GeoSpread: an Epidemic Spread Modeling Tool for COVID-19 Using Mobility Data

ANNA SCHMEDDING, William & Mary, USA
LISHAN YANG, William & Mary, USA
RICCARDO PINCIROLI, Gran Sasso Science Institute, Italy
EVGENIA SMIRNI, William & Mary, USA

We present an individual-centric agent-based model and a flexible tool, GeoSpread, for studying and predicting the spread of viruses and diseases in urban settings. Using COVID-19 data collected by the Korean Center for Disease Control & Prevention (KCDC), we analyze patient and route data of infected people from January 20, 2020, to May 31, 2020, and discover how infection clusters develop as a function of time. This analysis offers a statistical characterization of population mobility and is used to parameterize GeoSpread to capture the spread of the disease. We validate simulation predictions from GeoSpread with ground truth and we evaluate different *what-if* counter-measure scenarios to illustrate the usefulness and flexibility of the tool for epidemic modeling.

Additional Key Words and Phrases: Data Analysis, Simulation Models, Individual-Centric Models, COVID-19, Disease Spread Modeling

ACM Reference Format:
Anna Schmedding, Lishan Yang, Riccardo Pinciroli, and Evgenia Smirni. 2022. GeoSpread: an Epidemic Spread Modeling Tool for COVID-19 Using Mobility Data . 1, 1 (August 2022), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

On March 11, 2020, the WHO declared COVID-19 the first pandemic caused by a coronavirus [29]. 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 [22] (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. Due to the increasing rate of novel viral outbreaks [2], such as recent outbreaks of hepatitis in children [38] and monkeypox [37], using the growing amount of available data to predict and limit such spreads is vital.

In this paper, 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 [25]. The KCDC records patient

Authors' addresses: Anna Schmedding, William & Mary, Williamsburg, USA, akschmed@cs.wm.edu; Lishan Yang, William & Mary, Williamsburg, USA, lyang11@cs.wm.edu; Riccardo Pinciroli, Gran Sasso Science Institute, L'Aquila, Italy, riccardo.pinciroli@gssi.it; Evgenia Smirni, William & Mary, Williamsburg, USA, esmirni@cs.wm.edu.

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 [23], Kakao Map [24], or Naver Map [26]) to extract geographical coordinates (i.e., latitude and longitude) and other data.

To the best of our knowledge, the KCDC logs are the only 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. The KCDC data set is a valuable resource, yet it presents limitations:

- The last version of the KCDC data set contains data collected up to May 31, 2020 By that date approximately 11,500 COVID-19 cases were confirmed in South Korea [16, 25], but only 35% of them have been logged.
- Some locations visited by patients are not recorded due to privacy concerns. Consequently, patient infection information and route data do not always coincide.
- Patient and route data may be incomplete (i.e., some location attributes are occasionally missing) and require manual completion before analyzing the data set.
- There is route data information for only the 15% of all confirmed cases by May 31.

We adopt different data discovery 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.

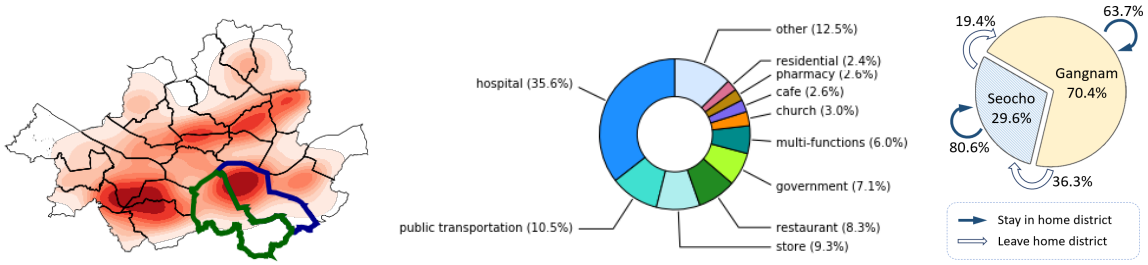
Regretfully, some missing data are not possible to recover. Specifically, provided that the mobility of only the 15% of confirmed patients are logged in detail, we can only "guess" the mobility of the remaining patients assuming it is independent and identically distributed to the 15% of patients with detailed logs. We contend 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. Here, we use this processed data in the form of histograms (and also make them available to the community [31]).

We use logs and histograms to feed GeoSpread, an extended version of the GeoMason [33] 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 [7]). 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. Our contributions and outline of this work are:

- **Data Discovery**. We analyze and connect data from various KCDC logs to extract information on patient movements (Sections 2 and 3). Missing information is manually retrieved, when possible.
- **Statistical Analysis**. We provide statistical analysis of population movements and habits.
- **Agent-based Model and Flexible Tool**. We create a tool, GeoSpread, and parameterize an agent-based model that uses the KCDC data as input, see Section 4, and outline its flexibility to capture a variety of conditions as well as new viral outbreaks. The simulation tool and processed data are publicly available [31].
- **Model Validation with Real Data** . We use the ground truth to validate the model in Section 5. Model limitations are discussed in Section 6.

2 THE KCDC DATA SET

The data sets [14] used in this paper 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.



(a) Seoul. Gangnam district is outlined in blue. Seocho district is outlined in green.

(b) Top-10 most visited location types in Seoul.

(c) Movement in Gangnam and Seocho.

Fig. 1. Location-related information from the KCDC data sets for Seoul.

PatientInfo data set. This data set provides epidemiological data. It contains 4004 different entries, each entry represents a different patient identified by a unique 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*).

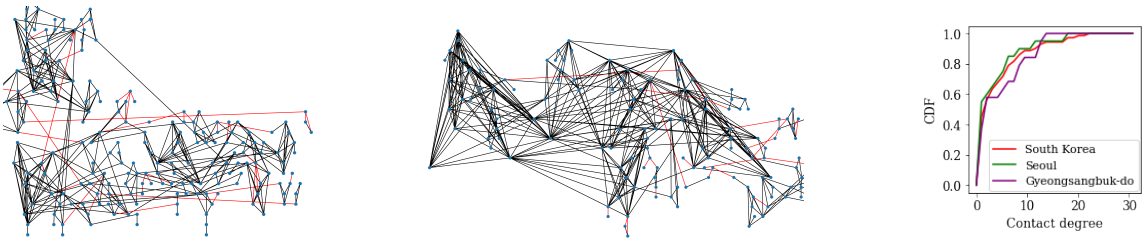
PatientRoute data set. This data set contains 8092 entries, each one reporting a visit (to one of 2992 unique locations) of 1472 (out of 4004) unique South Korean COVID-19 patients logged in the PatientInfo data set. A location is unequivocally identified by its *latitude* and *longitude*. *Province*, *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 [14]). We manually look for their type using their geographical coordinates and OpenStreetMap [27] 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 chronologically.

3 DATA DISCOVERY

Although the information contained in the KCDC data sets are not as accurate as one would like, it still allows for the analysis of patient movements and interactions with high accuracy. In this section, we discuss information that we extract from the data sets and how it is used to parameterize GeoSpread, our extension of the GeoMason ABM tool [33].

3.1 Visited Locations

Fig. 1(a) depicts a heat map of the most visited locations in Seoul. 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 and Seocho districts, outlined in blue and green, respectively, in Fig. 1(a). Many shopping and entertainment centers are located in Gangnam. Fig. 1(b) shows the ten most visited facility types in Seoul, with *Hospital* being the first one. No information about schools is available since this data set monitors only people in their 20s through 70s. The scarcity of logged residential facilities is due to privacy concerns. Finally, Fig. 1(c) illustrates the movement of population between two neighboring districts, Gangnam and Seocho that we use later in our model.



(a) Patient connections (partial 1). (b) Patient connections (partial 2). (c) Contact degree CDF.

Fig. 2. Patient contacts.

3.2 Patient Connections

Figs. 2(a) and 2(b) present subgraphs of patient connections discovered by linking the PatientRoute and PatientInfo data sets. To improve visibility, we only present a small portion (i.e., 13% and 9% respectively) of the entire graph. Here, nodes depict patients, black edges connect patients that visited the same place during the same day from the PatientRoute data set, 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 due to missing data, i.e., 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. 2(a) and 2(b) shows the contact degree among patients and illustrates visually the complexity of the problem.

Fig. 2(c) 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 they were sick or healthy when the contact occurred.

3.3 Super Spreaders

Fig. 3 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. The large “fans” in this figure are indicative of super spreaders.

Using this classification of patients based on the number of people they infect, we discover different behaviors of super/low spreaders, shown in Fig. 4. Super spreaders account for 3.59% and low spreaders for the rest of 96.41% of patients. Fig. 4 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.

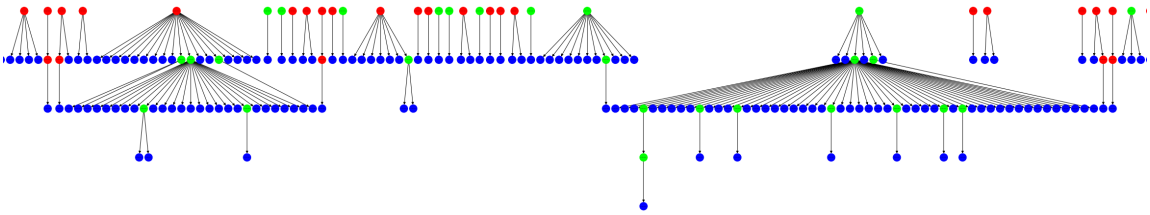
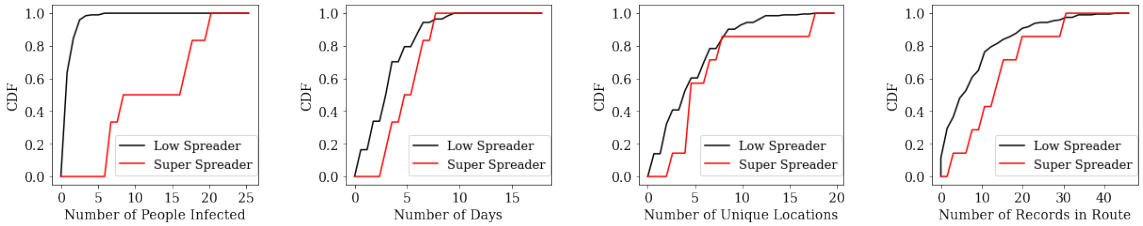


Fig. 3. Infection spread subgraph: Red nodes indicate patients with route information who infected others. Green nodes indicate patients who infected others but do not have any route information. Blue nodes indicate patients who did not infect anyone else.



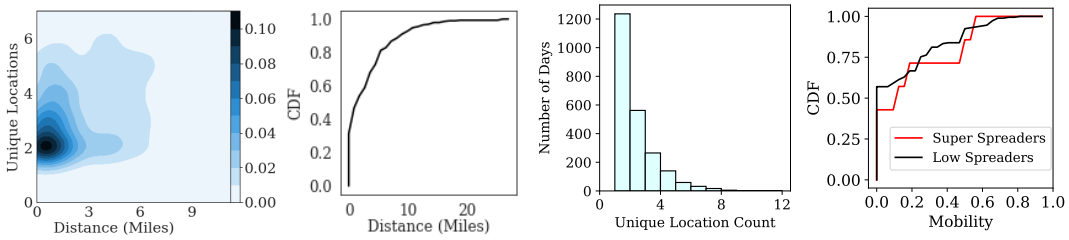
(a) People infected. (b) Logged days. (c) Unique locations visited. (d) Total locations visited.

Fig. 4. Super spreader analysis.

3.4 Mobility

Fig. 5(a) plots the density heat map of distance traveled by patients in Seoul and the number of locations visited in a day. 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. 5(b). Intuitively, the more places a patient visits, the higher their mobility is. Looking at the mobility of individual patients, there are days where they exhibit high mobility and days where they move significantly less. This leads us to a more usable definition of mobility as a function of different time periods (days). Fig. 5(c) 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 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 (Section 4), this definition provides a practical way to capture mobility with a probability. Based on this definition, Fig. 5(d) shows the difference in mobility between low and super spreaders.



(a) Density heat map. (b) Distance CDF. (c) Unique locations per day. (d) Low/super spreaders.

Fig. 5. Mobility: Daily traveled distance and visited locations.

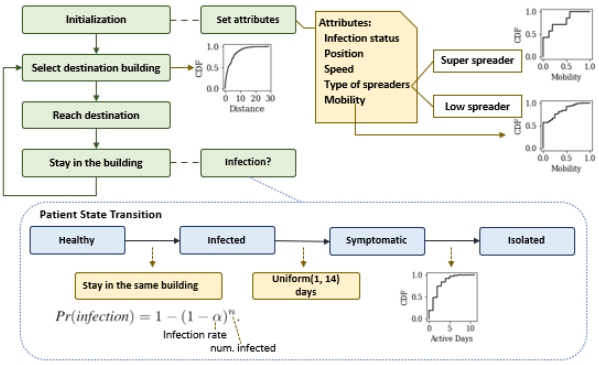


Fig. 6. Life cycle of an agent.

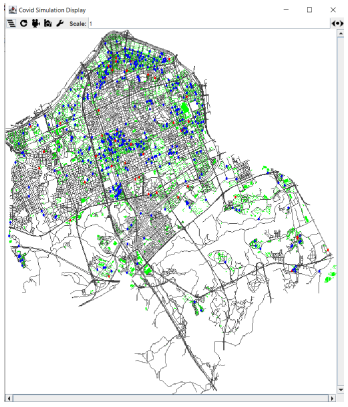


Fig. 7. Gangnam and Seocho Simulation

4 AGENT-BASED MODEL

In this section, we show how to parameterize a simulation based on our tool, GeoSpread, an expanded version of GeoMason [33] using the characterization presented in Section 3. A replication package containing the expanded version of GeoMason is available [31]. Attributes, life cycle, and states of each agent are shown in Fig. 6. 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/Distance.* Agents' speed and distance traveled. Both are fed to GeoSpread as distributions to model different movement habits. The CDF of daily traveled distance in Fig. 5(b) is used to determine the distance traveled.
(4) *Type of spreaders.* We define two classes of spreaders: 3.59% of patients are super spreaders and 96.41% are low spreaders (see Section 3.3).
(5) *Mobility.* We use the mobility of super spreaders and low spreaders depicted in Fig. 5(d) to model patient mobility.

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 [4].

If multiple agents are inside the same building, they may infect each other with a certain probability. When an infection happens, the agent state changes from healthy to infected. We assume the outdoor infection probability to be negligible. Given the probability of infection inside a building, α, 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:

$$Pr(infection) = 1 - (1 - \alpha)^n. \tag{1}$$

Note that the probability of infection defined by Eq. (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 [18] for examples on how to adjust Eq. (1).

It takes 1–14 days for patients to show COVID-19 symptoms after infection according to the WHO [28]. GeoSpread supports any distribution (e.g., Uniform, Exponential, Log-normal) to define the transition of an individual status from

infected to symptomatic. This allows capturing different scenarios and model future variants of SARS-CoV-2, different pathogens, or new viral outbreaks.

Consistent with infectious disease simulation studies [15], we set the simulation cycle to 5 minutes. The simulation terminates either when all agents are infected or after a number of cycles defined by the user.[1]

We simulate the COVID-19 outbreak in the neighboring Seocho and Gangnam districts, see Fig. 1(a). Roads and buildings are placed in the simulated area as described in [21], a collection of GIS data with regard to Seoul. GeoSpread loads the GIS data (e.g., roads, road intersections, buildings) stored in a shapefile format, i.e., a file containing geometric locations and their attribute information. We do not have any information on 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 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 Section 5.

A screenshot of the GeoSpread simulation execution that focuses on the Gangnam and Seocho districts can be seen in Fig. 7. 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., healthy (blue dots) or infected (red dots).

5 MODEL VALIDATION AND CASE STUDY

After presenting the generic model in Section 4, we showcase its flexibility. We use real data to validate GeoSpread, then we simulate different mitigation measures to assess their effectiveness when applied to Seocho and Gagnam districts.

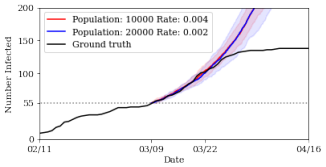
5.1 Validation

We focus on agents moving between Seocho and Gagnam. Fig. 1(c) 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 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. 1(c).

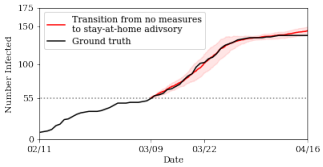
Since two districts are considered in this simulation, starting with only one infected agent in one of the two areas could bias the results. Here, we start the simulation with 55 infected agents, i.e., the number of infections observed from the data set on March 9, 2020, proportionally assigned to agents 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. 8(a) 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 two weeks

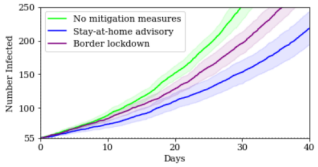
[1]In this simulation, we do not explicitly model agent recovery: a recovered agent that resumes its mobility is considered immune and non-contagious, therefore does not contribute to the disease spread. The simulation can be trivially extended to model recovered agents re-entering the simulation cycle.



(a) Simulation (no mitigation measures) vs. ground truth.



(b) Simulation (stay-at-home advisory) vs. ground truth.



(c) Comparison of mitigation measures.

Fig. 8. Infected population in the validation simulations. 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.

at the beginning of April. Our simulation in Fig. 8(a) does not model the effect of social distancing campaign so it is expected not to capture the knee of the ground truth curve.

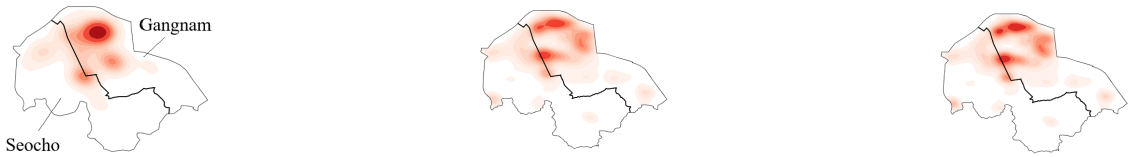
We align the beginning of simulation data to the time of 55 infection cases in the ground truth, since this is the starting point of the simulation. The two simulation lines in Fig. 8(a) (whose 95% confidence interval is represented by the shaded areas) closely follow the ground truth: the simulation 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. 8(a) 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.

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. Fig. 8(b) depicts the results of these simulations (with 95% confidence interval) 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. After applying the stay-at-home advisory mid-simulation, the simulation also exhibits a flattening trend, which is consistent with the ground truth.

Fig. 8(c) shows the effect of different mitigation measures on the number of infections in the two Seoul districts. The stay-at-home advisory is the most efficient counter-measure that keeps the number of cases below 250 people during the first 40 days. The number of infections is mildly contained when people cannot leave their home-district (i.e., border lockdown), while it sharply increases when no mitigation measures are taken. This highlights the ability of the model to capture what-if scenarios defined by different mitigation measures and patterns of population movement.

The accuracy of GeoSpread is also assessed through hotspot locations. In Fig. 9(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. Figs. 9(b) and 9(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.



(a) Ground truth (b) Simulation: 10K (c) Simulation: 20K

Fig. 9. Hotspots in the data set (ground truth) and model.

6 MODEL LIMITATIONS

Although the model is validated using ground truth, missing data limit its generalization. Limitations include:
First wave data. This data is from the first wave of the disease in South Korea. With South Korea having one of the best responses to the disease globally, the mobility patterns inevitably reflect cultural and demographic characteristics as well as policy decisions. We have no way to evaluate how mobility statistics changed during other waves of COVID-19.
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.
Transportation assumptions. The KCDC data set does not show the transportation mode of patients and we do not have data to parameterize this aspect of the simulation. Nevertheless, simulation results are in agreement with the ground truth. As for other input parameters, GeoSpread allows users to fully customize the ratio of pedestrian to vehicles in case this parameter is crucial to study new SARS-CoV-2 variants or future disease outbreaks.

7 RELATED WORK

The COVID-19 pandemic has been largely studied from different perspectives due to its disruptive effects. New habits forced by the unprecedented situation (i.e., usage of contact tracing apps, live streaming learning, and social media platforms for distress disclosure) are analyzed in [6, 17, 34, 40].

ABMs are a simulation-based alternative of mathematical models that incorporate human interactions [13]. ABMs are typically used for modeling pedestrian movements, human mobility during rare events (e.g., natural disasters), resource usage, and to successfully study the spread of diseases [7, 10, 20, 35].

Ferguson et al. [8] model the spread of influenza in British and American households, schools, and workplaces. Their simulations are parameterized using census and land use data. They use air travel patterns (i.e., large scale international population movements) to model people mobility. ABMs parameterized by census data have been used to capture the spread of COVID-19 in Australia [5, 30]. Using census and age-distribution data from Germany and Poland, Bock et al. [3] 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 [11], evaluate the efficiency of contact tracing [1], face masks [12], and testing strategies [36]. Kim et al. [15] 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. Souza et al. [32] use geo-located data from social networks (i.e., Twitter) to identify hotspots that facilitate the spread of infectious diseases (i.e., Dengue). ABMs are used also to model the spread of SARS-CoV-2 in small areas: crowded areas of supermarkets [39] and university campuses [9]. *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. [18, 19] 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. These are the closest to our work but they do not provide any detailed statistics on agent mobility during the pandemic as we do here.

Summarizing, in this paper 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.

8 CONCLUSIONS

In this paper, we extract human movement habits and dynamics from the KCDC data sets of real COVID-19 patients. Mobility statistics are used to tune an ABM used by our tool, GeoSpread, and to investigate the COVID-19 outbreak in two districts of Seoul. The proposed approach allows investigating scenarios under different circumstances to identify mitigation strategies. Simulation results are in excellent agreement with ground truth and show that this model can be used to flexibly examine and evaluate the spread of COVID-19 (and new disease outbreaks) in an urban setting. While we do not claim that it is a definitive COVID-19 spread model, it can be used to investigate useful *what-if* scenarios (e.g., mitigation measures) and future infectious diseases (e.g., more aggressive SARS-CoV-2 variants or new pathogens). We also plan to extract new information from the data set used in [18, 19] to investigate the impact of public transport.

ACKNOWLEDGMENTS

The authors would like to thank David Dowdy for his constructive feedback on the manuscript. This work is supported by the following grants: National Science Foundation IIS-2130681, IIS-1838022, and MIUR PRIN project SEDUCE 2017TWRCNB.

REFERENCES

[1] Jonatan Almagor and Stefano Picascia. 2020. Exploring the effectiveness of a COVID-19 contact tracing app using an agent-based model. *Scientific reports* 10, 1 (2020), 1–11.

[2] Aaron S Bernstein, Amy W Ando, Ted Loch-Temzelides, Mariana M Vale, Binbin V Li, Hongying Li, Jonah Busch, Colin A Chapman, Margaret Kinnaird, Katarzyna Nowak, et al. 2022. The costs and benefits of primary prevention of zoonotic pandemics. *Science advances* 8, 5 (2022), eabl4183.

[3] Wolfgang Bock, Barbara Adamik, Marek Bawiec, Viktor Bezborodov, Marcin Bodych, Jan Pablo Burgard, Thomas Goetz, Tyll Krueger, Agata Migalska, Barbara Pabjan, et al. 2020. Mitigation and herd immunity strategy for COVID-19 is likely to fail. *medRxiv* (2020).

[4] CDC. 2020. Quarantine and Isolation. https://www.cdc.gov/coronavirus/2019-ncov/your-health/quarantine-isolation.html. [Online; 2022-05-23].

[5] Sheryl L Chang, Nathan Harding, Cameron Zachreson, Oliver M Cliff, and Mikhail Prokopenko. 2020. Modelling transmission and control of the COVID-19 pandemic in Australia. *Nature communications* 11, 1 (2020), 1–13.

[6] Zhilong Chen, Hancheng Cao, Yuting Deng, Xuan Gao, Jinghua Piao, Fengli Xu, Yu Zhang, and Yong Li. 2021. Learning from Home: A Mixed-Methods Analysis of Live Streaming Based Remote Education Experience in Chinese Colleges during the COVID-19 Pandemic. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[7] Andrew Crooks and Atesmachew Hailegiorgis. 2014. An agent-based modeling approach applied to the spread of cholera. *Environmental Modelling & Software* 62 (2014), 164–177.

[8] Neil M Ferguson, Derek AT Cummings, Christophe Fraser, James C Cajka, Philip C Cooley, and Donald S Burke. 2006. Strategies for mitigating an influenza pandemic. *Nature* 442, 7101 (2006), 448–452.

[9] Philip T Gressman and Jennifer R Peck. 2020. Simulating COVID-19 in a university environment. *Mathematical Biosciences* 328 (2020).

[10] 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. 2016. Lessons from the Ebola outbreak: action items for emerging infectious disease preparedness and response. *EcoHealth* 13, 1 (2016), 200–212.

[11] Masoud Jalayer, Carlotta Orsenigo, and Carlo Vercellis. 2020. CoV-ABM: A stochastic discrete-event agent-based framework to simulate spatiotemporal dynamics of COVID-19. *arXiv preprint arXiv:2007.13231* (2020).

[12] De Kai, Guy-Philippe Goldstein, Alexey Morgunov, Vishal Nangalia, and Anna Rotkirch. 2020. Universal masking is urgent in the covid-19 pandemic: Seir and agent based models, empirical validation, policy recommendations. *arXiv preprint arXiv:2004.13553* (2020).

[13] 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. 2013. Selecting among five common modelling approaches for integrated environmental assessment and management. *Environmental modelling & software* 47 (2013), 159–181.

[14] Jihoo Kim and JoongKun Lee. 2020. Data Science for COVID-19 (DS4C). https://www.kaggle.com/kimjihoo/coronavirusdataset. [Online; 2022-05-23].

[15] Joon-Seok Kim, Hamdi Kavak, Chris Ovi Rouly, Hyunjee Jin, Andrew Crooks, Dieter Pfoser, Carola Wenk, and Andreas Züfle. 2020. Location-based social simulation for prescriptive analytics of disease spread. *SIGSPATIAL Special* 12, 1 (2020), 53–61.

[16] Sun Kim and Marcia C Castro. 2020. Spatiotemporal pattern of COVID-19 and government response in South Korea (as of May 31, 2020). *International Journal of Infectious Diseases* 98 (2020), 328–333.

[17] Xi Lu, Tera L. Reynolds, Eunkyung Jo, Hwajung Hong, Xinru Page, Yunan Chen, and Daniel A. Epstein. 2021. Comparing Perspectives Around Human and Technology Support for Contact Tracing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[18] Sebastian A Müller, Michael Balmer, William Charlton, Ricardo Ewert, Andreas Neumann, Christian Rakow, Tilmann Schlenther, and Kai Nagel. 2020. A realistic agent-based simulation model for COVID-19 based on a traffic simulation and mobile phone data. *arXiv preprint arXiv:2011.11453* (2020).

[19] Sebastian A Müller, Michael Balmer, William Charlton, Ricardo Ewert, Andreas Neumann, Christian Rakow, Tilmann Schlenther, and Kai Nagel. 2021. Predicting the effects of COVID-19 related interventions in urban settings by combining activity-based modelling, agent-based simulation, and mobile phone data. *medRxiv* (2021).

[20] Yanbo Pang, Kota Tsubouchi, Takahiro Yabe, and Yoshihide Sekimoto. 2020. Intercity Simulation of Human Mobility at Rare Events via Reinforcement Learning. In *Proceedings of the International Conference on Advances in Geographic Information Systems*. 293–302.

[21] BBBike.org. 2020. OSM extracts for Seoul. https://download.bbbike.org/osm/bbbike/Seoul/Seoul.osm.shp.zip. [Online; 2022-05-23].

[22] BigQuery Public Datasets Program. 2020. COVID-19 Open Data. https://console.cloud.google.com/marketplace/product/bigquery-public-datasets/covid19-open-data. [Online; 2022-05-23].

[23] Google. 2020. Google Maps. https://www.google.com/maps/. [Online; 2022-05-23].

[24] Kakao Corporation. 2020. Kakao Map. https://map.kakao.com/. [Online; 2022-05-23].

[25] Ministry of Health and Welfare of South Korea. 2020. Coronavirus Disease-19, Republic of Korea. http://ncov.mohw.go.kr/en/. [Online; 2022-05-23].

[26] Naver Corporation. 2020. Naver Map. https://m.map.naver.com/. [Online; 2022-05-23].

[27] OpenStreetMap. 2020. OpenStreetMap. https://www.openstreetmap.org/. [Online; 2022-05-23].

[28] WHO. 2020. Coronavirus disease (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses. [Online; 2022-05-23].

[29] WHO. 2020. 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. [Online; 2022-05-23].

[30] 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. 2020. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nature medicine* (2020), 1–7.

[31] Anna Schmedding, Lishan Yang, Riccardo Pinciroli, and Evgenia Smirni. 2022. Replication Package: GeoSpread: an Epidemic Spread Modeling Tool for COVID-19 Using Mobility Data. https://github.com/akschmedding/GeoSpread.

[32] Roberto CSNP Souza, Renato M Assunção, Daniel B Neill, and Wagner Meira Jr. 2019. Detecting spatial clusters of disease infection risk using sparsely sampled social media mobility patterns. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 359–368.

[33] Keith Sullivan, Mark Coletti, and Sean Luke. 2010. *GeoMason: Geospatial support for MASON.* Technical Report. Department of Computer Science, George Mason University.

[34] Christine Utz, Steffen Becker, Theodor Schnitzler, Florian M Farke, Franziska Herbert, Leonie Schaewitz, Martin Degeling, and Markus Dürmuth. 2021. Apps against the spread: Privacy implications and user acceptance of COVID-19-related smartphone apps on three continents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.

[35] Srinivasan Venkatramanan, Bryan Lewis, Jiangzhuo Chen, Dave Higdon, Anil Vullikanti, and Madhav Marathe. 2018. Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics* 22 (2018), 43–49.

[36] Yingfei Wang, Inbal Yahav, and Balaji Padmanabhan. 2020. Whom to Test? Active Sampling Strategies for Managing COVID-19. *arXiv preprint arXiv:2012.13483* (2020).

[37] WHO. 2022. Multi-country monkeypox outbreak in non-endemic countries: Update. https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON388. [Online; 2022-05-30].

[38] WHO. 2022. Multi-Country – Acute, severe hepatitis of unknown origin in children. https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON376. [Online; 2022-05-30].

[39] Fabian Ying and Neave O'Clery. 2021. Modelling COVID-19 transmission in supermarkets using an agent-based model. *Plos one* 16, 4 (2021).

[40] Renwen Zhang, Natalya N. Bazarova, and Madhu Reddy. 2021. Distress disclosure across social media platforms during the COVID-19 pandemic: Untangling the effects of platforms, affordances, and audiences. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.