Abhinav Anand, Saurabh Kumar and Palash Ghosh*

Dynamic data-driven algorithm to predict cumulative COVID-19 infected cases using susceptible-infected-susceptible model

<https://doi.org/10.1515/em-2021-0034>

Received April 20, 2021; accepted September 27, 2021; published online October 15, 2021

Abstract

Objectives: In recent times, researchers have used Susceptible-Infected-Susceptible (SIS) model to understand the spread of the COVID-19 pandemic. The SIS model has two compartments, susceptible and infected. In this model, the interest is to determine the number of infected cases at a given time point. However, it is also essential to know the cumulative number of infected cases at a given time point, which is not directly available from the SIS model's present structure. The objective is to provide a modified SIS model to address that gap.

Methods: In this work, we propose a modified structure of the SIS model to determine the cumulative number of infected cases at a given time point. We develop a dynamic data-driven algorithm to estimate the model parameters based on an optimally chosen training phase to predict the number of cumulative infected cases.

Results: We demonstrate the proposed algorithm's prediction performance using COVID-19 data from Delhi, India's capital city. Considering different time periods, we observed the proposed algorithm's performance using the modified SIS model is well to predict the cumulative infected cases with two different prediction periods 30 and 40. Our study supports the idea of estimating the modified SIS model's parameters based on the optimal training phase instead of the entire history as the training phase.

Conclusions: Here, we have provided a modified SIS model that accounts for deaths due to disease and predicts cumulative infected cases based on an optimally chosen training phase. The proposed estimation process is beneficial when the disease under study changes its spreading pattern over time. We have developed the modified SIS model considering COVID-19 as the disease under focus. However, the model and algorithms can be applied to predict the cumulative cases of other infectious diseases.

Keywords: COVID-19; cumulative infected; dynamic data-driven algorithm; infected; optimal training data.

Introduction

The use of epidemiological models to control the spread of disease and predict the course of an outbreak has a long history. In 1760, Daniel Bernoulli proposed a mathematical model for smallpox (Hethcote 2000). At the beginning of the 20th century, William Hamer and Ronald Ross studied the epidemic behavior using the law of mass action (Hamer et al. 1928; Ross 1911). In recent times, the use of epidemiological models is inevitable for better management of infectious diseases (Agarwal and Jhajharia 2021; Gumel et al. 2021; Ramezani, Amirlatifi, and Rahimi 2021).

We have seen the use of various epidemiological models to combat the recent outbreak of Coronavirus disease 2019 (COVID-19). COVID-19 was first reported in Wuhan city of China, but soon spread to other parts of

---

*Corresponding author: Palash Ghosh, PhD, Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati, India; and Centre for Quantitative Medicine, Duke-NUS Medical School, National University of Singapore, Singapore, Singapore, E-mail: palash.ghosh@iitg.ac.in

Abhinav Anand and Saurabh Kumar, Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati, India

the world (Al-Raeei, El-Daher, and Solieva 2021; Gagliardi et al. 2020). Many authors have used some version of Susceptible-Infected-Recovered (SIR) models to predict the COVID-19 outbreak in different countries or regions (Gounane et al. 2021; Leonardi et al. 2020; Magnoni 2021; Ray et al. 2020; Wangping et al. 2020). The basic SIR model assumes that infected individuals are either recovered (and immune) from the disease or died (Keeling and Rohani 2008). It also assumes the number of deaths from the disease is negligible compared to the total population. However, the World Health Organization (WHO) mentioned that “there is currently no evidence that people who have recovered from COVID-19 and have antibodies are protected from a second infection” (WHO 2020). For example, health authorities in South Korea noticed that 163 patients became COVID-19 positive again after a full recovery (NPR 2020). Brouqui et al. 2021 stated that “in COVID-19, it quickly became apparent that naturally acquired immunity would not, in all cases, provide protection for the months following the first infection”. Iwasaki 2021 claimed that “reinfection cases tell us that we cannot rely on immunity acquired by natural infection to confer herd immunity”. Several other studies have found that individuals who are infected by the COVID-19 may build short-term immunity against the disease, and there is no long-lasting guaranteed protection (Edridge et al. 2020; Liu et al. 2020; Tillett et al. 2021). In this context, when there is no long-term protection from the disease after infection, the Susceptible-Infected-Susceptible (SIS) model is appropriate. In an SIS model, people who recover from the disease are added to the susceptible compartment as they can be infected again. In this work, we consider the SIS model to predict the COVID-19 outbreak.

In an SIS model, the main focus is to determine the number of infected people at a given time point. However, it is also essential for planning purposes to know the cumulative number of infected people at a given time point. One cannot directly find the cumulative number of infected people from the SIS model’s present structure. In this work, our main contribution is to provide an SIS model-structure which can give the cumulative number of infected people easily. We incorporate a death due to disease compartment in the SIS model to estimate the model parameters accurately. We develop a dynamic data-driven algorithm to estimate the model parameters efficiently to predict the cumulative infected cases. In this process, we show how to select the optimal training phase to build the model. Finally, the developed algorithm is implemented using COVID-19 data from Delhi, India’s capital city. We also provide an R-package so that users can easily implement the developed model with their data.

Susceptible-Infected-Susceptible (SIS) model

In an SIS model (Hethcote 1989), there are only two compartments, Susceptible and Infected. An SIS model assumes that an individual has not developed any long-term immunity against the disease after infection and thus is at risk of re-infection; hence, it gets added back to the susceptible population. In other words, as shown in Figure 1, after recovering from an infection, an individual again becomes susceptible. Examples of such infections are the common cold and influenza.

These equations can well describe an SIS model (Keeling and Rohani 2008),

$$\frac{dS}{dt} = -\beta \frac{SI}{N} + \gamma I, \quad (1)$$

$$\frac{dI}{dt} = +\beta \frac{SI}{N} - \gamma I. \quad (2)$$

Here t denotes time. In this work, a day is the smallest unit of time t . However, one can choose other suitable units as necessary. S and I are the susceptible and the infected number of people in the population, respectively. The total population size is N , which is the sum of susceptible (S) and infected (I) populations (see Table 1). From Eqs. (1) and (2), it is evident that $\frac{dS}{dt} + \frac{dI}{dt} = 0$. The parameter β , the transmission rate, is

Figure 1: Pictorial representation of an SIS model, where S stands for susceptible and I for infected.

Table 1: Description of key terms and parameters.

Key term/parameter	Description
N	Total population size
S	Susceptible population
I	Infected populations
C	Cumulative infected cases (re-infection is also a count)
D	Deceased population due to the disease
β	Transmission rate
γ	Recovery rate
μ	Mortality rate of infection

the product of the contact rates among infected, and transmission probability (Bjørnstad 2019; Keeling and Rohani 2008). In other words, the parameter β is the average number of individuals infected per unit time (a day) from an infected person. It can also be interpreted as contact rates between infected and susceptible. Here, by the assumption, I infected individuals can contact some individuals randomly; a fraction of $\frac{S}{N}$ of them will be susceptible. The parameter γ is the recovery rate. It is assumed to be $\frac{1}{T}$ where T is the average duration for which infection lasts in an infected person (Ferguson et al. 2020; Ghosh, Ghosh, and Chakraborty 2020). Equation (1) shows the rate of change in the susceptible (S) population; whereas Eq. (2) depicts the rate of change in the infected (I) population. The term $\beta \frac{SI}{N}$ denotes the number of susceptible people infected daily and is removed from the susceptible compartment and added to the infected compartment. The γI denotes the number of people recovered daily and is added back to the susceptible compartment and removed from the infected compartment. Figure 2 shows a simulated SIS model with $S + I = N$ at all time points. Notice that one cannot get the cumulative infected cases directly from the above SIS model. The death due to the disease is not adjusted into this model. It may affect the efficiency of estimating model parameters when the number of deaths due to disease is not negligible (as observed in COVID-19). In the next section, we consider a modified SIS model to address these two issues.

Model equations for modified SIS model

The proposed model (Figure 3) can be well described in these equations,

$$\frac{dS}{dt} = -\beta \frac{SI}{N} + \gamma I, \quad (3)$$

Figure 2: A simulated SIS model with the initial values of S, I as $S(0) = 1,000$ and $I(0) = 1$, respectively. Here, $\beta = 0.2$ and $\gamma = 0.05$.

Figure 3: Pictorial representation of a modified SIS model, where S stands for susceptible, I for infected and D for deceased population due to the disease.

$$\frac{dI}{dt} = +\beta \frac{SI}{N} - \gamma I - \mu I, \quad (4)$$

$$\frac{dC}{dt} = \beta \frac{SI}{N}, \quad (5)$$

$$\frac{dD}{dt} = \mu I, \quad (6)$$

$$\text{with, } \frac{dS}{dt} + \frac{dI}{dt} + \frac{dD}{dt} = 0, \quad (7)$$

where S, I, N, β and γ are the same as defined earlier. The C is the cumulative infected cases from the beginning (see Table 1). It includes every person who is infected or was infected. Note that if a person is infected twice (re-infection), it will be counted as 2 cases rather than 1 in C . The D is the deceased population due to the disease. Note that D does not include death counts from other causes. We assume that the death rate from other causes not involving the concerned disease is the same as the birth rate.

Equation (3) is the same as (1). Equation (4) represents the effective change in the infected compartment. As explained earlier, $\beta \frac{SI}{N}$ number of susceptible individuals get infected daily and are added to this compartment. The μ is the mortality rate of infection. Thus, γI infected individuals are recovered from infection daily, and μI infected cases resulting in death, are removed from this compartment. Equation (5) represents the rate of change in the cumulative infected cases, which is equal to daily infected cases ($\beta \frac{SI}{N}$). Equation (6) represents the rate of change in the deceased compartment which is equal to μI . Note that, $\frac{dC}{dt}$ is not included in Eq. (7) as the addition of $\beta \frac{SI}{N}$ is already done using $\frac{dI}{dt}$. In other words, $\frac{dC}{dt}$ is the first term of $\frac{dI}{dt}$.

Dynamic data-driven algorithm to estimate model parameters

In general, the SIS model parameters are constant for the entire duration of the study period. When the disease under consideration is present in the community for a longer time, the estimated parameter based on the entire study period may not give the right picture. For example, the COVID-19 disease outspread is highly unpredictable in the long term because contact rates and transmission probabilities are changing over time. They vary due to various reasons like control measures implemented by respective governments. Therefore, it may be appropriate to train an SIS model with a shorter training phase and make short-term predictions. Here, the “*dynamic data-driven algorithm*” means the training phase, used to estimate the model parameters, is dynamic (not fixed) and optimally chosen based on the appropriate historical data.

The two phases of the study period are the training and the prediction phases. Figure 4 shows how the study period is divided into different parts for estimation purpose. We define the four-time variables as follows:

- T_{Current} : Denotes the *date* when the training phase ends. After this date, the prediction phase starts.
- T_{Start} : Denotes the length of the default training phase (in days). The default training phase is the interval, $[T_{\text{Current}} - T_{\text{Start}} + 1, T_{\text{Current}}]$.
- T_{Pred} : Denotes the length of the prediction phase (in days). The prediction phase is the interval, $[T_{\text{Current}} + 1, T_{\text{Current}} + T_{\text{Pred}}]$.
- T_{Limit} : Denotes the upper limit of the number of additional days that can be added to the default training phase to optimally choose the training phase. The length of the training phase keeps increasing with a step of 1 day. Therefore, the maximum training phase interval can be, $[T_{\text{Current}} - T_{\text{Start}} - T_{\text{Limit}} + 1, T_{\text{Current}}]$.

Figure 4: Different time points of training and prediction phases.

Here, our objective is to choose an optimal training phase that can be used to predict the near future accurately. Note that for each value of $0 \leq t \leq T_{\text{Limit}}$, there is a different training phase denoted by $[T_{\text{Current}} - T_{\text{Start}} + 1 - t, T_{\text{Current}}]$. The default training phase, for $t = 0$, is $[T_{\text{Current}} - T_{\text{Start}} + 1, T_{\text{Current}}]$, whereas the maximum training phase, for $t = T_{\text{Limit}}$, is $[T_{\text{Current}} - T_{\text{Start}} - T_{\text{Limit}} + 1, T_{\text{Current}}]$. Note that a user can choose an appropriate length of the default training phase by taking a specific value of T_{Start} . We consider the default training phase as the fixed in-sample assessment period to compare different models based on different training phases. In general, the training phase and the assessment period are disjoint sets to compare several models' performance based on a fixed training phase data. However, here, we aim to choose an optimal training phase first and then select the model fitted on that phase as optimal. Here, "in-sample" refers that the assessment period is a subset of the considered training phase. The optimal criterion to choose the appropriate training phase is defined by the root mean squared error,

$$L(t, \beta, \mu) = \sqrt{\frac{1}{T_{\text{Start}}} \sum_i (C_p[i; t, \beta, \mu] - C_o[i])^2}, \quad (8)$$

where $T_{\text{Current}} - T_{\text{Start}} + 1 \leq i \leq T_{\text{Current}}$; $C_o[i]$ denotes the observed cumulative infected cases on i th day and $C_p[i; t, \beta, \mu]$ denotes the predicted cumulative infected cases on i th day considering $[T_{\text{Current}} - T_{\text{Start}} + 1 - t, T_{\text{Current}}]$ as the training phase. The optimal value of t is obtained (using (8)) as

$$t_{\text{opt}} = \arg \min_t L(t, \beta, \mu). \quad (9)$$

Finally, we obtain the optimal training phase that can be used for future prediction as $[T_{\text{Current}} - T_{\text{Start}} + 1 - t_{\text{opt}}, T_{\text{Current}}]$.

There are three parameters in the modified SIS model, namely, β , γ , and μ . As argued earlier, the recovery rate, $\gamma = \frac{1}{T}$, where T is the average duration for which infection lasts in an infected person. In other words, T refers to the average duration of recovery. In the case of COVID-19, T is taken as 14 (Ferguson et al. 2020; Ghosh, Ghosh, and Chakraborty 2020). However, in some specific cases, T can be more than 14 also (Arifin et al. 2020). Given a training phase, Algorithm 1 is used to find the estimates of β and μ by minimizing $L(t, \beta, \mu)$.

Prediction of cumulative infected cases

The $\hat{\beta}_{\text{opt}}$ and $\hat{\mu}_{\text{opt}}$ are the optimum values of β and μ , respectively, using Algorithm 1. Using Algorithm 2, predicted values of cumulative infected cases (C) are obtained for every day starting from $T_{\text{Current}+1}$ to $T_{\text{Current}+T_{\text{Pred}}}$. Here, the prediction period's length depends on the user-supplied value of T_{Pred} . For $1 \leq i \leq T_{\text{Pred}}$, the root mean squared error for prediction is

$$\sqrt{\frac{1}{T_{\text{Pred}}} \sum_i (C_p[i; t_{\text{opt}}, \hat{\beta}_{\text{opt}}, \hat{\mu}_{\text{opt}}] - C_o[i])^2}. \quad (10)$$

Algorithm 1: Dynamic data-driven algorithm to estimate β and μ .

Set N .
while $0 \leq t \leq T_{Limit}$ **do**
  Training Phase: $[T_{Current} - T_{Start} + 1 - t, T_{Current}]$;
  Set C_0, I_0, D_0 being the starting values of cumulative infected cases, active infected cases and deceased cases, respectively, on date $T_{Current} - T_{Start} + 1 - t$;
  for $\beta_{min} \leq \beta \leq \beta_{max}$ and $\mu_{min} \leq \mu \leq \mu_{max}$ **do**
     $(\hat{\beta}_t, \hat{\mu}_t) = \arg \min_{\beta, \mu} L(t, \beta, \mu)$;
    Calculate $L(t, \hat{\beta}_t, \hat{\mu}_t)$;
  end
end

$$\begin{aligned} t_{opt} &= \arg \min_t L(t, \hat{\beta}_t, \hat{\mu}_t); \\ \hat{\mu}_{opt} &= \hat{\mu}_{t_{opt}} \\ \hat{\beta}_{opt} &= \hat{\beta}_{t_{opt}} \end{aligned}$$

Algorithm 2: Prediction algorithm for the cumulative infected cases (C).

Prediction Phase: $[T_{Current} + 1, T_{Current} + T_{Pred}]$;
Obtain $\hat{\mu}_{opt}$ and $\hat{\beta}_{opt}$ from Algorithm 1;
Set $C_{T_{Current}}, S_{T_{Current}}, I_{T_{Current}}, D_{T_{Current}}$ being the starting values of cumulative infected cases, Susceptible cases, active infected cases and deceased cases, respectively, on date $T_{Current}$;
for $1 \leq j \leq T_{Pred}$ **do**

$$C_{T_{Current}+j} = C_{T_{Current}+j-1} + \hat{\beta}_{opt} \frac{S_{T_{Current}+j-1} I_{T_{Current}+j-1}}{N};$$

end

$$C = [C_{T_{Current}+1}, \dots, C_{T_{Current}+T_{Pred}}]$$

Equation (8) in Section 3 refers to the root mean squared error for the optimal criterion to choose the appropriate training phase. Whereas Eq. (10) refers to the root mean squared error for prediction. Denominators and parameter values in these two equations are different. Note that the equation of cumulative cases in the Algorithm 2 is approximated.

Application on real data

R-package

An R-package has been developed to help users easily implement the developed methodology with their data. The R-package is available from <https://github.com/abh2k/sisd>, with detailed instructions for its use. The

package is highly flexible in terms of different user-supplied parameters like T_{Current} , T_{Start} , T_{Limit} and T_{Pred} etc. Given the appropriate data and other required input parameter values, the R-package will provide $\hat{\beta}_{\text{opt}}$ and $\hat{\mu}_{\text{opt}}$, root mean squared error (based on “in-sample” assessment), predicted cumulative infected cases.

Predicting cumulative infected COVID-19 cases for Delhi

We consider the COVID-19 data from Delhi, India’s capital city with a population size of around 20 million, to demonstrate the proposed algorithm’s prediction performance. Delhi observed more than 600 thousands of cumulative COVID-19 infected cases at the end of 2020. The data is publicly available from <https://www.covid19india.org/>.

In Figure 5, we have considered four different T_{Current} as 29 May 2020 (in (A)), 24 July 2020 (in (B)), 29 December 2020 (in (C)) and 15 January 2021 (in (D)). This set-up can check the proposed algorithm’s prediction performance using the modified SIS model concerning different time periods. Table 2 shows the $\hat{\beta}_{\text{opt}}$, $\hat{\mu}_{\text{opt}}$, root mean squared error (RMSE) for prediction and other related information. From all the four graphs in Figure 5, it is evident that the proposed algorithm is working well to predict the cumulative infected cases with two different prediction periods 30 (for (A) and (B)) and 40 (for (C) and (D)). From Table 2, we see that the chosen optimal training periods’ lengths can be different with different values of $\hat{\beta}_{\text{opt}}$, $\hat{\mu}_{\text{opt}}$. Notice that $\hat{\beta}_{\text{opt}}$ is decreasing over time from 0.12 to 0.09 for Delhi, whereas $\hat{\mu}_{\text{opt}}$ increases for the first three scenarios (from 0.007 to 0.091) then dropped a little to 0.070. The reported RMSEs, specially in the last two scenarios, are quite good considering the scale of cumulative cases above 600,000. These observations support the idea of estimating the modified SIS model’s parameters based on the optimal training phase instead of the entire history as the training phase.

Figure 6 shows what could happen if we include the entire history as a training phase to estimate the model parameters. The blue line is the fitted trained line based on the estimated parameters from Algorithm 1 and 2. The red line indicates the fitted trained line based on the estimated parameters from the entire history data, not using Algorithm 1 and 2. The 30-day prediction curve based on the entire history (125 days) is exponentially higher than the observed curve of the cumulative infected cases (root mean squared error=175,884.1). The difference between the two curves is getting much bigger for the latter part of the prediction period. However, the prediction curve based on the optimal training phase (a total 23 days with 15 days of the assessment phase) is closer to the curve of observed cumulative infected cases (root mean squared error=3,090.25). The estimated model parameter $\hat{\beta} = 0.24$, $\hat{\mu} = 0.09$ based on the entire history, whereas the same parameter estimates are 0.10 and 0.06 using the optimal training phase, respectively. The $\hat{\beta}$ is quite higher in the case of entire history compared to the same ($\hat{\beta}_{\text{opt}}$) based on the optimal training period. It suggests that the estimation of β should be based on an optimal training period to capture the most recent trend rather than the overall trend using entire history.

Importance of the deceased compartment

Incorporating the deceased compartment into the modified SIS model is crucial because death due to disease may not be negligible. For example, in COVID-19, the number of deaths to the number of people infected is significant in many countries. Figure 7 shows the importance of the deceased compartment in the modified SIS model in terms of μ for Delhi. The prediction curve (purple line) with $\hat{\beta}_{\text{opt}} = 0.08$ and pre-fixed $\mu = 0$ (no deaths due to disease) shows a large disparity from the observed cumulative infected cases, and the difference between the two curves keeps increasing over time, with root mean squared error 12,990.46. Here, the prediction is based on the optimal training period denoted by blue line. The prediction curve (green line) with $\hat{\beta}_{\text{opt}} = 0.11$ and $\hat{\mu}_{\text{opt}} = 0.071$ is much closer to the observed cumulative infected cases’ curve with root mean squared error 5,187.47. Here, the prediction is based on the optimal training period denoted by red line. In both scenarios, the prediction phase and the assessment phase are of 40 and 15 days, respectively.

Figure 5: Modified SIS model predictions based on optimal training phases for Delhi at different periods.

Table 2: Summary of the prediction process for Delhi.

Scenarios of T_{Current}	$\hat{\beta}_{\text{opt}}$	$\hat{\mu}_{\text{opt}}$	T_{Current} (YYYY-MM-DD)	Assessment period (days)	Optimal training length (days)	Prediction length (days)	Root mean squared error
A	0.12	0.007	2020-05-29	15	36	30	1,639.27
B	0.12	0.086	2020-07-24	15	31	30	2,678.58
C	0.11	0.091	2020-12-29	30	31	40	245.43
D	0.09	0.070	2021-01-15	30	34	40	976.25

Figure 6: Comparison of the modified SIS model predictions for Delhi based on the optimal training phase and the entire history period. Based on the optimal training period (fitted by the blue line), the green line shows a better prediction than the purple line's prediction using the entire history as the training period (fitted by the red line).

Discussion

This work has provided a modified SIS model that accounts for deaths due to disease and predicts cumulative infected cases based on an optimally chosen training phase. The estimation process described in this work is beneficial when the disease under study changes its spreading pattern over time. We have developed the modified SIS model considering COVID-19 as the disease under focus. However, the model and algorithms can be applied to predict the cumulative cases of other infectious diseases.

Even though one can predict for any period-length in the future using the developed model, we recommend restricting the prediction to the short-term only. Any prediction with more than 30 days may not be reliable due to continuous changes in the COVID-19 virus' characteristics and human behavior (e.g., how social distancing norms followed from time to time). For example, in Delhi, considering the current day as 24 July

Figure 7: Shows the importance of the deceased compartment (μ) in the modified SIS model for better prediction performance. The red line represents the optimal training period with optimal estimates of parameters β and μ . The blue line represents the optimal training period with an optimal estimate of parameter β but $\mu = 0$, fixed.

20 with 30 days assessment period, the root mean square errors (RMSE) were 2,166.65, 6,046.05, 16,305.69, and 32,412.88, corresponding to the number of prediction days 30, 40, 50, and 60 from the current day, respectively. Similarly, with the same setup, for the current day, 29 December 2020, the root mean square errors (RMSE) were 159.16, 245.43, 468.84, and 896.93, respectively. Therefore, an increase of 30 days of prediction phase from 30 to 60 days can increase the RMSE substantially (almost 15 times and 6 times in two examples, respectively).

Note that the dynamic data estimation of the parameters, using Algorithm 1 and 2, is approximated using only one variable. Therefore, computational errors may occur when there are more sources of variation in active cases are present in reality. In this work, we have taken a fixed value of $\gamma = 1/14$. However, some studies also reported other values of γ (Arifin et al. 2020). For the current day, 24 July 2020, we have applied Algorithm 1 and 2 considering different values of γ as $\frac{1}{5}, \frac{1}{7}, \frac{1}{10}, \frac{1}{12}, \frac{1}{14}, \frac{1}{15}, \frac{1}{20}$, with fixed 30 days assessment period and a prediction length of 30 days. The RMSE values are 14,157.61, 4,134.77, 4,161.06, 8,208.95, 2,678.59, 4,119.57, 4,167.57, respectively for different γ values. In this scenario, the lowest RMSE value is observed when $\gamma = \frac{1}{14}$. However, in general, this might not be the case.

The developed open-access R-package (<https://github.com/abh2k/sisd>) can be helpful to implement the modified SIS model without dealing with mathematical details of the model. One only needs to prepare the input data set as described in the R-package documentation.

The objective of infectious disease prediction is to give the respective Governments an idea of what can happen in the near future (say 30 days) so that they can act promptly to avoid more difficult situations. Depending on the Government approach and the participation of the common people in the next 30 days, the accuracy of the predicted numbers may vary. For example, suppose we predict an increase of 100,000 cases in the next 30 days, and the Government imposes a complete lockdown from tomorrow. In that case, no model can be able to predict accurately based on the history data.

Acknowledgments: Not Applicable

Research funding: None declared.

Author contribution: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: Authors state no conflict of interest.

Informed consent: Not Applicable

Ethical approval: Not Applicable

References

- Agarwal, P., and K. Jhahharia. 2021. "Data Analysis and Modeling of COVID-19." *Journal of Statistics & Management Systems* 24 (1): 1–16.
- Al-Raei, M., M. S. El-Daher, and O. Solieva. 2021. "Applying SEIR Model without Vaccination for COVID-19 in Case of the United States, Russia, the United Kingdom, Brazil, France, and India." *Epidemiologic Methods* 10: s1.
- Arifin, W. N., W. H. Chan, S. Amaran, and K. I. Musa. 2020. "A Susceptible-Infected-Removed (SIR) Model of COVID-19 Epidemic Trend in Malaysia under Movement Control Order (MCO) Using a Data Fitting Approach." *MedRxiv*.
- Bjørnstad, O. N. 2019. "Population Dynamics of Pathogens." In *Handbook of Infectious Disease Data Analysis*, Vol. 13. Chapman and Hall/CRC.
- Brouqui, P., P. Colson, C. Melenotte, L. Houhamdi, M. Bedotto, C. Devaux, P. Gautret, M. Million, P. Parola, D. Stoupan, B. La Scola, J.-C. Lagier, and D. Raoult. 2021. "COVID-19 Re-infection." *European Journal of Clinical Investigation* 51 (5): e13537.
- Edrige, A. W. D., J. Kaczorowska, A. C. R. Hoste, M. Bakker, M. Klein, K. Loens, M. F. Jebbink, A. Matser, C. M. Kinsella, P. Rueda, M. Ieven, H. Goossens, M. Prins, P. Sastre, M. Deijs, and L. van der Hoek. 2020. "Seasonal Coronavirus Protective Immunity is Short-Lasting." *Nature Medicine* 26: 1691–3.
- Ferguson, N., D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, A. Dighe, I. Dorigatti, H. Fu, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, L. C. Okell, S. Elstrand, H. Thompson, R. Verity, E. Volz, H. Wang, Y. Wang, P. G. T. Walker, C. Walters, P. Winskill, C. Whittaker, C. A. Donnelly, S. Riley, and A. C. Ghani. 2020. *Report 9: Impact of Non-pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand*, Vol. 10, 77482. Imperial College London.
- Gagliardi, I., G. Patella, A. Michael, R. Serra, M. Provenzano, and M. Andreucci. 2020. "COVID-19 and the Kidney: From Epidemiology to Clinical Practice." *Journal of Clinical Medicine* 9 (8): 2506.
- Ghosh, P., R. Ghosh, and B. Chakraborty. 2020. "COVID-19 in India: Statewise Analysis and Prediction." *JMIR Public Health Surveill* 6 (3): e20341.
- Gounane, S., Y. Barkouch, A. Atlas, M. Bendahmane, F. Karami, and D. Mesquine. 2021. "An Adaptive Social Distancing SIR Model for COVID-19 Disease Spreading and Forecasting." *Epidemiologic Methods* 10: s1.
- Gumel, A. B., E. A. Iboi, C. N. Ngonghala, and E. H. Elbasha. 2021. "A Primer on Using Mathematics to Understand COVID-19 Dynamics: Modeling, Analysis and Simulations." *Infectious Disease Modelling* 6: 148–68.
- Hamer, W. 1928. *Epidemiology Old and New*. Kegan Paul, Trench, Trubner & Co., Ltd.
- Hethcote, H. W. 1989. "Three Basic Epidemiological Models." In *Applied Mathematical Ecology*, 119–44. Berlin, Heidelberg: Springer.
- Hethcote, H. W. 2000. "The Mathematics of Infectious Diseases." *SIAM Review* 42 (4): 599–653.
- Iwasaki, A. 2021. "What Reinfections Mean for COVID-19." *The Lancet Infectious Diseases* 21 (1): 3–5.
- Keeling, M. J., and P. Rohani. 2008. *Introduction to Simple Epidemic Models*, 15–53. Princeton University Press.
- Leonardi, M., A. W. Horne, K. Vincent, J. Sinclair, K. A. Sherman, D. Ciccio, G. Condous, N. P. Johnson, and M. Armour. 2020. "Self-management Strategies to Consider to Combat Endometriosis Symptoms during the COVID-19 Pandemic." *Human Reproduction Open* 2020 (2): hoaa028.
- Liu, T., S. Wu, H. Tao, G. Zeng, F. Zhou, F. Guo, and X. Wang. 2020. "Prevalence of IgG Antibodies to Sars-Cov-2 in Wuhan-Implications for the Ability to Produce Long-Lasting Protective Antibodies against Sars-Cov-2." *MedRxiv*.
- Magnoni, M. 2021. "The First Diffusion of the COVID-19 Outbreak in Northern Italy: an Analysis Based on a Simplified Version of the SIR Model." *Epidemiologic Methods* 10: s1.
- WHO. 2020. "Immunity Passports" in the Context of COVID-19. Also available at <https://www.who.int/news-room/commentaries/detail/immunity-passports-in-the-context-of-covid-19>.
- NPR. 2020. In South Korea, a Growing Number of COVID-19 Patients Test Positive after Recovery. Also available at <https://www.npr.org/sections/coronavirus-live-updates/2020/04/17/836747242/in-south-korea-a-growing-number-of-covid-19-patients-test-positive-after-recover>.
- Ramezani, S. B., A. Amirlatifi, and S. Rahimi. 2021. "A Novel Compartmental Model to Capture the Nonlinear Trend of COVID-19." *Computers in Biology and Medicine* 134: 104421.

- Ray, D., M. Salvatore, R. Bhattacharyya, L. Wang, J. Du, S. Mohammed, S. Purkayastha, A. Halder, A. Rix, D. Barker, M. Kleinsasser, Y. Zhou, D. Bose, P. Song, M. Banerjee, V. Baladandayuthapani, P. Ghosh, and B. Mukherjee. 2020. "Predictions, Role of Interventions and Effects of a Historic National Lockdown in India's Response to the COVID-19 Pandemic: Data Science Call to Arms." *Harvard Data Science Review* 6: 1–45, <https://hdsr.mitpress.mit.edu/pub/r1qq01kw>.
- Ross, R. 1911. *The Prevention of Malaria*. John Murray.
- Tillett, R. L., J. R. Sevinsky, P. D. Hartley, H. Kerwin, N. Crawford, A. Gorzalski, C. Laverdure, S. C. Verma, C. C. Rossetto, D. Jackson, M. J. Farrell, S. Van Hooser, and M. Pandori. 2021. "Genomic Evidence for Reinfection with Sars-Cov-2: A Case Study." *The Lancet Infectious Diseases* 21 (1): 52–8.
- Wangping, J., H. Ke, S. Yang, C. Wenzhe, W. Shengshu, Y. Shanshan, W. Jianwei, K. Fuyin, T. Penggang, Li. Jing, L. Miao, and H. Yao. 2020. "Extended SIR Prediction of the Epidemics Trend of COVID-19 in Italy and Compared with Hunan, China." *Frontiers of Medicine* 7: 169.