-cov-1. *New England Journal of Medicine*, *382*(16), 1564–1567.

Villmann, T., & Bauer, H.-U. (1998). Applications of the growing self-organizing map. *Neurocomputing*, *21*(1-3), 91–100.

Wang, P., Fu, Y., Liu, G., Hu, W., & Aggarwal, C. Human mobility synchronization and trip purpose detection with mixture of hawkes processes. In: *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 2017, 495–503.

Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., Engø-Monsen, K., & Buckee, C. O. (2015). Impact of human mobility on the emergence of dengue epidemics in pakistan. *Proceedings of the National Academy of Sciences*, *112*(38), 11887–11892.

West, D. B. et al. (1996). *Introduction to graph theory* (Vol. 2). Prentice hall Upper Saddle River, NJ.

Westerlund, M. L. (2005). Classification with kohonen self-organizing maps. *Soft Computing, Haskoli Islands*, *24*.

WHO. (2021). WHO Coronavirus Disease (COVID-19) Dashboard | WHO Coronavirus Disease (COVID-19) Dashboard. Retrieved January 31, 2021, from https://covid19.who.int/

Yager, R. R., & Kacprzyk, J. (2012). *The ordered weighted averaging operators: Theory and applications*. Springer Science & Business Media.

Yamana, T., Pei, S., & Shaman, J. (2020). Projection of covid-19 cases and deaths in the us as individual states re-open may 4, 2020. *medRxiv*.

Yang, W., Cao, Q., Qin, L., Wang, X., Cheng, Z., Pan, A., Dai, J., Sun, Q., Zhao, F., Qu, J., et al. (2020). Clinical characteristics and imaging manifestations of the 2019 novel coronavirus disease (covid-19): A multi-center study in wenzhou city, zhejiang, china. *Journal of Infection*.

Zhang, J., Shi, H., & Zhang, Y. Self-organizing map methodology and google maps services for geographical epidemiology mapping. In: *2009 digital image computing: Techniques and applications*. IEEE. 2009, 229–235.

Zheng, Y, Fu, H, Xie, X, Ma, W., & Li, Q. (2011). Geolife gps trajectory dataset-user guide. microsoft research.

Zheng, Y. (2015). Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *6*(3), 1–41.

Zhou, C., Yuan, W., Wang, J., Xu, H., Jiang, Y., Wang, X., Wen, Q. H., & Zhang, P. (2020). Detecting suspected epidemic cases using trajectory big data. *arXiv preprint arXiv:2004.00908*.