Analysis of COVID-19 in Tokyo with Extended SEIR model and ensemble Kalman filter

T. Miyoshi[a], S. Richard[a,b], Q. Sun[a,b]✉

a) Data Assimilation Research Team, RIKEN Center for Computational Science (R-CCS), Japan

b) Graduate School of Mathematics, Nagoya University, Chikusa-ku, Nagoya 464-8602, Japan

E-mails: *takemasa.miyoshi@riken.jp, richard@math.nagoya-u.ac.jp, qiwen.sun@riken.jp*

1 The continuous comportmental model

We introduce the framework of the continuous model, and the various parameters involved.

1.1 The extended SEIR model

The existence of COVID-19 infectious spreaders who do not show (yet or at all) any symptom already brings a lot of uncertainties to health services. In addition, whenever the symptoms are relatively mild and because of some external pressure, some symptomatic individuals do hesitate to report the health service [12, p. 11]. As a consequence, daily new cases are under-reported. Combining this effect with some delayed information, some inaccurate tests (especially in the first phase of the epidemic), and other reasons that we are not aware of, it turns out that that the reported data are not very accurate. In such a situation one needs to use proper strategies to analyze the observation data.

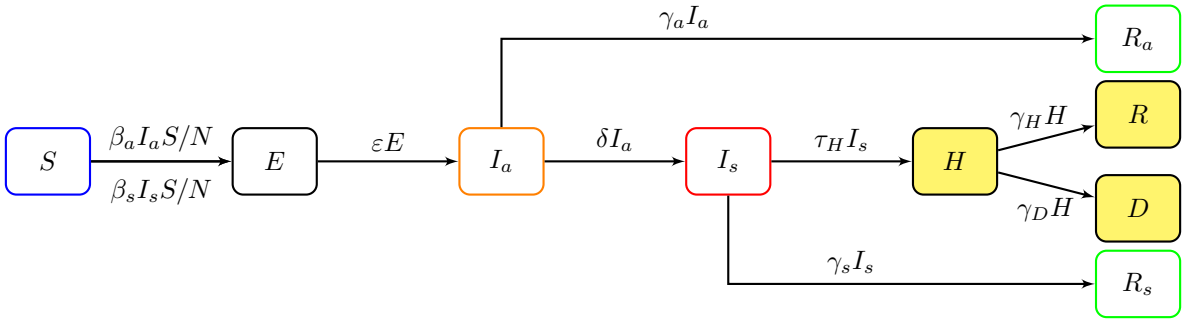

Figure 1: Transfer diagram for the extended SEIR model. Compartment I of the SEIR model is divided into two compartments I_a (asymptomatic/pre-symptomatic) and I_s (symptomatic)

In order to meet the special features of COVID-19, we separate the compartment I of the SEIR model into two compartments I_a and I_s. These compartments correspond to asymptomatic / pre-symptomatic and to symptomatic states, respectively. Both I_a and I_s can infect S. As shown in Figure 1, the transmission processes related to I_a and to I_s have transmission coefficients β_a and β_s respectively. Once an individual in S gets infected, this person becomes a member of E, and then moves to I_a when it becomes infectious. In I_a, part of these individuals (the asymptomatic) will never generate any symptoms. They will thus spend

some time in I_a, and then recover. The corresponding compartment is denoted by R_a. In contrast, the other part of individuals in I_a will develop symptoms, and then move to I_s. In I_s, a majority of individuals will be identified by health services, but as mentioned above a minority will recover without being identified by any health service. Compartment R_s corresponds to those recovered individuals who has not been registered. The identified persons in I_s will then move to the compartment H, which corresponds to hospitalized or treated patients. The ones staying at home but under medical control, or the ones isolated in a hotel, are all included in the compartment H. Finally, the patients in H will recover, and thus move to the compartment R, or will pass away and end up in the compartment D. As for the SIR or SEIR models, we assume a permanent immunity, which means that there is no flow from R_a, R_s, or R to S. Also, the total number N of individual is constant, namely at all time one has

$$N = S + E + I_a + I_s + H + R + D + R_a + R_s. \tag{1}$$

For Tokyo, the value $N = 13'955'000$ has been chosen (approximate mean value between the population of 2020 and 2021).

Based on the Figure 1 and on the explanations provided above, the differential system of the extended SEIR model reads:

$$\begin{aligned}
\frac{dS}{dt} &= -\beta_a I_a S/N - \beta_s I_s S/N & \qquad \frac{dR}{dt} &= \gamma_H H \\
\frac{dE}{dt} &= \beta_a I_a S/N + \beta_s I_s S/N - \varepsilon E & \qquad \frac{dD}{dt} &= \gamma_D H \\
\frac{dI_a}{dt} &= \varepsilon E - \delta I_a - \gamma_a I_a & \qquad \frac{dR_a}{dt} &= \gamma_a I_a \\
\frac{dI_s}{dt} &= \delta I_a - \tau_H I_s - \gamma_s I_s & \qquad \frac{dR_s}{dt} &= \gamma_s I_s \\
\frac{dH}{dt} &= \tau_H I_s - \gamma_H H - \gamma_D H
\end{aligned} \tag{2}$$

where $\beta_a, \beta_s, \varepsilon, \delta, \gamma_a, \gamma_s, \gamma_D$ and γ_H are some medical parameters. Note that some of them will be time dependent.

1.2 The reproduction number

The basic reproduction number, denoted by R_0, can be interpreted as the average number of secondary cases generated by on primary case in a susceptible population. It is commonly admitted that $R_0 = 1$ is a threshold value, as explained below. We also refer to [5, 8] for more explanations and more precise statements.

To study R_0 for different models, a general definition of the basic reproduction number is introduced based on the notion of disease free equilibrium (DFE). In a DFE the population remains in the absence of disease. For example, in the SIR or SEIR models, it means that $I = 0$ while in the extended SEIR model it means that $I_a = I_s = 0$. In this context, the basic reproduction number can be defined as the number of new infections produced by a typical infectious individual in a population at a DFE. The main feature of R_0 is that it corresponds to a threshold parameter, namely if $R_0 < 1$, the DFE is locally asymptotically stable, while if $R_0 > 1$, the DFE is unstable and an outbreak is possible.

The precise expression for R_0 is clearly model dependent, but numerous examples are available in the literature. For example, let us consider a general compartmental disease transmission model. Such models are represented by a system of ordinary differential equations, and under natural assumptions it can be shown that these models have a DFE, see [5]. In this reference, the precise expression for R_0 is then provided for some classes of models, and the extended SEIR model meets the assumptions of the staged progression model, as presented in [5, Sec. 4.3]. For the extended SEIR model, the expression for R_0 then reads:

$$R_0 = \frac{\beta_a}{\delta + \gamma_a} + \frac{\beta_s \delta}{(\delta + \gamma_a)(\gamma_s + \tau_H)}, \tag{3}$$

where β_a and β_s are the initial transmission coefficients at time 0.

To understand the above expression, observe that $\delta/(\delta+\gamma_a)$ corresponds to the fraction of individuals of I_a going to the compartment I_s, while $1/(\delta+\gamma_a)$ and $1/(\gamma_s+\tau_H)$ define the average times an infected individual spends in compartments I_a and I_s respectively. Thus, each term on the R.H.S. of (3) represents the number of new infections generated by an infectious individual during the time spent in the compartments I_a and I_s.

In contrast, the definition of the effective reproduction number R_t at time time t is much more delicate. The various challenges and possible errors have recently been discussed in [7], to which we refer for a thorough discussion. In our setting, we shall keep the interpretation of R_t as the average number of secondary cases generated by one primary case. This approach is possible because the transmission coefficients at time t are available in our approach, and therefore one can compute R_t with (3) and the corresponding β_a and β_s at time t.

1.3 The medical parameters

It clearly appears in Figure 1 and in the corresponding system (2) that several parameters control the flows between the compartments. The values of these parameters may result in very different behaviors of the model. We refer for example to [10, Sec. 1.4] and to [2, Chap. 2] for general discussions of model behaviors and the role of the parameters. In our setting, some parameters are easy to evaluate, as for example the recovery rate γ_H or the death rate γ_D, but others are difficult to estimate, as for example the transmission coefficients β_a and β_s. For that reason, some experiments are performed in order to study the stability of the model with respect to the parameters.

In Table 1 we list the estimated values of some parameters for the extended SEIR model, and provide the sources of information. Several parameters in the table have the form of the product of a percentage and the inverse of a time length. A similar setting for the construction of the parameters can be found for example in [6, 9]. For δ and γ_a, the percentage parts should sum up to 1. For the parameters τ_H and γ_s, some information can be found in the surveys [12, 14] and the health services website [15]. For parameter γ_H and γ_D, instead of using constant value estimated by health services, we shall use the observation data to estimate their values on a daily basis. The details will be discussed in Section 2.

Let us stress that the value assigned to some of these parameters has evolved during the last 12 months. For example, the ratio of asymptomatic individuals was thought to be very high at the beginning of the epidemic (up to 80%), but some recent research [1, 3, 13] have revised this ratio to 17% to 20%. Our knowledge about the length of infectious periods has also evolved, and the possible values are spread over a rather long intervals. In Table 1 we list some mean values, but in the simulations additional uncertainties are implemented. Since pre-symptomatic patients become infected before the appearance of symptoms, the incubation period (encoded in ε) has been shorten a little bit, and the last 2 days of this incubation period have been moved to I_a.

One very delicate question is the relation between β_a and β_s, namely between the transmission coefficient for asymptomatic / pre-symptomatic and the transmission coefficient for symptomatic. For our investigations, we shall rely on the result of the systematic review [1] which asserts that the relative risk of asymptomatic transmission is 42% lower than that for symptomatic transmission. As a consequence, we shall fix

$$\beta_a = 0.58\beta_s \quad \text{ or equivalently } \quad \beta_s = 1.72\beta_a. \tag{4}$$

This factor 0.58 is slightly smaller but of a comparable scale compared to earlier investigations, see for example [3]. In order to perform some experiment with different factors, let us set more generally

$$k := \frac{\beta_a}{\beta_s}$$

and call it *the relative infectivity coefficient.*

parameter	estimation	source	remark
ε	$(3 \text{ days})^{-1}$	[4]	Incubation period, not contagious
δ	$83\% \times (2 \text{ days})^{-1}$ (95% CI 80% to 86%)	[1, 4]	proportion of pre-symptomatic $\times$ (duration of pre-symptomatic)$^{-1}$
τ_H	$78\% \times (8.3 \text{ days})^{-1}$ (until May 31) $78\% \times (5.2 \text{ days})^{-1}$ (from June 1)	[12, 11]	proportion of detected symptomatic $\times$ (days of symptoms onset)$^{-1}$
γ_a	$17\% \times (9 \text{ days})^{-1}$ (95% CI 14% to 20%)	[1, 13]	proportion of asymptomatic $\times$ (duration of asymptomatic)$^{-1}$
γ_s	$22\% \times (7 \text{ days})^{-1}$	[12, 14]	proportion of undetected symptomatic $\times$ (duration of symptomatic)$^{-1}$
γ_H	computed by observations	[16]	discussed in Section 2
γ_D	computed by observations	[16]	discussed in Section 2

Table 1: Medical parameters

1.4 The observations

One of the aims of this study is to estimate the unknown parameters β_a and β_s, and the unobservable compartments I_a and I_s. Because of relation (4), it is clear that for the parameters only β_s has to be evaluated. The Ensemble Kalman filter (EnKF) requires observations of some compartments in the extended SEIR model. In Figure 1, we highlight the three compartments with observations available in yellow, namely H, R and D. The data corresponding to these compartments may not be perfect, and for the analysis based on these data we shall take some uncertainties into consideration. More precisely, since the source of uncertainties may not be clear, and since the scale of uncertainties may be difficult to evaluate, some random noise will be added subsequently.

For our experiment, we shall firstly and mainly use the data about Tokyo, with a total population N of nearly 14 million residents. As already mentioned, the compartment H corresponds to the identified individuals either hospitalized or treated at home or in a hotel. On the other hand, R and D describe the accumulated number of recovered and deceased individuals. The information about these three compartments are very easy to find for Tokyo. One can check for example the website of Ministry of Health, Labour and Welfare, prefecture's websites or some mass communication companies' websites to get more details. The information is provided on a daily basis. Note however, that these records were not very accurate at the beginning of the outbreak. This was caused by the delay of reports, but also by some changes in the policy for the collect of information. Our analysis will take this source of uncertainties into consideration.

2 Experiments

Health officials started providing data from February 2020 but they included some uncertainties caused by delay or by policy changes. Our experiments start with the observations from March 6^{th} 2020, when recorded

data became more reliable.

For the state space model mentioned in Section 1, the corresponding dynamical system is described by the differential system (2). To estimate the parameter β_s, we use the augmented state, which means we add one more equation $\frac{d\beta_s}{dt} = 0$ to the system (2) but we assume that there is no dynamics for β_s. This means that β_s will not be updated during the integration process, but this parameter will nevertheless be updated during the assimilation process. In other words, if β_s has a correlation with the observations, then β_s will be updated together with the other states. In system (2), the unit of time $t = 1$ represents one day. Thus, the one day forecast from day n is obtained by integrating the system (2) on $[n, n+1]$ with the initial values equal to the analysis on day n. To avoid negative values when the data assimilation (DA) is performed, all the compartments are transformed to log scale with base e. The lower case will be used for these new variables, as for example $s(t) := \log S(t)$, and the corresponding equation becomes $\frac{ds(t)}{dt} = \frac{1}{S(t)}\frac{dS(t)}{dt}$. The DA update directly applies to the log-transformed values.

Now the forecast value of x_t is a 9-dimensional vector

$$\big(e(t), i_a(t), i_s(t), h(t), r(t), d(t), r_a(t), r_s(t), \log \beta_s\big)^T. \tag{5}$$

Note that the compartment $s(t)$ is not explicitly taken into account since it can be deduced from the conservation relation (1). Before performing the DA update, the above vector has to be projected to the observation space, namely the forecast vector (5) is projected on $\big(h(t), r(t), d(t)\big)^T$ which corresponds to the three compartments with observations.

To initiate the experiments, initial values for all compartments and parameters have to be provided. When these initial values are far away from the observations, the system goes through an unstable transition period. To avoid or reduce this unstable analysis, preliminary experiments were performed for determining suitable initial values. Namely, we run the system for a few days, starting with a few individuals in some compartments and a presumed value for β_s. Different values of β_s were tested, until values comparable to the data of the compartments H, D, R on day zero (March 6^{th} 2020) were obtained. The values obtained for the different compartments were then chosen as initial conditions. However, independent normal distributed random errors are also added to create 15 ensemble members $\{x_0^{a(i)}, i = 1, \ldots, 15\}$. Note that each member $x_0^{a(i)}$ is a 9-dimensional vector containing the initial conditions for the compartments and for the parameter β_s.

The ensemble on day zero are then integrated by the model (2) for one day and the EnKF will update the one day forecast by using the observations. The procedure of the experiments is shown in Figure 2.

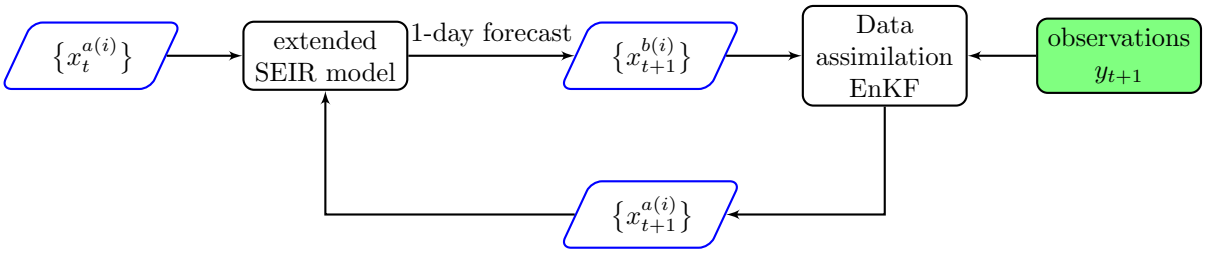

Figure 2: Data assimilation flow chart

As mentioned in Section 1.3, parameter γ_H and γ_D are estimated by using the observations. Consider equation $\frac{dR}{dt} = \gamma_H H$ in model (2). Since the time is discrete, we can rewrite this relation as $R(t+1) - R(t) = \gamma_H(t)H(t)$ which leads to and

$$\gamma_H(t) = \frac{R(t+1) - R(t)}{H(t)}. \tag{6}$$

Similar computation can be generated for γ_D also, namely,

$$\gamma_D(t) = \frac{D(t+1) - D(t)}{H(t)}. \tag{7}$$

For the implementation of these parameters for the computation of the 1-day forecast, we have used a slightly smoothed version obtained by a 7-day convolution with weights $\frac{1}{64}(1, 6, 15, 20, 15, 6, 1)$. The effect is to decrease a little bit the size of the weekly oscillations.

For the integration process, we have also included some uncertainty to all parameters: to the ones presented in Table 1, but also to the values of the parameters γ_H, γ_D. Thus, the 1-day forecast is obtained with the parameters of the previous day, each of them perturbed by a normal distribution $N\big(0, (M/10)^2\big)$, where M corresponds to the value of this parameter. These perturbations are independent and randomly generated for each 1-day forecast.

For simplicity, we assume that the observation error covariance is independent of time. The diagonal elements are chosen as $\big(\log(1.3)\big)^2$ for any time t. One can interpret this as the observation errors for h, r and d have $\big[-\log(1.3), \log(1.3)\big]$ as 68% CI and $\big[-2\log(1.3), 2\log(1.3)\big]$ as 95% CI. Considering that all the observations have been transformed into log scale, it means that the observations in original scale are distributed in $\big[(x/(1.3), (1.3)x\big]$ as 68% CI and $\big[x/(1.69), (1.69)x\big]$ as 95% CI.

References

[1] O. Byambasuren, M. Cardona, K. Bell, J. Clark, M.-L. McLaws, P. Glasziou, *Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis*, J. Association of Medical Microbiology and Infectious Disease Canada Volume 5 Issue 4, December 2020, 223–234.

[2] F. Brauer, P. v. d. Driessche, J. Wu, *Mathematical epidemiology*, Lecture Notes in Mathematics 1945, Springer, 2008.

[3] D. Buitrago-Garcia, D. Egli-Gany, M. J. Counotte, S. Hossmann, H. Imeri, A, M. Ipekci, et al., *Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis*, PLoS Med 17(9): e1003346, 2020.

[4] C. McAloon, A. Collins, K. Hunt, et al., *Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research*, BMJ Open 2020;10:e039652.

[5] P. v. d. Driessche, J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Mathematical Biosciences 180 (2002) 29–48.

[6] G. Evensen, J. Amezcua, M. Bocquet, A. Carrassi, A. Farchi, A. Fowler, P. L. Houtekamer, C. K. Jones, R. J. de Moraes, M. Pulido, C. Sampson, F. C. Vossepoel, *An international initiative of predicting the SARS-CoV-2 pandemic using ensemble data assimilation*, Foundations of Data Science (2020), American Institute of Mathematical Sciences.

[7] K.M. Gostic, L. McGough , E.B. Baskerville, S. Abbott, K. Joshi, C. Tedijanto, et al., *Practical considerations for measuring the effective reproductive number, R_t*, PLoS Comput Biol 16 No. 12 (2020), e1008409, 21 pages.

[8] H.W. Hethcote, *The mathematics of infectious diseases*, SIAM Rev. 42 No. 4 (2000) 599–653.

[9] T. Kuniya, H. Inaba, *Possible effects of mixed prevention strategy for COVID-19 epidemic: massive testing, quarantine and social distancing*, AIMS Public Health 7 No 3 (2020) 490–503.

[10] M. Li, *An Introduction to Mathematical Modeling of Infectious Diseases*, Mathematics of Planet Earth 2, Springer 2018.

[11] https://www.m3.com/open/iryoIshin/article/849820/

[12] Osaka prefecture government, *Citizens awareness and behavior change of measures against COVID-19*, http://www.pref.osaka.lg.jp/hodo/attach/hodo-40479_4.pdf

[13] A. M. Pollock, J. Lancaster, *Asymptomatic transmission of covid-19*, BMJ 2020;371:m4851.

[14] Bureau of social welfare and public health, *About death cases due to COVID-19 in Tokyo*, https://www.fukushihoken.metro.tokyo.lg.jp.

[15] Tokyo metropolitan government, *COVID-19 The information website*, https://stopcovid19.metro.tokyo.lg.jp.

[16] Toyokeizai, *Coronavirus disease (COVID-19) situation report in Japan*, https://toyokeizai.net.