

Epidemic Spread Modeling for COVID-19 Using Hard Data

Anna Schmedding

Williamsburg, VA, USA

Master of Science in Mathematics, Syracuse University, 2019  
Bachelor of Science in Computer Science and Mathematics, York College of  
Pennsylvania, 2016

A Thesis presented to the Graduate Faculty of  
The College of William & Mary in Candidacy for the Degree of  
Master of Science

Department of Computer Science

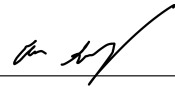
College of William & Mary  
May 2021



## APPROVAL PAGE

This Thesis is submitted in partial fulfillment of  
the requirements for the degree of

Master of Science



---

Anna Schmedding

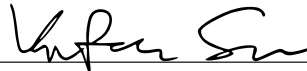
Approved by the Committee, May 2021



---

Committee Chair

Evgenia Smirni, Sidney P. Chockley Professor, Computer Science  
College of William & Mary



---

Yifan Sun, Assistant Professor, Computer Science  
College of William & Mary



---

Adwait Nadkarni, Assistant Professor, Computer Science  
College of William & Mary

## ABSTRACT

We present an individual-centric model for COVID-19 spread in an urban setting. We first analyze patient and route data of infected patients from January 20, 2020, to May 31, 2020, collected by the Korean Center for Disease Control & Prevention (KCDC) and illustrate how infection clusters develop as a function of time. This analysis offers a statistical characterization of mobility habits and patterns of individuals. We use this characterization to parameterize agent-based simulations that capture the spread of the disease, we evaluate simulation predictions with ground truth, and we evaluate different *what-if* counter-measure scenarios. Although the presented agent-based model is not a definitive model of how COVID-19 spreads in a population, its usefulness, limitations, and flexibility are illustrated and validated using hard data.

## TABLE OF CONTENTS

Acknowledgments	iii
List of Tables	iv
List of Figures	v
1 Introduction	2
2 The KCDC Data Set	6
3 Data Analysis	8
3.1 Visited Locations . . . . .	8
3.2 Seoul Population . . . . .	9
3.3 Patient Connections . . . . .	10
3.4 Super Spreaders . . . . .	11
3.5 Daily Traveled Distance . . . . .	12
3.6 Patient Mobility . . . . .	13
3.7 Irresponsible Behaviors . . . . .	15
4 Agent-based Model	16
5 Model Validation and Case Study	21
5.1 Validation . . . . .	21
5.2 Applying mitigation measures . . . . .	24

6	Discussion and Limitations	26
7	Related Work	28
8	Conclusions and Ongoing Work	31
	Bibliography	33

## ACKNOWLEDGMENTS

I would like to thank everyone who helped me with this thesis. In particular, the guidance and support of my advisor Professor Evgenia Smirni has been indispensable to my success. My collaborators on this project, Dr. Riccardo Pincioli and Lishan Yang have also provided me with valuable advice and assistance throughout our time working on this research, and none of this would have been possible without them. I would also like to thank my other committee members, Professor Yifan Sun and Professor Adwait Nadkarni for their time and assistance on this thesis.

Next, I would like to thank the Computer Science Department and administrative team, especially Vanessa Godwin, Jacquelyn Johnson, and Dale Hayes for their care and support, especially through these difficult times in the COVID-19 pandemic.

Finally, I would like to thank my family and friends for the immeasurable love and support they have provided me.

## LIST OF TABLES

2.1	Number of (unique) entries of PatientInfo and PatientRoute, two of the three data sets used in this thesis. . . . .	7
-----	---	---



## LIST OF FIGURES

3.1	Heat maps of most visited locations. . . . .	9
3.2	Mobility of Seoul population over time by age group according to cell-phone data provided by SK telecom. . . . .	9
3.3	Patient contacts. . . . .	10
3.4	Contact Degree CDF. . . . .	11
3.5	Infection spread subgraph: Red nodes indicate patients with route information who infected others. Green nodes indicate patients who infected others but do not have any route information. Blue nodes indicate patients who did not infect anyone else. . . . .	12
3.6	Super spreader analysis. . . . .	12
3.7	Daily traveled distance and visited locations. . . . .	13
3.8	Patient unique locations. . . . .	14
3.9	Patient mobility. . . . .	14
3.10	Irresponsible behavior of sick patients. . . . .	15
4.1	Life cycle of an agent. . . . .	18
4.2	Simulation screenshot. . . . .	18
4.3	Simulating patient isolation. . . . .	19
4.4	Percentage of active agents while infected. . . . .	19
4.5	Percentage of isolated population. . . . .	20
5.1	Movements of Gangnam and Seocho residents. . . . .	21

5.2	Infected population in the validation simulation. The overlap of two simulation cases with the ground truth retrieved from the data set validates the simulation settings. Results are presented with 95% confidence intervals (error margins are give by the colored ranges).	22
5.3	Hotspots in the data set (ground truth) and model.	23
5.4	Effect of different counter-measures. Results are presented with 95% confidence intervals (shaded areas).	24

## Epidemic Spread Modeling for COVID-19 Using Hard Data

# Chapter 1

## Introduction

On March 11, 2020, the WHO declared COVID-19 the first pandemic caused by a coronavirus [6]. Since then, prediction of the spread of the disease became a critical guide of public health policy. A tremendous amount of data is collected to help policy decisions that can limit the spread of COVID-19. For example, Google provides time-series data of infections at a coarse granularity<sup>1</sup> (i.e., as a function of the area's population, no information is provided at the granularity of single individuals). Epidemiological simulation and mathematical models have been used to predict the spread of the disease. Typically, model effectiveness is tied to its input parameterization.

In this thesis, we use data provided by the Korean Center for Disease Control (KCDC) and local governments during the first wave of the disease in South Korea. In contrast to the Google data, the KCDC data focus on individual patients and allow the development of an individual-centric model of the COVID-19 epidemic. Infected individuals are monitored and their movements are logged using CCTV, cellphones, and credit card transactions [17]. The KCDC records patient movements in plain text (i.e., natural language) without any unified rule. These logs are parsed through automated code and rule-based methods to extract keywords that are then used with web mapping service APIs (e.g., Google Maps [1], Kakao Map [2], or Naver Map [3]) to extract geographical coordinates (i.e., latitude and

---

<sup>1</sup><https://console.cloud.google.com/marketplace/product/bigquery-public-datasets/covid19-open-data>

longitude) and other data. The parsed logs are made publicly available [25] and being collected by KCDC are deemed trustworthy.

To the best of our knowledge, the KCDC logs are the only publicly available data that contain patient-centric information in great detail: they report on the patient mobility, i.e., traveled distance and the sequence of locations visited on a daily basis, the date of the onset of symptoms, whether and when the patient got in contact with other patients that are also diagnosed. The KCDC data set remains a valuable resource for studying the spread of COVID-19, yet it presents some limitations:

- South Korea has a small number of COVID-19 cases (i.e., 81,185 on February 7, 2021) compared to other countries, and the last version of the KCDC data set contains data collected up to May 31, 2020 (the KCDC data set has not been updated since then). By May 31, approximately 11,500 COVID-19 cases were confirmed in South Korea [17, 28], but only 35% of them have been logged into the data set.
- Some locations visited by patients are not recorded due to privacy concerns. Consequently, patient infection information and route data do not always coincide. For example, there are patients that infect each other even if their routes do not cross. This may happen when patients belong to the same household (locations where people live are rarely logged).
- Patient and route data may be incomplete (i.e., some attributes are occasionally missing, such as the type of locations visited by some patients) and require manual completion before analyzing the data set.
- There is route data information for only a portion of the patients. Patient movement has been logged only for the 15% of all confirmed cases by May 31. Because of privacy concerns, this data set is no longer publicly available.

We adopt different strategies to address the above challenges. We have manually retrieved certain missing attributes: in the case of patient routes with missing location type (e.g.,

store, school, hospital, airport), we use the provided geographical coordinates to retrieve the visited location and identify its type. Regretfully, some missing data are not possible to recover.

Specifically, provided that the mobility of only the 15% of confirmed patients are logged in detail, we can only “guess” the pertinent information of the remaining patients assuming that their mobility is independent and identically distributed to the 15% of patients with detailed logs. We content that while detailed logs provide data of statistical significance, their usage introduces some unavoidable bias towards the percentage of patients who voluntarily shared more information than others. Yet, statistical information derived from histograms (i.e., processed data) fill-in the gaps of missing information and can be used as input of patient activity in the simulation. We point out that our analysis and processing of this portion of the data was made *before* the detailed movement logs became unavailable. Here, we use this processed data in the form of histograms (and also make them available to the community), see the supplement for information on the simulation tool and its input data.

We use logs and histograms to feed a patched version of GeoMason [35], a tool that uses agent-based models (ABM) and geographic information systems (GIS). GeoMason has been used to study disease outbreaks (e.g., a cholera outbreak is studied using this tool in [14]). We simulate interactions of thousands of people in the Gangnam and Seocho districts of Seoul on roads and in buildings to investigate the COVID-19 outbreak in the largest metropolis of South Korea and evaluate different *what-if* mitigation scenarios. We validate the results of the simulation with the ground truth derived from the KCDC logs. This tool offers a flexible model based on real-world COVID-19 spread information and can be used to facilitate evaluation of different mitigation measures and patient behaviors. Our contributions and outline of this thesis are:

- We analyze and connect hard data from various KCDC logs to extract information on detailed patient movements (Chapters 2 and 3). Missing information is manually retrieved, when possible.

- We provide statistical analysis of population movements and habits in the form of histograms.

- We parameterize an agent-based model that uses the KCDC data as input, see Chapter 4, and its flexibility to capture a variety of conditions is outlined. The simulation tool and processed data will be open sourced, see the appendix for details.

- The simulation model is validated in Chapter 5 and its usage and limitations are discussed in Chapter 6.

## Chapter 2

# The KCDC Data Set

Here, the KCDC data sets are described. The data sets [25] used in this thesis contain data collected by the KCDC and local governments from January 20, 2020, to May 31, 2020. The PatientInfo and PatientRoute data sets contain information and routes of COVID-19 patients in Seoul, respectively. The amount of data in each data set is shown in Table 2.1. The number of (healthy and sick) people moving across Seoul districts are also provided in the SeoulFloating data set. This data has been collected using the Big Data Hub of SK Telecom, a Korean wireless telecommunications operator.

**PatientInfo data set.** This data set provides epidemiological data of COVID-19 patients. It contains 4004 different entries, each entry represents a different patient identified by an ID (*patient\_id*). Other attributes include their gender and age, their provenance (*country*, *province*, and *city*), whether they have been infected in a known case (*infection\_case*, e.g., overseas inflow or contact with patient) and the ID of the patient that infected them (*infected\_by*), the number of people that the patient came in contact with (*contact\_number*), and the date of their first symptoms (*symptom\_onset\_date*). This data set is also described in [26].

**PatientRoute data set.** This data set is no longer publicly available. We retrieved this data set from the Kaggle repository [25] that contained trace data collected up to May 31, 2020. This data set contains 8092 entries, each one reporting a visit (to one of



**Table 2.1:** Number of (unique) entries of PatientInfo and PatientRoute, two of the three data sets used in this thesis.

	PatientInfo	PatientRoute
Total entries	4004	8092
Unique patients	4004	1472
Unique locations	–	2992
Unknown location <i>type</i>	–	2341

2992 unique locations) of 1472 (out of 4004) unique South Korean COVID-19 patients logged in the PatientInfo data set. A location is unequivocally identified by its *latitude* and *longitude*. *Province*, *city*, and *type* (e.g., airport, hospital, store) of each location are also provided. The attribute *type* of almost 30% of entries is set to *etc* (i.e., locations that cannot be identified using the rule-based approach of [25]). We manually look for their type using their geographical coordinates and OpenStreetMap [4] to compensate for this lack of data. Each entry also contains the patient (identified by *patient\_id*, the same as in the PatientInfo data set, and by *global\_num*, another ID used only in this data set) that visited the location on a specific *date*. The time spent in the location is not available. Locations visited by a patient in a single day are logged in chronological order.

**SeoulFloating data set.** This data set provides hourly data of people moving across Seoul districts. Data are collected from January 1 to May 31, 2020, by SK Telecom. Collected data are grouped by *gender*, *age*, and *district* and allows visualizing the movement of people in Seoul during this period. Age is provided at the decade granularity for people in their 20s through 70s. No information is provided for children or for people who are 80 or older. As a result, it is not possible to conclude on infections at education facilities or directly model mitigation measures that include school closings. This data set reports data on the *entire* Seoul population, not just the COVID-19 patients, and only considers those with cell phones.

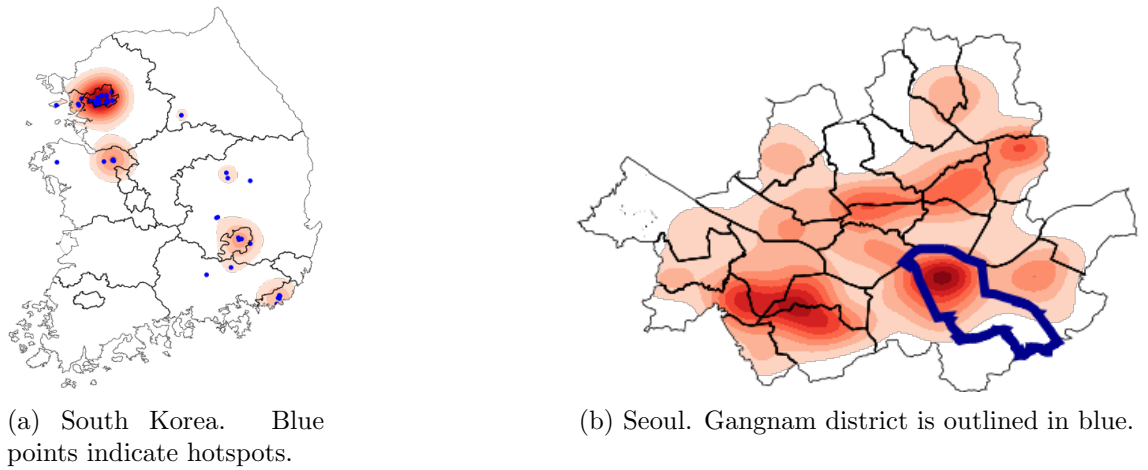
## Chapter 3

# Data Analysis

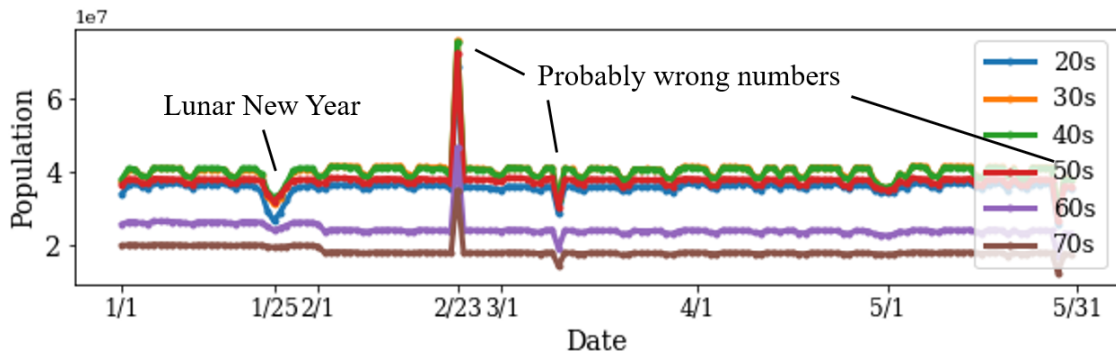
Although the information contained in the KCDC data sets is not as accurate as one would like, it still allows for the analysis of patient movements and interactions with high accuracy. In this chapter, we discuss information that we extract from the data sets and how it is used to parameterize the GeoMason ABM tool [35].

### 3.1 Visited Locations

Figs. 3.1(a) and 3.1(b) depict a heat map of the most visited locations in South Korea and Seoul, respectively, showing where COVID-19 outbreaks are more likely to happen. Heat maps in Fig. 3.1 also show the South Korean cities for which movement data are recorded. Visibly, Seoul is the city with the most visited locations. Within Seoul, the south-west and south-east areas are those with more patient routes. The financial district and company head-quarters are located in the south-west part of the city. The south-east region corresponds to the Gangnam district, outlined in blue in Fig. 3.1(b). Many shopping and entertainment centers are located in Gangnam.



**Figure 3.1:** Heat maps of most visited locations.



**Figure 3.2:** Mobility of Seoul population over time by age group according to cell-phone data provided by SK telecom.

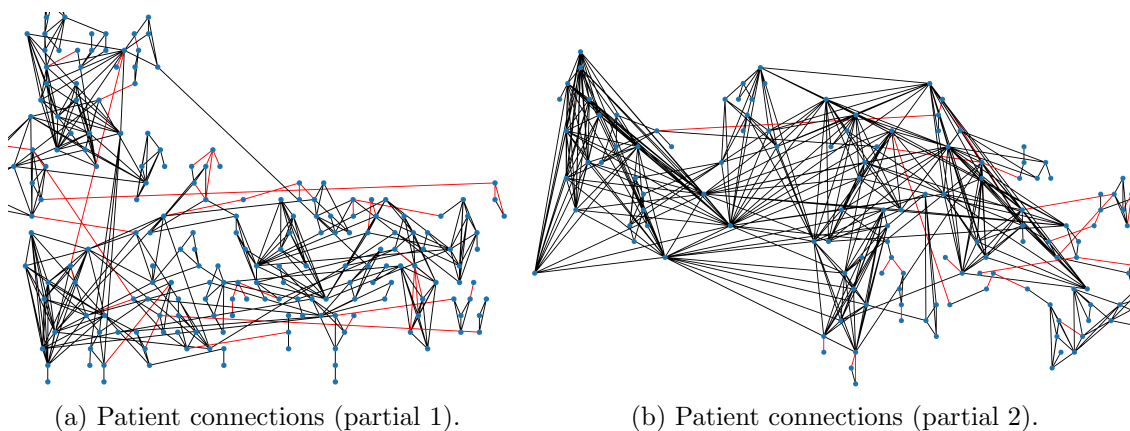
## 3.2 Seoul Population

Since Seoul has more logs in PatientRoute as shown in Fig. 3.1a, we analyze its population habits from January 1, 2020, to May 31, 2020, and extract information to determine how to put residents in different classes to model population movements. Fig. 3.2 depicts the population (grouped by age) of both healthy and sick people moving in Seoul on a per-day basis. Two clear classes of people are identified depending on their mobility: people that are 20 – 50 years old (adults) and those that are 60 – 70 (seniors). The first group has higher mobility within the city during week days, but this mobility decreases during weekends. The second group (seniors) does not have any discernible change in mobility

patterns during the week. A dip for the adult class observed on January 25 corresponds to the lunar new year day, no such dip is observed for the senior class. Perhaps because of the pandemic onset in South Korea and KCDC advice, we observe the mobility of seniors to decrease starting at the beginning of February.

### 3.3 Patient Connections

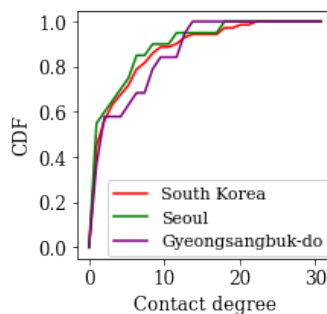
Figs. 3.3(a) and 3.3(b) present a subgraph of patient connections (to improve visibility, we only present a small portion of the entire graph). Here, nodes depict patients, black edges connect patients that visited the same place during the same day, and red edges represent the virus spreading information obtained from the PatientInfo data set (i.e., *infected\_by* attribute). Some red edges do not overlap with black edges. This means that, even if one of the two nodes connected by the red edge infected the other, no connections (i.e., visits to the same location during the same day) have been recorded in the data set. The node degree in Figs. 3.3(a) and 3.3(b) shows the contact degree among patients and illustrates visually the complexity of the problem.



**Figure 3.3:** Patient contacts.

Fig. 3.4 shows a summary view of patient connections: the contact degree CDF of all patients for the entire dataset. Three CDFs are shown: one for the whole South Korea, one for Seoul, and another one for the Gyeongsangbuk-do province. Interestingly, all

CDFs have a similar shape. High contact degrees indicate potential super spreaders (i.e., patients that infect many other people). People who come into contact with many others are not necessarily super spreaders since it is unknown whether or not they were sick or healthy when contact occurred. Because of this, further analysis is required to determine whether or not a patient is a super spreader.



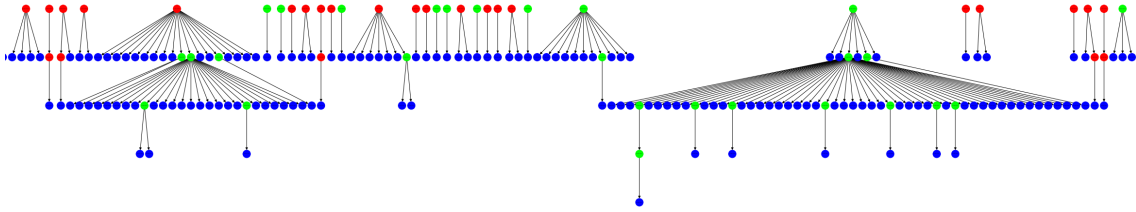
**Figure 3.4:** Contact Degree CDF.

### 3.4 Super Spreaders

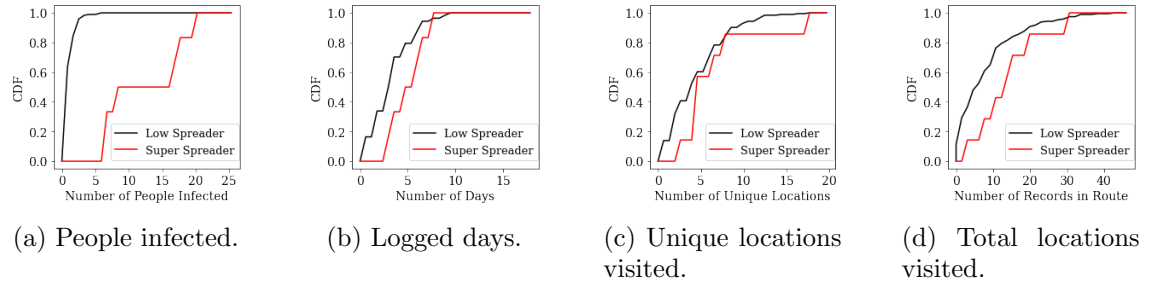
Fig. 3.5 illustrates a subset of patients where the *infected\_by* relationship (i.e., patient A is infected by patient B) is *known* from the PatientInfo data set. The entire graph contains 1052 patient nodes and 822 edges representing the known infection spread. For the sake of visibility, we present just a data subset. Red nodes correspond to individuals with available route information who are known to have infected others, green nodes correspond to individuals who infected others but have no available route information, and blue nodes correspond to patients who are not known to have infected others. This particular subset shows a mix of super spreaders (i.e., people who infected more than six people) and low spreaders, who infected six or fewer people<sup>1</sup>. The large “fans” in this figure are indicative of super spreaders. The different behaviors of super/low spreaders are shown in Fig. 3.6. Super spreaders account for 3.59% and low spreaders account for

<sup>1</sup>We define a “super spreader” as someone who infects at least 6 people. This allows us to divide the data set to obtain the most noticeable difference in patient behavior (number of locations, number of days, number of records).

the remaining 96.41% of patients.



**Figure 3.5:** Infection spread subgraph: Red nodes indicate patients with route information who infected others. Green nodes indicate patients who infected others but do not have any route information. Blue nodes indicate patients who did not infect anyone else.

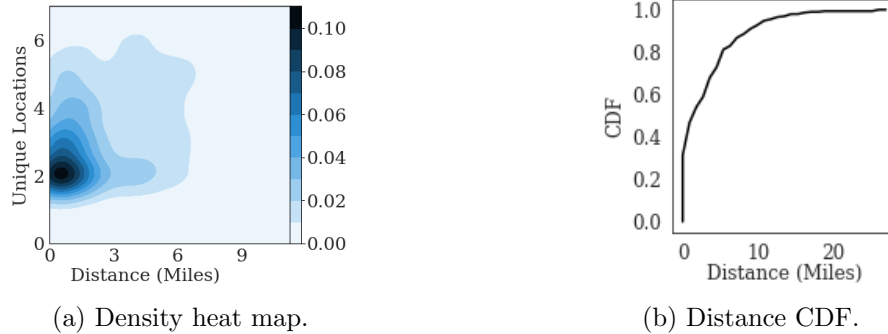


**Figure 3.6:** Super spreader analysis.

Fig. 3.6 presents CDFs of the number of people infected by an individual, the number of days in the log that the individual appears, the unique visited locations, and the total number of visited locations. The CDFs in this figure indicate that, in general, super spreaders tend to be active for more days, visit more unique locations, and have longer routes than low spreaders. The figure shows that all super spreaders in the data set are active for three or more days and visit three or more unique locations. Some of these super spreaders are active for up to 19 days and visit up to 18 unique locations with route lengths of up to 31 locations.

### 3.5 Daily Traveled Distance

Fig. 3.7(a) plots the density heat map of distance traveled by patients in Seoul and the number of locations visited in a day, two important features due to the vital nature of



**Figure 3.7:** Daily traveled distance and visited locations.

patient movement to spread COVID-19. The darker the area, the more patients have the same traveled distance and visited locations. With some exceptions, people mostly travel short distances and visit only a few locations each day. The CDF of the daily traveled distance is shown in Fig. 3.7(b).

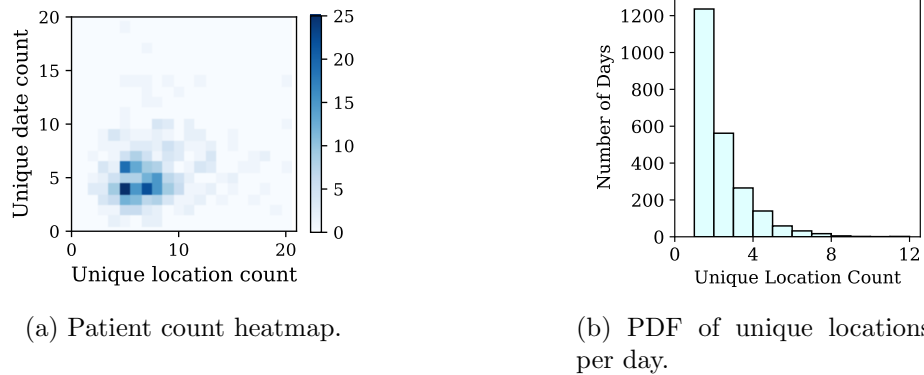
### 3.6 Patient Mobility

Patient mobility is another important attribute to consider. Intuitively, the more places a patient visits, the higher their mobility is. Fig. 3.8(a) depicts the number of patients that are seen on a specific number of unique locations (x-axis) for a specific number of days (y-axis). Note that this graph does not distinguish patient mobility across *different* days. Indeed, looking at the mobility of individual patients, there are days where they exhibit high mobility and days where they move significantly less. This points to a more usable definition of mobility as a function of different time periods (days). Fig. 3.8(b) shows the day count of unique locations reached by the patients in the data set: for 2,063 days (88.9% of days) a typical patient visits 1–3 locations, while for 258 days (11.1%) more than 3 unique locations are visited.

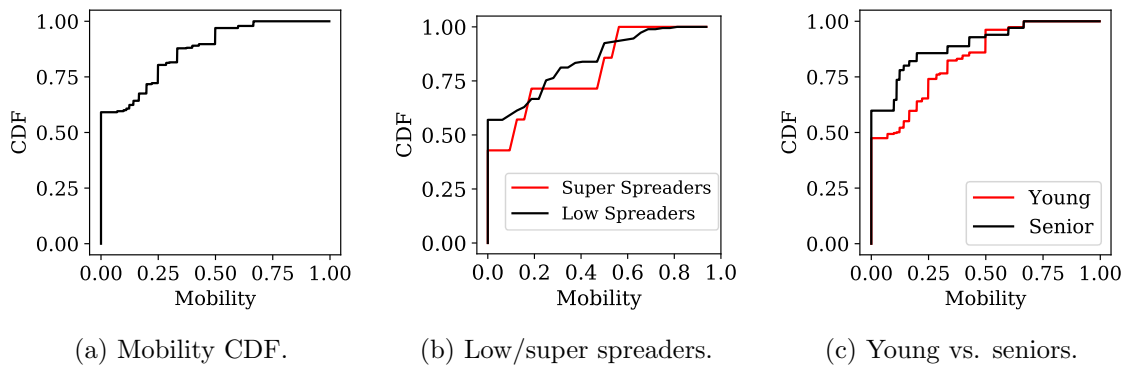
Defining a *high mobility day* as a day during which a patient visits at least  $L$  locations, the *mobility of a patient* is given as the ratio of the patient high mobility days to all logged days for this specific individual. Note that this is not the only way to define mobility. For

simulation purposes (see Chapter 4), this definition provides a practical way to capture mobility with a probability. Based on the histogram shown in Fig. 3.8(b), days with  $L \leq 3$  are considered of low mobility. The CDF of patient mobility using the above definition is depicted in Fig. 3.9(a). The figure shows that 57.6% of patients never visit more than 4 locations in a day.

Different classes of patients have different mobility. Fig. 3.9(b) shows the difference in mobility between super spreaders and low spreaders, while Fig. 3.9(c) illustrates mobility by age groups. Super spreaders and young people have higher mobility compared to low spreaders and seniors, respectively. For higher percentiles, the low spreaders have larger mobility than super spreaders due to the small number of super spreader agents in the KCDC data set.

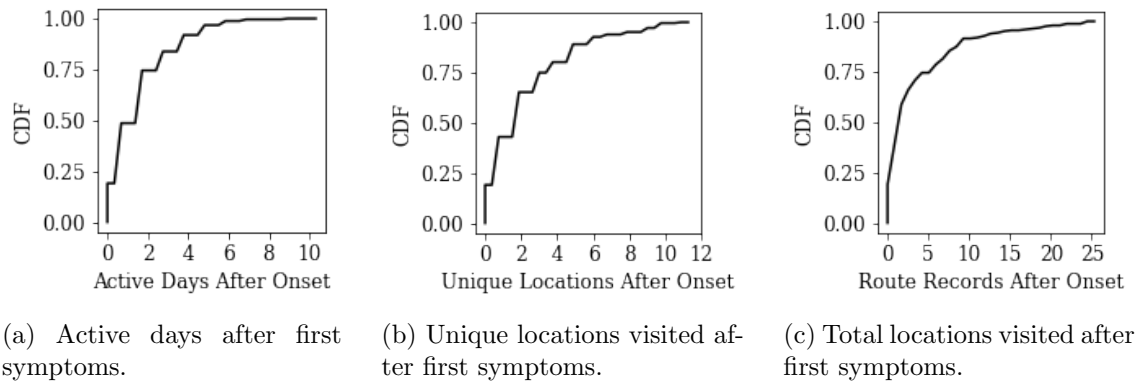


**Figure 3.8:** Patient unique locations.



**Figure 3.9:** Patient mobility.





**Figure 3.10:** Irresponsible behavior of sick patients.

### 3.7 Irresponsible Behaviors

Patients behave irresponsibly when they keep moving after the onset of their first COVID-19 symptoms, which facilitates the diffusion of the disease. We present how long sick people continue to show mobility after exhibiting symptoms, see Fig. 3.10. The figure shows that only the 20% of patients stop moving and isolate immediately after initial symptoms are observed. Some patients keep moving for more than a week after the onset of symptoms, see Fig. 3.10(a). They also visit many locations; Figs. 3.10(b) and 3.10(c) show the number of unique and total locations that sick patients visit after initial symptoms are observed.

## Chapter 4

# Agent-based Model

In this chapter, we show how to parameterize a simulation based on a patched version of GeoMason [35] using the characterization presented in Chapter 3. The attributes, life cycle, and states of an agent are shown in Figure 4.1. The following attributes are set during the initialization phase:

1. *Infection status.* One or more random agents are selected as the initial case(s).
2. *Position.* Agents are randomly placed on a road in the simulated area.
3. *Speed.* There are two types of agents: 50% of agents are considered pedestrian and walk at a speed of 3 MPH before reaching their destination; other agents drive a vehicle and their speed is uniformly distributed between 10 and 25 MPH.<sup>1</sup>
4. *Type of spreaders.* We define two classes of spreaders: 3.59% of patients are super spreaders and 96.41% are low spreaders (see Chapter 3.4).
5. *Mobility.* We use the mobility of super spreaders and low spreaders depicted in Fig. 3.9(b) to model different types of patient mobility.

In addition to the mobility distribution of super spreaders and low spreaders, the CDF of daily traveled distance in Fig. 3.7(a) is also used to determine the distance to a destination.

---

<sup>1</sup>We stress that these are nominal choices: any pedestrian to vehicles ratios can be used as input to the model.

Simulation time is defined by cycles. In each simulation cycle, agents outside a building move along the road towards their destination; agents inside a building can choose to stay or leave, based on their mobility. Agents with high mobility have a high probability to leave the building. Note that agents stay in a building for at least 15 minutes in order to meet the definition of close contact [11]. If multiple agents are inside the same building, they may infect each other with a certain probability.

When infection happens, the agent state changes from healthy to infected, as the state transition shown in Fig. 4.1. We assume the outdoor infection probability to be negligible. Given the probability of infection inside a building,  $\alpha$ , and the number of infected agents in the building,  $n$ , the probability of a healthy agent to be infected by a contact within the building is:

$$Pr(\text{infection}) = 1 - (1 - \alpha)^n. \quad (4.1)$$

Note that the probability of infection defined by Eq. 4.1 is nominal. Any model can be used here to capture the viral load: the total number of people in the location, the duration of interaction among individuals, the square footage of the room, its air circulation, wearing a mask or not, see [29] for examples on how to adjust Eq. (4.1).

It takes 1–14 days for patients to show symptoms after infection according to the WHO [38]. We therefore use a Uniform distribution between 1 and 14 days to transition from infected to symptomatic. A uniform distribution is again nominal here, one could easily use any distribution, e.g., a lognormal distribution with its peak set to 5 to capture a more realistic scenario consistent with hard data.

Since there exist patients who continue to move even after showing symptoms, as seen in Fig. 3.10, we use the CDF in Fig. 3.10(a) to determine the number of active days after their first symptoms. We do not distinguish the behavior of super and low spreaders because of lack of data (there are only two super spreaders with *symptom\_onset* information available). After each infected person exhausts their active days after infection, they are isolated.

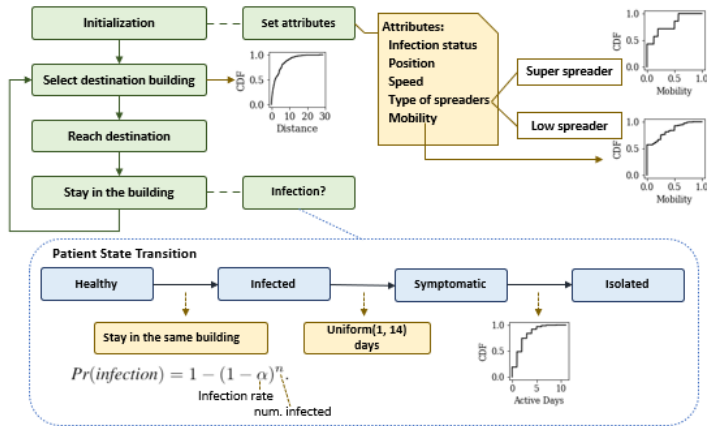


Figure 4.1: Life cycle of an agent.

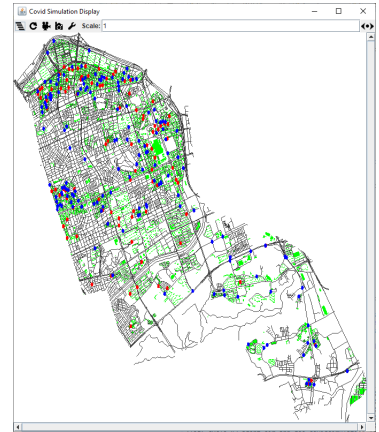
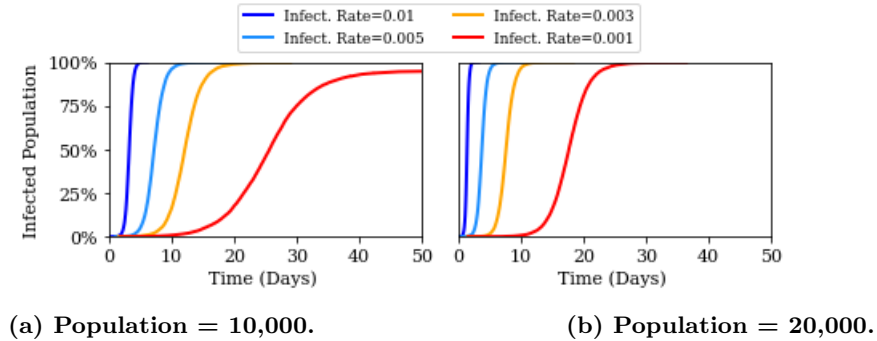


Figure 4.2: Simulation screenshot.

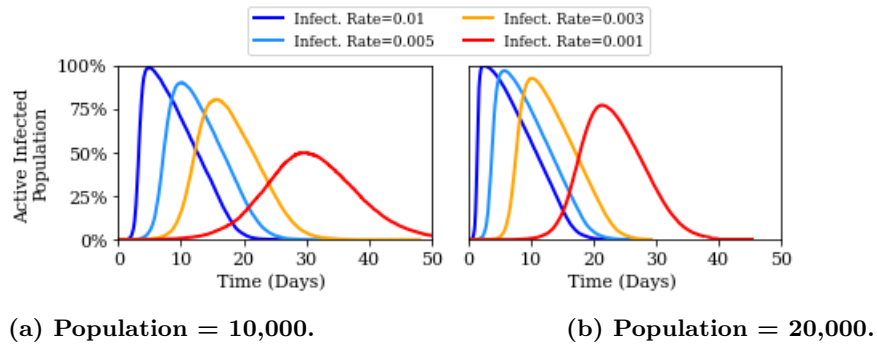
Consistent with infectious disease simulation studies [27], we set the simulation cycle to 5 minutes. The simulation stops either when all agents are infected or after a number of cycles defined by the user.<sup>2</sup>

We simulate the COVID-19 outbreak in the Gangnam district, i.e., the municipality of Seoul with the most hotspots, see Fig. 3.1(b). This area has 11,438 road intersections and 7,043 buildings. Roads and buildings are placed in the simulated area as described in [5], a collection of GIS data with regard to Seoul. GeoMason loads the GIS data (e.g., roads, road intersections, buildings) stored in a shapefile format, i.e., a file that stores geometric locations and their attribute information. Although the longest distance we observe in PatientRoute data set in Seoul is 30 miles, the longest distance between two buildings in the simulated Gangnam district is 7.06 miles. Therefore, we normalize the maximum distance to 3.53, which is half of the longest distance in the simulated area, to ensure a valid building selection as the agent’s destination. In the Gangnam district there are 604,586 people and a total of 7,043 buildings. We do not have any information on the building stories, entries, or number of rooms. This information is crucial, especially for apartment buildings, where multiple people can be inside the same building at the same

<sup>2</sup>In this simulation, we do not explicitly model agent recovery: a recovered agent that resumes its mobility is considered immune and non-contagious, therefore does not contribute to the disease spread. The simulation can be trivially extended to model recovered agents re-entering the simulation cycle.



**Figure 4.3:** Simulating patient isolation.



**Figure 4.4:** Percentage of active agents while infected.

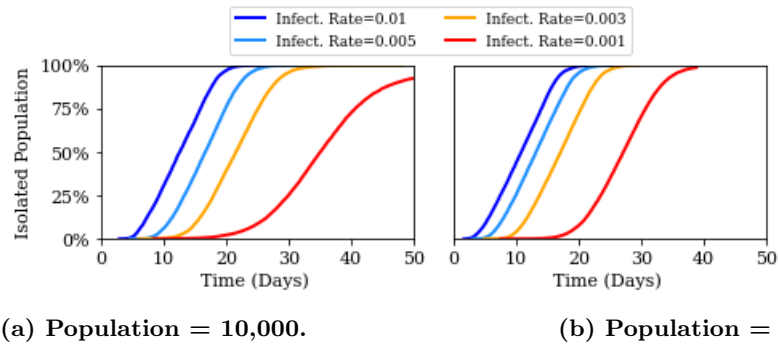
time without contact. To address this lack of information, we limit the population in our simulations. We validate parameter choices against ground truth data in Chapter 5.

A screenshot of the GeoMason simulation execution can be seen in Fig. 4.2. Black lines are roads that agents travel on and green areas are buildings where agents stop. Agents only have two states in terms of infection, i.e., infected (red dots) or healthy (blue dots).

Fig. 4.3 depicts the percentage of infected population as a function of time. The simulation begins with one infected agent and stops after 50 days. The graph illustrates how quickly the entire population is infected for four infection rates that correspond to measures such as mask wearing and social distancing. The figure includes results for two population sizes and shows the speed of the disease spread as a function of population density, infection in Fig. 4.3(b) is faster than Fig. 4.3(a). For simulation scalability reasons, we limit the entire population to a manageable number. We illustrate in Chapter 5 that

a smaller population can still capture observed trends with appropriate parameterization.

As a companion to Fig. 4.3, we also present the portions of “active while infected” and isolated agents, see Fig. 4.4. In Fig. 4.4, the benefit of patient isolation can be seen clearly: the percentage of active infected population is decreasing after showing a peak, which limits the speed of the spread of the disease. The percentage of isolated population shown in Fig. 4.5 explains the decrease of active infected population. After more agents show symptoms and are isolated, the active infected population starts dropping.



**Figure 4.5:** Percentage of isolated population.

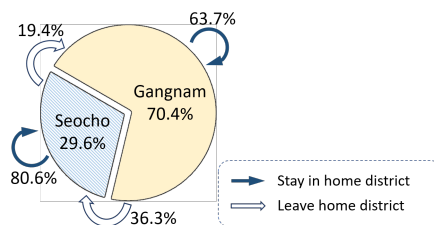
## Chapter 5

# Model Validation and Case Study

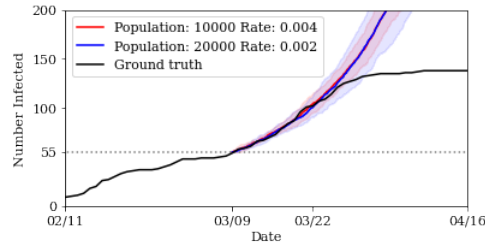
After presenting the generic results in Chapter 4, we showcase the flexibility of this simulation model. We first validate the simulation using the ground truth, then we simulate different mitigation measures to assess their effectiveness.

### 5.1 Validation

In this simulation, we include the Seocho district, a neighboring district of Gangnam, to study the effect of moving agents across different districts. Fig. 5.1 shows the percentage of residents in these two districts that have been infected, the figure also illustrates the frequency of residents visiting buildings in their home district, as well as visiting the other district. We use this information to parameterize the simulation. During the initialization phase, we separate the agents into Gangnam residents (70.4% of the population) and Seocho residents (29.6% of the population). Next, we retrieve the distributions of agent



**Figure 5.1:** Movements of Gangnam and Seocho residents.



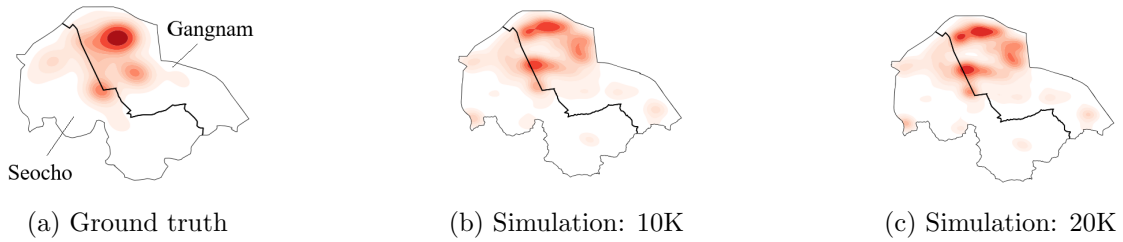
**Figure 5.2:** Infected population in the validation simulation. The overlap of two simulation cases with the ground truth retrieved from the data set validates the simulation settings. Results are presented with 95% confidence intervals (error margins are given by the colored ranges).

mobility and spreader types from the data set for residents of each district to set their attributes. After initialization, when selecting destination buildings, the probability of a resident staying or leaving their home district follows Fig. 5.1.

Since two districts are considered in this simulation, starting with only one infected agent in one of the two areas could bias the results. Here, we start the simulation with 55 infected agents, i.e., the number of infections observed from the data set on March 9, 2020, proportionally assigned to agents in the two districts (29.6% in Seocho, 70.4% in Gangnam). We selected March 9, 2020 because mitigation efforts in Seoul have yet to produce a noticeable effect on disease spread, while also allowing us to clearly see trends. Simulations starting at any time earlier or around March 9, result in similar infection trends.

Fig. 5.2 depicts the number of COVID-19 cases in the Gangnam and Seocho districts observed from the data set (black line) and simulation (red and blue lines). The ground truth line illustrates the COVID-19 outbreak in the two districts. At the beginning of April, the curve flattens. This is likely due to effective counter-measures executed in Seoul, especially the Strong Social Distancing Campaign which began on March 22. Consistent with the COVID-19 incubation timeline, the effectiveness of the Strong Social Distancing Campaign does not show immediately, but after the beginning of April. We align the beginning of simulation data to the time of 55 infection cases in the ground truth, since



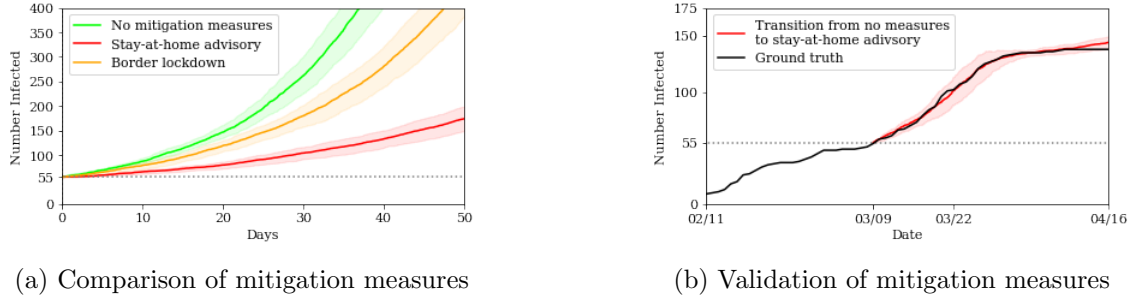


**Figure 5.3:** Hotspots in the data set (ground truth) and model.

this is the starting point of the simulation. The two simulation lines in Fig. 5.2 closely follow the ground truth: the simulation of population 10,000 with infection rate 0.004 and the simulation of population 20,000 with infection rate 0.002 are in excellent agreement with the ground truth from March 26, 2020 to April 5, 2020, when the effects of any counter-measures are not discernible yet. The overlap of two simulation cases with the ground truth validates the simulation.

We note in Fig. 5.2 an interesting relationship between population and infection rate: when the population is doubled, dividing the infection rate in half gives similar simulation outcomes. This observation also meets the results in the generic simulation that higher population leads to faster spreading of the COVID-19 virus, while lowering the infection rate slows down the virus spreading. We conclude that we can use a “limited” population with an adjusted infection rate to efficiently (yet accurately) model the expected behavior of larger populations.

Next, we focus on hotspot locations. In Fig. 5.3(a), we present the heat map of most visited locations in the Gangnam and Seocho districts from the data set (ground truth). The most visited areas are in the northern part of Gangnam and across the border between the two districts. These hotspots correspond to the density of commercial buildings in these areas, which results in higher traffic areas. Fig. 5.3(b) and (c) show the heat map of visits in the first week for simulated populations of 10,000 and 20,000, accordingly. From both simulations, we observe similar hotspots, consistent with the ground truth heat map. This similarity further validates the accuracy of the simulation.



**Figure 5.4:** Effect of different counter-measures. Results are presented with 95% confidence intervals (shaded areas).

## 5.2 Applying mitigation measures

We now turn to the evaluation of the effectiveness of counter-measures. We first consider stay-at-home advisory that allows for only essential activity outside of the agent’s domicile. On average, agents stay home for longer periods time under the advisory, but are permitted to leave periodically. The probability of leaving home is set to 20% of the agent’s mobility. This can be tuned to simulate a stricter (or more relaxed) stay-at-home advisory. Once the agent arrives at the destination building, the probability of leaving the building is defined by the mobility without any additional scaling (i.e., the time spent outside the domicile is not affected).

In addition to this counter-measure, we also consider strict district border control between the Gangnam and Seocho districts, i.e., forbid movements between these two areas entirely. With a strict border control between these two districts, agents can only stay in their home district: the probability of leaving their home district is set to 0. We simulate these two mitigation measures under population 10,000, see Fig. 5.4(a) for results. First, the application of a stay-at-home advisory decreases the rate of virus spread in comparison to the baseline scenario where no counter-measures are applied. The strict border control offers a mild mitigation measure comparing to the baseline scenario.

As further validation, we simulate the effects of applying a stay-at-home advisory mid-simulation in order to capture the effects of the mitigation measures taken in Seoul on

March 22 – the Strong Social Distancing Campaign. Figure 5.4(b) depicts the results of these simulations against the ground truth. In this simulation case, we begin with no mitigation measures and apply a stay-at-home advisory once we reach a certain threshold number of infections. Here we select this threshold based on the number of infections in the ground truth data when the Strong Social Distancing campaign was enacted, however, this threshold is a parameter and we can choose to transition between no measures and a stay-at-home advisory at any given number of infections. This further highlights the ability of the model to capture what-if scenarios of different patterns of population movement.

## Chapter 6

# Discussion and Limitations

The proposed model captures the spread of COVID-19 in an urban setting. Although the model is validated using ground truth, incomplete and/or missing data may limit its generalization and make it far from being the definitive COVID-19 spreading model. Main limitations of our approach include:

**First wave data.** This data is from the first wave in the disease in South Korea. With South Korea having one of the best responses to the disease globally, the mobility patterns reflect inevitably cultural and demographic characteristics as well as policy decisions.

**Scarcity of data.** We continue to seek additional data sets on COVID-19 outbreaks. The current lack of substantial data is an unfortunate limitation. For example, the data on super-spreader mobility are not of statistical significance, there is no exact information on the elapsed time in each location by each agent but only the sequence of locations, we do not have exact information on the movements *inside* buildings. In addition, the data on patient mobility was removed from Kaggle on May 31, 2020. While we did analyze the mobility data, we can make available to the community all information presented in this thesis in the form of histograms and CDFs (not in their raw form, the appendix presents how such data can be retrieved).

**Privacy concerns.** The KCDC data set is anonymized and no sensitive data of monitored patients can be retrieved. No data about the underage population is provided as well as

movements of patients from/to their private homes. This limits the scenarios that can be analyzed, e.g., the impact of school closures or the spreading of the virus within households. Note also that the per-patient mobility information (and its statistics) were retrieved from the PatientRoute data set while it was available to the public. Since June 2020, this data set became unavailable. We have no way to evaluate how the mobility statistics changed during the second wave in Fall 2020.

**Transportation assumptions.** The KCDC data set does not show the transportation mode of patients. We overcome this limitation by assuming a pedestrian:vehicles ratio of 1:1, this ratio can be adjusted as needed. Input parameters can be fully customized and other researchers using our approach can easily change these values.

## Chapter 7

# Related Work

The COVID-19 pandemic has been studied extensively in recent months due to its disruptive effects. Different approaches have been adopted to increase our knowledge on the pandemic. Pung et al. [32] interview COVID-19 patients in Singapore to collect epidemiological/clinical data to study the spread of the virus in three different Singapore *clusters*, this approach by its nature can be applied to populations of a small scale only. Epidemiological models allow studying how an infection spread on a larger scale and are classified as mathematical or agent-based.

**Mathematical models** are defined by a set of equations that allow describing the evolution of the disease [30]. Bi et al. [9] use conditional logistic regression to study the transmission of COVID-19 in Shenzhen, China. Using data from contact-based surveillance and accurate infector-infectee relationships, they confirm that, on average, COVID-19 has an incubation period of less than a week and a long clinical course. Rader et al. [33] use regression models to evaluate how the socio-economic and environmental aspects of a region affect the spreading of COVID-19. Garg et al. [18] predict hospitalization rates from clinical data (e.g., age, ethnicity, medical conditions, clinical course) of COVID-19 patients in 14 states of the USA. Note that the above works do not focus on the SARS-CoV-2 spread in a community.

Pej3 and Bicz3k [31] use game theory to evaluate the efficiency of face masks and social

distancing in limiting the spread of COVID-19 when there are selfish patients who do not use any counter-measures. Similarly, Bhattacharyya and Bauch [8] use game theory to evaluate the efficiency of protective vaccines (the safest way to achieve herd immunity [15]).

Grossmann et al. [20] propose a stochastic network-based to model COVID-19 spread, and compare its results with those obtained through an ordinary differential equations (ODE) model. Their network-based model leverages random graph models to represent interaction structures and human connections. They observe that ODE models struggle to correctly represent heterogeneity of interaction structures, a feature that profoundly affects the spread of the virus. While this work does focus on human interactions, it does not explicitly model spatial population movements.

**Agent-based models (ABMs)** are a simulation-based alternative of mathematical models that incorporate human interactions [24]. ABMs are typically used for modeling pedestrian movements, resource usage, and to successfully study the spread of diseases [14, 21, 36].

Ferguson et al. [16] model the spread of influenza in British and American households, schools, and workplaces. Their simulations are parameterized using census and land use data as well as air travel patterns. Note that the above work considers only large scale (international) population movements. ABMs parameterized by census data have been used to capture the spread of COVID-19 in Australia [34, 13]. Using census and age-distribution data from Germany and Poland, Bock et al. [10] investigate the efficiency of mitigation strategies by accounting for interactions within households where it is hard to social distance. Census ABM-based frameworks have been used to simulate the COVID-19 outbreak [22], evaluate the efficiency of contact tracing [7], face masks [23], and testing strategies [37]. Kim et al. [27] use synthetic, location-based social network data to study outbreaks and evaluate the effectiveness of different mitigation strategies, especially how social behaviors affect the virus spread. ABMs are used also to model the spread of SARS-CoV-2 in small areas: crowded areas of supermarkets [40] and university campuses [19].

*Differently from our approach, no fine-grained movement data is used in any of the above works. The above models are parameterized using census or synthetic data while population movement habits are captured at a coarse granularity.*

Müller et al. [29] use an ABM parameterized with synthetic mobility traces (originally generated from mobile phone data for public transportation applications) to study the COVID-19 outbreak in Berlin and analyze how mitigation measures result in reduction of activity in public. [29] is the closest to our work but it does not provide any detailed statistics on agent mobility during the pandemic as we do here.

*Summarizing, in this thesis we extract human movement habits and dynamics from the KCDC data set of real COVID-19 patients. The mobility information (i.e., patient mobility, traveled distance, visited locations) and statistics are used to tune an ABM and investigate the COVID-19 outbreak in two districts of Seoul. Agent movements and behaviors are simulated using the statistics of actual human movements, other structures (e.g., networks or graphs) are not required. The proposed approach allows investigating scenarios under different circumstances to identifying mitigation strategies.*



## Chapter 8

# Conclusions and Ongoing Work

Information and routes of South Korean COVID-19 patients are analyzed to study the disease outbreak in the Gangnam and Seocho districts of Seoul. Movement habits in South Korea are extracted from available data sets to parameterize simulations, based on ABM and GIS, and study interactions among people. Simulation results are in excellent agreement with ground truth and show that this model can be used to flexibly examine and evaluate a wide variety of different scenarios based on different human mobility patterns from real-world data. While we do not claim that it is a definitive COVID-19 spread model, it can be used to investigate useful *what-if* scenarios.

We are currently working on expanding the simulation model to create a prediction ecosystem for evaluating detailed scenarios: geographical restrictions of mobility, work from home orders/advisories, school closures (and partial openings under different conditions), points of interest operating under various capacities, time in quarantine, and vaccination priority. We propose to enrich the existing data that currently drive the model via cross-fertilization of datasets: correlate the sojourn at points-of-interest from Safegraph with Google mobility data in the U.S. as done in [12] to provide informative guesses for sojourn times in Seoul (note that Google mobility data exist for Seoul but there are no Safegraph data). We will also compare the Seoul KCDC mobility with stochastic mobility

models of Berlin [29]<sup>1</sup> to identify similarities and differences among population mobility in two urban settings. Focusing on the KCDC data set again, we will use hypergraphs to identify “patient bubbles” that frequent within the same points-of-interest, explore different ways to create “bubbles,” and how these “bubbles” evolve across time. Our study of the KCDC logs is an example of data that are incomplete, a common problem in tracing datasets. We will use the ABM model to “generate” mobility data for larger populations (note that the KCDC trace logs are for a relative small set as South Korea successfully dealt with the disease early on). The generated mobility logs will create a “ground truth”. We will then introduce “gaps” in this set of logs, use machine learning to fill them [39] and explore whether we can use this mechanism to enrich missing data in the KCDC logs.

---

<sup>1</sup>The Berlin logs are not publicly available, but stochastic models of mobility are.

# Bibliography

- [1] Google Maps. <https://www.google.com/maps/>, 2020. [Online; 2021-01-13].
- [2] Kakao Map. <https://map.kakao.com/>, 2020. [Online; 2021-01-13].
- [3] Naver Map. <https://m.map.naver.com/>, 2020. [Online; 2021-01-13].
- [4] OpenStreetMap. <https://www.openstreetmap.org/>, 2020. [Online; 2021-01-13].
- [5] OSM extracts for Seoul. <https://download.bbbike.org/osm/bbbike/Seoul/>, 2020. [Online; 2021-01-13].
- [6] WHO Director-General’s opening remarks at the media briefing on COVID-19 – 11 March 2020. <https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>, 2020. [Online; 2021-01-13].
- [7] JONATAN ALMAGOR AND STEFANO PICASCIA. Can the app contain the spread? An agent-based model of COVID-19 and the effectiveness of smartphone-based contact tracing. *arXiv preprint arXiv:2008.07336*, 2020.
- [8] SAMIT BHATTACHARYYA AND CHRIS BAUCH. “Wait and see” vaccinating behaviour during a pandemic: a game theoretic analysis. *Vaccine*, 29(33):5519–5525, 2011.
- [9] QIFANG BI, YONGSHENG WU, SHUJIANG MEI, CHENFEI YE, XUAN ZOU, ZHEN ZHANG, XIAOJIAN LIU, LAN WEI, SHAUN A TRUELOVE, TONG ZHANG, ET AL.

- Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1,286 of their close contacts. *MedRxiv*, 2020.
- [10] WOLFGANG BOCK, BARBARA ADAMIK, MAREK BAWIEC, VIKTOR BEZBORODOV, MARCIN BODYCH, JAN PABLO BURGARD, THOMAS GOETZ, TYLL KRUEGER, AGATA MIGALSKA, BARBARA PABJAN, ET AL. Mitigation and herd immunity strategy for COVID-19 is likely to fail. *medRxiv*, 2020.
- [11] CDC. Public Health Guidance for Community-Related Exposure. <https://www.cdc.gov/coronavirus/2019-ncov/php/public-health-recommendations.html>, 2020. [Online; 2021-01-13].
- [12] SERINA CHANG, EMMA PIERSON, PANG WEI KOH, JALINE GERARDIN, BETH REDBIRD, DAVID GRUSKY, AND JURE LESKOVEC. Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.
- [13] SHERYL L CHANG, NATHAN HARDING, CAMERON ZACHRESON, OLIVER M CLIFF, AND MIKHAIL PROKOPENKO. Modelling transmission and control of the COVID-19 pandemic in Australia. *arXiv preprint arXiv:2003.10218*, 2020.
- [14] ANDREW CROOKS AND ATESMACHEW HAILEGIORGIS. An agent-based modeling approach applied to the spread of cholera. *Environmental Modelling & Software*, 62:164–177, 2014.
- [15] GYPSYAMBER D’SOUZA AND DAVID DOWDY. What’s herd immunity and how can we achieve it with COVID-19. <https://www.jhsph.edu/covid-19/articles/achieving-herd-immunity-with-covid19.html>, 2020. [Online; 2021-01-13].
- [16] NEIL M FERGUSON, DEREK AT CUMMINGS, CHRISTOPHE FRASER, JAMES C CAJKA, PHILIP C COOLEY, AND DONALD S BURKE. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448–452, 2006.

- [17] KOREA CENTERS FOR DISEASE CONTROL & PREVENTION. Coronavirus Disease-19, Republic of Korea. <http://ncov.mohw.go.kr/en/>, 2020. [Online; 2021-01-13].
- [18] SHIKHA GARG. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—COVID-NET, 14 States, March 1–30, 2020. *MMWR. Morbidity and mortality weekly report*, 69, 2020.
- [19] PHILIP T GRESSMAN AND JENNIFER R PECK. Simulating COVID-19 in a University Environment. *arXiv preprint arXiv:2006.03175*, 2020.
- [20] GERRIT GROSSMANN, MICHAEL BACKENKOEHLER, AND VERENA WOLF. Importance of Interaction Structure and Stochasticity for Epidemic Spreading: A COVID-19 Case Study. *medRxiv*, 2020.
- [21] KATHRYN H JACOBSEN, A ALONSO AGUIRRE, CHARLES L BAILEY, ANCHA V BARANOVA, ANDREW T CROOKS, ARIE CROITORU, PAUL L DELAMATER, JHUMKA GUPTA, KYLENE KEHN-HALL, AARTHI NARAYANAN, ET AL. Lessons from the Ebola outbreak: action items for emerging infectious disease preparedness and response. *EcoHealth*, 13(1):200–212, 2016.
- [22] MASOUD JALAYER, CARLOTTA ORSENIGO, AND CARLO VERCELLIS. CoV-ABM: A stochastic discrete-event agent-based framework to simulate spatiotemporal dynamics of COVID-19. *arXiv preprint arXiv:2007.13231*, 2020.
- [23] DE KAI, GUY-PHILIPPE GOLDSTEIN, ALEXEY MORGUNOV, VISHAL NANGALIA, AND ANNA ROTKIRCH. Universal masking is urgent in the covid-19 pandemic: Seir and agent based models, empirical validation, policy recommendations. *arXiv preprint arXiv:2004.13553*, 2020.
- [24] REBECCA A KELLY, ANTHONY J JAKEMAN, OLIVIER BARRETEAU, MARK E BORSUK, SONDOSS ELSAWAH, SERENA H HAMILTON, HANS JØRGEN HENRIKSEN, SAKARI KUIKKA, HOLGER R MAIER, ANDREA EMILIO RIZZOLI, ET AL. Selecting

- among five common modelling approaches for integrated environmental assessment and management. *Environmental modelling & software*, 47:159–181, 2013.
- [25] JIHOO KIM AND JOONGKUN LEE. Data Science for COVID-19 (DS4C). <https://www.kaggle.com/kimjihoo/coronavirusdataset>, 2020. [Online; 2021-01-13].
- [26] JIMI KIM, SEOJIN JANG, WONCHEOL LEE, JOONG KUN LEE, AND DONG-HWAN JANG. DS4C Patient Policy Province Dataset: a Comprehensive COVID-19 Dataset for Causal and Epidemiological Analysis. In *Advances in Neural Information Processing Systems*, 2020.
- [27] JOON-SEOK KIM, HAMDİ KAVAK, CHRIS OVI ROULY, HYUNJEE JIN, ANDREW CROOKS, DIETER PFOSE, CAROLA WENK, AND ANDREAS ZÜFLE. Location-based social simulation for prescriptive analytics of disease spread. *SIGSPATIAL Special*, 12(1):53–61, 2020.
- [28] SUN KIM AND MARCIA C CASTRO. Spatiotemporal pattern of COVID-19 and government response in South Korea (as of May 31, 2020). *International Journal of Infectious Diseases*, 98:328–333, 2020.
- [29] SEBASTIAN A MÜLLER, MICHAEL BALMER, WILLIAM CHARLTON, RICARDO EWERT, ANDREAS NEUMANN, CHRISTIAN RAKOW, TILMANN SCHLENTHER, AND KAI NAGEL. A realistic agent-based simulation model for COVID-19 based on a traffic simulation and mobile phone data. *arXiv preprint arXiv:2011.11453*, 2020.
- [30] H VAN DYKE PARUNAK, ROBERT SAVIT, AND RICK L RIOLO. Agent-based modeling vs. equation-based modeling: A case study and users’ guide. In *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, pages 10–25. Springer, 1998.
- [31] BALÁZS PEJÓ AND GERGELY BICZÓK. Corona Games: Masks, Social Distancing and Mechanism Design. In *Proceedings of the 1st ACM SIGSPATIAL International*

- Workshop on Modeling and Understanding the Spread of COVID-19*, pages 24–31, 2020.
- [32] RACHAEL PUNG, CALVIN J CHIEW, BARNABY E YOUNG, SARAH CHIN, MARK IC CHEN, HANNAH E CLAPHAM, ALEX R COOK, SEBASTIAN MAURER-STROH, MATTHIAS PHS TOH, CUIQIN POH, ET AL. Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. *The Lancet*, 2020.
- [33] BENJAMIN RADER, SAMUEL SCARPINO, ANJALIKA NANDE, ALISON HILL, ROBERT REINER, DAVID PIGOTT, BERNARDO GUTIERREZ, MUNIK SHRESTHA, JOHN BROWNSTEIN, MARCIA CASTRO, ET AL. Crowding and the epidemic intensity of COVID-19 transmission. *medRxiv*, 2020.
- [34] REBECCA J ROCKETT, ALICIA ARNOTT, CONNIE LAM, ROSEMARIE SADSAD, VERLAINE TIMMS, KAREN-ANN GRAY, JOHN-SEBASTIAN EDEN, SHERYL CHANG, MAILIE GALL, JENNY DRAPER, ET AL. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nature medicine*, pages 1–7, 2020.
- [35] KEITH SULLIVAN, MARK COLETTI, AND SEAN LUKE. GeoMason: Geospatial support for MASON. Technical report, Department of Computer Science, George Mason University, 2010.
- [36] SRINIVASAN VENKATRAMANAN, BRYAN LEWIS, JIANGZHUO CHEN, DAVE HIGDON, ANIL VULLIKANTI, AND MADHAV MARATHE. Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics*, 22:43–49, 2018.
- [37] YINGFEI WANG, INBAL YAHAV, AND BALAJI PADMANABHAN. Whom to Test? Active Sampling Strategies for Managing COVID-19. *arXiv preprint arXiv:2012.13483*, 2020.

- [38] WHO. Q&A on coronaviruses (COVID-19). <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses>, 2020. [Online; 2021-01-13].
- [39] JI XUE, BIN NIE, AND EVGENIA SMIRNI. Fill-in the gaps: Spatial-temporal models for missing data. In *13th International Conference on Network and Service Management, CNSM 2017, Tokyo, Japan, November 26-30, 2017*, pages 1–9, 2017.
- [40] FABIAN YING AND NEAVE O’CLERY. Modelling COVID-19 transmission in supermarkets using an agent-based model. *arXiv preprint arXiv:2010.07868*, 2020.