Annals of Data Science (2024) 11(3):959–978
https://doi.org/10.1007/s40745-023-00467-3

Jump-Drop Adjusted Prediction of Cumulative Infected Cases Using the Modified SIS Model

Rashi Mohta[1] · Sravya Prathapani[1] · Palash Ghosh[1,2,3]

Received: 16 June 2022 / Revised: 18 January 2023 / Accepted: 21 March 2023 /
Published online: 15 May 2023

Abstract

Accurate prediction of cumulative COVID-19 infected cases is essential for effectively managing the limited healthcare resources in India. Historically, epidemiological models have helped in controlling such epidemics. Models require accurate historical data to predict future outcomes. In our data, there were days exhibiting erratic, apparently anomalous jumps and drops in the number of daily reported COVID-19 infected cases that did not conform with the overall trend. Including those observations in the training data would most likely worsen model predictive accuracy. However, with existing epidemiological models it is not straightforward to determine, for a specific day, whether or not an outcome should be considered anomalous. In this work, we propose an algorithm to automatically identify anomalous 'jump' and 'drop' days, and then based upon the overall trend, the number of daily infected cases for those days is adjusted and the training data is amended using the adjusted observations. We applied the algorithm in conjunction with a recently proposed, modified Susceptible-Infected-Susceptible (SIS) model to demonstrate that prediction accuracy is improved after adjusting training data counts for apparent erratic anomalous jumps and drops.

Rashi Mohta and Sravya Prathapani have contributed equally to this work.

✉ Palash Ghosh
palash.ghosh@iitg.ac.in

Rashi Mohta
rashi@alumni.iitg.ac.in

Sravya Prathapani
prathapa@alumni.iitg.ac.in

[1] Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati, Assam, India

[2] Jyoti and Bhupat Mehta School of Health Sciences and Technology, Indian Institute of Technology Guwahati, Guwahati, Assam, India

[3] Centre for Quantitative Medicine, Duke-NUS Medical School, National University of Singapore, Singapore, Singapore

Keywords Adjustment · COVID-19 · Epidemiological models · Jumps and drops · Modified SIS model

1 Introduction

Researchers have used different epidemiological models to understand the dynamics of the Coronavirus disease 2019 (COVID-19) since it was first reported in the city of Wuhan, China [1]. These models focus on prediction [2], estimation of the basic reproduction number (R_0) [3], trend detection [4], reinfection [5] and the effect of preventive measures such as lockdown [6], social distancing [7], etc. All of these objectives require high-quality training data to develop meaningful and effective models. During the last two and half years we have observed several waves of COVID-19 with different patterns, at distinct points of time in different parts of the world. Due to many variants of the COVID-19 virus, it may not be wise to consider the entire data history to estimate model parameters. However, relatively few attempts have been made to assess the appropriateness of the available training data for model building. Anand et al. [8] proposed an algorithm using a modified susceptible-infected-susceptible (SIS) model to determine the optimal training period length for predicting the number of cumulative infected cases.

The daily number of reported COVID-19 infected cases may exhibit high fluctuations due to reduced testing during festivals (drops) or sudden movement of a group (student/workers) of people (jumps), etc. These anomalous, or *false*, jumps and drops are not aligned with the actual trend of the reported infected cases. To the contrary, we observe genuine declines (drops) and increases (jumps) in the reported infected cases due to the actual trend in COVID-19 spread. This phenomenon is quite common in a large country like India, with highly dense populations in cities and diversity in terms of climate and behavior of people. It is essential to be able to distinguish between actual trends in the number of reported infected cases and the false drops and jumps. A few false drops and jumps in the training data may result in poor estimates of model parameters with subsequent dismal prediction performance.

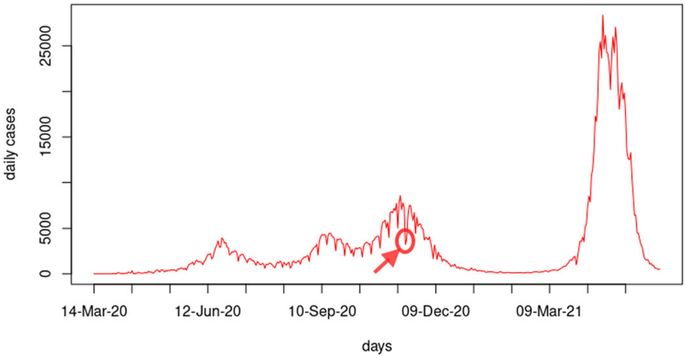

Fig. 1 Daily infected COVID-19 cases from 14 March 2020 to 4 June 2021 in Delhi highlighting the drop on 14 November 2020

Figure 1 shows daily COVID-19 infected cases from 14 March 2020 to 4 June 2021 in Delhi, the capital city of India. We observe many jumps and drops in the number of infected cases from the overall trend of the curve. There is a visibly sharp drop from 7750 to 3330 infected cases on 14 November 2020. On this day (in 2020), Diwali, a major Indian festival, was celebrated. A possible explanation for the drop would be that people with minor symptoms did not get tested on Diwali, or some testing centers were closed or operated for only a few hours. Therefore, this drop can be considered a false drop that is not aligned with the actual trend of the daily infected curve. Such events are usually one-time occurrences and seldom have any impact after they are over. However, if we allow these false drops (or jumps) into the training data, they have the potential to pull the fitted curve towards them, biasing the actual overall trend. Therefore, it is essential to adjust for these drops and jumps and mitigate their impact on model development. In other words, these drops and jumps are similar to the notion of outliers or influential points, which may alter the predictive capability of a model. However, it is not straightforward to use existing epidemiological models to decide a specific day as a sudden drop or jump in the training data.

In this era of data science and machine learning, we have numerous statistical approaches and optimization algorithms that we can use to model an epidemic such as COVID-19 [9–11]. In COVID-19, the major challenge is real-time decision making and quick information sharing with various stakeholders [12]. In this work, depending on the evidence-based data science approach, we propose an algorithm to automatically determine a drop or jump day in the training data. Based on the overall trend in the training data, we then adjust the number of daily infected cases for those days. The effective training data becomes the adjusted daily COVID-19 infected cases. To show that adjusted training data provides improved prediction accuracy compared to ignoring jumps and drops in the training data, we have applied the proposed algorithm in a recently proposed modified Susceptible-Infected-Susceptible (SIS) model [8]. However, this algorithm is useful for any model in cleaning the training data for better prediction accuracy. Note, however, that the algorithm will only use the adjusted training data if there is potential to improve prediction accuracy. Otherwise, it will use the actual training data-with no adjustments-to build the model.

2 Methodology

2.1 Detection Methods

Three detection methods-referred to as C1, C2 and C3-are used in the early aberration reporting system (EARS) of the syndromic surveillance system of the Centers for Disease Control and Prevention (CDC), USA to detect deviations in current data compared to the historical mean [13, 14]. These methods are also used in the CDC's BioSense system to combat 'health situational awareness' and 'event recognition and response' [14], and are also referred to as the cumulative sum (CUSUM) methods [15, 16]. In this work, we use these three methods to determine false drop or jump days in the training data. Once confirmed, the infected COVID-19 cases for each identified drop or jump day will be replaced by an adjusted value. The training data for any model will then be based on this adjusted data.

Let $Y(t)$ denotes the number of confirmed infected cases on the t^{th} day, then the C1 statistic is defined as

$$C_1(t) = \frac{Y(t) - \overline{Y_1}(t)}{S_1(t)}, \tag{1}$$

where $\overline{Y_1}(t)$ is the moving average and $S_1(t)$ is the standard deviation of the number of infected cases over the past seven days. They are calculated as

$$\overline{Y_1}(t) = \frac{1}{7}\sum_{i=1}^{7} Y(t-i), \text{ and } {S_1}^2(t) = \frac{1}{6}\sum_{i=1}^{7}(Y(t-i) - \overline{Y_1}(t-i))^2.$$

The sample average and the sample standard deviation used in C1 are based on the 7 days prior to the current observation. Therefore, C1 indicates the normalized value of the present day infected cases ($Y(t)$) considering the last 7 days' average of infected cases ($\overline{Y_1}(t)$) and the corresponding variation ($S_1^2(t)$). In the EARS system, $C_1(t) > 3$ indicates a signal at time t. In other words, a signal means the $Y(t)$ exceeds the three sample standard deviations above the sample mean. Note that, the EARS only concerns about the higher reporting cases (jumps). However, in this work, we are interested in both drop and jump. Therefore, in this context, $C_1(t) > 3$ denotes a possible jump and $C_1(t) < -3$ refers to a possible drop, where 'possible' implies that a drop (or jump) could be either a natural trend or a false drop (or false jump). We are only interested in detecting false drops or false jumps.

The C2 statistic is similar to the C1, but it also includes a 1-day lag while calculating the moving average and standard deviation. That is

$$C_2(t) = \frac{Y(t) - \overline{Y_2}(t)}{S_2(t)}, \tag{2}$$

$$\text{where, } \overline{Y_2}(t) = \frac{1}{7}\sum_{i=2}^{8} Y(t-i), \text{ and } {S_2}^2(t) = \frac{1}{6}\sum_{i=2}^{8}(Y(t-i) - \overline{Y_2}(t-i))^2.$$

Note that [14] defined the C2 statistic with a 2-day lag. However, we observe better results (empirically) if a 1-day lag is considered instead of 2 days. Hence, we have considered a 1-day lag in the C2 statistic. Similar to the C1 statistic, if $\mid C_2(t) \mid > 3$, we will check subsequent days for a possible jump or drop.

The C3 statistic for the t^{th} day uses the C2 statistics for the t^{th} day and the previous two days, as

$$C_3(t) = \sum_{i=t}^{t-2} max[0, \mid C_2(i) \mid -1]. \tag{3}$$

We conclude there is a possible jump or drop when $C_3(t) > 2$. In EARS, the expression of $C_3(t)$ has $C_2(i)$ instead of its absolute value, as the objective is to check for a jump (an increase in infected reported cases), not a drop [14].

2.2 Adjustment

The detection methods described in the previous section are used to detect any jump or drop in the training data. This section distinguishes between the actual trend and false drop/jump in the reported infected cases. In identifying a jump or drop, we constrain that the length of a jump or drop be categorized as 'sharp'. The idea behind this is that if the trend of the jump or drop continues for an extended period (the number of days), it would be considered an actual trend. Hence, we do not require any adjustment in the existing training data to eliminate its impact. Based on empirical results, we have set the upper bound for the duration of a jump or drop at five days. In other words, if the extended period is more than five days, then we consider it an actual trend.

Once a detection method confirms a possible jump or drop for the i^{th} day, we check the same for the subsequent five days and identify the duration of the jump or drop as $j - i$, where j^{th} day is the first day after i^{th} day without a drop or jump. Then, we propose using the mean of the moving average ($\overline{Y}_1(i)$: as defined in the previous section) of the number of daily confirmed cases in the last seven days and the number of infected cases just after the jump or drop ($Y(j+1)$), to adjust a sharp jump or drop. If the number of daily confirmed cases is outside the specific cut-offs (with respect to C-statistics) from the i^{th} day to the j^{th} day, and j-i $\leq$ 5, then $Y(t)$ for t $\in$ [i, j] would be replaced by

$$Y_{adjusted}(t) = \frac{\overline{Y_1}(i) + Y(j+1)}{2}.$$

After the adjustment, we consider the root mean square error (RMSE) calculated over the assessment period and compare the values before and after adjustment. If the adjustment has resulted in a decreased RMSE value for the assessment period, then we predict using adjusted data–else we use the original prediction.

Algorithm 1 shows automatic detection and adjustment of the jumps and drops in the training data using the C1, C2 or C3 detection methods.

3 Jump-Drop Adjusted Prediction Using Modified SIS Model

In this section, we show the use of jump-drop adjusted training data for prediction using a recently proposed modified Susceptible-Infected-Susceptible (SIS) model [8]. However, the proposed jump-drop adjustment algorithm can be applied to training data for any epidemiological model to improve the prediction accuracy. The main motivation for considering an SIS model for COVID-19 prediction is because it accommodates the possibility of reinfection–an essential consideration inasmuch as multiple studies have shown reinfection to be quite common in COVID-19 [17–20].

Algorithm 1 Algorithm to identify and adjust jumps and drops:

```
$n$ = length (in days) of history data
upper-bound = 5
while $1 \leq i \leq n$ do
    if $| CMetric(i) | > threshold$ then
        t = i + 1
        while $| CMetric(t) | > threshold$ do
            t = t + 1
        end while
        duration = t - i
        if duration ≤ upper-bound then
            while 1 ≤ j ≤ duration do
                $$Y[i + j] = \frac{Y_1(i) + Y[i + duration + 1]}{2}$$
            end while
        end if
    end if
    i = i + 1
end while
```

The modified SIS model determines the cumulative number of infected cases (COVID-19) at a specific time point. Anand et al. [8] developed a dynamic data-driven algorithm that estimates the model parameters based on an optimally chosen training phase for the prediction of cumulative infected cases. The model also accounts for deaths due to disease. Their estimation process is useful when the disease under study (like COVID-19) alters its spreading pattern over time. The following equations describe the modified SIS model proposed in [8],

$$\frac{dS}{dt} = -\beta \frac{SI}{N} + \gamma I \tag{4}$$

$$\frac{dI}{dt} = +\beta \frac{SI}{N} - \gamma I - \mu I \tag{5}$$

$$\frac{dC}{dt} = \beta \frac{SI}{N} \tag{6}$$

$$\frac{dD}{dt} = \mu I \tag{7}$$

with,

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dD}{dt} = 0$$

Here t denotes the time for which the smallest unit is a day. S and I denote the susceptible and infected numbers of people in the population, respectively. C denotes the cumulative infected cases (reinfection is also a count) and D denotes the total deaths caused by the disease. We assume that the death rate from any other cause is the same as the birth rate. The total population size is N. The parameter β denotes the average number of individuals infected per day from an infected person [21]. In other

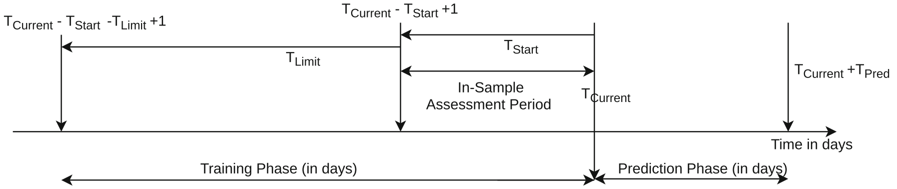
T_{Current} - T_{Start} -T_{Limit} +1
T_{Current} - T_{Start} +1
T_{Limit}
T_{Start}
In-Sample
Assessment Period
T_{Current}
T_{Current} +T_{Pred}
Time in days
Training Phase (in days)
Prediction Phase (in days)

Fig. 2 Partition of the study period into different training and prediction time points [8]

words, it is the contact rates between infected and susceptible [22]. Here, γ stands for the recovery rate.

The equation (4) represents the rate of change in the population of the susceptible compartment. Here, $\beta\frac{SI}{N}$ is the number of people infected daily and subsequently removed from the S compartment. The γI is the number of infected individuals recovering daily from the illness and subsequently added back into the S compartment. The recovery rate, γ is assumed to be $1/T$, where T is the average duration of infection in an infected person [2, 23]. Equations (5) and (6) represent the population rate of change in the infected compartment and the cumulative number of cases, respectively. μ is the infection mortality rate and μI is the daily number of fatal infected cases, as shown in the equation (7).

As argued by [8], SIS model parameters are generally assumed constant over the entire duration of the study period. However, since the contact rate can change substantially over time in the case of COVID-19, it would be more reasonable to make predictions based on a shorter training phase. Hence, the training phase has been dynamically chosen based on the historical data. As illustrated in Fig. 2, the study period is divided into two phases: training and prediction. The current date is denoted by $T_{Current}$. The minimum length of the training period is T_{Start} and the corresponding training interval is $[T_{Current} - T_{Start} + 1, T_{Current}]$. In selecting the optimal training phase, at most T_{Limit} number of days (set by the user) can be added to the minimum training period. Therefore, the training period is $[T_{Current} - T_{Start} + 1 - t, T_{Current}]$, where $0 \leq t \leq T_{Limit}$. T_{Pred} denotes the length of the prediction phase.

Hence, the objective is to optimally choose a training phase that allows accurate prediction of the near future, using proper model parameters. To compare different models, we use the minimum training phase as the fixed in-sample assessment period. A user can decide on the assessment period. Generally, it can range from 7 to 30 days for a prediction period of 30 to 50 days. The root mean square error (RMSE) has been used as a criterion to choose the optimal training period,

$$L(t, \beta, \mu) = \sqrt{\frac{1}{T_{Start}} \sum_{i} (C_p[i; t, \beta, \mu] - C_0[i])^2} \tag{8}$$

where $T_{Current}$ - $T_{Start} + 1 \leq i \leq T_{Current}$. The observed cumulative infected cases on the i^{th} day is given by $C_o[i]$ and $C_p[i; t, \beta, \mu]$ denotes the predicted cumulative infected cases on the i^{th} day from the modified SIS model. t_{opt} is the value of t for which (8) is minimized. Similarly, $\hat{\beta}_{opt}$ and $\hat{\mu}_{opt}$ are the values of $\hat{\beta}_{t_{opt}}$ and $\hat{\mu}_{t_{opt}}$, respectively. The assessment of model prediction accuracy is accomplished using (8) but replacing T_{Start}, t, β and μ, by T_{Pred}, t_{opt}, $\hat{\beta}_{t_{opt}}$ and $\hat{\mu}_{t_{opt}}$, respectively, where $0 \leq i \leq T_{Pred}$. The recovery period from COVID-19 infection in India is 14 days on average [3]. γ can thus be calculated as 1/(duration for recovery), that is, 1/14.

4 Prediction of Cumulative Infected Cases

4.1 R-Package

For easy use of the jump-drop adjusted prediction model, an R package has been developed and is available at https://github.com/RashiMohta/COVID-19-cases-prediction. Given the daily infected COVID-19 data (say for a state) and other input parameter values, the R-package would predict the cumulative number of cases for both the original data and the adjusted data using the given bound metric and return the prediction with the smaller root mean squared error over the in-sample assessment period.

4.2 Predicting the Cumulative Number of COVID-19 Cases

To demonstrate the performance of the proposed jump-drop adjustment algorithm, we used the R-package to predict cumulative infected cases in the capital of India, Delhi. The data containing the number of daily confirmed cases of COVID-19 is publicly available at https://www.covid19india.org/. In Table 1, we consider six different $T_{current}$ as 12 May 2020 (in (A)), 12 June 2020 (in (B)), 12 July 2020 (in (C)), 14 September 2020 (in (D)), 12 April 2021 (in (E)) and 27 April 2021 (in (F)), to observe the impact of adjusting for jumps and drops using the proposed algorithm. We set the upper limit (T_{Limit}) of the number of additional days that can be added to the minimum training period at 30. The prediction was made using the original and the jump-drop adjusted data for the period $T_{Current} + 1$ to $T_{Current+T_{Pred}}$ where T_{Pred} was taken as 20 days. Table 1 shows the optimal values of μ and β calculated for different prediction periods and the RMSE values corresponding to each assessment and the prediction period. Note that in all the scenarios, we used the modified SIS model described in Sect. 3.

In practice, in predicting the future, there are no observed infected cases for the prediction period. Therefore, as discussed earlier, the RSME of the in-sample assessment period is used to choose between the adjusted and unadjusted prediction and the corresponding adjustment method (C1, C2 or C3). From Table 1, we observe that in five of six scenarios, the adjustment with C3 provided the best results. An increasing trend in jump-drop adjusted prediction in terms of RMSE values over time (calendar days) is also noted. This is due to the prediction of cumulative infected cases, which always increase over time. For example, in the first scenario (A), the cumulative infected cases in the training period were less than 10,000. Thus, the corresponding RMSE values (for C3, RMSE-assessment: 248.02 and RMSE-prediction: 1171.24) are expected to be less. Parameters $\hat{\beta}_{opt}$, $\hat{\mu}_{opt}$ and the optimal training period also change for the different scenarios justifying the use of the dynamic data-driven modified SIS model.

Table 1 Using the data from Delhi, the estimated optimal parameters of the modified SIS model, RMSE of assessment and prediction periods with original training data and with adjusted training data using C1, C2, C3 statistics

Graphs No. in Fig. 3	$T_{Current}$	Variable parameter	Original	With adjustment		
				C1	C2	C3
(A)	12 May 2020	$\hat{\mu}_{t_{opt}}$	0.063	0.053	0.048	0.06
		$\hat{\beta}_{t_{opt}}$	0.19	0.19	0.19	0.18
		Opt. training period	41	40	41	35
		No. of Outliers Adjusted		9	7	9
		RMSE (Assessment)	296.70	316.10	311.60	**248.02**
		RMSE (Prediction)	10984.22	8870.75	7395.23	**1171.24**
(B)	12 Jun 2020	$\hat{\mu}_{t_{opt}}$	0.002	0.018	0.001	0.008
		$\hat{\beta}_{t_{opt}}$	0.12	0.13	0.12	0.13
		Opt. training period	20	20	20	24
		No. of Outliers Adjusted		5	5	9
		RMSE (Assessment)	1417.65	1374.42	1291.85	**1148.56**
		RMSE (Prediction)	12654.64	11648.56	8285.07	**5105.93**
(C)	12 Jul 2020	$\hat{\mu}_{t_{opt}}$	0.097	0.095	0.099	0.098
		$\hat{\beta}_{t_{opt}}$	0.14	0.14	0.14	0.15
		Opt. training period	27	27	26	24
		No. of Outliers Adjusted		2	3	8
		RMSE (Assessment)	2649.52	2430.44	**2193.65**	2381.78
		RMSE (Prediction)	11092.92	10206.85	**7865.84**	12247.35

Table 1 continued

Graphs No. in Fig. 3	$T_{Current}$	Variable parameter	Original	With adjustment		
				C1	C2	C3
(D)	14 Sep 2020	$\hat{\mu}_{topt}$	0.013	0.013	0.013	0.001
		$\hat{\beta}_{topt}$	0.14	0.14	0.14	0.13
		Opt. training period	20	20	20	22
		No. of Outliers Adjusted		0	0	1
		RMSE (Assessment)	3350.79	3350.79	3350.79	**2886.15**
		RMSE (Prediction)	31161.71	31161.71	31161.71	**25972.73**
(E)	12 Apr 2021	$\hat{\mu}_{topt}$	0.008	0.006	0.007	0.069
		$\hat{\beta}_{topt}$	0.22	0.22	0.22	0.28
		Opt. training period	20	20	20	20
		No of Outliers Adjusted (Training)		3	3	18
		RMSE (Assessment)	6140.81	6162.93	6367.32	**5374.55**
		RMSE (Prediction)	96356.66	101354.75	92301.01	**73298.98**
(F)	27 Apr 2021	$\hat{\mu}_{topt}$	0.089	0.089	0.089	0.069
		$\hat{\beta}_{topt}$	0.26	0.26	0.26	0.24
		Opt. training period	49	49	49	49
		No. of Outliers Adjusted		0	1	17
		RMSE (Assessment)	25354.74	25354.74	26088.72	**23272.09**
		RMSE (Prediction)	443016.61	443016.61	436018.05	**162015.09**

In all the six scenarios, the assessment and prediction periods are taken as 20 days. Note that the number of outliers (jumps/drops) adjusted considers the maximum training period, as based on that period, the optimal training period is chosen

The lowest RMSE value (in a row) is denoted in bold

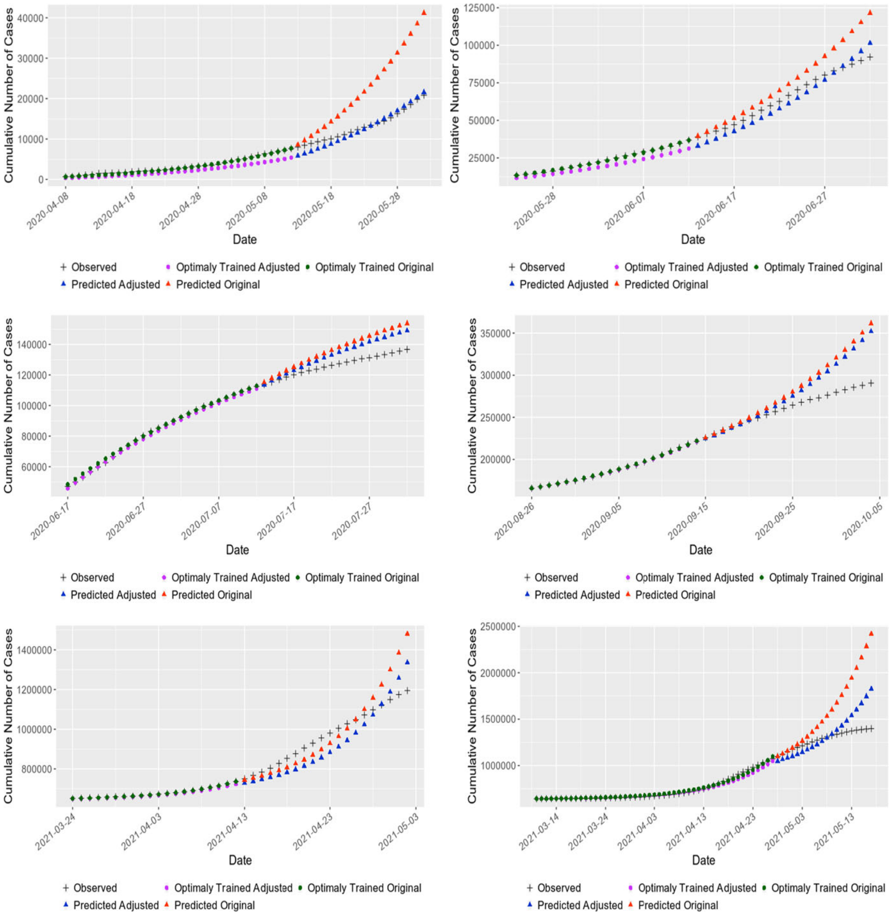

Fig. 3 Predictions made with and without adjustments for Delhi at different periods using best performing metric. The graphs show the observed (unadjusted), and adjusted training periods along with predictions based on original (unadjusted) and adjusted training data

Figure 3 provides a graphical comparison of the scenarios presented in Table 1. In Fig. 3, all the cases, the observed curve and the adjusted prediction curve are close to each other. Table 2 (in Appendix) shows the prediction of Maharashtra and Madhya Pradesh, two other Indian states.

5 Discussion

This work establishes the importance of adjusting for erratic, anomalous jumps and drops observed in infected COVID-19 cases during the training phase, in anticipation of fitting a predictive model. The proposed algorithm used the C1, C2, and C3 methods to automatically detect these jumps and drops and adjust the corresponding values. Model predictions based upon the adjusted training data were then obtained using all three methods and compared to results using the original unadjusted training data. An assessment of training phase RMSE values supported our proposition that a decrease in RMSE will result in improved predictive accuracy.

Upon detection of a possible anomalous jump or drop, we propose a single method in Sect. 2.2 to adjust that jump or drop. The method replaces the anomalous jump or drop with the mean number of daily confirmed cases during the prior seven days and the number of infected cases just after the jump or drop. However, one can use other methods to replace a detected jump or drop, e.g., linear interpolation, end points mean, or a percentile. The linear interpolation method uses infected cases just before and after the jump or drop and adjusts by joining a straight line between those points. The end points mean method uses infected cases just before and after jump or drop and adjusts outliers with the mean of those two points. The percentile method uses data points involved in metrics and outliers and finds the 10^{th} and 90^{th} percentile values of those data points; the anomalous jump or drop with values less than the 10^{th} percentile or greater than the 90^{th} percentile are then replaced with the corresponding percentile values. Based on empirical studies presented in Table 3, the proposed adjustment method performs well in most scenarios.

Note that the C2 statistic is similar to C1, but it includes a 1-day lag while calculating the moving average and standard deviation. However, C1 uses no lag in computing the moving average. Consider a situation where we observed similar trends, say high value (may not be a jump) on $(t-1)^{th}$ day as on the t^{th} day (may be a jump) and normal values from $(t-2)$ and earlier days. In this situation, the moving average $\overline{Y_1}(t)$ used in C1 may be influenced by the high value on $(t-1)^{th}$ day. However, C2 will not be affected by the high value on $(t-1)^{th}$ day as it uses a 1-day lag that discards the value of $(t-1)^{th}$ day. Here, C2 can detect a possible jump better than C1. On the other hand, the C3 statistic for the t^{th} day uses the C2 statistics for the t^{th} day and the previous two days. Thus, if the run of jump or drop days is longer, then C3 is more effective than C2 in detecting a possible anomalous value. We observed that the C3 method performed well in most of the Table 1 scenarios.

For a specific date (current day), the proposed algorithm will check whether or not an adjustment is required based on the ‘in sample assessment period.’ If required, the prediction will be made using adjusted training data. Otherwise, the algorithm will use original training data with no adjustment. In other words, the proposed algorithm is sufficiently adaptive to automatically choose the appropriate training data–with or without adjustment. However, it is possible that the assessment period RMSE is minimized using one detection method (say C3), but the prediction period RMSE–assuming the observed data are available–is minimized using a different method (say C1). In fact, we observe that phenomenon in the final scenario of Table 2 (Appendix), which indicates that the infection trend during the prediction period changed substantially from that of the training period.

In Table 4, we consider data from Delhi with $T_{Current}$ as 6 May 2020. The objective is to assess the performance of the proposed method when artificial jumps are added to the original data. In the first panel, with no artificially added jumps, the C3 method gives the minimum root mean square error (RMSE) of 2321.54 for the prediction period of 20 days. Here, the number of adjusted outliers (anomalous jump and drop days) are 9, 10, and 20 for the C1, C2, and C3 methods, respectively. In the next panel (separated by thick lines), in addition to the existing jumps and drops, we have artificially added two jump days on 18 and 19 April 2020 by replacing the respective original values of 186 and 110 with artificial values of 350 and 380. Again, the C3 method gives the best prediction RMSE of 1232.09. In this case, the number of outliers adjusted is 11, 12, and 20 for the C1, C2, and C3 methods, respectively. Note that, the addition of artificial jumps may result in an increase, decrease, or no change in the number of outliers adjusted by the respective methods C1, C2, and C3. This is because, each time we alter the original data by adding artificial values, analysis results for the training data sets and the adjusted data sets may change with respect to a specific method. Addition of artificial jumps can have an impact on the normalized values (based on the moving averages) and hence an earlier adjusted jump may not be required to be adjusted once original values are changed. In subsequent panels, we added a few more jumps in consecutive panels. Finally, in the last panel, the C2 method gives the lowest prediction RMSE of 4617.16. Notice that, due to the artificial addition of jumps, the prediction RMSE values corresponding to the original method on unadjusted data increase from 5969.82 to 18075.04. However, in the same comparison, RMSE values improved from 6029.63 to 4617.16 for the C2 method. In summary, in all the panels, the C1, C2, and C3 methods performed better than the original method in terms of prediction RMSE. From this empirical study, we conclude that the proposed method works well even when the number of jumps of various magnitudes in the original data is increased.

Acknowledgements The authors thank Dr. John Allen for some helpful discussion.

Author Contributions The corresponding author designed and supervised the research. The first two authors have contributed in terms of coding and data analysis, interpretation, and development of the R-package. All authors jointly wrote and reviewed the whole manuscript.

Funding Rashi Mohta, Sravya Prathapani would like to acknowledge the Samsung Fellowship for this work. Palash Ghosh would like to acknowledge support by ICMR Centre for Excellence, Grant no. 5/3/8/20/2019-ITR.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Statements The authors follow all the relevant ethical rules.

Data Availability The data containing the number of daily confirmed cases of COVID-19 is publicly available at https://www.covid19india.org/

Code Availability: For easy use of the jump-drop adjusted prediction model, an R package has been developed and is available at https://github.com/RashiMohta/COVID-19-casesprediction.

Appendix A: Jump-drop Adjusted Prediction for Maharashtra and Madhya Pradesh

Table 2 Using the data from Maharashtra and Madhya Pradesh, the estimated optimal parameters of the modified SIS model, RMSE of assessment and prediction periods with original training data and with adjusted training data using C1, C2, C3 statistics.

State	$T_{Current}$	Variable parameter	Original	With adjustment		
				C1	C2	C3
Maharashtra	12 May 2020	$\hat{\mu}_{topt}$	0.096	0.036	0.095	0.033
		$\hat{\beta}_{topt}$	0.23	0.16	0.22	0.16
		Opt. training period	36	28	32	25
		No of Outliers Adjusted		6	5	11
		RMSE (Assessment)	1012.60	971.56	955.99	**945.98**
		RMSE (Prediction)	56252.22	20378.43	38767.01	**18147.96**
Maharashtra	21 Jun 2020	$\hat{\mu}_{topt}$	0.04	0.02	0.038	0.003
		$\hat{\beta}_{topt}$	0.13	0.11	0.13	0.09
		Opt. training period	48	42	48	34
		No of Outliers Adjusted		3	4	6
		RMSE (Assessment)	**3055.70**	3107.53	3171.17	3181.29
		RMSE (Prediction)	42851.86	21425.95	37001.71	**6794.92**
Maharashtra	19 Sep 2020	$\hat{\mu}_{topt}$	0.018	0.018	0.018	0.001
		$\hat{\beta}_{topt}$	0.10	0.10	0.10	0.09
		Opt. training period	20	20	20	21
		No of Outliers Adjusted		0	0	2
		RMSE (Assessment)	23770.72	23770.72	23770.72	**21489.21**
		RMSE (Prediction)	171609.02	171609.02	171609.02	**134484.32**
Maharashtra	14 Oct 2020	$\hat{\mu}_{topt}$	0.027	0.027	0.027	0.031

Table 2 continued

State	$T_{Current}$	Variable parameter	Original	With adjustment		
				C1	C2	C3
		$\hat{\beta}_{t_{opt}}$	0.07	0.07	0.07	0.07
		Opt. training period	25	25	25	21
		No of Outliers Adjusted		0	0	1
		RMSE (Assessment)	13876.04	13876.04	13876.04	**13741.02**
		RMSE (Prediction)	43017.37	43017.37	43017.37	**29206.76**
Maharashtra	17 Apr 2021	$\hat{\mu}_{t_{opt}}$	0.069	0.069	0.069	0.043
		$\hat{\beta}_{t_{opt}}$	0.18	0.18	0.18	0.15
		Opt. training period	33	33	33	25
		No of Outliers Adjusted		0	0	4
		RMSE (Assessment)	57767.96	57767.96	57767.96	**55974.33**
		RMSE (Prediction)	884749.37	884749.37	884749.37	**560469.33**
Madhya Pradesh	19 Oct 2020	$\hat{\mu}_{t_{opt}}$	0.089	0.055	0.055	0.056
		$\hat{\beta}_{t_{opt}}$	0.15	0.10	0.10	0.10
		Opt. training period	43	26	25	22
		No of Outliers Adjusted		1	1	5
		RMSE (Assessment)	2141.95	**1521.49**	1545.69	1597.72
		RMSE (Prediction)	11708.52	**5065.79**	5051.13	2525.85

Table 2 continued

State	$T_{Current}$	Variable parameter	Original	With adjustment		
				C1	C2	C3
Madhya Pradesh	15 Aug 2021	$\hat{\mu}_{t_{opt}}$	0.085	0.003	0.003	0.002
		$\hat{\beta}_{t_{opt}}$	0.13	0.05	0.05	0.05
		Opt. training period	20	48	48	48
		No of Outliers Adjusted		1	1	1
		RMSE (Assessment)	**16.20**	21.55	20.81	18.92
		RMSE (Prediction)	**29.74**	2066.62	2065.15	2049.2
Madhya Pradesh	12 May 2020	$\hat{\mu}_{t_{opt}}$	0.040	0.063	0.083	0.087
		$\hat{\beta}_{t_{opt}}$	0.12	0.15	0.17	0.20
		Opt. training period	28	28	28	40
		No of Outliers Adjusted		8	7	12
		RMSE (Assessment)	129.24	139.67	142.37	**88.62**
		RMSE (Prediction)	316.52	**307.81**	430.65	3126

In all seven scenarios, the assessment and prediction periods are taken as 20 days. Note that the number of outliers (jumps/drops) adjusted considers the maximum training period, as based on that period, the optimal training period is chosen

The lowest RMSE value (in a row) is denoted in bold

Appendix B: Jump-drop Adjusted Prediction for Maharashtra and Madhya Pradesh Using Different Outlier Adjustment Methods

Table 3 Using the data from Maharashtra and Madhya Pradesh, the estimated optimal parameters of the modified SIS model, RMSE of assessment and prediction periods with original training data and with adjusted training data using C3 statistics and different methods to adjust outliers (jumps/drops)

State	$T_{Current}$	Variable parameter	Original	Different methods to adjust outliers			
				Mean (Proposed method)	Linear interpolation	End points mean	Percentile
Maharashtra	12 May 2020	$\hat{\mu}_{t_{opt}}$	0.096	0.033	0.072	0.091	0.068
		$\hat{\beta}_{t_{opt}}$	0.23	0.16	0.2	0.22	0.2
		Opt. training period	36	25	34	35	32
		No. of Outliers Adjusted		11	9	9	12
		RMSE (Assessment)	1012.60	945.98	989.45	944.28	1034.46
		RMSE (Prediction)	56252.22	**18147.96**	36958.00	43835.26	40353.02
Maharashtra	19 Sep 2020	$\hat{\mu}_{t_{opt}}$	0.018	0.001	0.019	0.019	0.019
		$\hat{\beta}_{t_{opt}}$	0.10	0.09	0.10	0.10	0.10
		Opt. training period	20	21	20	20	20
		No. of Outliers Adjusted		2	2	2	3
		RMSE (Assessment)	23770.72	21489.21	23766.34	23766.34	23593.98
		RMSE (Prediction)	171609.02	**134484.32**	155377.43	155377.43	158473.83

Table 3 continued

State	$T_{Current}$	Variable parameter	Original	Different methods to adjust outliers			
				Mean (Proposed method)	Linear interpolation	End points mean	Percentile
Maharashtra	14 Oct 2020	$\hat{\mu}_{t_{opt}}$	0.027	0.031	0.032	0.032	0.033
		$\hat{\beta}_{t_{opt}}$	0.07	0.07	0.07	0.07	0.07
		Opt. training period	25	21	21	21	21
		No. of Outliers Adjusted		1	1	1	1
		RMSE (Assessment)	13876.04	13741.02	14084.72	14084.72	13790.12
		RMSE (Prediction)	43017.37	**29206.76**	31564.83	31564.83	32241.33
Madhya Pradesh	19 Oct 2020	$\hat{\mu}_{t_{opt}}$	0.089	0.056	0.054	0.065	0.083
		$\hat{\beta}_{t_{opt}}$	0.15	0.10	0.10	0.11	0.13
		Opt. training period	43	22	26	29	35
		No. of Outliers Adjusted		5	5	5	7
		RMSE (Assessment)	2141.95	1597.72	1501.49	1451.34	1612.37
		RMSE (Prediction)	11708.52	2525.85	4374.84	4193.26	**2140.34**

In all the four scenarios, the assessment and prediction periods are taken as 20 days

The lowest RMSE (Prediction) value (in a row) is denoted in bold

Appendix C: Performance of the Proposed Method as the Number of Jump-drop Increases

Table 4 Performance of jump-drop adjusted model with artificially added (cumulatively) jumps of different magnitudes.

Artificially added (cumulatively) jumps			Variable parameter	Original	With adjustment		
Dates	Origi- nal value	Artifi- cial value					
					C1	C2	C3
Data with current date 6 May 2020			$\hat{\mu}_{topt}$	0.022	0.022	0.066	0.004
	(No added jumps)		$\hat{\beta}_{topt}$	0.15	0.16	0.21	0.12
			Opt. training period	34	33	36	27
			No. of Outliers Adjusted		9	10	20
			RMSE (Assessment)	274.22	242.94	213.91	170.59
			RMSE (Prediction)	5969.82	4568.53	6029.63	**2321.54**
17 April 2020	67	67	$\hat{\mu}_{topt}$	0.096	0.083	0.094	0.006
18 April 2020	186	**350**	$\hat{\beta}_{topt}$	0.21	0.22	0.24	0.13
19 April 2020	110	**380**	Opt. training period	34	35	37	27
			No. of Outliers Adjusted		11	12	20
20 April 2020	78	78	RMSE (Assessment)	270.96	261.12	223.59	228.84
			RMSE (Prediction)	10173.82	7929.90	6955.23	**1232.09**
22 April 2020	92	92	$\hat{\mu}_{topt}$	0.085	0.06	0.097	0.062
23 April 2020	128	**400**	$\hat{\beta}_{topt}$	0.2	0.2	0.24	0.17
24 April 2020	138	**450**	Opt. training period	32	34	37	26
			No. of Outliers Adjusted		10	15	21
25 April 2020	111	111	RMSE (Assessment)	302.31	263.04	217.50	228.58
			RMSE (Prediction)	11518.21	8528.50	5116.52	**1924.22**
26 April 2020	293	**450**	$\hat{\mu}_{topt}$	0.09	0.091	0.097	0.072
27 April 2020	190	**480**	$\hat{\beta}_{topt}$	0.2	0.23	0.24	0.18
28 April 2020	206	206	Opt. training period	28	35	37	20
			No. of Outliers Adjusted		12	15	16
			RMSE (Assessment)	319.20	237.94	217.50	313.57
			RMSE (Prediction)	12297.00	10184.88	**4883.55**	6823.83
1 May 2020	223	223	$\hat{\mu}_{topt}$	0.068	0.091	0.097	0.071
2 May 2020	284	**700**	$\hat{\beta}_{topt}$	0.2	0.23	0.24	0.18
3 May 2020	427	**720**	Opt. training period	34	35	37	20
			No. of Outliers Adjusted		14	15	18
4 May 2020	349	349	RMSE (Assessment)	369.46	254.37	217.50	314.72
			RMSE (Prediction)	18075.04	9328.98	**4617.16**	6417.06

We use the data from Delhi with $T_{Current}$ = 6 May 2020, and both assessment and prediction periods are 20 days

The lowest RMSE (Prediction) value (in a row) is denoted in bold

References

1. Al-Raeei M, El-Daher MS, Solieva O (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)
2. Ghosh P, Ghosh R, Chakraborty B (2020) COVID-19 in India: statewise analysis and prediction. JMIR Public Health Surveill 6(3):e20,341. https://doi.org/10.2196/20341
3. Qaffou A, Zbair M, El Maroufy H (2022) Bayesian inference of the basic reproduction number for a sir epidemic model. Biostat Epidemiol 6(1):128–143
4. Ramezani SB, Amirlatifi A, Rahimi S (2021) A novel compartmental model to capture the nonlinear trend of COVID-19. Comput Biol Med 134(104):421
5. Iwasaki A (2021) What reinfections mean for COVID-19. Lancet Infect Dis 21(1):3–5
6. Ray D, Salvatore M, Bhattacharyya R et al (2020) Predictions, role of interventions and effects of a historic national lockdown in India's response to the the COVID-19 pandemic: data science call to arms. Harv Data Sci Rev. https://doi.org/10.1162/99608f92.60e08ed5
7. Gounane S, Barkouch Y, Atlas A, et al (2021) An adaptive social distancing SIR model for COVID-19 disease spreading and forecasting. Epidemiol Method 10(s1)
8. Anand A, Kumar S, Ghosh P (2021) Dynamic data-driven algorithm to predict cumulative covid-19 infected cases using susceptible-infected-susceptible model. Epidemiol Method 10(s1):20210,034. https://doi.org/10.1515/em-2021-0034
9. Shi Y (2022) Advances in big data analytics: theory, algorithms and practices. Springer Nature
10. Olson DL, Shi Y, Shi Y (2007) Introduction to business data mining, vol 10. McGraw-Hill, Irwin New York
11. Shi Y, Tian Y, Kou G et al (2011) Optimization based data mining: theory and applications. Springer, Science & Business Media
12. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. Ann Data Sci 4(2):149–178
13. Lawson B, Fitzhugh E, Hall S et al (2004) From implementation to automation-a step-by-step approach to developing syndromic surveillance systems from a public health perspective. MMWR Morb Mortal Wkly Rep Syndr Surveill 53:248
14. Fricker RD Jr, Hegler BL, Dunfee DA (2008) Comparing syndromic surveillance detection methods: ears'versus a cusum-based methodology. Stat Med 27(17):3407–3429
15. Zhu Y, Wang W, Atrubin D et al (2005) Initial evaluation of the early aberration reporting system. Morb Mortal Wkly Rep 54(Suppl):123–130
16. Hutwagner L, Thompson W, Seeman GM et al (2003) The bioterrorism preparedness and response early aberration reporting system (ears). J Urban Health 80(1):i89–i96
17. WHO (2020) "Immunity passports" in the context of COVID-19. https://www.who.int/news-room/commentaries/detail/immunity-passports-in-the-context-of-covid-19
18. NPR (2020) In south korea, a growing number of COVID-19 patients test positive after recovery. 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
19. Tillett RL, Sevinsky JR, Hartley PD, et al (2020) Genomic evidence for reinfection with sars-cov-2: a case study. The Lancet Infectious Diseases
20. Brouqui P, Colson P, Melenotte C, et al (2021) COVID-19 re-infection. Eur J Clin Investig 51(5)
21. Bjørnstad ON (2019) Population dynamics of pathogens. Handb Infect Dis Data Anal 13
22. Keeling MJ, Rohani P (2011) Introduction to simple epidemic models. Modeling infectious diseases in humans and animals. Princeton University Press, Princeton, pp 15–53
23. Ferguson N, Laydon D, Nedjati-Gilani G et al (2020) Report 9: impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imp Coll Lond 10(77):482

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.