Individual-based epidemiological model of COVID19 using location data

Yoriyuki Yamagata*, Shunki Takami†, Keisuke Yamazaki‡ Tomoki Nakaya§ Masaki Onishi†.
*National Institute of Advanced Industrial Science and Technology (AIST), Ikeda, Japan
email: yoriyuki.yamagata@aist.go.jp
†National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan
s-takami@aist.go.jp; onishi-masaki@aist.go.jp
‡National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
k.yamazaki@aist.go.jp
§Graduate School of Environmental Studies, Tohoku University, Miyagi, Japan
email: tomoki.nakaya.c8@tohoku.ac.jp

***Abstract*—Because human movement spreads infection, and mobility is a good proxy for other social distancing measures, human mobility has been an important factor in the COVID19 epidemic. Therefore, the control of human mobility is one of the countermeasures used to suppress an epidemic.**

As a notable feature, COVID19 has had multiple waves (subepidemics). Understanding the causes of the start and end of each wave has important implications for a policy evaluation and the timely implementation of countermeasures. Some of the waves have been correlated with the changes in mobility, and some can be attributed to the emergence of new variants. However, the start and end of some of the waves are difficult to explain through known factors.

To evaluate the effect of human mobility, we built a stochastic model incorporating individual movements of 500,000 people obtained from anonymized, user-approved location data of smartphones throughout Japan. Instead of using aggregate values of human mobility, our model tracks the movements of individuals and predicts the infection of all persons within the entire country. Although the model only has a single static parameter, it successfully reproduced the occurrence of three waves of the number of confirmed cases within the study period of March 01 to December 31, 2020 in Japan. It was previously difficult to explain the end of the second wave and the start of the third wave in the study period by human mobility alone. Our results suggest the importance of tracking individual movements instead of relaying the aggregate values of human mobility.

***Index Terms*—Big data, location data, COVID-19, epidemiology, simulation.**

I. Introduction

the outbreak of COVID-19 (caused by SARS-CoV-2) has become a pandemic with a scale comparable to that of the Spanish flu. COVID-19 has caused 6.41 million confirmed deaths worldwide as of August 4, 2022, [1]. An excess mortality of 18.2 million [2] individuals, attributable to COVID-19, was estimated from January 1, 2020 to December 31, 2021.

Facing this crisis, governments implemented unprecedented measures such as lockdowns, border closures, and mask mandates. However, these countermeasures have greatly restricted the lives of citizens and placed a heavy burden on national economies. Therefore, an efficient and timely implementation of such countermeasures is extremely important.

A mathematical model of the epidemic can be an important tool for modeling the efficiency and timeliness of a countermeasure policy before it is implemented. As another benefit of a mathematical model, it can be used to predict the future course of an epidemic under a specific scenario, allowing us to allocate the necessary healthcare resources in advance.

To address these demands, many mathematical models have been proposed, ranging from simple SIR models to complex models that incorporate factors representing human mobility [3], [4], [5]. However, most of these models suffer from serious shortcomings. One of the notable features of the current COVID-19 pandemic is that it occurred in multiple waves. Understanding the causes of the start and end of each wave and predicting the timing and size of the next wave are important with direct implications on the choice and timing of an intervention. Unfortunately, most mathematical models proposed thus far only predict a single wave of an epidemic or require changes to the model parameters to explain multiple waves. Because required changes to the model parameters are unknowable before a wave arrives, we cannot use such models to access the efficiency of a policy and predict the timing and scale of a future wave.

In this paper, we propose a stochastic model using anonymized, user-approved location data from 500,000 smartphones throughout Japan. Unlike other models that rely on the aggregate values of human mobility such as Google's Community Mobility Reports and Apple's mobility data, our model tracks the movement of individual persons and predicts the infection of each individual. Applying our model to the epidemic in Japan from March 1, 2020 to December 31, 2020, our model successfully reproduced the occurrence of three waves in terms of the number of confirmed cases using mobility data alone, without changing a model parameter. The end of the second wave and the start of the third wave seem difficult to explain using the aggregate mobility index, and our results suggest the importance of the movement of individuals. It is surprising that a model with only a single parameter, which does not change over time or location, can predict the long-term (10 month) behavior of the epidemic.

Because our model has only a single parameter, its prediction does not fully fit the observed data. Nevertheless, our study shows the effectiveness of approaches using large-scale location data and possible future enhancements.

The remainder of this paper is organized as follows. Section II discusses previous related studies. Section III introduces status of COVID-19 and countermeasures implemented in Japan. In Section IV, we introduce the proposed method. Section V presents the experimental results. Finally, Section VI provides some concluding remarks and discusses possible future areas of research.

II. Related work

Because our study focuses on a model based on mobility, we discuss previous studies concerning mobility with regard to COVID-19. Because moving people spread the infection of the virus, mobility is a direct cause of epidemic, as well as a good proxy of social distancing in general.

The correlation between the epidemic and the aggregate mobility index has been discussed [6], [7], [8], [9], [10]. Kraemer et al. [6] showed that human mobility from Wuhan accurately predicted the epidemic in other parts of China during the early stage of the outbreak, whereas the correlation decreased after the mobility from Wuhan was restricted. Yabe et al. [7] found that Google's COVID-19 Community Mobility Reports [11] correlated the effective reproduction rate R_t in Tokyo during the early stage of the epidemic (March 22 to April 15), strong enough for containment of the epidemic. Nagata et al. found that changes in mobility, particularly in nighttime activities, was effective for containment in Japan during the early stage of the epidemic (March to July). These studies suggest that human mobility is an important factor in the COVID-19 pandemic.

By contrast, Nouvellet et al. [9] and Gatalo et al. [10] found that the correlation between the mobility index and epidemic was reduced after the initial stage of the epidemic in many countries and concluded that social distancing measures other than mobility restriction were important in controlling the epidemic.

In our study, the model predicted that the number of cases would fluctuate (Figs. 3, 4, 5 and 6) despite the relatively high mobility index after July 2020 in Japan (Fig. 1). This suggests that despite Nouvellet et al. and Gatalo et al. finding a decoupling between the mobility index and epidemic, the mobility at a finer granularity may still be an important factor.

There are many studies that have used mathematical models incorporating human mobility [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30] With the exception of [12], [13], [23], [24], most of these studies do not address the existence of multiple waves. Whereas Kexiras and Neofotistos [13] and Silva et al. [23] explained multiple waves by changing the model parameters, our model can explain the multiple waves of an epidemic using a single constant parameter throughout the entire period under consideration. Liu and Yamamoto [24] also explained multiple waves in Japan by changing parameters. Their model used an aggregate mobility index whereas our model uses the movement of individuals. Rahmandad et al. built a model in which mobility is endogenous and explains multi-wave behavior. By contrast, our model uses the observed movement of individuals as an exogenous variable. Therefore, our model is more suitable for evaluating the impact of mobility changes, observed or planned.

As our approach, some studies incorporate fine-scale mobility data. However, none of them performed long-term (10 moths) simulation and reconstructed the multiple waves of the epidemic. Therefore, our study would be more useful to access the effect of fine-scale mobility to the epidemic, in particular, the occurrence of the multiple wages. Fan et al. [30] and Yang et al. [29] proposed a city-wide individual-based model. Both models consists of the contagion and movement stages. In contagion stage, diseases propagate in each geographic grid. In movement state, the human movement spreads diseases to different grids. In our model, meta-populations are people having similar movements, while Fu et al. and Yang et al. used geological grids as meta-populations. Further, we used real-time mobility data while they used the past data. Chang et al. [31] uses a meta-population SIR model, in which each meta-population consists of a census block group and its contact with another meta-population is derived from POS visit data. Chang et al. used the different definition of meta-populations (census block groups) then us. Aleta et al. [15] proposed an individual-based model which tracks each synthetic individual by detailed POI visit data. Kerr et al. [26] also proposed an individual-based model using synthetic population and contact network to evaluate the effect of different interventions. Chiba [21] used a synthetic population and its movement based on the census and location data of mobile phones and evaluated the effectiveness of government interventions. Both used synthetic individuals, not real individuals, Some works used fine-scale, individual-based large-scale mobility data probided by Jiang et al. [28] and Cao et al. [27] used real-time data provided Blogwatcher Inc., the same company as ours. Unlike us, they treated each prefecture as a single meta-population and the mobility data are used to compute the movement between prefectures.

Numerous theoretical studies have focused on the spatiotemporal behavior of infectious diseases. For example, Weng and Zao [32] considered wave-like propagation, and Sun [33] considered the formation of spatial patterns. Such studies might be related to the wave-like behavior of our model.

III. Context

Because our research focuses on 2020, we only discuss the status in Japan during this period. We exclude the years 2021 and onward because new variants and vaccinations complicated the modeling.

Figure 1 shows the number of new infections based on the infection dates estimated in Section IV-A2 and the changes in mobility in transit stations according to Google's COVID-19 community mobility report [11], both for Tokyo, a prefecture with largest population in Japan. Three waves were observed

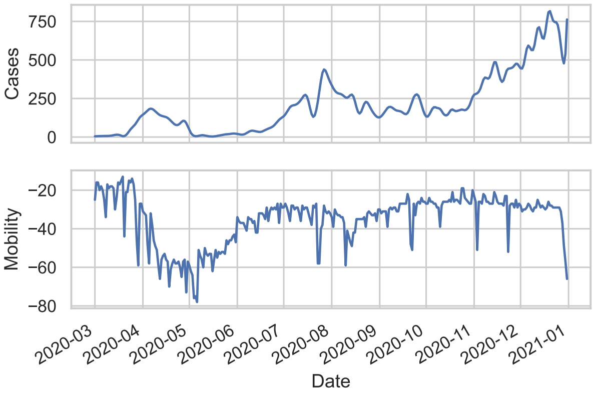


Fig. 1. New infection and mobility in Tokyo

in terms of the number of infections. The first wave started in the middle of March and ended at the end of April. The second wave began around the beginning of June and ended at the end of August. The third wave began at the beginning of November.

A large decrease in mobility was observed from April to June. After the end of June, mobility remained relatively high, except for a small drop around the first half of August, which corresponds to the summer vacation period.

From April 7 to May 25, the central government declared a state of emergency. During this state of emergency, the government asked citizens to stay at home and businesses to close or shorten their operating times. These measures were not enforced because there is no legal system to enable their enforcement. However, these non-compulsory measures significantly reduced the mobility [7], which was sufficient for containing the epidemic. The state of emergency corresponds to the drop in mobility from April to June. After the end of the emergency period, the mobility slowly recovered. During the study period, the government asked the citizen to avoid crowded places, wear masks, and refrain from long trips, particularly during vacation periods.

The start and end of the first wave can be easily explained by the introduction of the virus by people returning from Europe and the declaration of the state of emergency. In addition, the start of the second wave can easily be explained by the recovery of mobility. However, the end of the second wave is difficult to explain because mobility remained relatively high after July. Furthermore, it is difficult to explain the relatively low infection rate around September and October and the start of the third wave by mobility alone.

IV. Methods

A. Data acquisition and preprocessing

1) Location data: We obtained anonymized, commercially available location data for mobile devices, collected by Blogwatcher Inc., from the user-approved location functions of mobile applications. The location data consist of an encrypted *ADID* using the cryptographic hash function, latitude, longitude, and timestamp of an observed smartphone location. An ADID represents either a *Google Advertising ID (ADID)* or *Identifier For Advertising (IDFA)* specified by Apple, which is unique to each smartphone but can be randomly reset. Only first two authors handled the raw location data. The institute implements a strict measure of privacy protection, based on the Japanese Act on the Protection of Personal Information.

We have location data for more than five million smartphones per day, during the study period. However, not all devices are suitable because some ADIDs may have changed during the study period. Therefore, we only used devices whose ADIDs were observed in the first and last months of the study period. Furthermore, the locations of many smartphones are infrequently observed. We therefore only used smartphones whose locations were observed once per hour on average. To reduce the memory consumption, we sampled 500,000 devices from the selected devices. Furthermore, we assumed that each participant had only a single smartphone.

Our model uses a *contact matrix*, which indicates the number of hours each pair of individuals i and j stayed within the same grid cell of approximately $200\mathrm{m} \times 200\mathrm{m}$ in size based on the i-th column and j-th row of a matrix of a specified day. To obtain the contact matrices, we computed the grid cell and time, in 1 h units, for each person. If the location of a person was not observed for a particular hour, we assumed that the person stayed at the last observed location. If the location of a person was observed more than twice, the location was randomly chosen. We then tallied the number of hours in which the i-th and j-th persons were in the same grid cell for i and j.

We assumed that the i-th person lived in prefecture p during the study period if i was observed most frequently in prefecture p.

2) Number of infection incidents: We used the number of infections published by the public broadcaster NHK [34], which contains the reported number of cases as counted by the reported date for each prefecture. However, we wanted to apply our model to the number of infection incidents as counted by the infection dates, not the reported number of cases, because the reported number is delayed by the incubation period and the reporting process. Furthermore, the reported number fluctuated spuriously on weekdays because of testing and reporting practices.

To remove the weekday bias of the reported numbers, we took seven-day moving averages of the reported numbers.

We then estimated the number of cases counted based on the infection date using the standard back-projection technique [35]. For this purpose, we used the dataset containing the onset and confirmed and reported dates of individual patients, obtained from expert members of the National COVID-19 Cluster Task Force in Japan, who compiled publicly available information on positive polymerase chain reaction cases released by each local authority. We estimated the distribution of delays from onset to reporting by fitting a log-normal distribution to delays in the dataset. We applied a back-projection using the log-normal distribution and obtained the number of cases of onset on each day. We conducted a further back-projection based on the incubation period [36] to estimate the number of infections each day.

We referred to the number of cases counted by the infection date as the observed number of cases.

3) Reporting rate: Because unreported infections are known to be important for the spread of infection [14], our model incorporates an under-reporting.

We estimated the reporting rate based on the number of cumulative deaths and the infection fatality rate, because the mortality would be more accurately reported. We assumed that the reporting rate remained constant during the study period. The number of cumulative deaths caused by COVID19 was 3414 by the end of 2020, whereas the cumulative number of reported cases was 228 418 [37]. The infection fatality rate was estimated to be 0.68% [38]. Using these figures, the true number of cumulative infections was estimated to be approximately 502 058, indicating a reporting rate of approximately 50%. This 50% reporting rate is consistent with the ratio of asymptomatic to total number of cases [39].

4) Epidemiological parameters: Our model used the epidemiological parameters determined in previous studies.

The generation time is the interval between when person A is infected and when A infects another person B. We assumed a distribution of generation time as specified in EpiNow2 [40], which is based on Ganyani et al.'s approach [41] but modified using the incubation period obtained by Lauer et al. [36].

The viral load of each patient differs significantly by several orders of magnitude. We modeled the distribution of the viral load based on a gamma distribution, the dispersion of which was determined by Endo et al. [42].

B. Mathematical model

Because we sampled approximately 500 000 devices out of a total population of approximately 125 800 000, there are approximately 252 people represented by each sampled device, which we call the meta-population of the sample. Contacts were assessed on a meta-population basis; that is, we assumed that all persons in the same meta-population had the same contact time with any given person in any other meta-population (Fig. 2).

Our model is a stochastic epidemiology model that explicitly addresses individual transmissions. Variables in bold indicate vectors or matrices. For example, $\boldsymbol{I}$ is a vector whose ith element is $\boldsymbol{I}[i]$. Let S be the ratio of the total population of Japan to the number of sampled smartphones.

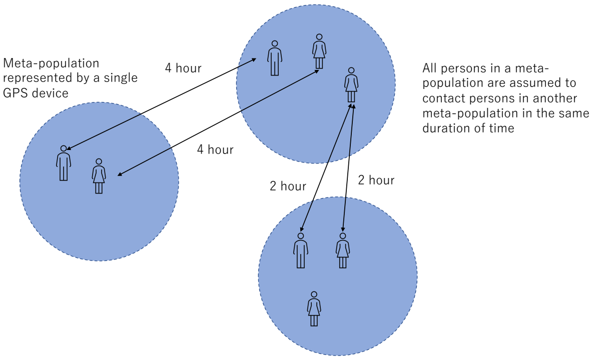

Meta-population represented by a single GPS device
4 hour
4 hour
All persons in a meta-population are assumed to contact persons in another meta-population in the same duration of time
2 hour
2 hour

Fig. 2. All persons in a meta-population as represented by a single device contacting persons in another meta-population for the same duration of time.

$$\boldsymbol{L}[d] \sim \mathrm{Gamma}(D, D/\boldsymbol{I}[d])$$

$$\boldsymbol{V}[d] = \sum_{k=1}^{K} g[k]\boldsymbol{L}[d-k]$$

$$\boldsymbol{F}[d] = \boldsymbol{M}[d]\boldsymbol{V}[d]$$

$$\boldsymbol{I}[d+1] \sim \mathrm{Binomial}(S - \boldsymbol{C}[d], 1 - e^{\beta \boldsymbol{F}[d]})$$

$$\boldsymbol{C}[d+1] = \boldsymbol{C}[d] + \boldsymbol{I}[d]$$

$$\boldsymbol{I}_p[d] \sim \mathrm{Binomial}\left(\sum_{i \text{ lived } p} \boldsymbol{I}[d][i], q\right)$$

Here, $\boldsymbol{I}[d][i]$ is the number of new infections occurring in meta-population i on day d, and $\boldsymbol{L}[d][i]$ is the sum of the viral loads of the people in meta-population i who were infected on day d. Note that the sum of the gamma distributions with dispersion D is the gamma distribution with dispersion D. In addition, $\boldsymbol{V}[d][i]$ is the total viral emission of meta-population i on day d; $g[k]$ is the distribution of the generation time; and $\boldsymbol{M}[d][i,j]$ is the contact matrix on day d, indicating the number of hours that i and j remained in the same location grid cell. $\boldsymbol{M}$ is precomputed from the location data, and thus is given outside the model. Next, $\boldsymbol{F}[d][i]$ is the viral emission to which a person in a meta-population i is exposed on day d. $\boldsymbol{I}[d+1][i]$, the number of infected people in meta-population i on day $d+1$, is determined through a sampling of people who are not yet infected in meta-population i on day n based on the probability determined through $\boldsymbol{F}[d][i]$ and the infection rate β. $\boldsymbol{C}[d]$ is the number of cumulative infections. Finally, p is a prefecture, and $\boldsymbol{I}_p[d]$ is the observed number of infected people in prefecture p on day d. q denotes the reporting rate.

The only unknown parameter in the model is β. The values of D and $g[k]$ were determined based on previous studies, and $\boldsymbol{M}$ was determined using location data. Other variables, $\boldsymbol{L}, \boldsymbol{V}, \boldsymbol{F}, \boldsymbol{I}, \boldsymbol{C}$, are endogenous. We assigned $\boldsymbol{I}[-K], \boldsymbol{I}[-K+1], \cdots, \boldsymbol{I}[-1]$ as the initialization conditions for the model. The value of $\boldsymbol{I}[-k]$ is determined as follows:

$$\boldsymbol{I}[-k][i] \sim \mathrm{Binomial}(S, \boldsymbol{I}_p[-k]/P_p) \tag{1}$$

if i lives in prefecture p, and P_p is the population of p.

C. Parameter estimation

In our model, only the parameter β is unknown. We assume that β is constant across the study period and throughout Japan. We determine β by fitting the observed number of infection cases in each prefecture p using a Bayesian optimization based on the tree Parzen estimator (TPE) [43].

To define a loss function, we assume that $I[d][i]$ and $I[d][j]$ are probabilistically independent, although this assumption is unlikely to be correct. Subsequently, the number of new infections in prefecture p on day d follows a Poisson distribution.

$$\boldsymbol{I}_p[d] \sim \text{Poisson}\left(\sum_{i \text{ lives } p} E[\boldsymbol{I}[d][i]]\right) \tag{2}$$

$$= \text{Poisson}(E[\boldsymbol{I}_p[d]]) \tag{3}$$

However, it was impossible to determine $E[\boldsymbol{I}_p[d]]$. Therefore, we estimated $E[\boldsymbol{I}_p[d]]$ by running the simulations N times and using Bayesian inference. Let $\boldsymbol{I}_p[k][d]$ be the number of new infections in prefecture p on day d for the k-th simulation, and NB be a negative binomial distribution. We used $\text{Gamma}(1, 1)$ as the prior, and the posterior was

$$\boldsymbol{I}_p[d] \sim \text{NB}\left(1 + \sum_{k=1}^{N} \boldsymbol{I}_p[k][d], \frac{1}{1+N}\right). \tag{4}$$

Let $\boldsymbol{I}_p^{\text{obs}}[d]$ be the observed number of new infections in prefecture p on day d. We then define the loss function as

$$\sum_{p,d} \log P\left(\boldsymbol{I}_p^{\text{obs}}[d] \;\middle|\; \text{NB}\left(1 + \sum_{k=1}^{N} \boldsymbol{I}_p[k][d], \frac{1}{1+N}\right)\right). \tag{5}$$

We call a simulation run a *trial*, and an optimization process a *study*.

D. Implementation and experiment setup

The contact matrices from the location data were computed using Apache Spark [44]. The epidemiological model was implemented using Pytorch [45], which utilized a GPU for the matrix computations. Optuna [46] was used for the parameter optimization.

To run Spark, we used a cluster of 16 machines equipped with two 2.4-GHz Intel Xeon Gold 6148 20 core CPUs with 384 GB of RAM. The computation took approximately 1 d.

Optimization was conducted on a single machine with the same CPU and memory as above, along with four GPUs (NVIDIA V100 for NVLink 15 GiB HBM). There were four processes, each conducting an optimization using a single GPU communicating through a database (SQLite). We limited the time for optimization to 2 d.

For each simulation, we set $K = 14$ to compute $\boldsymbol{V}[d]$ and for the initialization. The possibility of a generation interval of more than 14 d was considered negligible. We used $N = 120$ simultaneous simulations, as described in Section IV-C, running 30 simulations as a batch using a GPU at once. We searched for the optimal β for $0 \leq \beta \leq 1$.

We estimated β by fitting the model to three different periods, March 01 to April 30, March 01 to August 31, and finally, March 01 to December 31, 2020. The period of March 01 to April 30 corresponds to the first wave, whereas the period of March 01 to August 31 corresponds to the first and second waves. The period of March 01 to December 31 is the entire study period, which contains the first to third waves.

Using β obtained by the optimization, we reproduced or predict the number of cases, running 120 simulations simultaneously. As the optimization, the model was initialized by the number of cases from 14 to 1 d before March 01, 2020. After initialization, the model was run without any input from the observed data.

V. Results

We performed two kinds of experiments. First, we fitted the model to the observation during the entire observed period, March 01, 2020 to December 31, 2020 and estimated the model parameter. A simulation, starting March 01, 2020, performed using the parameter above and compared to the observed values. This experiment shows how well our model can reproduce the observation, thus how well our model reflects the underling mechanism of the epidemic. The result of the first experiment is presented in Section V-A.

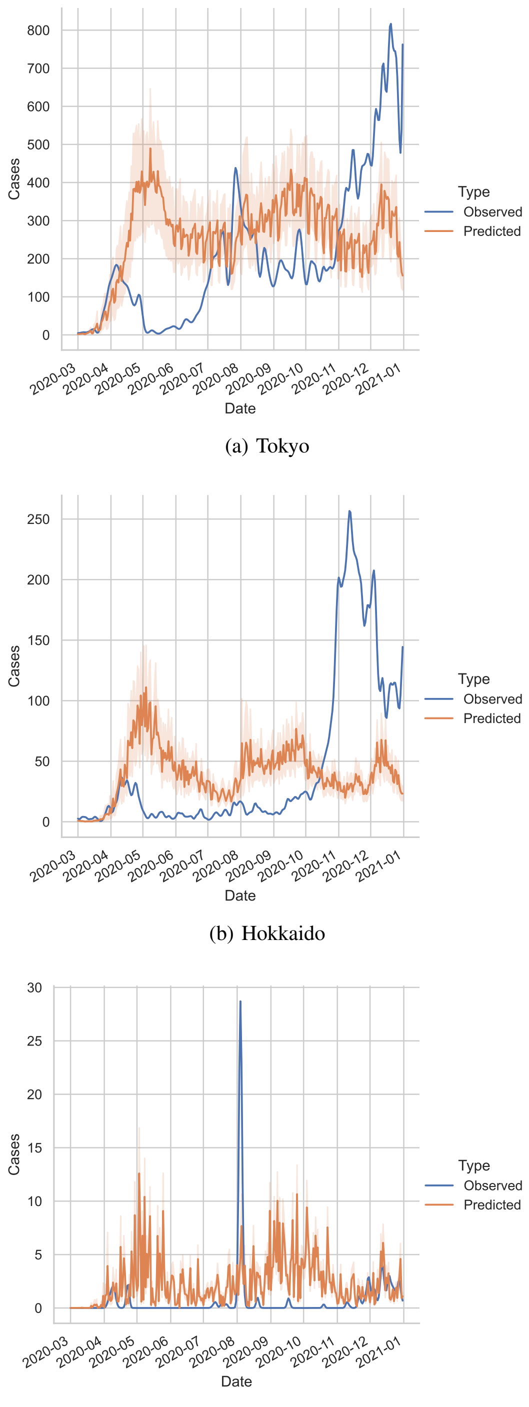
Second, we fitted the model to the observation during the first wave (March 01 to April 30) and the second wave (March 01 to August 31) and estimated the model parameter. Using the parameter obtained, we performed the simulation the number of cases from March 01, 2020 to December 31, 2020 and compared to the observation. Because this model only used the part of observation, this experiment shows how useful our model is for prediction. Because there seems the history effect in the model, we cannot start the simulation from the arbitrary point. Instead, we started the simulation from March 01, 2020, when the cases would be randomly scattered across the country.

A. Reproduction

In this section, we fitted the model to the entire study period. We simulated the number cases using the model parameter obtained, and compared to the observation. We ran five independent studies to obtain the optimized β, and then chose the best value among those obtained. The optimal beta is $\beta = 0.000995$. A total of 7301 trials were conducted.

Fig. 3 shows the observed and simulated number of cases. Our model simulated all cases in Japan, and we selected Tokyo, Hokkaido, and Shimane from among the 47 prefectures in Japan. Tokyo is the largest population center in Japan and has experienced a sustained epidemic. Hokkaido is another population center that has experienced an outbreak. Shimane is a less-densely populated area. The blue line represents the number of observed cases. The orange line is the average of all outputs of the simulations, and the orange band shows the standard deviation.

We note the following:



(a) Tokyo

(b) Hokkaido

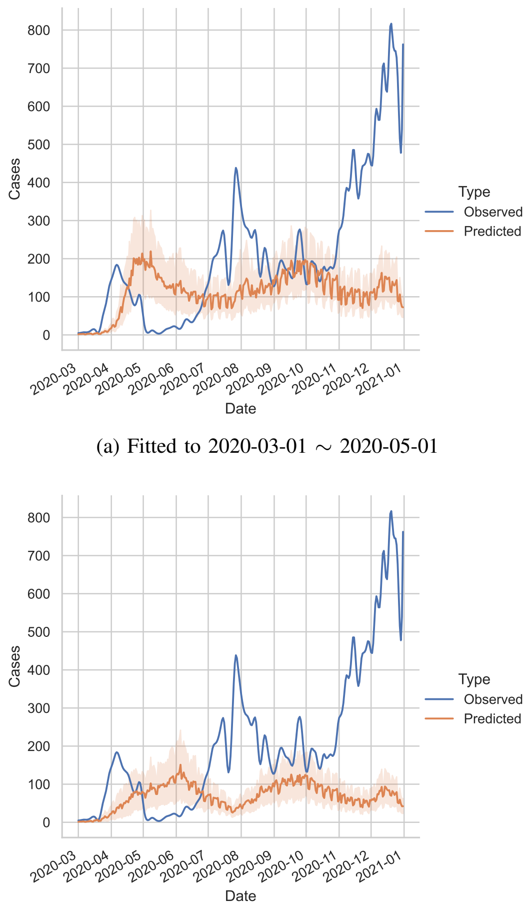
(c) Shimane

Fig. 3. Fit with the observed data

- Regional differences: The model predicted a three-wave pattern in all prefectures, whereas the observed cases in Hokkaido had only two peaks, and a single outbreak occurred in Shimane.
- The period of each wave: The tail of the first wave is longer, and the start of the second wave is delayed in the model prediction compared with the observed data.
- Intensity: Although all wave peaks are predicted to be almost equal, the peaks were increased in the observed data.

B. Prediction

We also investigated the ability of our model for prediction. First, we fitted β for the period of March 01 to April 30, namely, the first wave, and the period from March 01 to August 31, 2020, namely, the first and second waves. The, a value of $\beta = 0.000946$ was obtained by fitting the period from March 01 to April 30, 2020 through 32 044 trials. Similarly a value of $\beta = 0.000858$ was obtained by fitting the period from March 01 to August 31, 2020 using 12 488 trials. Using these β, we ran the simulations from March 01, 2020 to December 31, 2020 and compared to the observed values. We started the simulations from March 01, 2020, not from the end of the periods we used for training, because the history effect seems to exist in our model.

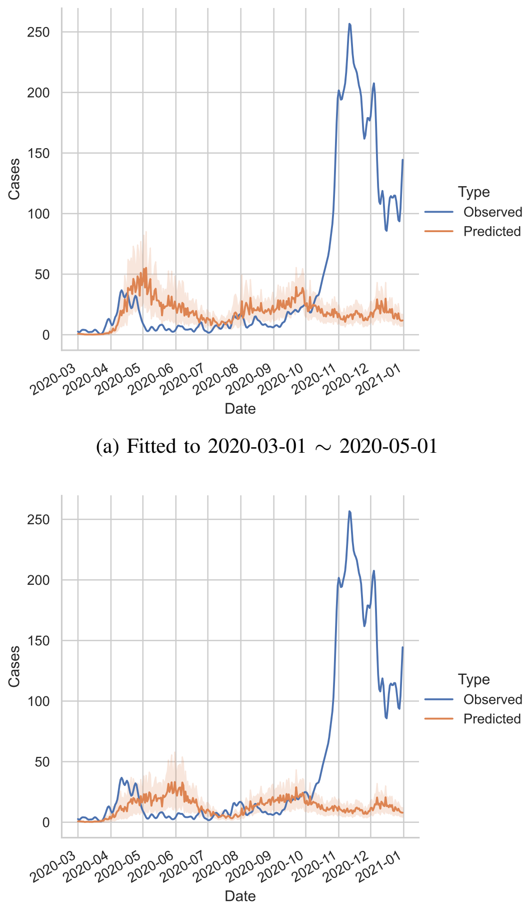


(a) Fitted to 2020-03-01 ∼ 2020-05-01

(b) Fitted to 2020-03-01 ∼ 2020-08-31

Fig. 4. Prediction using parameters fitted from 2020-03-01 to 2020-04-30 and from 2020-03-01 to 2020-08-31 in Tokyo

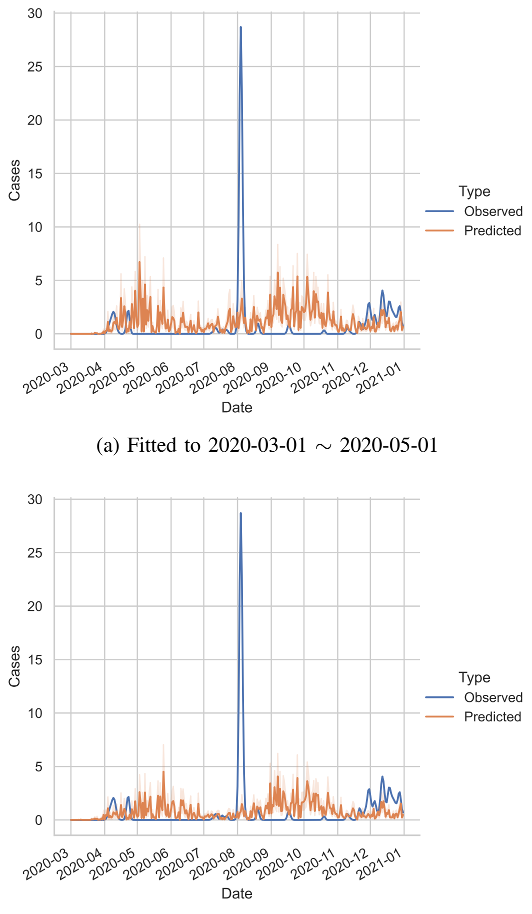
Figs. 4, 5 and 6 show the prediction using β optimized for March 01 to April 30 and March 01 to August 31, 2020. The predictions of all three prefectures has three wave, as the case fitted to the entire study period. Because β is different, the intensity of the predicted epidemic is different.



(a) Fitted to 2020-03-01 ∼ 2020-05-01

(b) Fitted to 2020-03-01 ∼ 2020-08-31

Fig. 5. Prediction using parameters fitted from 2020-03-01 to 2020-04-30 and from 2020-03-01 to 2020-08-31 in Hokkaido



(a) Fitted to 2020-03-01 ∼ 2020-05-01

(b) Fitted to 2020-03-01 ∼ 2020-08-31

Fig. 6. Prediction using parameters fitted from 2020-03-01 to 2020-04-30 and from 2020-03-01 to 2020-08-31 in Shimane

VI. Discussion and Future Research

Our model used a single static parameter constant across time and location. Nevertheless, using the movement data of individuals, our model can reproduce multiple waves that previous models using aggregate mobility data did not explain. This suggests that the use of aggregate mobility data is insufficient and tracking the movement of individuals is therefore important.

As expected, the model did not fit the observations because it had few parameters. However, a simple model is useful for understanding the underlying mechanisms and possible improvements.

There are several directions for future research in this area.

We will consider two ways to make our model better fit and predict the observed data. First, additional parameters will be introduced into the model. The value of β optimized for the first and second waves was smaller than that of the entire period. This suggests that β might have increased during the study period, possibly through an adherence fatigue not captured by the mobility. We can introduce a gradual change of β into the model.

As suggested by Nagata et al. [8], β can differ in different places, which may partially explain the regional differences described in Section V-A. We can therefore incorporate the differences in land usage into the model.

We have also improved the epidemiological mechanisms assumed in the model. In our model, we assume random mixing in each grid cell; however, this assumption is too simplistic. For example, in residential areas, people mostly interact with their family members. Therefore, random mixing does not occur. Therefore, We may combine the mobility data with census data such as by Chang [16] for a better simulation of infection within a family. We may also introduce a multi-tiered model, such as Atela et al.'s and Chiba's approachs [15], [21], which distinguishes four tiers: family, workplace, community, and school. Furthermore, we may distinguish the POI inside each grid cell. We can introduce age and other personal characteristics that are considered important in epidemiology [47].

Another direction is to make the model genuinely predictive. Currently, location data are obtained by observing real movements. To predict future epidemics, it will be necessary to predict the movements of individuals. We can combine our epidemiological model with an agent-based mobility model.

Because our model does not include variants and vaccinations, it is only applicable to the year 2020. We wanted to refine our model to incorporate such factors, which would be useful for investigating the impacts of new variants and vaccination strategies.

Because our model tracks the movement of individuals, it can predict the geographic pattern of future epidemics. Therefore, our model can be useful in determining regional interventions.

Finally, extending our model to other infectious diseases is important for preparing for future pandemics.

Acknowledgments

The authors extend their gratitude to the expert members of the National COVID-19 Cluster Task Force in Japan for integrating the published data, as well as to the following institutions for their contributions in supporting our data collection: public health centers, prefectural governments, public health institutes, and medical facilities. We used the computational resources of the AI Bridging Cloud Infrastructure provided by the National Institute of Advanced Industrial Science and Technology.

References

[1] Our World in Data. Cumulative confirmed covid-19 deaths.

[2] H. Wang, K. R. Paulson, S. A. Pease, S. Watson, H. Comfort, P. Zheng, A. Y. Aravkin, C. Bisignano, R. M. Barber, T. Alam, J. E. Fuller, E. A. May, D. P. Jones, M. E. Frisch, C. Abbafati, C. Adolph, A. Allorant, J. O. Amlag, B. Bang-Jensen, G. J. Bertolacci, S. S. Bloom, A. Carter, E. Castro, S. Chakrabarti, J. Chattopadhyay, R. M. Cogen, J. K. Collins, K. Cooperrider, X. Dai, W. J. Dangel, F. Daoud, C. Dapper, A. Deen, B. B. Duncan, M. Erickson, S. B. Ewald, T. Fedosseeva, A. J. Ferrari, J. J. Frostad, N. Fullman, J. Gallagher, A. Gamkrelidze, G. Guo, J. He, M. Helak, N. J. Henry, E. N. Hulland, B. M. Huntley, M. Kereselidze, A. Lazzar-Atwood, K. E. LeGrand, A. Lindstrom, E. Linebarger, P. A. Lotufo, R. Lozano, B. Magistro, D. C. Malta, J. Månsson, A. M. Mantilla Herrera, F. Marinho, A. H. Mirkuzie, A. T. Misganaw, L. Monasta, P. Naik, S. Nomura, E. G. O'Brien, J. K. O'Halloran, L. T. Olana, S. M. Ostroff, L. Penberthy, R. C. Reiner, Jr, G. Reinke, A. L. P. Ribeiro, D. F. Santomauro, M. I. Schmidt, D. H. Shaw, B. S. Sheena, A. Sholokhov, N. Skhvitaridze, R. J. D. Sorensen, E. E. Spurlock, R. Syailendrawati, R. Topor-Madry, C. E. Troeger, R. Walcott, A. Walker, C. S. Wiysonge, N. A. Worku, B. Zigler, D. M. Pigott, M. Naghavi, A. H. Mokdad, S. S. Lim, S. I. Hay, E. Gakidou, and C. J. L. Murray, "Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21," *Lancet*, vol. 399, no. 10334, pp. 1513–1536, Apr. 2022.

[3] US Centers for Disease Control and Prevention. COVID-19 Forecasting and Mathematical Modeling. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/mathematical-modeling.html

[4] C. C. John, V. Ponnusamy, S. Krishnan Chandrasekaran, and N. R, "A survey on mathematical, machine learning and deep learning models for COVID-19 transmission and diagnosis," *IEEE Rev. Biomed. Eng.*, vol. 15, pp. 325–340, Jan. 2022.

[5] R. Padmanabhan, H. S. Abed, N. Meskin, T. Khattab, M. Shraim, and M. A. Al-Hitmi, "A review of mathematical model-based scenario analysis and interventions for COVID-19," *Comput. Methods Programs Biomed.*, vol. 209, p. 106301, Sep. 2021.

[6] M. U. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, L. Du Plessis, N. R. Faria, R. Li, W. P. Hanage *et al.*, "The effect of human mobility and control measures on the covid-19 epidemic in china," *Science*, vol. 368, no. 6490, pp. 493–497, 2020.

[7] T. Yabe, K. Tsubouchi, N. Fujiwara, T. Wada, Y. Sekimoto, and S. V. Ukkusuri, "Non-compulsory measures sufficiently reduced human mobility in tokyo during the COVID-19 epidemic," *Sci. Rep.*, vol. 10, no. 1, p. 18053, Oct. 2020.

[8] S. Nagata, T. Nakaya, Y. Adachi, T. Inamori, K. Nakamura, D. Arima, and H. Nishiura, "Mobility change and COVID-19 in japan: Mobile data analysis of locations of infection," *Journal of Epidemiology*, pp. 1–5, 2021.

[9] P. Nouvellet, S. Bhatia, A. Cori, K. E. C. Ainslie, M. Baguelin, S. Bhatt, A. Boonyasiri, N. F. Brazeau, L. Cattarino, L. V. Cooper, H. Coupland, Z. M. Cucunuba, G. Cuomo-Dannenburg, A. Dighe, B. A. Djaafara, I. Dorigatti, O. D. Eales, S. L. van Elsland, F. F. Nascimento, R. G. FitzJohn, K. A. M. Gaythorpe, L. Geidelberg, W. D. Green, A. Hamlet, K. Hauck, W. Hinsley, N. Imai, B. Jeffrey, E. Knock, D. J. Laydon, J. A. Lees, T. Mangal, T. A. Mellan, G. Nedjati-Gilani, K. V. Parag, M. Pons-Salort, M. Ragonnet-Cronin, S. Riley, H. J. T. Unwin, R. Verity, M. A. C. Vollmer, E. Volz, P. G. T. Walker, C. E. Walters, H. Wang, O. J. Watson, C. Whittaker, L. K. Whittles, X. Xi, N. M. Ferguson, and C. A. Donnelly, "Reduction in mobility and COVID-19 transmission," *Nat. Commun.*, vol. 12, no. 1, pp. 1–9, 2021.

[10] O. Gatalo, K. Tseng, A. Hamilton, G. Lin, E. Klein, and CDC MInD-Healthcare Program, "Associations between phone mobility data and COVID-19 cases," *Lancet Infect. Dis.*, vol. 21, no. 5, p. e111, May 2021.

[11] Google LLC. (Accessed: Apr. 8, 2021) Google covid-19 community mobility reports. online. [Online]. Available: https://www.google.com/covid19/mobility/

[12] H. Rahmandad, T. Y. Lim, and J. Sterman, "Behavioral dynamics of COVID-19: estimating underreporting, multiple waves, and adherence fatigue across 92 nations," *Syst Dyn Rev*, vol. 37, no. 1, pp. 5–31, Jan. 2021.

[13] E. Kaxiras and G. Neofotistos, "Multiple epidemic wave model of the COVID-19 pandemic: Modeling study," *J. Med. Internet Res.*, vol. 22, no. 7, p. e20912, Jul. 2020.

[14] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman, "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2)," *Science*, vol. 368, no. 6490, pp. 489–493, 2020.

[15] A. Aleta, D. Martín-Corral, A. Pastore y Piontti, M. Ajelli, M. Litvinova, M. Chinazzi, N. E. Dean, M. E. Halloran, I. M. Longini, S. Merler, A. Pentland, A. Vespignani, E. Moro, and Y. Moreno, "Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19," *Nature Human Behaviour*, vol. 4, no. September, 2020.

[16] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec, "Mobility network models of COVID-19 explain inequities and inform reopening," *Nature*, vol. 589, no. January, 2020.

[17] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. P. y Piontti, K. Mu, L. Rossi, K. Sun *et al.*, "The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak," *Science*, vol. 368, no. 6489, pp. 395–400, 2020.

[18] K. Linka, M. Peirlinck, F. Sahli Costabal, and E. Kuhl, "Outbreak dynamics of COVID-19 in europe and the effect of travel restrictions," *Comput. Methods Biomech. Biomed. Engin.*, vol. 23, no. 11, pp. 710–717, 2020.

[19] O. El Deeb and M. Jalloul, "The dynamics of COVID-19 spread: Evidence from lebanon," *Math. Biosci. Eng.*, vol. 17, no. 5, pp. 5618–5632, 2020.

[20] T. Yang, Y. Liu, W. Deng, W. Zhao, and J. Deng, "SARS-Cov-2 trajectory predictions and scenario simulations from a global perspective: a modelling study," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, 2020.

[21] A. Chiba, "The effectiveness of mobility control, shortening of restaurants' opening hours, and working from home on control of COVID-19 spread in japan," *Health Place*, vol. 70, p. 102622, Jul. 2021.

[22] H. Zhang, P. Li, Z. Zhang, W. Li, J. Chen, X. Song, R. Shibasaki, and J. Yan, "Epidemic versus economic performances of the COVID-19 lockdown: A big data driven analysis," *Cities*, vol. 120, p. 103502, Jan. 2022.

[23] C. J. Silva, G. Cantin, C. Cruz, R. Fonseca-Pinto, R. Passadouro, E. Soares Dos Santos, and D. F. M. Torres, "Complex network model for COVID-19: Human behavior, pseudo-periodic solutions and multiple epidemic waves," *J. Math. Anal. Appl.*, vol. 514, no. 2, p. 125171, Oct. 2022.

[24] S. Liu and T. Yamamoto, “Role of stay-at-home requests and travel restrictions in preventing the spread of COVID-19 in japan,” *Transp. Res. Part A: Policy Pract.*, Mar. 2022.

[25] A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, and Centre for Mathematical Modelling of Infectious Diseases COVID-19 working group, “Early dynamics of transmission and control of COVID-19: a mathematical modelling study,” *Lancet Infect. Dis.*, vol. 20, no. 5, pp. 553–558, May 2020.

[26] C. C. Kerr, R. M. Stuart, D. Mistry, R. G. Abeysuriya, K. Rosenfeld, G. R. Hart, R. C. Núñez, J. A. Cohen, P. Selvaraj, B. Hagedorn, L. George, M. G. Fowler, A. Palmer, D. Delport, N. Scott, S. L. Kelly, C. S. Bennette, B. G. Wagner, S. T. Chang, A. P. Oron, E. A. Wenger, J. Panovska-Griffiths, M. Famulare, and D. J. Klein, “Covasim: An agent-based model of COVID-19 dynamics and interventions,” *PLoS Comput. Biol.*, vol. 17, no. 7, p. e1009149, Jul. 2021.

[27] Q. Cao, R. Jiang, C. Yang, Z. Fan, X. Song, and R. Shibasaki, “MepoGNN: Metapopulation epidemic forecasting with graph neural networks,” *2022.ecmlpkdd.org*, 2022.

[28] R. Jiang, Z. Wang, Z. Cai, C. Yang, Z. Fan, T. Xia, G. Matsubara, H. Mizuseki, X. Song, and R. Shibasaki, “Countrywide Origin-Destination matrix prediction and its application for COVID-19,” in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*. Springer International Publishing, 2021, pp. 319–334.

[29] C. Yang, Z. Zhang, Z. Fan, R. Jiang, Q. Chen, X. Song, and R. Shibasaki, “EpiMob: Interactive visual analytics of citywide human mobility restrictions for epidemic control,” *IEEE Trans. Vis. Comput. Graph.*, vol. PP, Apr. 2022.

[30] Z. Fan, X. Song, Y. Liu, Z. Zhang, C. Yang, Q. Chen, R. Jiang, and R. Shibasaki, “Human mobility based individual-level epidemic simulation platform,” *SIGSPATIAL Special*, vol. 12, no. 1, pp. 34–40, Jun. 2020.

[31] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec, “Mobility network models of COVID-19 explain inequities and inform reopening,” *Nature*, vol. 589, no. January, 2020.

[32] P. Weng and X.-Q. Zhao, “Spreading speed and traveling waves for a multi-type SIS epidemic model,” *J. Differ. Equ.*, vol. 229, no. 1, pp. 270–296, Oct. 2006.

[33] G.-Q. Sun, “Pattern formation of an epidemic model with diffusion,” *Nonlinear Dyn.*, vol. 69, no. 3, pp. 1097–1104, Jan. 2012.

[34] NHK. List of Data on New Coronavirus (Japanese). [Online]. Available: https://www3.nhk.or.jp/news/special/coronavirus/data-widget/

[35] N. G. Becker, L. F. Watson, and J. B. Carlin, “A method of non-parametric back-projection and its application to aids data,” *Statistics in Medicine*, vol. 10, no. 10, pp. 1527–1542, 1991.

[36] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler, “The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application,” *Annals of Internal Medicine*, vol. 172, no. 9, pp. 577–582, 2020.

[37] Ministry of Health, Labour and Welfare. Press Relase, 2020-12-31, (Japanese). [Online]. Available: https://www.mhlw.go.jp/stf/newpage_15828.html

[38] G. Meyerowitz-Katz and L. Merone, “A systematic review and meta-analysis of published research data on COVID-19 infection fatality rates,” *Int. J. Infect. Dis.*, vol. 101, pp. 138–148, 2020.

[39] Z. Gao, Y. Xu, C. Sun, X. Wang, Y. Guo, S. Qiu, and K. Ma, “A systematic review of asymptomatic infections with COVID-19,” *J. Microbiol. Immunol. Infect.*, vol. 54, no. 1, pp. 12–16, Feb. 2021.

[40] S. Abbott, J. Hellewell, K. Sherratt, K. Gostic, J. Hickson, H. S. Badr, M. DeWitt, R. Thompson, EpiForecasts, and S. Funk, *EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters*, 2020.

[41] T. Ganyani, C. Kremer, D. Chen, A. Torneri, C. Faes, J. Wallinga, and N. Hens, “Estimating the generation interval for coronavirus disease (covid-19) based on symptom onset data, march 2020,” *Eurosurveillance*, vol. 25, no. 17, p. 2000257, 2020.

[42] A. Endo, S. Abbott, A. J. Kucharski, and S. Funk, “Estimating the overdispersion in covid-19 transmission using outbreak sizes outside china,” *Wellcome Open Research*, vol. 5, 2020.

[43] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, “Algorithms for hyper-parameter optimization,” *Advances in neural information processing systems*, vol. 24, 2011.

[44] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin *et al.*, “Apache spark: a unified engine for big data processing,” *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.

[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.

[46] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[47] E. Vynnycky and R. White, *An introduction to infectious disease modelling*. OUP oxford, 2010.