Mathematical Modeling of COVID-19

Dongming Li
dongmingli@link.cuhk.edu.cn

Abstract

COVID-19 modeling, a special case in the field of epistemology modeling, is becoming increasingly significant as the epidemic situation worsens. To investigate and potentially predict the trend of the pandemic, mathematical models from compartmental models and o.d.e. systems and agent-based models have been developed. We develop deterministic models of o.d.e. system from compartmental models and stochastic models of Markov chain from agent-based approaches and show the close relationship between the deterministic models and the stochastic models. Empirical results on predicting the future evolution of the pandemic indicate the effectiveness of our models in most cases. Additional investigation on the outbreak scenario and a policy called "dynamic clearing" also implies the validness of our approach. The code is available at https://github.com/dongmingli-Ben/covid-19-modeling.

1 Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus, whose initial outbreak happened in China, late 2019. Later, the disease was transmitted to other countries and became a worldwide pandemic. To date, over 5 million lives have been taken by COVID-19 and economies suffer greatly by the government policies to control the pandemic situation, such as lockdown and limited transportation. Modeling the transmission of the disease can be a great help to prognostication and policy devising and evaluation.

Through the literature of epidemic modeling, stochastic and deterministic models have been proposed. Among them, compartmental models and the corresponding systems of ordinary differential equations (o.d.e.) are prevailing in modeling COVID-19. Compartmental models divide the population into several classes with specified dynamics between the classes and derive the evolution of the disease as an o.d.e. system (Sameni, 2020; Mandal et al., 2020; forecasting team, 2020; Mokhtari et al., 2021). With the o.d.e. system, one can investigate the "what if" scenarios by simply changing the underlying parameters of the o.d.e. system, which can be helpful for governments to devise non-pharmaceutical intervention policies. The o.d.e. system can be viewed as a deterministic overall description of the system. Another popular approach is agent-based method, where the behavior of each individual is specified (Gu, 2020; Mokhtari et al., 2021). In contrast to the top-down view of the o.d.e. system, agent-based method adopts a bottom-up approach to recover the overall dynamics from behavior of individuals, which grants the model more flexibility but potentially suffering from computational difficulty.

In this work, we present deterministic models derived from compartmental model and stochastic models originated from agent-based method and show the link between them both theoretically and empirically. Compared with previous work, where the models are validated by predicting future epidemic situation, we also validate our models in a specific scenario, i.e. the evolution of the regional outbreak under the dynamic clearing strategy (perform throughout COVID-19 RNA test and strict quarantine policy once an indigenous case is found).

Report for MAT3300 final project, 2021 Fall, the Chinese University of Hong Kong, Shenzhen.

2 Literature Review

Compartmental Models Compartmental models (Roberts and Heesterbeek, 2003; Brauer, 2008) model the system by dividing individuals into distinct groups, i.e. compartments and assume that people in the same group are homogeneous. Typically, an o.d.e. system is used to describe the system dynamics (Sameni, 2020; Sharov, 2020; Mandal et al., 2020; forecasting team, 2020; Mokhtari et al., 2021). The first mathematical model of epistemology develop by Bernoulli (1760), the SIR model, can be viewed as a compartmental model with three compartments, i.e. susceptible, infected, and recovered. Compartmental models essentially model the transition from one compartment to another, hence also called as *mass transport* (Rideout, 1991) or *mass action* (Ingalls, 2012) in other fields. Compartmental models can be easily adapted to different scenarios by adding new compartments or substituting existing compartments, which is one possible reason for its prevalence in epistemology modeling. Sameni (2020) introduced a "exposed" group which is infectious but does not show symptoms. Mokhtari et al. (2021) constructed a sophisticated compartmental model to better match the reality, where addition groups such as asymptomatic infection, hospitalization, and critical infection are included into the naive SEIR model. Mandal et al. (2020) added a quarantine group upon the SEIR model and include the effect of media by affecting the transition from susceptible to recovered group to study the effect of quarantine and media propaganda. forecasting team (2020) distinguished infected people into two groups, i.e. pre-symptomatic stage and the later stage.

Agent-based Models Agent-based models are a new approach to model the complex dynamics of the system comprised of relative simple and interacting components, i.e. agents, given the recent advancement in computational resources (Macal and North, 2005). One application of agent-based models is epistemology modeling (Auchincloss and Diez Roux, 2008; El-Sayed et al., 2012; Marshall and Galea, 2015). In agent-based models, typically the system is investigated via simulation instead of solving the system. In terms of calibration of the model, the work closest to ours is probably Gu (2020). Gu (2020) started with dynamic of individuals and ran a grid search to search for the most applicable parameters by maximizing the fitting loss. However, different from ours, their model used carefully hand-crafted dynamics, such as the immunity of recovered people expire after some days and the immigration from other regions, which are not Markovian and require lots of efforts to design these dynamics. Furthermore, in order to mitigate the high computational requirement of grid search, given the large amount of parameters (dozens), Gu (2020) limited the search range to a rather limited preset range according to experience, which can lead to inferior fitting results.

Stochastic Models Given the inherit randomness in the real world, stochastic models are a natural choice for epistemology modeling. Typical stochastic models for epistemology include Poisson process (Kiouach and Sabbar, 2020), Markov chains (West and Thompson, 1997; Ball, 1983), branching process (Ball and Donnelly, 1993; Farrington et al., 2003), and stochastic differential equations (Gray et al., 2011; Mahrouf et al., 2021). Stochastic models are able to tackle some problems that deterministic models cannot. For example, branching process can be used to find the probability whether a small number of initial infected people will start an outbreak or not and to find the outbreak period length (Allen, 2017).

3 Proposed Model

In this section, we first present the deterministic and stochastic models and later prove the link between them both theoretically and empirically. In this work, we seek to develop models that model the evolution of COVID-19 epidemic in a short range (within one to two years). Therefore, the effect of births and deaths due to other factors are neglected.

3.1 Deterministic Models

Susceptible-Infected-Recovered (SIR) model is one of the most well known epidemic model based on compartmental model. Because for COVID-19, it has been widely known that there is a roughly two or three week period of incubation, when people are infected with the disease and are able to infect other susceptible people but show no symptom, we add another group "exposed" to distinguish people who are infected but have not shown symptom yet, following Sameni (2020). To account for the deaths, following Sameni (2020), we add a "death" group to represent the mentalities.

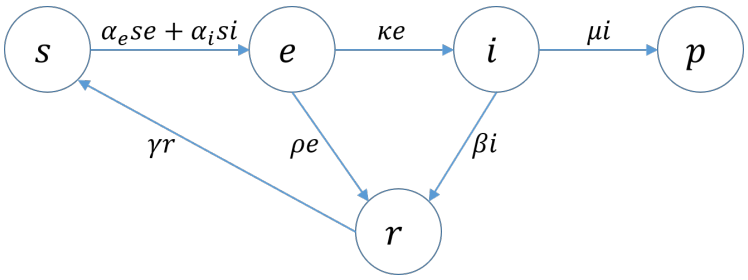


Figure 1: The graphical representation of the SEIR model. Each circle represents a group in the SEIR model and each arrow means the transition from one group to another.

Formally, we define the following quantities:

- $s(t)$: the number of people who are susceptible (can be infected by an exposed or infected person) to COVID-19 at time t.
- $e(t)$: the number of people who are exposed (infected but not *recognized* as infected people, possibly because they have not shown symptom yet) at time t.
- $i(t)$: the number of people who are infected and *recognized* as infected to COVID-19 (have symptoms or tested positive in COVID-19 RNA tests) at time t.
- $r(t)$: the number of people who have recovered from COVID-19 and maintain immunity to COVID-19 at time t.
- $p(t)$: the number of people who are dead due to COVID-19 at time t.

The proposed SEIR model, following Sameni (2020), and its compartmental representation are presented in equation 1 and Figure 1.

$$\begin{aligned} s'(t) &= -\alpha_e s(t)e(t) - \alpha_i s(t)i(t) + \gamma r(t) \\ e'(t) &= \alpha_e s(t)e(t) + \alpha_i s(t)i(t) - \kappa e(t) - \rho e(t) \\ i'(t) &= \kappa e(t) - \beta i(t) - \mu i(t) \\ r'(t) &= \rho e(t) + \beta i(t) - \gamma r(t) \\ p'(t) &= \mu i(t) \end{aligned} \tag{1}$$

It can be seen that the system is close, i.e.

$$s'(t) + e'(t) + i'(t) + r'(t) + p'(t) = 0$$

which indicates that the number of people in the system will not change.

Because in real world practice of measurement, the number of cumulative recovered people is recorded, which is different from the number of recovered people because some people may lose immunity to COVID-19 and become susceptible. To match the real world data, we denote the cumulative recovered people at time t as $cr(t)$. The relationship between $cr(t)$ and the system is

$$cr'(t) = \rho e(t) + \beta i(t) \tag{2}$$

Linear Approximate For a region with millions of people and the epidemic situation is not severe, the number of people that are not susceptible is small. In this scenario, we can assume that $s(t) \approx s_0$ and equation 1 reduces to a linear o.d.e. system, i.e.

$$\begin{aligned} \begin{bmatrix} e(t) \\ i(t) \\ r(t) \end{bmatrix}' &= \begin{bmatrix} \bar{\alpha_e} - \kappa - \rho & \bar{\alpha_i} & 0 \\ \kappa & -\beta - \mu & 0 \\ \rho & \beta & -\gamma \end{bmatrix} \begin{bmatrix} e(t) \\ i(t) \\ r(t) \end{bmatrix} \\ s(t) &= s_0 - e(t) - i(t) - r(t) - p(t) \\ p'(t) &= \mu i(t) \end{aligned} \tag{3}$$

where $\bar{\alpha_e} = s_0\alpha_e, \bar{\alpha_i} = s_0\alpha_i$.

Similar to the analysis in Sameni (2020), the eigenvalues and eigenvectors corresponding to the system of linear o.d.e. are

$$\lambda_1 = \frac{\delta - \beta - \mu + \sqrt{(\delta+\beta+\mu)^2 + 4\kappa\bar{\alpha_i}}}{2}, \mathbf{v}_1 = [1, \frac{\lambda_1 - \delta}{\bar{\alpha_i}}, \frac{\rho\bar{\alpha_i} + \beta(\lambda_1 - \delta)}{\bar{\alpha_i}(\lambda_1 + \gamma)}]^\top$$
$$\lambda_2 = \frac{\delta - \beta - \mu - \sqrt{(\delta+\beta+\mu)^2 + 4\kappa\bar{\alpha_i}}}{2}, \mathbf{v}_2 = [1, \frac{\lambda_2 - \delta}{\bar{\alpha_i}}, \frac{\rho\bar{\alpha_i} + \beta(\lambda_2 - \delta)}{\bar{\alpha_i}(\lambda_2 + \gamma)}]^\top$$
$$\lambda_3 = -\gamma, \mathbf{v}_3 = [0, 0, 1]^\top$$

where $\delta = \bar{\alpha_e} - \kappa - \rho$. With initial condition $e(0) = e_0, i(0) = i_0, r(0) = r_0$, the solution is

$$\begin{aligned} e(t) &= \frac{\bar{\alpha_i}i_0 + (\delta - \lambda_2)e_0}{\lambda_1 - \lambda_2}e^{\lambda_1 t} + \frac{\bar{\alpha_i}i_0 + (\delta - \lambda_1)e_0}{\lambda_2 - \lambda_1}e^{\lambda_2 t} \\ i(t) &= \frac{\bar{\alpha_i}i_0 + (\delta - \lambda_2)e_0}{\lambda_1 - \lambda_2}\frac{\lambda_1 - \delta}{\bar{\alpha_i}}e^{\lambda_1 t} + \frac{\bar{\alpha_i}i_0 + (\delta - \lambda_1)e_0}{\lambda_2 - \lambda_1}\frac{\lambda_2 - \delta}{\bar{\alpha_i}}e^{\lambda_2 t} \end{aligned} \tag{4}$$

3.2 Stochastic Models

We specify the behavior of each individual following agent-based approaches and derive the dynamics of the overall system. Similar to the setting of the SEIR model, we classify each individual into 5 distinctive states, i.e. susceptible, exposed, infected, recovered, death. The behavior of each individual is specified as follows:

- Each exposed people infects a susceptible person independently of other susceptible people, exposed people, infected people, and dead people (or *other people* for brevity) with rate α_e.
- Each infected people infects a susceptible person independently of *other people* with rate α_i.
- Each exposed people becomes infected independently of *other people* with rate κ.
- Each exposed people recovered independently of *other people* with rate ρ.
- Each infected people recovered independently of *other people* with rate β.
- Each infected people dies independently of *other people* with rate μ.
- Each recovered people loses immunity to COVID-19 and becomes susceptible independently of *other people* with rate γ.

Because for each susceptible person, the rate of infection by each exposed person and each infected person are α_e and α_i respectively, the rate that a susceptible person becomes exposed is $\alpha_e e(t) + \alpha_i i(t)$. For $s(t)$ susceptible people, the rate that one person become exposed is $\alpha_e e(t)s(t) + \alpha_i i(t)s(t)$. Similarly, the rate that an exposed person becomes infected is $\kappa e(t)$. The rate that an infected person died is $\mu i(t)$. The rate that an exposed person recovered is $\rho e(t)$. The rate that an infected person recovered is $\beta i(t)$. The rate that an recovered person loses immunity to COVID-19 and becomes susceptible is $\gamma r(t)$. With the specified behavior above, the system follows a continuous time Markov chain (CTMC) with state $(s(t), e(t), i(t), r(t), p(t))$ and the following transition rates

$$q_{(s(t),e(t),i(t),r(t),p(t)),S} = \begin{cases} \alpha_e s(t)e(t) + \alpha_i s(t)i(t) & S = (s(t)-1, e(t)+1, i(t), r(t), p(t)) \\ \kappa e(t) & S = (s(t), e(t)-1, i(t)+1, r(t), p(t)) \\ \mu i(t) & S = (s(t), e(t), i(t)-1, r(t), p(t)+1) \\ \rho e(t) & S = (s(t), e(t)-1, i(t), r(t)+1, p(t)) \\ \beta i(t) & S = (s(t), e(t), i(t)-1, r(t)+1, p(t)) \\ \gamma i(t) & S = (s(t)+1, e(t), i(t), r(t)-1, p(t)) \end{cases} \tag{5}$$

The graphical illustration of the CTMC is in Figure 2.

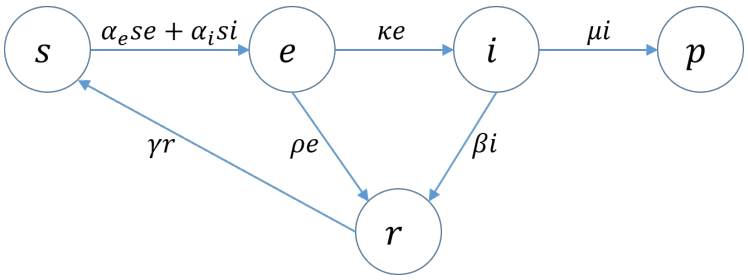


Figure 2: A graphical illustration of the CTMC model. The circles represent people in different classes and the arrow between two circle means the rate that one people transforms from one class to another in the system.

Discrete Approximate To simulate a CTMC, one need to simulate the inter-arrival time between events, which is computationally heavy. In practice, since the COVID-19 cases are updated daily, one may find using a discrete time Markov chain (DTMC) appealing for both its computational efficiency and natural fit to the data. In principle, one can first find the transition probability $p_{x,y}(1)$ (i.e. $P(S(1) = y|S(0) = x)$) and the DTMC resulting from discretizing the CTMC on time grid $\{0, 1, 2, ...\}$ has transition probability $p^*_{x,y} = p_{x,y}(1)$. Theoretically, the transition probability can be calculated with Kolmogorov backward or forward equation (Kolmogoroff, 1931), that is

$$p'_{x,y}(t) = \sum_{k\in S, k\neq x} q_{x,k}p_{k,y}(t) - \sum_{k\in S,k\neq x} q_{x,k}p_{x,y}(t) \tag{6}$$

or

$$p'_{x,y}(t) = \sum_{k\in S, k\neq y} q_{k,y}p_{x,k}(t) - \sum_{k\in S,k\neq y} q_{y,k}p_{x,y}(t) \tag{7}$$

Because of the enormous amount of states in the CTMC, the Kolmogorov backward or forward equation are theoretically solvable but practically intractable.

To circumvent this issue, we consider an approximate of the CTMC instead of an exact simulation. Because for each susceptible person, the rate of being infected by an exposed or infected person is $\alpha_e e(t) + \alpha_i i(t)$, which is by definition $P(\text{Infected by time } t) = (\alpha_e e(t) + \alpha_i i(t))t + o(t)$, we approximate it using $P(\text{Infected in a day}) \approx (\alpha_e e(t) + \alpha_i i(t)) \cdot 1$. Further because the infection of each susceptible person is independent of other people, the number or susceptible people becoming infected on day n can be approximated by a binomial distribution $b(s(n), \alpha_e e(n) + \alpha_i i(n))$. For each exposed person, because the rate of becoming an infected person and the rate of recovering are κ and ρ respectively, $P(\text{Become infected in a day}) = \kappa \cdot 1 + o(h) \approx \kappa$ and $P(\text{Become recovered in a day}) = \rho \cdot + o(h) \approx \rho$. Because of the independence of people, the transition of exposed people in a day can be approximated by a multinomial distribution with κ probability of becoming infected, ρ probability of becoming recovered, and $1 - \kappa - \rho$ probability of still being exposed.

Similarly, the transition of infected and recovered people can be derived and the transition probability of the approximate DTMC is summarized in Figure 3. Note that in order to enforce the natural constraint that $\alpha_e e(t) + \alpha_i i(t) \leq 1$, we replace it with $\alpha_e \min(e(t), e_{max}) + \alpha_i \min(i(t), i_{max})$. The alternative term also has plausible physical meaning, which is each susceptible person can at most get in touch with e_{max} exposed and i_{max} infected people at the same time. This is intuitive since a person can only get in touch with people around him/her spatially.

To see how well the DTMC approximate the CTMC, we run simulations of DTMC and CTMC under the same parameter setting. The result is shown in Figure 4, where the approximation matches the global trend and is close to the CTMC.

3.3 Closeness Between Deterministic and Stochastic Models

In this section, we prove the o.d.e. system model and the CTMC model are close to each other both theoretically and empirically. In particular, the o.d.e. system is close to the mean of the CTMC model.

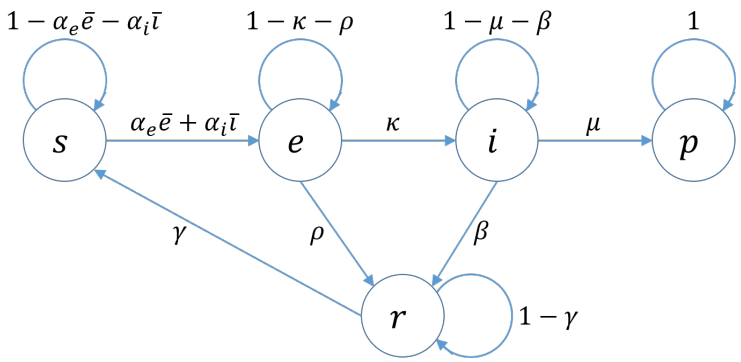


Figure 3: A graphical illustration of the DTMC approximate model. Each circle represents people in a class and each arrow from class A to class B means that the probability for each person in class A to become class B independently in one day. In the graph, $\bar{e} = \max(e(t), e_{max})$ and $\bar{i} = \max(i(t), i_{max})$.

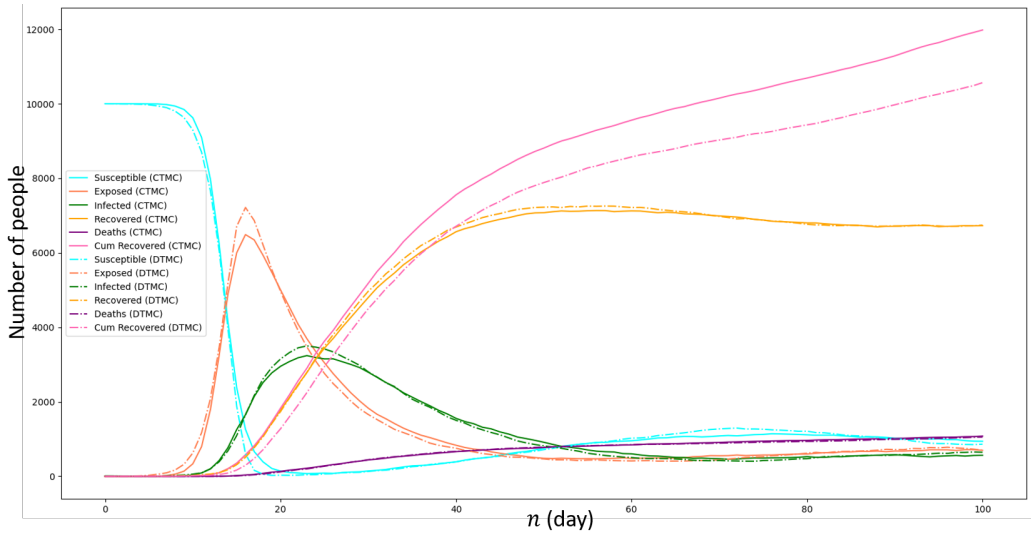


Figure 4: Comparison between CTMC and the approximate DTMC model under the same parameter setting, i.e. $s_0 = 1e4, i_0 = 10, e_0 = r_0 = p_0 = 0$ and $\alpha_e = 1e-4, \alpha_i = 1e-5, \gamma = \rho = \mu = 0.01, \kappa = \beta = 0.1, e_{max} = i_{max} = 1e9$. Different classes are highlighted in different colors while result from CTMC is plotted with solid line and result from DTMC is plotted with dotted line.

As demonstrated by Wang (2021), though incorrectly, the expectation of the state at time t for a SIR model is closely related to the solution of the o.d.e. system. To show the same result holds for the SEIR model, first define $S(t) = \mathbb{E}[s(t)|A(0) = A_0], E(t) = \mathbb{E}[e(t)|A(0) = A_0], I(t) = \mathbb{E}[i(t)|A(0) = A_0], R(t) = \mathbb{E}[r(t)|A(0) = A_0], P(t) = \mathbb{E}[p(t)|A(0) = A_0]$ and state $A(t) = (s(t), e(t), i(t), r(t), p(t))$. Then, we have

$$
\begin{aligned}
S(t+h) &= \mathbb{E}[s(t+h)|A(0) = A_0] = \mathbb{E}_{A_t}[\mathbb{E}[s(t+h)|A(0) = A_0, A(t) = A_t]] \\
&= \mathbb{E}_{A_t}[(s(t)+1)\cdot(\gamma r(t)h + o(h)) + (s(t)-1)\cdot(\alpha_i s(t)i(t)h + \alpha_e s(t)e(t)h + o(h)) \\
&+ s(t)(1-\gamma r(t)h - \alpha_i s(t)i(t)h - \alpha_e s(t)e(t)h + o(h)) + o(h)] \\
&= \mathbb{E}_{A_t}[s(t) + \gamma r(t)h - \alpha_i s(t)i(t)h - \alpha_e s(t)e(t)h + o(h)] \\
&= \mathbb{E}_{A_t}[s(t)] + \gamma \mathbb{E}_{A_t}[r(t)]h - \alpha_i \mathbb{E}_{A_t}[s(t)i(t)]h - \alpha_e \mathbb{E}_{A_t}[s(t)e(t)]h + \mathbb{E}_{A_t}[o(h)]] \\
&= S(t) + \gamma R(t)h - \alpha_i S(t)I(t)h - \alpha Cov(s(t), i(t))h \\
&- \alpha_e S(t)E(t)h - \alpha_e Cov(s(t), e(t))h + o(h)
\end{aligned}
\tag{8}
$$

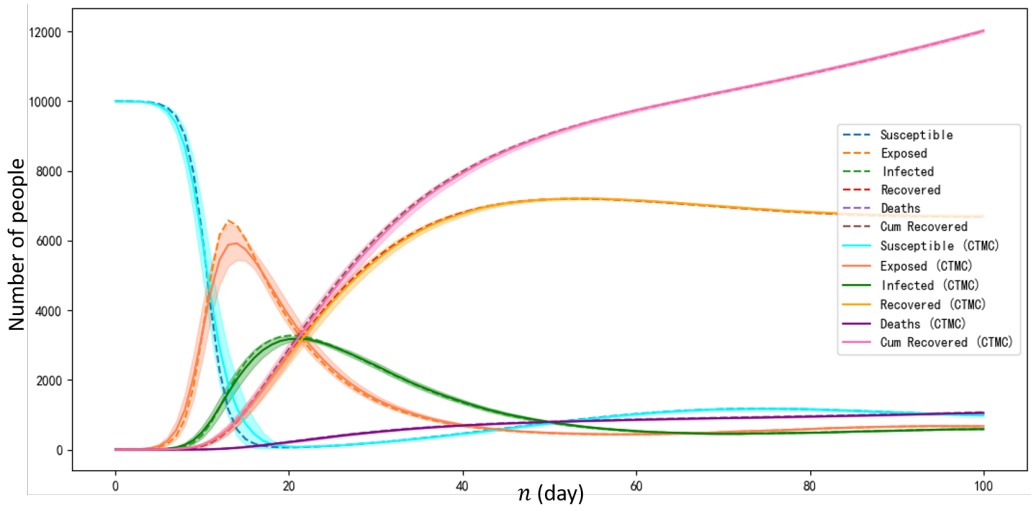


Figure 5: The comparison of the o.d.e. system and the CTMC model under the same parameter setting, i.e. $s_0 = 1e4, i_0 = 10, e_0 = r_0 = p_0 = 0$ and $\alpha_e = 1e-4, \alpha_i = 1e-5, \gamma = \rho = \mu = 0.01, \kappa = \beta = 0.1, e_{max} = i_{max} = 1e9$. The dotted lines are the results of the o.d.e. system and the solid lines are the mean of 20 independent simulation of the CTMC model. The shadow areas are the 95% confidence interval calculated from the 20 simulations.

Equation 8 is because

$$s(t+h) = \begin{cases} s(t)+1 & p = \gamma r(t)h + o(h) \\ s(t)-1 & p = (\alpha_i s(t)i(t) + \alpha_e s(t)e(t))h + o(h) \\ s(t) & p = 1 - \gamma r(t)h - \alpha_i s(t)i(t) - \alpha_e s(t)e(t) + o(h) \\ \text{other cases} & p = o(h) \end{cases}$$

$$\mathbb{E}_{A_t}[s(t)i(t)] - \mathbb{E}_{A_t}[s(t)]\mathbb{E}_{A_t}[i(t)] = Cov(s(t), i(t))$$
$$\mathbb{E}_{A_t}[s(t)e(t)] - \mathbb{E}_{A_t}[s(t)]\mathbb{E}_{A_t}[e(t)] = Cov(s(t), e(t))$$

After some algebraic operations to Equation 8, we have

$$S'(t) = \lim_{h\to 0} \frac{S(t+h) - S(t)}{h} = \gamma R(t) - \alpha_i S(t)I(t) - \alpha_e S(t)E(t) \\ -\alpha_i Cov(s(t), i(t)) - \alpha_e Cov(s(t), e(t))$$

Similarly, we have the following system of o.d.e. for the CTMC

$$\begin{aligned} S'(t) &= \gamma R(t) - \alpha_i S(t)I(t) - \alpha_e S(t)E(t) \\ &\quad - \alpha_i Cov(s(t), i(t)) - \alpha_e Cov(s(t), e(t)) \\ E'(t) &= \alpha_i S(t)I(t) + \alpha_e S(t)E(t) - \kappa E(t) - \rho E(t) \\ &\quad + \alpha_i Cov(s(t), i(t)) + \alpha_e Cov(s(t), e(t)) \\ I'(t) &= \kappa E(t) - \mu I(t) - \beta I(t) \\ R'(t) &= \rho E(t) + \beta I(t) - \gamma R(t) \\ P'(t) &= \mu I(t) \end{aligned} \tag{9}$$

Observe that Equation 9 and Equation 1 are identical except for the covariance terms. While it is non-trivial to prove that $Cov(s(t), e(t))$ and $Cov(s(t), i(t))$ are negligible, we empirically show that they are. A comparison of an o.d.e. system model and the CTMC model under the same parameter setting is shown in Figure 5. The close gap between the two model indicates that $Cov(s(t), e(t))$ and $Cov(s(t), i(t))$ are small.

4 Validation Results

We validate our models in two approaches. One is to evaluate the fitness of the model to real world data and the accuracy of future prediction. The other, in contrast to previous work, is to apply the model to the outbreak scenario and establish some specific results under the dynamic clearing policy.

4.1 Fit and Predict

To validate our models, we evaluate the fitness and prediction power of our models on real world data comprehensively. Because our models take only the simplest assumptions, which are implied by the fixed parameters of the models, any natural processes that may cause the change in the dynamics of the transmission will make our models perform badly. The change of dynamic, i.e. change of parameters of the models, can be due to new variants of virus, more effective medication, change in pandemic regulation policies, e.t.c. Therefore, we fit the models in a "wave" of pandemic and evaluate the models' prediction accuracy of predicting the "wave" evolution. In particular, we fit our models on data from March 1, 2021 to May 1, 2021 and evaluate the prediction from May 1, 2021 to May 31, 2021, which are times when a new COVID-19 variant, delta, became increasingly predominant. Note that, even though most countries were affected by delta variant in that period, there were still countries which were not hit by this "wave" or hit after the evaluation period. Therefore, it would be plausible that for most countries, our models perform much better than the remaining countries.

In the data[1] provided by John Hopkins University (Dong et al., 2020), only the number of active cases (i.e. infected cases), cumulative confirmed cases, cumulative recovered cases, and deaths can be observed (see Appendix A for more details). Since the number of exposed people, recovered people, and susceptible people are not observable, a regression-based method to calculate the parameters is not feasible. Instead, inspired by Gu (2020), we directly optimize the goodness of fit to the real data by changing the underlying model parameters. Specifically, given the number of infected people i_t, number of cumulative recovered people cr_t, and the number of deaths p_t, we define the loss as a weighted average of the mean square error (MSE) and attempt to minimize the loss[2]

$$L(param) = w_i \cdot MSE(i(t), i_t) + w_r \cdot MSE(cr(t), cr_t) + w_p \cdot MSE(p(t), p_t) \tag{10}$$

where $cr(t)$ is the number of cumulative recovered people by time t, which can be calculated using the cumulative number of people who has transformed from infected to recovered (exposed people are not registered as patient hence not counted in cumulative recovered). The parameters include the parameters for transmission dynamic and the unknown initial number of people in the unobservable classes.

To optimize the problem, which is intractable, we resort to numerical heuristic method, simulated annealing. Due to the large computational efforts required by simulated annealing, we only evaluate our most efficient model, i.e. the discrete approximate model[3]. The DTMC model is fitted on all 280 countries and regions worldwide with the same initial parameters for simulated annealing[4] and the predictions are evaluated by MSE and relative rooted MSE (rRMSE) defined as follows

$$\text{rRMSE} = \frac{\sqrt{\text{MSE}}}{\max_t x_t} \tag{11}$$

where x_t is the target sequence.

An overview of the evaluation results is presented in Table 1. The predictions are calculated by first run a simulation from March 1, 2021 to May 1, 2021 with optimized parameters, then reset the number of infected people, recovered people, deaths, cumulative recovered to the actual observation i_t, r_t, p_t, r_t (the number of recovered people is set to r_t because it is believed that it is extremely unlikely for recovered people to be re-infected and recovered again) and continue the simulation to May 31, 2021.

[1] `https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series`

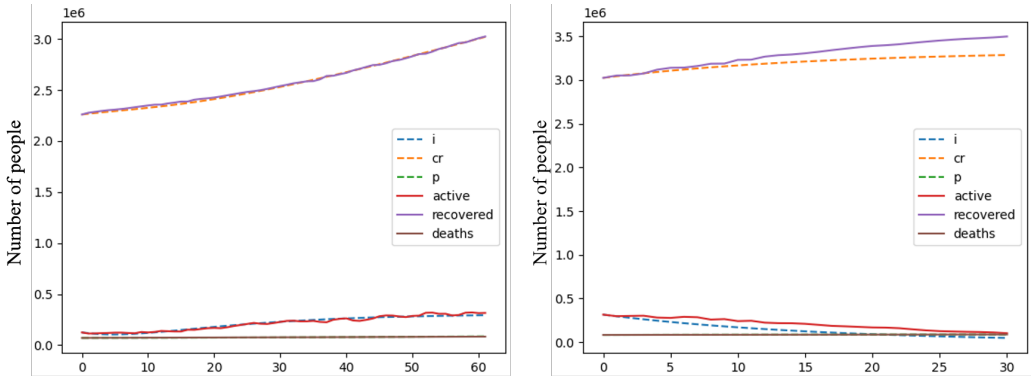
[2] In our experiments, we use $w_i = 0.2, w_r = 0.3, w_p = 0.5$ because we believe the reported deaths is the most reliable and the current infected is the least reliable.

[3] A comparison of the efficiency of the models can be seen in Appendix B.

[4] See Appendix B for more details.

	Fitting					Predicting				
	Loss	rRMSE	rRMSE_i	rRMSE_r	rRMSE_p	Loss	rRMSE	rRMSE_i	rRMSE_r	rRMSE_p
Mean	$6.45 \cdot 10^8$	0.72	0.39	0.11	1.22	$2.94 \cdot 10^{10}$	0.45	0.46	0.09	0.67
Std	$6.10 \cdot 10^9$	3.44	1.39	0.71	5.98	$3.92 \cdot 10^{11}$	2.14	0.22	0.29	4.14
Min	0.55	0.01	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00
25%	$7.06 \cdot 10^3$	0.06	0.10	0.00	0.04	$2.65 \cdot 10^3$	0.09	0.26	0.01	0.01
50%	$3.58 \cdot 10^5$	0.14	0.25	0.01	0.14	$1.88 \cdot 10^6$	0.15	0.51	0.02	0.07
75%	$9.89 \cdot 10^6$	0.36	0.43	0.03	0.51	$7.12 \cdot 10^7$	0.25	0.63	0.04	0.24
Max	$8.42 \cdot 10^{10}$	44.22	19.40	9.39	75.05	$5.49 \cdot 10^{12}$	29.38	0.98	2.77	56.81

Table 1: A statistic summary of the fitting and prediction results. rRMSE_i, rRMSE_r, and rRMSE_p denote the rRMSE of the number of infected people, the number of cumulative recovered people, and the number of deaths respectively. The overall rRMSE is the same weighted average as the loss, i.e. $\text{rRMSE} = w_i \text{rRMSE}_i + w_r \text{rRMSE}_r + w_p \text{rRMSE}_p$. Note that in some countries or regions, there might be no infected people during the period, which will result in null values in rRMSE. The results here are from the remaining 196 countries or regions which do not have null values.



(a) Fitting (b) Predicting

Figure 6: The result of fitting and predicting on data of Germany. (a): The fitted DTMC model's one simulation and the actual data from March 1, 2021 to May 1, 2021. (b): The fitted DTMC model's prediction and the actual data from May 1, 2021 to May 31, 2021. The dotted lines are results of the model (either fitted simulation or predictions) while the solid lines are real data.

From Table 1, the loss and rRMSE are heavy-tail for both fitting and predicting, which is as expected because our model perform badly for data that does not satisfy the assumptions and make the mean error much larger than median error. From the low error of majority of countries or regions, it implies that our model fits well to the natural process of COVID-19 transmission.

We also present a fit and prediction to a specific country or region to demonstrate the result graphically. The result of Germany is in Figure 6 and Table 2. The model Successfully predicts a downward trend for number of infected people even though it is fitted to data with upward trend of number of infected people only. The fitted parameters also reveal interesting results. From $\kappa = 0.092$ and $1/\kappa \approx 11$, it means approximately, exposed people (infected but not revealed symptoms or recognized yet) show symptoms or somehow identified as infected people in 11 days, which is shorter than the generally acknowledged 14 to 21 days of incubation. This can imply that the COVID-19 tests were effective and some people were identified as infected before they showed symptons. Moreover, from $\mu = 1.27 \cdot 10^{-3}$, the case-fatality rate is 0.1% from our model, which is smaller than the reported 1.6%[5]. This difference may not be significant considering that the reported case-fatality rate is the average starting from the origin outbreak in late 2019, when the fatality rate was higher due to lack of experience.

[5]Source at `https://coronavirus.jhu.edu/data/mortality`

	MSE	rRMSE	rRMSE_i	rRMSE_r	rRMSE_p
Fitting	$7.59 \cdot 10^7$	0.019	0.040	0.004	0.020
Predicting	$5.33 \cdot 10^9$	0.057	0.214	0.035	0.007

Table 2: The result of fitting on data of Germany from March 1, 2021 to May 1, 2021 and predicting from May 1, 2021 to May 31, 2021. The fitted parameters for Germany in this period is $\alpha_e = 3.92 \cdot 10^{-8}, \alpha_i = 4.82 \cdot 10^{-8}, \gamma = 8.35 \cdot 10^{-3}, \kappa = 0.092, \beta = 0.059, \rho = 2.90 \cdot 10^{-4}, \mu = 1.27 \cdot 10^{-3}, s_0 = 5.23 \cdot 10^7, e_0 = 5081, e_{max} = 2357, i_{max} = 5610$.

4.2 Outbreak Analysis

In the post-COVID-19 period, different governmental policies are adopted to maintain epidemic situation. One of the controversial but effective policy to the so called "dynamic clearing" adopted in China, which means adopting relative loose regulation when no indigenous case is found and imposing throughout COVID-19 RNA tests and strict quarantine policy (e.g. students are encourage to stay on campus unless for some indispensable reasons). Given its unexceptionable success in controlling regional outbreaks, we use our models to establish some results in this specific scenario.

Given that in the cases where outbreaks are controlled by the "dynamic cleaning" strategy, the number of total exposed and infected people is modicum, which means the linear approximate o.d.e. system (i.e. Equation 3) behaves closely to the exact SEIR model (i.e. Equation 1). Because of the intractability of the exact SEIR model, we instead investigate the behavior of the linear approximate.

As in Sameni (2020), in the linear model, we have

$$\frac{i(t)}{e(t)} \to \frac{\lambda_1 - \delta}{\alpha_i} \text{ for } t \gg \frac{1}{\lambda_2 - \delta}, s(t) \gg i(t) \tag{12}$$

which means that in the initial stage of the pandemic where infections grow at an exponential rate, the ratio between exposed and infected people tend to become a fixed number, i.e. $e(t) \approx ki(t)$.

We empirically validate this result on real world data. To begin with, given the solution in Equation 4, the cumulative confirmed cases can be recovered, i.e.

$$\begin{aligned} c'(t) = \kappa e(t) \Rightarrow c(t) &= \frac{\alpha_i i_0 + (\delta - \lambda_2)e_0}{\lambda_1 - \lambda_2} \cdot \frac{\kappa}{\lambda_1}(e^{\lambda_1 t} - 1) + \frac{\alpha_i i_0 + (\delta - \lambda_1)e_0}{\lambda_2 - \lambda_1} \cdot \frac{\kappa}{\lambda_2}(e^{\lambda_2 t} - 1) \\ &= a_1 e^{\lambda_1 t} + a_2 e^{\lambda_2 t} - a_1 - a_2 \end{aligned} \tag{13}$$

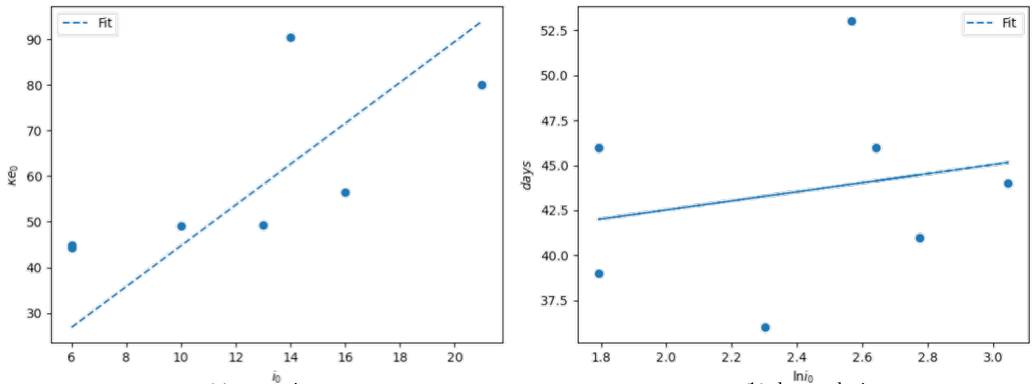
where $c(t)$ is the number of cumulative confirmed cases up to time t and $a_1 = \frac{\alpha_i i_0 + (\delta - \lambda_2)e_0}{\lambda_1 - \lambda_2} \cdot \frac{\kappa}{\lambda_1}, a_2 = \frac{\alpha_i i_0 + (\delta - \lambda_1)e_0}{\lambda_2 - \lambda_1} \cdot \frac{\kappa}{\lambda_2}$. Further notice that $e_0 = \frac{\lambda_1 a_1 + \lambda_2 a_2}{\kappa} \propto \lambda_1 a_1 + \lambda_2 a_2$. It is equivalent to validate that $\kappa e_0 = \lambda_1 a_1 + \lambda_2 a_2 \propto i_0$.

Data Processing We extract data of outbreak and subsequent "dynamic clearing" from John Hopkins University's data[6] (Dong et al., 2020). Specifically, we first extract candidate outbreak data by searching for periods that have a sudden consistent increase in confirmed cases in China (detailed rules can be seen in Appendix A) and manually filter periods that does not belong to the initial outbreak of COVID-19 in early 2020 (because "dynamic clearing" policy was not applied until mid 2020) and those where the influence of non-indigenous cases is significant. In total, we obtain 7 periods of outbreak from 7 different cities/provinces. A summary of the extracted data can be found in Appendix A.

Evaluation To fit Equation 13 to real world data, we adopt the least square criteria, i.e. $\min_{a_1, a_2, \lambda_1, \lambda_2} \sum_{t=0}^{n} (c_t - a_1 e^{\lambda_1 t} - a_2 e^{\lambda_2 t} + a_1 + a_2)^2$. Note that for each λ_1, λ_2, the best a_1, a_2 can be uniquely determined by

$$\begin{bmatrix} \sum_{t=0}^{n} (e^{\lambda_1 t} - 1)^2 & \sum_{t=0}^{n} (e^{\lambda_1 t} - 1)(e^{\lambda_2 t} - 1) \\ \sum_{t=0}^{n} (e^{\lambda_1 t} - 1)(e^{\lambda_2 t} - 1) & \sum_{t=0}^{n} (e^{\lambda_2 t} - 1)^2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum_{t=0}^{n} (e^{\lambda_1 t} - 1)c_t \\ \sum_{t=0}^{n} (e^{\lambda_2 t} - 1)c_t \end{bmatrix} \tag{14}$$

[6] https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series



(a) $\kappa e_0 \sim i_0$ (b) days $\sim \ln i_0$

Figure 7: Evaluation results of the linear approximate model. (a): The scatter plot of κe_0 against i_0 and the fitted proportional line. (b): The scatter plot of the duration of outbreaks against $\ln i_0$ and the fitted linear line.

A simulated annealing algorithm is then used to search for the best λ_1, λ_2. Because empirically, we find that $\lambda_1 \approx \lambda_2$, which will make the coefficient matrix of Equation 14 ill-conditioned and lead to large variance on the estimation of a_1, a_2, we instead set $\lambda_1 = \lambda_2 = \lambda, a_1 = a_2 = a$ and search for optimal λ within the range $[-0.5, 0.5]$. The fit is good given average $R^2 = 95.6\%$ and minimum $R^2 = 91.2\%$, which implies the linear approximate model fits the outbreak setting.

The scatter plot of the recovered κe_0 and i_0 is in Figure 7 and the estimated relationship is $e_0 = \frac{4.47}{\kappa} i_0$. Given the usual value of $\kappa \approx 0.1$, $e_0 \approx 45 i_0$, and the usual i_0 within the range of $[1, 10]$, it implies that in an usual outbreak, the number of exposed people is within the range from several dozens to several hundreds, which matches with previous outbreaks. The R^2 of the fit is 5.9%.

Given that $e_0 \propto i_0$ and suppose $\lambda_1 \approx \lambda_2 \approx \lambda$, the solution in Equation 4 can be written as $i(t) \approx C \cdot e^{\ln i_0 + \lambda t}$ and the duration of a outbreak given initially found cases is

$$i(t) = c \Rightarrow t = a \ln i_0 + b \tag{15}$$

where c is some small constant, which marks the end of a period. The scatter plot of the duration of the outbreak against $\ln i_0$ is shown in Figure 7. The fitted line is $t = 2.52 \ln i_0 + +37.48$ and $R^2 = 4.7\%$. This result indicates that an outbreak would end in one month or one and a half months, which matches to the real world experiences. The low R^2 indicates that there are still gaps between our model and the real world dynamics.

5 Conclusion

In this work, we developed 4 different models from either top-down view or bottom up view and demonstrated the closeness between the models. We fitted our model to the real data and performed comprehensive investigation to the applicability to the real world situation. Compared with previous work, we additionally study the evolution of the pandemic situation in an outbreak scenario. Results from fitting to real world data indicate that our model is applicable to most of the cases. The results produced from the outbreak scenario imply that our model generate reasonable results but still has a gap to real world data. The gap can be from factors not included in the models, such as immigration, transportation, inoculation, spatial distribution of different group of people, e.t.c.. Future work can focus on including factors to make the model more realistic.

References

L. J. Allen. A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling*, 2(2):128–142, 2017.

A. H. Auchincloss and A. V. Diez Roux. A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health. *American journal of epidemiology*, 168(1): 1–8, 2008.

F. Ball. The threshold behaviour of epidemic models. *Journal of Applied Probability*, 20(2):227–241, 1983.

F. Ball and P. Donnelly. Branching process approximation of epidemic models. *Theory of Probability & Its Applications*, 37(1):119–121, 1993.

D. Bernoulli. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem*, pages 1–45, 1760.

F. Brauer. Compartmental models in epidemiology. In *Mathematical epidemiology*, pages 19–79. Springer, 2008.

E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

A. M. El-Sayed, P. Scarborough, L. Seemann, and S. Galea. Social network analysis and agent-based modeling in social epidemiology. *Epidemiologic Perspectives & Innovations*, 9(1):1–9, 2012.

C. Farrington, M. Kanaan, and N. Gay. Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, 4(2):279–295, 2003.

I. C.-. forecasting team. Modeling covid-19 scenarios for the united states. *Nature medicine*, 2020.

A. Gray, D. Greenhalgh, L. Hu, X. Mao, and J. Pan. A stochastic differential equation sis epidemic model. *SIAM Journal on Applied Mathematics*, 71(3):876–902, 2011.

Y. Gu. Covid-19 projections using machine learning. `https://covid19-projections.com`, 2020. Accessed: 2021-12-01.

B. Ingalls. An introduction to mathematical modelling in molecular systems biology, 2012.

D. Kiouach and Y. Sabbar. Ergodic stationary distribution of a stochastic hepatitis b epidemic model with interval-valued parameters and compensated poisson process. *Computational and mathematical methods in medicine*, 2020, 2020.

A. Kolmogoroff. Über die analytischen methoden in der wahrscheinlichkeitsrechnung. *Mathematische Annalen*, 104(1):415–458, 1931.

C. M. Macal and M. J. North. Tutorial on agent-based modeling and simulation. In *Proceedings of the Winter Simulation Conference, 2005.*, pages 14–pp. IEEE, 2005.

M. Mahrouf, A. Boukhouima, H. Zine, E. M. Lotfi, D. F. Torres, and N. Yousfi. Modeling and forecasting of covid-19 spreading by delayed stochastic differential equations. *Axioms*, 10(1):18, 2021.

M. Mandal, S. Jana, S. K. Nandi, A. Khatua, S. Adak, and T. Kar. A model based study on the dynamics of covid-19: Prediction and control. *Chaos, Solitons & Fractals*, 136:109889, 2020.

B. D. Marshall and S. Galea. Formalizing the role of agent-based modeling in causal inference and epidemiology. *American journal of epidemiology*, 181(2):92–99, 2015.

A. Mokhtari, C. Mineo, J. Kriseman, P. Kremer, L. Neal, and J. Larson. A multi-method approach to modeling covid-19 disease dynamics in the united states. *Scientific Reports*, 11(1):1–16, 2021.

V. C. Rideout. *Mathematical and computer modeling of physiological systems*. Prentice Hall Englewood Cliffs, NJ:, 1991.

M. Roberts and J. Heesterbeek. *Mathematical models in epidemiology*, volume 215. EOLSS, 2003.

R. Sameni. Mathematical modeling of epidemic diseases; a case study of the covid-19 coronavirus. *arXiv preprint arXiv:2003.11371*, 2020.

K. S. Sharov. Creating and applying sir modified compartmental model for calculation of covid-19 lockdown efficiency. *Chaos, Solitons & Fractals*, 141:110295, 2020.

Z. Wang. Stochastic models for epidemics. https://www.mathstat.dal.ca/~tsusko/honours-theses/ziwei-wang.pdf, 2021. Accessed: 2021-12-02.

R. W. West and J. R. Thompson. Models for the simple epidemic. *Mathematical biosciences*, 141(1): 29–39, 1997.

Province/City	Start date	End date
Beijing	6/11/20	7/19/20
Fujian	9/10/21	10/23/21
Hubei	1/3/21	2/17/21
Heilongjiang	1/11/21	2/20/21
Jiangsu	7/20/21	9/10/21
Jilin	1/15/21	2/19/21
Xinjiang	7/15/20	8/29/20

Table 3: A brief description of the extracted outbreaks (after manual selection).

Appendix A. Data Description

In this work, all real world data used for validation is from John Hopkins University[7] (Dong et al., 2020). The original data contains daily reported cumulative confirmed cases $ci(t)$, cumulative recovered cases $cr(t)$, and total deaths $p(t)$ in 280 regions, including countries, provinces, and states, from Jan 22, 2020 to Nov 15, 2021[8]. Because John Hopkins University stopped updating cumulative recovered data after Aug 4, 2021[9], we therefore use data from Jan 22, 2020 to Aug 4, 2021. With above data, we are able to recover the active cases, i.e. infected cases using

$$i(t) = ci(t) - cr(t) - p(t) \tag{16}$$

Due to some errors in John Hopkins University's data, we find that for some countries, there exists some time t when $i(t) < 0$. To ensure the data is reasonable and correct, we filter data of countries which have $i(t) < 0$ for some t.

Extracting Outbreak Data

Since the "dynamic clearing" policy is mainly implemented in China, we only extract outbreak periods from China. The rules for extracting candidate outbreak periods are as follows:

- The first day of a period has newly infected cases while the day before it does not.
- There are at most 14 days in a period which has no newly infected cases.

For provinces that have frequent international travellers, there are usually some non-indigenous cases every day and the rules above would capture a period covering the exact outbreak period. To this end, we perform post-processing to remove the period before the actual outbreak by the following rules:

- The number of newly infected cases in the first day of outbreak should be at least twice the number of newly infected cases in the day before the outbreak period.
- The number of newly infected cases in the second day of outbreak should be at least twice the number of newly infected cases in the day before the outbreak period.
- The number of newly infected cases in the first day of outbreak should be at least 5.

A brief description of the extracted outbreak period is in Table 3.

Appendix B. Implementation Details

Simulated Annealing For all countries or regions, we use the same parameters for the simulated annealing algorithm, that is initial temperature $T = 1$, decaying factor $\beta = 0.999$, and minimum temperature $t_{min} = 0.001$. Instead of using a fixed number of iterations, we keep the algorithm running until it did not find a better solution in consecutive 3,000 steps. The initial parameters for the DTMC model and the step sizes for next state in simulated annealing are in Table 4

[7] `https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series`

[8] The data is updated daily. We did not update our data after we downloaded the data.

[9] See the official response at `https://github.com/CSSEGISandData/COVID-19/issues/4465`.

param	initial value	step size
α_e	10^{-8}	10^{-8}
α_i	10^{-8}	10^{-8}
γ	0.01	10^{-3}
κ	0.1	10^{-3}
β	0.1	10^{-3}
ρ	0.01	10^{-3}
s_0	$5 \cdot 10^7$	10^6
e_0	10^3	10^3
e_{max}	10^3	10^3
i_{max}	10^3	10^3

Table 4: The initial values of model parameters and the step size. For each step in simulated annealing, the next value for the parameter is the current value add a standard normal random variable times the step size. For negative values, we set them to 0.

Model	Finished	Time
SEIR	Yes	0.11s
CTMC	No	300s
DTMC	Yes	0.04s

Table 5: The time for each model to simulate 100 steps. The SEIR model is simulated with step size being 10^{-2} in the Euler scheme. The CTMC did not finish in 300s.

Computational Efficiency The time for the three models, i.e. SEIR, CTMC, DTMC, to simulate 100 days with initial state $i_0 = p_0 = 10^5, r_0 = cr_0 = 2 \cdot 10^6$ and the model parameters in Table 4 is shown in Table 5.