International Journal of Interdisciplinary Research Statistics,
Mathematics and Engineering
ISSN: 2992-9261 | Impact Factor: 6.46
Volume. 09, Number 2; April-June, 2022;
Published By: Scientific and Academic Development Institute (SADI)
8933 Willis Ave Los Angeles, California
https://sadipub.com/Journals/index.php/ijirsme | editorial@sadipub.com

COMBINING DATA ASSIMILATION AND THE SIR INFECTION EQUATION FOR BETTER UNDERSTANDING OF COVID-19

G. Kitagawa,
Institute of Mathematical Analysis Osaka, Japan

Abstract: The outbreak of the new coronavirus pandemic (COVID-19) has created a global health crisis, and a significant challenge to mitigate its spread and impact on global communities. The SIR equation has been widely used as a theoretical model to simulate the course of an infection, however, its parameters cannot be known precisely due to the time lag between infection and onset and infectivity in the not-onset state. In this document, we propose a solution to the SIR infection equation using data assimilation to tackle the issue of parameter estimation in the present case. We integrate the SIR equation and observation equation within the framework of data assimilation, and estimate the unknown parameters using the maximum likelihood method. Data assimilation is considered effective in estimating the true value of unknown parameters through Bayesian inference. The state-space model is used to improve the accuracy of parameter estimation, generating virtual numerical simulations under different conditions and comparing the observation data on newly infected individuals, dead individuals, and severely ill individuals. Our study evaluates the effectiveness of data assimilation in obtaining accurate estimates of the infection rate, recovery rate, initial number of infected people, observation coefficient, and standard deviation of observation data noise.

Keywords: SIR equation, COVID-19 pandemic, data assimilation, parameter estimation, maximum likelihood method, Bayesian inference, state-space model.

INTRODUCTION

The new coronavirus infection (COVID-19) is rampant. Even now, more than a year after the start of the infection, the infection has not yet shown signs of ending. It can be said to be an infection that remains in history.

The most troublesome part of this infection is that it not only takes several days from infection to onset, but it also infects for several days even in the not-onset state. Therefore, a considerable number of infected persons with infectivity are left unchecked. Therefore, even if the infection status is simulated by the SIR equation [12], the true values of the infection parameters and the number of infected persons cannot be grasped.

However, it is possible to observe the infection status as it is. The daily number of infected people and the cumulative number of infected people are announced. The numbers in these data are not true values, but they reflect true values. It is very useful for getting a rough idea of the infection status.

As mentioned above, it is impossible to grasp the true value only by the SIR equation [1-2]. However, it may be possible to estimate the true value by combining it with the observation equation. In short, the framework

of data assimilation or state-space model [3-5] is considered to be effective. The parameters of the SIR equation, such as the daily number of infected people, infection rate, and eviction rate, are unknown but are embedded in the observed values. Unknown parameters must consist with the observed values. Wouldn't it be possible to determine these unknowns by combining them using Bayesian inference and the method of least squares? We report this effectiveness because we were able to confirm this effectiveness from the numerical results.

In order to fully explain the infection phenomenon, it is necessary to explain spatial characteristics such as the effect of the population density distribution, but this is not covered here. We focus on elucidating recurring infection waves and hidden infections. Regarding the spatial characteristics of the SIR equation, we would appreciate it if you could refer to the author's paper [2] in References.

2. STATE SPACE MODEL OF SIR INFECTION EQUATION

2.1. Space State Model

Among the physical models we are targeting, there are many that we know the differential equations (system models) that describe the physical phenomena (systems) but cannot directly observe the physical phenomena themselves or observe them sufficiently. .. An earthquake-like phenomenon would be a good example. It is not fully understood what kind of process is going on underground, including the system model. There is a weather forecast around us. Practical level forecasts might be difficult with theory alone.

However, even in such a case, there might be observation data reflecting a physical phenomenon. Although it might be insufficient, it might help to solve the problem. A state-space model that makes predictions using both theory and observational data gives us an idea of such cases. Originally born in control engineering, it is now considered to be a means of improving the accuracy of numerical simulations, and has come to be widely used in many fields other than control as data assimilation. The best known would be the weather forecast.

Let the system variable $x\,t(\,)$ be a function of time t , and let the observed variable be $y\,t(\,)$. For simplicity, we consider one-dimensional case. In the state space model, the system model that describes the physical system is given by $\frac{dx}{dt} = ax + bu + v_x$, (1)

and the observation model by $y = cx + du + v_y$, (2) where $u\,t(\,)$ is an external input, $a\ b\ c\ d$, , , are parameters, and $v_x(\,)t$ and $v_y(\,)t$ are noises following normal distributions with mean 0 and variances τ_x^2 and τ_y^2 . Namely, in one-dimensional case, we have

$$p(v_x) = \frac{1}{\sqrt{2\pi\tau_x^2}} \exp\left(-\frac{v_x^2}{2\tau_x^2}\right), \quad p(v_y) = \frac{1}{\sqrt{2\pi\tau_y^2}} \exp\left(-\frac{v_y^2}{2\tau_y^2}\right). \tag{3}$$

The non-linear case may be considered, but the linear case may be sufficient for the essential discussion. However, the SIR infection equation actually discussed below is multidimensional and non-linear. When the noise v_x does not exist or can be negligible, we have

$$\frac{dx}{dt} = ax + bu, \tag{4}$$

$$y = cx + du + v_y. \tag{5}$$

Only the observed noise v_y exists as noise. In this case, since it is easy to handle, a mathematical model is made using

Eqs. (4) and (5). The application to the SIR infection equation described below is also considered in this direction.

In this paper, we use the likelihood method (a special case of Bayesian inference) for data assimilation to solve the above state-space model. The key statistical property at that time is Eq. (5).

2.2. State Space Model of SIR Equation

The infection phenomenon caused by the new coronavirus infection COVID-19 follows the mean-field theory called the SIR equation:

$$\frac{dS}{dt} = -\frac{\beta}{N} SI, \quad (6)$$

$$\frac{dI}{dt} = \frac{\beta}{N} SI - \gamma I, \quad (7)$$

$$\frac{dR}{dt} = \gamma I. \quad (8)$$

N, S, I, and R are the total population, the uninfected population (strictly speaking, the susceptible population), the currently infected population, and the recovered population (including the dead). From Eqs. (6), (7), and (8), the following population conservation equations are obtained:

$$N = S + I + R, \quad (9)$$

β/N and γ are infection and recovery rates, respectively. The first term on the right-hand side of equations (6) and (7) is generally given in the form of not dividing by N, but as will be described later, in order to reduce the population dependence of β, it should be considered in the form of dividing by N.

If persons found infected are quarantined, the infectivity can be contained. Hence, that amount must be subtracted from the presently infected persons I in the calculation of newly infected persons. Applying SIR theory in such cases seems problematic. However, the SIR theory could be considered sufficient for the purpose of theoretically examining the possibility of data assimilation.

The following four points give difficulties to handle the infection phenomenon by the SIR equation:

(1) Since it is a mean-field theory, it is not possible to express the influence of spatial distribution characteristics such as population density.

(2) Since there are unidentified infected persons, the true number of infected persons is unknown.

(3) There are waves in the infection phenomenon, but it is necessary to introduce the mechanism of the waves.

(4) The effect of people coming and going cannot be clearly expressed.

On the other hand, daily time-series data on newly infected persons, dead persons, severely ill persons, etc. are published daily. New daily infections are not true daily infections, as some of them are unidentified. Severely ill people might be fairly close to the true value, and dead people would be the almost true value. Regarding (1), it is sufficient to extend the theory of spatial discretization such as dividing a large population group into smaller population groups. Regarding (2), it is impossible to observe the true number of the infected person, but since there is observation data that reflects it, it is conceivable to use this. Namely, the infection phenomenon will be modeled within the framework of the state space model. In the following, we will consider from this point of view. Regarding (3), when the number of infections decreases, factors that promote infection such as fewer people wearing masks are created due to the relaxation of people, and when the number of infections increases, factors that suppress infection are created. Regarding (4), the more people contact, the more infections, and the fewer people contact the fewer infections. Therefore, since infections increase or decrease due to social influences, the infection coefficient should be regarded as an effective coefficient that reflects these effects.

If x in Eq. (4) is considered as a three-dimensional vector and extended to the non-linear case, it can be used as a system equation for a state-space model of the present case.

On the other hand, observational data include daily newly infected person O_{DlyI}, severely ill person O_{SrsI}, and dead persons O_{Dead}. The currently infected person O_{PrsI} (~ I) and the cumulative infected person O_{AccI} could also be used as observation data. Of course, it would be possible to use multiple observation data, but in this paper, we will limit it to one observation data. For simplicity, the observation model shall use only the cumulative number of infected individuals for example:

$O_{AccI\,n}$ $cI_{Acc\,n}$ $_{n}$,(n 1,2,, N) . (10) n is the time step, and time t is t $n\,t$ (t 1day) .

Even if only equation (10) is used as the observation model, the effects of unknown variables such as the number of initial infections $I(0)$, infection rate , recovery rate , and variance ^{2} are reflected in the observation data. Furthermore, very importantly, the observation coefficient c can also be an unknown variable in this model.

Even if only equation (10) is used as the observation model, the effects of unknown variables such as the number of initial infections $I(0)$, infection rate , recovery rate , and variance ^{2} are reflected in the observation data. Furthermore, very importantly, the observation coefficient c can also be an unknown variable in this model.

2.3. Solution of SIR State Space Model Using Bayes Inference

If the observed data is O and the unknown is U_j, (j 1,2, , J) , the conditional probability is given by Bayes' theorem:

$P\,O\,U(\,,\quad ^{j})P\,O\,U(\,,\quad ^{j})P\,O\,U(\,|\quad ^{j})\,(P\,U^{j})$

$P\,U(_j\,|\,O)$. (11) $P\,O(\,)\quad _j P\,O\,U(\,,_j)\quad _j P\,O\,U(\,|_j)\,(P\,U_j)$

$P\,U(_j)$, $P\,O\,U(\,|_j)$, and $P\,U(_j\,|\,O)$ are called prior probabilities, likelihood functions, and posterior probabilities, respectively. This makes it possible to calculate the probability of an unknown quantity that yields observational data. Bayesian inference infers that U_j, which maximizes this posterior probability, gives O .

If it is allowed to assume

$PU(_1)\,PU(_2)\,PU(_J)$, (12)

Eq. (11) becomes $P\,O\,U(\,|\quad ^{j})$

$P\,U(\quad _j|\,O)$. (13) $_j P\,O\,U(\,|\quad _j)$

This is nothing but the likelihood estimation. In the following, this assumption is usd.

In the calculation of likelihood function $P\,O\,U(\,|_j)$, in case of $O\,O_{Acc\,I}$ for example, Eq. (10) is used. Namely

$P\,O$ ($_{AccI\,n}$ $cI_{AccI\,n}$) ($cI_{AccI\,n}$ $N(0,\quad ^2)$ or $P\,O(\,_{AccI\,n}\,(cI_{AccI\,n}))$ $N(1,\quad ^2)$. (14)

The effects of the number of initial infections $I(0)$, infection rate and recovery rate are reflected in the accumulated number of infections $I_{AccI\,n}$, (n 1,2, , N) . Using Eq. (14), we have

$P\,O\,U(\,|\quad _j)$ $N\quad 1\quad (OAccI\,n\quad cIAccI\,n\,)2\,(cIAccI\,n\,)2$ / $2\exp\sqrt{\quad}\,_2 2$. (15) n 1 $_2$

The maximum likelihood method is used for the estimation of unknown parameters. That is, the one that maximizes the likelihood function is found by the hill-climbing method. However, as can be seen from Eq. (15), if $P\,O\,U(\,|_j)$ itself is used, there is a high possibility that underflow will occur, so we consider taking the natural logarithm $\log P\,O\,U(\,|_j)$. Namely, the maximum of

$\log P\,O\,U(\quad |_j)$ / $\sqrt{\quad}$

$N \log_2 1\ 2\ n^N\ 1\ (O^{AccI\ n}\ cI_2{}^n)^2 2\ (cI^{AccI\ n})^2$ (16) is searched using the hill-climbing method. Since the unknown parameter etc. do not appear explicitly, the numerical differentiation is used to obtain the derivatives due to parameters.

3. VIRTUAL NUMERICAL SIMULATION

3.1. Generation of Observation Model

First, a time series of observation data is generated using the system model and the observation model under the following conditions.

mode 1 When using the number of currently infected people:

$O_{PrsI\ n}\ \ cI_{PrsI\ n}(1\ \ \ _n),(n\ \ 1,2,,N)$. … (17) mode 2 When using the daily number of newly infected people:

$O_{DlyI\ n}\ \ cI_{DlyI\ n}(1\ \ \ _n),(n\ \ 1,2,,N)$. … (18) mode 3 When using the cumulative number of infected people:

$O_{AccI\ n}\ cI_{AccI\ n}(1\ _n),(n\ 1,2,,N)$. … (19) The data released include the number of new infections, the number of severe cases, and the number of deaths on a daily basis. We consider the following as data. The daily number of newly infected persons is the published data itself, and the cumulative number of infected persons can be calculated from the daily number of newly infected persons, but the current number of infected persons cannot be calculated from the published data. The daily and cumulative numbers of infected people reflect data on those who are infected, but not those who leave the infection. Both of these are reflected in the currently infected person. These things are not a problem in a purely theoretical examination, but they are problems in applying the theory to reality. Table 1 List of parameters

Name Code in	Definition	Value
mode	PrsI, DlyI, AccI (Currently, daily and Cumulative number of infected people is used as observation data)	1, 2, 3
Npop	N : Population	1,000
beta1	: Infection rate (When obs. data generated)	0.4
gam1	: Recovery rate (When obs. data generated)	0.04
c1	c: Observation coefft. (When obs. data generated: Ratio to the true data)	0.25
S1[0]	S_0 : Initial no of susceptible(When obs. data generated)	997
I1[0]	I_0 : Initial no of current infected (When obs. data generated)	3
R1[0]	R_0 : Ini. No of recovered (When obs. data generated)	0
sgmObs1	$_{Obs}$ Std. dev. of obs. (When obs. data generated)	0.25
T	Observation period	100
dt	Observation step	0.1
oSkp	Time step of sampling obs. data	10
dlmd	Parameter Differentiation of during numerical diff. (std. val.)	$1.0*10^{-7}$

lmd	Moving step at hill-climbing (std. val.)	$1.0*10^{-7}$
iEnd	No. of convergence cal. at hill-climbing (std. val.)	30,000
	The initial values of infection rate , recovery rate , initial number of infected people I_0 , observation coefficient c, and observation data std. dev. σ_{Obs} when searching for the maximum probability value by the hill-climbing method are 0.8 times the set value at the time of observation data generation. However, I_0 converges slowly, so use the set value.	

3.2. Data Assimilated Numerical Simulation

The calculation assumes five unknown parameters: infection rate , recovery rate , initial number of infected people I_0 , observation coefficient c , and standard deviation σ_{Obs} of observation data noise. Since the convergence of the hill-climbing method is slow, it would be realistic to make it given. And, it may not be necessary to make the standard deviation σ_{Obs} of the observed data unknown, but it will be supposed to be unknown below.

(A) The setting of virtual infection status

Figure 1 shows the actual infection status according to the settings in Table 1. A description of the symbols is in Table 1. is the basic reproduction number:

$$SN(\quad) . \qquad (20)$$

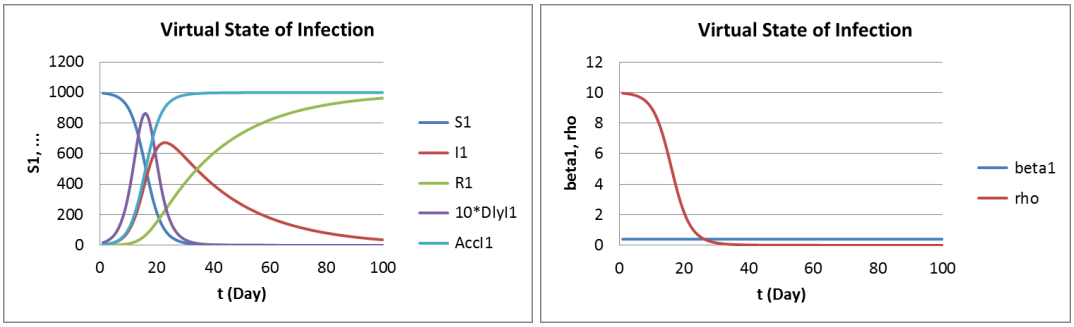


Fig. 1. Setting of Assumed Virtual State of Infection.

(B) The Effects of the Skipping of Observation Step

In order to ensure the accuracy of integration of the system model, we want to calculate in a 0.1-day time step. Time integration is performed by the Euler method. On the other hand, the actual observation data is given on a daily basis. Table 2 and Fig. 2 show the effects when the time step of integration of the system model and the time step of observation are not the same. The parameter I_0 was set to I_0 3 for the above reasons.

Table *2. Effects of Observation Skipping on Data Assimilation.

Parameter	Set Value	Skipping of Time Step for Observation	
		No Skipping	Observation Every 10 Steps
: Infection Rate	0.4	0.3980	0.3931
: Retired Ratio	0.04	0.0366	0.0399

c: Observation Coefft.	0.25	0.2699	0.2769
$\square_{Obs}^2$: Variance of Observed Data	0.25^2	0.2716^2	0.2590^2

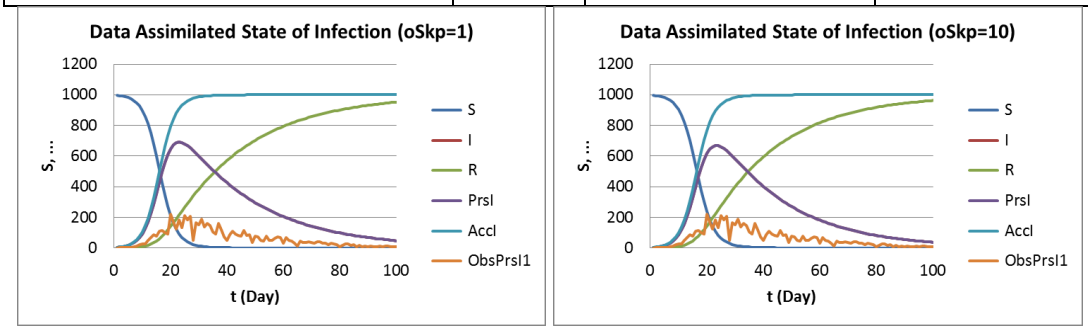


Fig. *2. Effects of Observation Skipping on Data Assimilation.

(C) Comparison of Observation Data

The observation data *PrsI* , *DlyI* and *AccI* are compared below. The comparison results are shown in Table 3 and Fig. 3. In the present calculation, there was a problem with data assimilation by *AccI* . As shown in Table 3, the recovery rate $\square$ is not accurate enough. Looking at Fig. 3, it is clear that the case of *AccI* is incorrect.

Table 3. Effects of Observation mode on Data Assimilation.

Mode	Set Value	*PrsI* : Present Infection	*DlyI* : Daily Infection	*AccI* : Accumulated Infection
$\square$: Infection Rate	0.4	0.3980	0.3982	0.3684
$\square$: Retired Ratio	0.04	0.0366	0.0402	0.0109
c: Observation Coefft.	0.25	0.2699	0.2653	0.2812
$\square_{Obs}^2$: Variance of Observed Data	0.25^2	0.2716^2	0.2584^2	0.2594^2

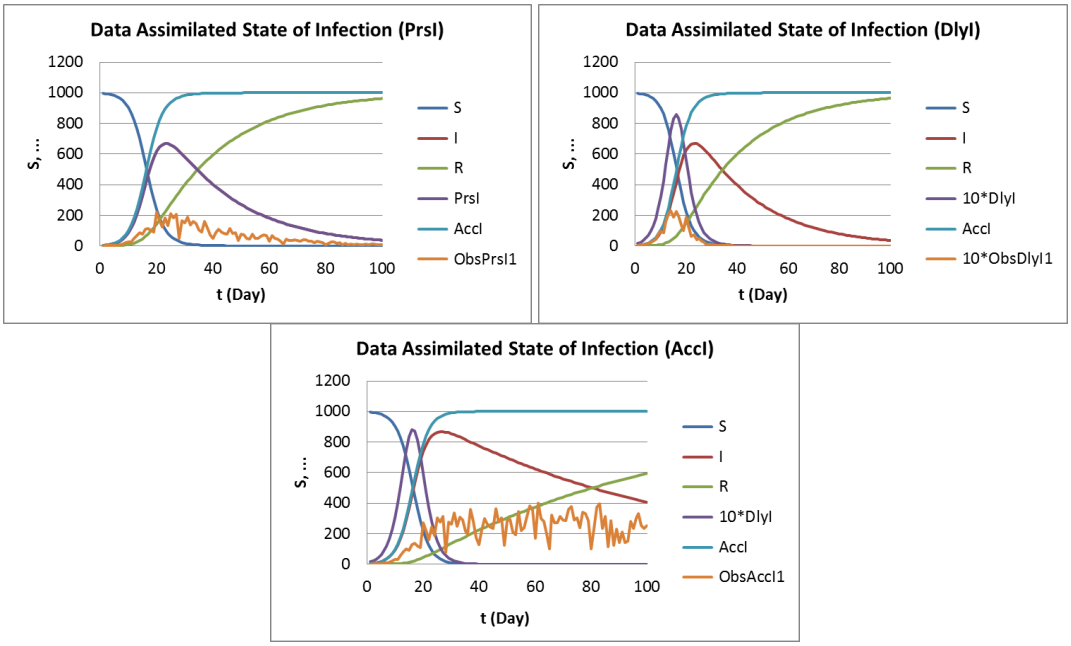


Fig. 3. Effects of Observation mode on Data Assimilation.

The daily and cumulative number of infected people reflects data on infected people but does not reflect data on those who leave the infection. *PrsI* , which reflects both the infection rate and the recovery rate , can be said to be the best observation data. However, the observation data cannot be obtained from the currently published data. In the case of *DlyI* and *AccI* , it is better to exclude the recovery rate from the estimation and to estimate a fixed value by some method. Table 4 and Fig. 4 show the results in that case.

Table 4. Effects of Fixing Retired Rate on Data Assimilation.

Mode	Set Value	*AccI* : Accumulated Infection
: Infection Rate	0.4	0.3915
: Retired Rate	0.04	0.04 (Given)
c: Observation Coefficient	0.25	0.2817
$_{Obs}^2$: Variance of Observed Data	0.25^2	0.2592^2

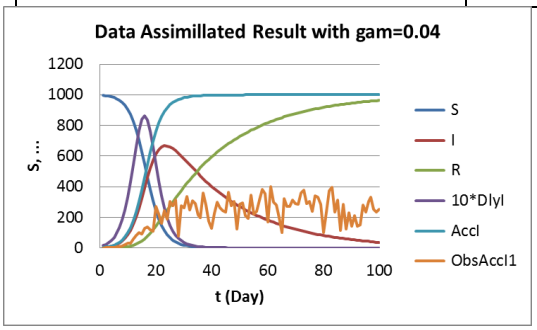


Fig. 4. Effects of Fixing Retired Rate on Data Assimilation.

(D) Predictive characteristics of data assimilated simulation

Table 5 and Fig. 5 show the results of data assimilated simulation using the number of currently infected people
PrsI as the observation data. It can be seen that even if the learning period is short, the infection state for the entire Period (100 days) can be predicted fairly accurately.

Table 5. Prediction of State of Infection (PrsI; TS: Days for Study).

Mode	Set Value	TS=10	TS=20	TS=50	TS=100
□: Infection Rate	0.4	0.4045	0.4035	0.3998	0.3933
□: Retired Ratio	0.04	0.0343	0.0342	0.0374	0.0398
c: Observation Coefft.	0.25	0.2463	0.2462	0.2592	0.2763
$\square_{Obs}^2$: Variance of Observed Data	0.25^2	0.2169^2	0.2171^2	0.2345^2	0.2587^2

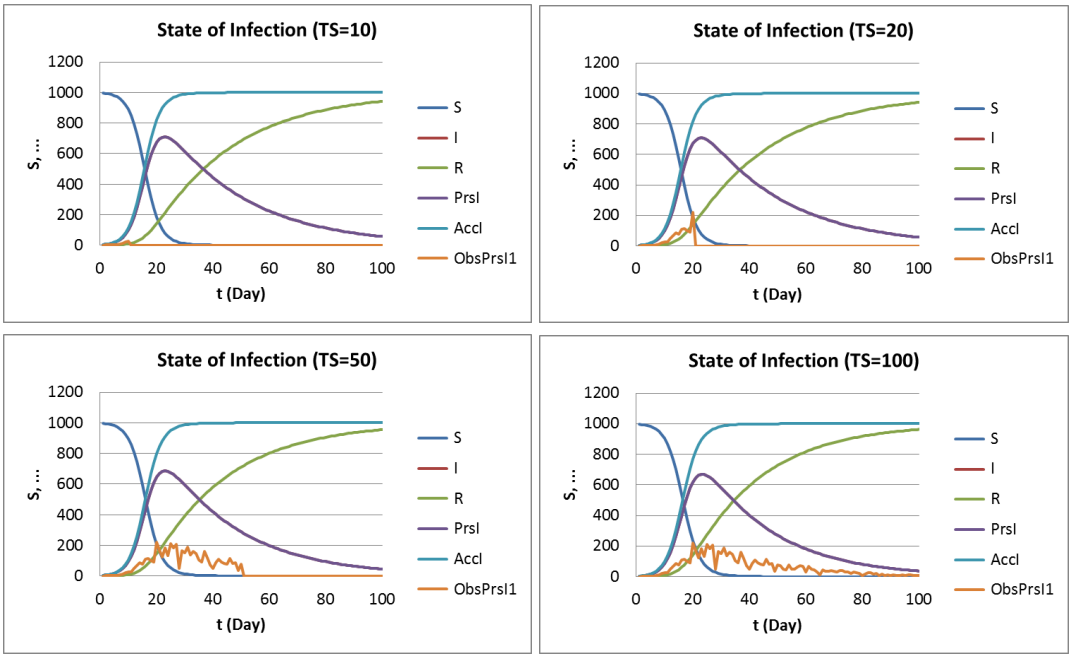


Fig. 5. Prediction of State of Infection (PrsI; TS: Days for Study)

(E) Effect of social infection control

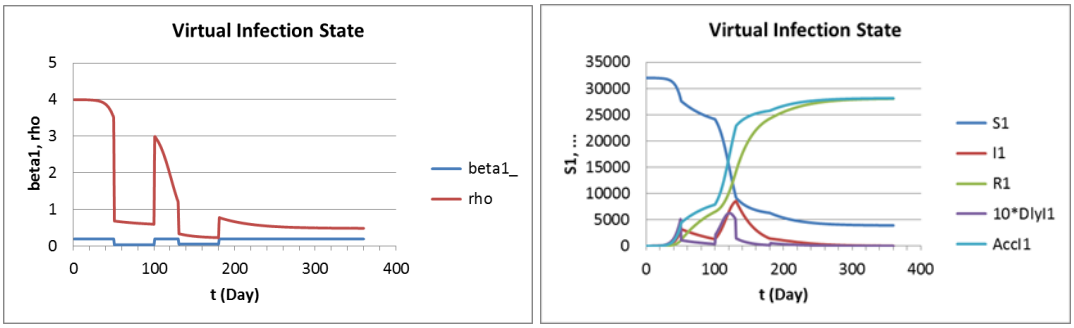
Infectious diseases can be controlled by social measures such as wearing masks, social distance, and controlling human flow. These effects manifest themselves as effective changes in infection rates. Table 6 shows the parameter settings during the numerical simulation, and Fig. 6 shows the virtual data observation results and the data assimilated simulation results. The number of currently infected people *PrsI* is used as observation data.

Fig. 6 (a) shows the virtual infection state performed to generate the observation data. It is assumed that the effective infection rate □was reduced from 0.2 to 0.04 with 80% control by two states of emergency declarations made to prevent the spread of infection. Looking at the data assimilated simulation results in Fig.

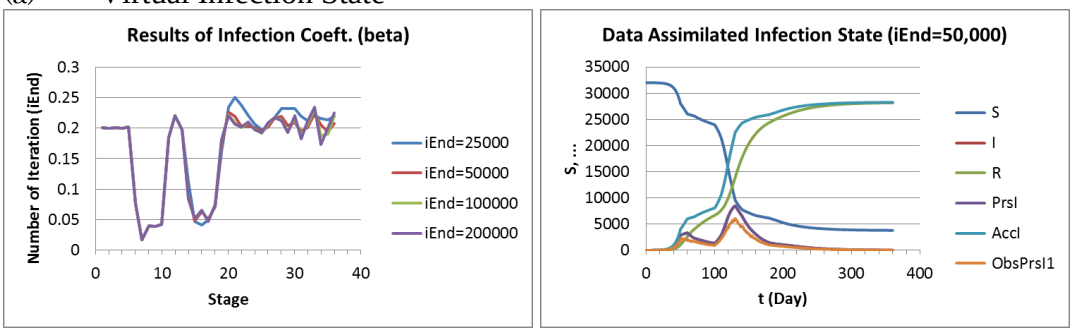
6 (b), it is clear that the virtual infection state is reproduced. However, in this data assimilated simulation, only the infection rate is set as an unknown variable in order to stably converge the parameters.

In this problem, the infection rate changes with time, so it is not possible to perform a batch calculation for the entire calculation time *T*. The calculation is performed assuming that 10 days is one stage and the parameters are constant in each stage. Table 6. Parameters for Numerical Simulation

Name in Code	Definition	Value
mode	*PrsI* (Currently number of infected people is used as observation data)	1
Npop	*N* : Population	32,000
beta1	: Infection rate (When obs. data generated)	0.2
gam1	: Recovery rate (When obs. data generated)	0.05
c1	*c*: Observation coefft. (When obs. data generated: Ratio to the true data)	0.7
S1[0]	S_0 : Initial no of susceptible(When obs. data generated)	31,998
I1[0]	I_0 : Initial no of current infected (When obs. data generated)	2
R1[0]	R_0 : Ini. No of recovered (When obs. data generated)	0
sgmObs1	$_{Obs}$ Std. dev. of obs. (When obs. data generated)	0.25
T	Observation period	360
dt	Observation step	0.1
oSkp	Time step of sampling obs. data	10
dlmd	Parameter Differentiation of during numerical diff. (std. val.)	$1.0*10^{-8}$
lmd	Moving step at hill-climbing (std. val.)	$2.0*10^{-7}$
iEnd	No. of convergence cal. at hill-climbing (std. val.)	50,000
	The entire period (360 days) is calculated by dividing it into 36 stages where 1 stage means 10 days. The maximum probability is searched by the hill-climbing method, using only the infection rate as an unknown parameter to make the convergence stable. The initial value of the infection rate is 0.8 times the set value at the	
	generation of the observation data.	



(a) Virtual Infection State



(b) Data Assimilatd Infection State

Fig. 6. Data Assimilatd Infection State.

4. DATA ASSIMILATED NUMERICAL SIMULATION USING REAL DATA

Unlike the case of virtual data assimilation using virtually created virtual observation data, real data assimilated simulation using actually observed real data is not straightforward. The study has just begun, as it seems to contain a variety of issues.

In the handling of COVID-19 in Japan, the persons found infected are quarantined and the infectivity is contained, so that amount must be subtracted from the presently infected persons I in the calculation of the newly infected person. Applying SIR theory to real data is problematic. However, if the spread of infection is not sufficiently isolated, the outline can be grasped by using SIR theory for the time being. As a result, the infection rate will be underestimated.

Although the observation data of Tokyo was used, the SIR equation, which is the mean-field theory, cannot be used as it is. The population of Tokyo is about 14 million, but the effective infection opportunity population involved in the infection at the time is only a small part of it. It increases with the spread of infection. In the following calculation, it is considered that the infection spreads in a ring shape at a constant rate, and it is considered that the infection increases with a linear function of time. As a result of trial and error, 16,000 people at 120 days and 200,000 people at 370 days. The numerical value connecting the two points with a straight line was taken as N .

As an observation model for real data, the following expression was used:

$$O_{AccI\,n}\ \ cI_{AccI\,n}\ \ _{n},(n\quad 1,2,,N)\ \square \qquad \square\square \qquad \square \qquad (21)$$

The parameters used in the data assimilated simulation are shown in Table 7, the data observation results are shown in Fig. 7, and the assimilated simulation results are shown in Fig. 8.

Until now, the phenomenon of infection wave could not be rationally reproduced in the calculation. If people's social activities are changed by a state of emergency, etc., the infection rate will change over time, causing wave motion. The infection waves have been successfully reproduced.

Comparing the 120-day analysis results in Fig. 8 with the 370-day analysis results, there is a difference in the results for the first 120 days, which should be the same. This is because, when applying the mean-field theory, the number of people who have the chance of infection is different from 16,000 and 200,000. The solution to this problem is a future problem, but it can be said that the current calculation results do not deny the validity of this calculation method.

In this calculation, only the infection rate is an unknown parameter. In order to make other parameters unknown, at least observation data that reflects the effect of the recovery rate will be required. Since only the infection rate is unknown, it seems that the temporal changes in other parameters are not sufficiently reflected in the changes in the infection rate .

In addition, the infection rate that appears in the actual infection phenomenon is not a pathologically defined infection rate with strictly defined conditions, but an effective infection rate that takes into account social impacts such as wearing masks, social distance, and human flow. The infection rate in Fig. 8 is such, and the effect of artificially suppressing infection by declaring an emergency is reflected some extent. Table 7 List of Parameters.

Name Code in	Definition	Value
mode	*AccI* (Cumulative number of infected people is used as observation data)	3
Npop	N : Effective Infective opportunity Population (Given at each stage)	16,000–200,000
beta1	: Infection rate (Initial value for convergence calculation)	0.2
gam1	: Recovery rate (Given at the beginning)	0.05
c1	c: Observation coefft. (Given at the beginning)	0.6
S1[0]	S_0 : Initial no of susceptible(Given at the beginning)	31,998
I1[0]	I_0 : Initial no of current infected (Given at the beginning)	15,999 – 199,999
R1[0]	R_0 : Ini. Number of recovered (Given at the beginning)	0
sgmObs1	$_{Obs}$ Std. dev. of obs. (Given at each stage)	240–4,000
T	Observation period	370
dt	Observation step	1
oSkp	Time step of sampling obs. data	1
dlmd	Parameter Differentiation of during numerical diff. (std. val.)	$1.0*10^{-8}$
lmd	Moving step at hill-climbing (std. val.)	$2.0*10^{-7}$
iEnd	No. of convergence cal. at hill-climbing (std. val.)	32,000

	The entire period (370 days) is calculated by dividing it into 37 stages where 1 stage means 10 days. The maximum probability is searched by the hill-climbing method, using only the infection rate as an unknown parameter to make the convergence stable. The initial value of the infection rate is 0.8 times the set value at the generation of the observation data.	

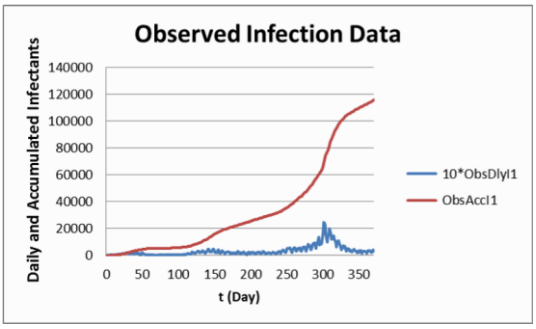


Fig. 7. Observed Infection Data.

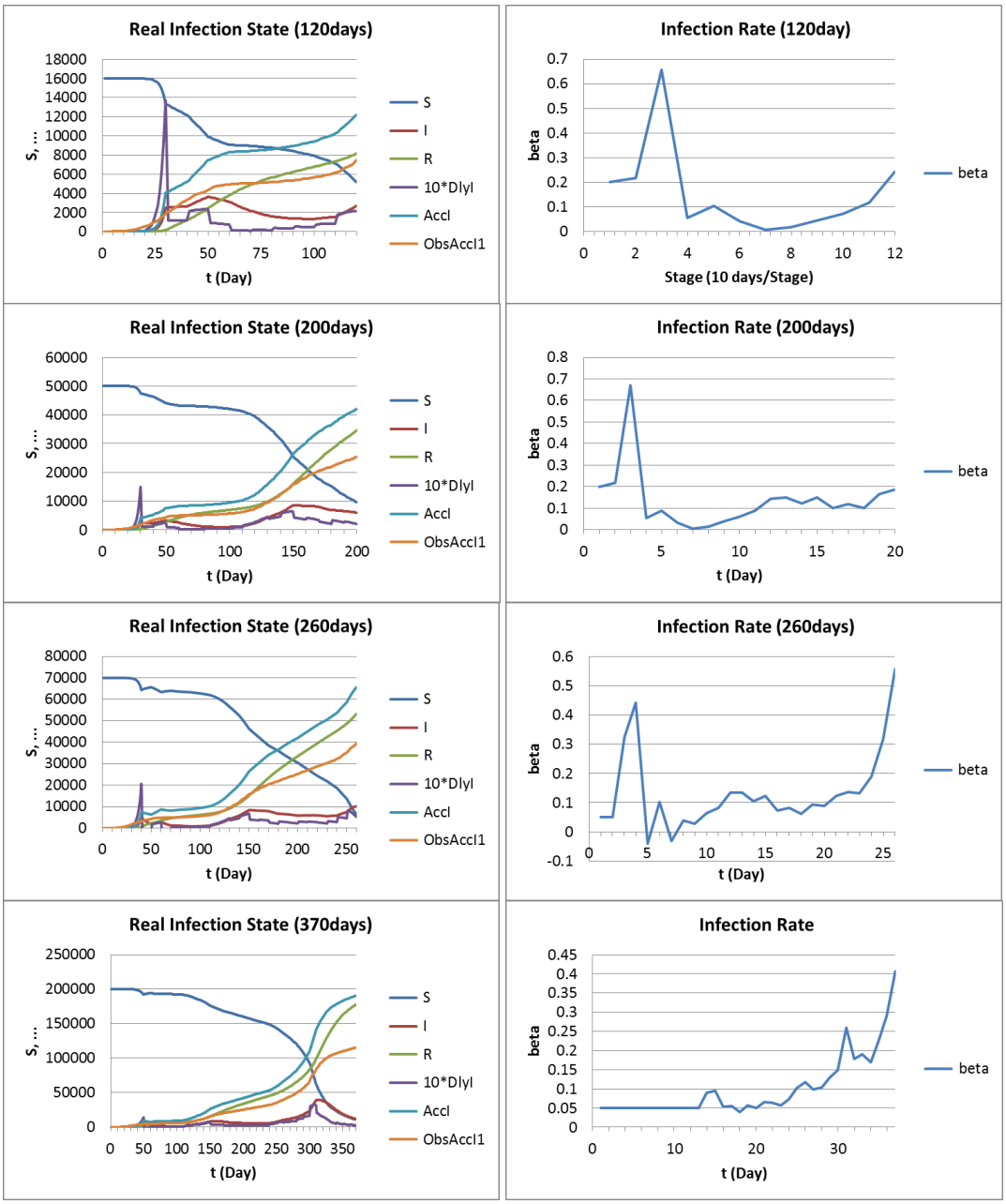


Fig. 8. Data Assimilation Results Using Observation Data of Tokyo (2020.03.13 - 2021.03.17)

5. CONCLUSIONS

The new coronavirus infection (COVID-19) is rampant. Even now, more than a year after the start of the infection, the infection has not yet shown signs of ending. It can be said to be an infection that remains in history.

The most troublesome part of this infection is that it not only takes several days from infection to onset, but it also infects for several days even in the not-onset state. Therefore, a considerable number of infected persons with infectivity are left unchecked. Therefore, even if the infection status is simulated by the SIR equation [13], the true values of the infection parameters and the number of infected persons cannot be grasped. However, it may be possible to estimate the true value by combining it with the observation equation. In short, a state-space model or a data assimilation framework is considered to be effective. The parameters of the SIR equation, such as the daily number of new infections, infection rate, and eviction rate, are unknown but

embedded in the observed values. Unknown parameters must match the observed values. Wouldn't it be possible to determine these unknowns by combining them using Bayesian inference and the method of least squares? This effectiveness could be confirmed by the numerical results.

Numerical calculations were performed not only when the observation data was artificially generated, but also when the actually published observation data was used. In the former case, consistent results were obtained for all observed data of the current number of infected persons, the number of newly infected persons on a daily basis, and the cumulative number of infected persons. Regarding the latter, the daily number of newly infected persons and the cumulative number of infected persons were used as observation data, and in this case as well, consistent results were obtained regardless of which observation data was used.

However, the daily number of newly infected persons and the cumulative number of infected persons reflect the infection rate, but do not reflect the eviction rate. As the observation data required for data assimilation, good data that reflects the eviction rate is absolutely necessary.

For virus mutant strains, the current concept of time-varying parameters is sufficient, but for the effect of the vaccines, it is necessary to subtract the vaccinated persons from the infected persons S. The author would also like to use the infection equation including the effect of the vaccine.

However, efforts to assimilate data that combine the mathematical theory and observational data of the time evolution of infectious diseases have only just begun, and many issues have not been fully considered. The current method of collecting and organizing statistical data on infectious diseases is not premised on being used for data assimilation. On the premise of introducing data assimilation, it will be necessary to rethink the way statistical data should be.

6. ACKNOWLEDGEMENT

The author would like to thank Professor Emeritus Hiroyuki Kajiwara of Kyushu University for providing the actual data of the new coronavirus infection (COVID-19).

7. REFERENCES

W. 0. Kermack, A. G. McKendrick, A Contribution to the Mathematical Theory of Epidemics, 1927, Proceedings of the Royal Society A (1927). https://doi.org/10.1098/rspa.1927.0118.

H. Isshiki, M. Namiki, T. Kinoshita, R. Yano : Effective Infection Opportunity Population (EOIP) Hypothesis in Applying SIR Infection Theory, cornell arXive, arXiv:2009.01837 (2020). https://arxiv.org/search/?query=Hiroshi+Isshiki&searchtype=all&source=header

G. Kitagawa, Use of a State Space Model in Time Series Analysis, Proceedings of the Institute of Statistical Mathematics, Vol. 67, No. 2 (2019) 181–192, in Japanese. https://www.ism.ac.jp/editsec/toukei/pdf/67-2-181.pdf

K. Fukaya, Time series analysis by state space models and its application in ecology, Japanese Journal of Ecology, Vol. 66, No. 2 (2016), 375-389 in Japanese. https://doi.org/10.18960/seitai.66.2_375

K. Law, A. Stuart, K. Zygalakis, Data Assimilation: A Mathematical Introduction, Springer (2015).