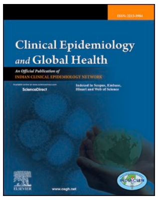
Clinical Epidemiology and Global Health 18 (2022) 101176

Contents lists available at ScienceDirect

Clinical Epidemiology and Global Health

journal homepage: www.elsevier.com/locate/cegh



Original article

Penalized splines model to estimate time-varying reproduction number for Covid-19 in India: A Bayesian semi-parametric approach

Ranjita Pandey, Himanshu Tolani *

Department of Statistics, University of Delhi, India

ARTICLE INFO

Keywords:
Penalized splines
Bayesian semiparametric regression
Reproduction number
COVID-19
Delay structures

ABSTRACT

Statistical modelling is pivotal in assessing intensity of a stochastic processes. Novel Corona virus disease demanded proactive measures to understand the severity of disease spread and to plan its control accordingly. We propose estimation of reproduction number as a crucial factor to monitor the random dynamics of Covid-19 in India. In the present paper, semi-parametric regression based on penalized splines embedded under Bayesian formulation is utilised to estimate reproduction number while incorporating effects of underreporting and delay in reporting for the actual number of daily occurrences. Monte Carlo Markov Chain approximations are utilised to perform simulation study and thereby to assess the impact of the reporting probability and misspecification of delay pattern on potential for further substance of the pandemic. For a cycle of reporting on weekly basis, the proposed penalized spline Bayesian framework fits closest to the empirical data drawn for a two-day delay in reporting with approximately half of the actual cases being reported. The present paper is a contribution towards estimation of the true daily reproduction number of Covid-19 incidences in its next generation cycle.

1. Introduction

It is difficult to capture the exact evolution and dynamics of any pandemic through deterministic modelling. Appearance of novel pathogen leading to sustained pandemic, therefore calls for accounting of temporal piecewise changes in short time-spans. Existing count of infectees and transmission time of infection are the two-key factors in assessment of growth rate of the infection in a specified population. Transmission and propagation of any novel communicable pathogen is quantified through statistical models which are validated through short-history data curated from the current time-chain of the pandemic. Average infection transmissions generated from a currently active single infected unit in a given specified time t is called reproduction number (R_t). In simple words, R_t represents number of secondary infections that one infected individual can spread further on time t. $R_t \leq 1$ indicates that the epidemic is under control and approaching a disease free state while $R_t > 1$ indicates that the control measures lack efficacy leading to the possibility of endemic transforming into epidemic. During progression of any epidemic the expected rise or decline in the reproduction number explains the time-based renewals. These estimates are therefore vital in future medical-assistance and planning in terms of manpower, medical machinery, medicines and health care infrastructure.

Spline is a flexible mathematical function that represents changing data through disaggregated polynomials. Splines have been an important key for addressing various mathematical problems in approximation theory and in numerical analysis.[1] Incorporating splines in modelling of R_t as an alternative to other popular mathematical models like SIR[2] and SIERD[3] portray challenges[4] in modelling of Covid-19 data and create essential need to appropriately trace the sample-path of its evolution. Rapid changes in trends on daily incidence of Covid- 19 is modelled using semiparametric regression in conjunction with versatile penalized splines to estimate R_t.[5]

Renewal equation-based estimates of R_t have been explored under closed population assumption which mirrors epidemic persistence under lockdown.[6] This approach is motivated by the fact that R_t is viewed as an autoregressive entity determined completely by its present value and the transmission period only. This idiom is also studied through likelihood approach.[7] R_t is also investigated under Bayesian lens for efficacy of control measures of communicable diseases in real time.[8] Under-reporting of R_t is undertaken for H1N1 pandemic using Bayesian data augmentation and renewal process.[9] Estimation of R_t based on various delay scenarios and misreported H1N1data is studied in context of Mexico and USA.[10]

During any epidemic outbreak, regional governmental bodies are the

* Corresponding author.
E-mail addresses: ranjitapandey111@gmail.com (R. Pandey), tolani.ht@gmail.com (H. Tolani).

https://doi.org/10.1016/j.cegh.2022.101176
Received 9 August 2022; Received in revised form 28 October 2022; Accepted 31 October 2022
Available online 5 November 2022

sole source of reporting of the related incidence data on infection and mortality. Potential for misreporting exists in such data sets due to uncertainties in the real-time reporting of the test outcome as well as due to the limitations of the capacity of the testing facilities usually when pandemic is close to its peak. These limitations have been classified as *delay in reporting* and *underreporting* respectively. Length of diagnosis method, limited testing centres, hospital holidays including its restricted-capacity functioning during weekends are multiple sources which cause delays in reporting, while misreporting may arise due to false negative test and ignorance among the population units during early stages of the pandemic. Distribution of the time-elapsed since onset of infection till its occurrence in the infectee is assumed to be known and is termed as serial (or generation) interval distribution. Use of Bayesian inference for semi-parametric regression model brings flexibility in the estimation process through sequential updation of the new infectee counts. In the present work, we adopt Bayesian semiparametric regression model with splines to estimate R_t along with delay and reporting probability. The obtained Bayes estimates are expected to determine the degree of success of the epidemic control strategies.

Our paper is organised as follows. Section 2 describes the statistical model for step-wise inclusion of the underreporting active cases and delay in the reporting process. In Section 3, we analyse corona virus disease (COVID-19) incidence data for India, estimate the daily reproduction number and the delay parameter for the duration of the outbreak from 15th March 2020 up to 13th April 2020(https://www.covid19india.org/). We conduct a simulation study to assess and validate the proposed method of Bayes estimation of R_t, after adjusting for the impact of misspecification of delay patterns and different reporting probabilities. Section 4 is attributed to discussion. Section 5 concludes the study and explores avenues for further research.

2. Methods

The manuscript comprises of majorly four broad methodological concepts. First, observed cases and actual cases are assumed to follow Poisson distribution. Second, the epidemic renewal equation is incorporated for reflecting the progression of disease spread. Third, semi-parametric spline regression which is used to model R_t incorporating effect of underreporting and delay structures fitted under Bayesian paradigm and different scenarios created are compared through Deviance Information Criterion (DIC). Fourth, the simulated R_t under various scenarios is compared with estimated R_t with Mean average square error (MASE) and its components.

2.1. Notations and assumptions

Let $\boldsymbol{A_t} = \{A_1, A_2, \ldots\ldots, A_T\}$ denote the actual cases of new disease counts during T days of an epidemic which consists of a reported part $\boldsymbol{S_t} = \{S_1, S_2, \ldots\ldots, S_T\}$ and an unreported part $\boldsymbol{C_t} = \{C_1, C_2, \ldots\ldots, C_T\}$ such that $\boldsymbol{A_t} = \boldsymbol{S_t} + \boldsymbol{C_t}$ for $t = 1, 2, 3 \ldots$ T. During any epidemic, the total count of reported cases (S_t) are always less than the actual counts (A_t). This underreporting at time t and delay in reporting are responsible for underestimation of R_t.

Reporting probability (τ) may be fixed or time dependent and is referred to as thinning parameter. Suppose maximum length of generation interval is 'r' then the distribution of time interval between the infection times of an infected case and its infector is represented by the ordered set $\boldsymbol{g} = \{g_1, g_2, \ldots, g_r\}$ and termed as the generation interval probability. Hence, count of individuals infected on day $t-$ i is given as $g_i A_{t-i}$ for $i = 1, 2, \ldots, r$.

Assuming that all the infected individuals on day $t-$ i have the same capacity of further transmission (R_t) and the transmission capability R_t changes with each generation, the average infections at time t is denoted by $\lambda_t = R_t(\sum_{i=1}^{t-1} g_i A_{t-i})$ for $t = 2, 3, \ldots, r$ while average infected in the first generation remains deterministic at λ_1. Additionally, $d_{i\to t}$ represents delay probabilities which account for the percentage of cases on day i reported on day t. More specifically, $d_{i\to t}$ captures the delay structure or delayed cases and is the proportion of total cases which are reported on day t but actually belong to day i or in other words if total number of reported or observed cases are 100 on day t and $d_{i\to t}$ is 0.4 then 40 cases actually belong to day i and are reported on day t. Observed count of infectees, including delaying and underreporting is denoted by $\boldsymbol{H_t} = \{H_1, H_2, \ldots\ldots, H_T\}$. Due to delay structures, H_t would exceed A_t for some t. For constant value of τ and $d_{i\to t}$ over time T, shape of epidemic curve remains same accompanied by its positional shift only. For dynamic τ and $d_{i\to t}$ over a time period, epidemic curve experiences change in its shape. Each of the two types of cases, described observed and actual are assumed to follow Poisson distribution with different means. Observed cases which comprise of underreporting and delay structures ultimately are of great analytical importance for this study and are considered for estimation and inferential purposes.

2.2. Building distributional structures

Since Poisson process is a renewal counting process, therefore, total count of infected individuals on day t is assumed to follow Poisson law with parameter λ_t. Probability of observing $A_t(t > r)$ conditional on the past prevalence $A_{tr} = (A_{t-1}, A_{t-2}, \ldots., A_{t-r})$ is given as,

$$P(A_t \,/\, A_{tr}, g, R_t) = \frac{exp\,(-\lambda_t)\lambda_t^{A_t}}{(A_t)!} \;, \tag{1}$$

More specifically A_{tr} represents the process prior to the current renewal point. Probability of reported cases is assumed to follow Binomial law, while accounting for data augmentation of actual prevalence due to underreporting is expressed as,

$$P(S_t \,/\, A_t\,, \tau) = \binom{A_t}{S_t}\tau^{S_t}(1-\tau)^{A_t - S_t} \tag{2}$$

Thus, the observed effective mean prevalence reduces to $\lambda_t\tau_t$ without disturbing the existing correlation dynamics in the chain of the infectee and the infected as follows,

$$P(S_t \,/\, A_{tr}, R_t\,, \tau\,, g) = \frac{exp\,(-\lambda_t\tau)(\lambda_t\tau)^{S_t}}{(S_t)!} \;, \tag{3}$$

Thus, (3) includes the impact of underreporting in the data. Further, incorporating impact caused by $d_{i\to t}$, in (3), we have

$$P(H_t \,/\, \lambda_t\tau, d_{i\to t}) = \frac{exp\,(-\eta_t)(\eta_t)^{H_t}}{(H_t)!} \tag{4}$$

where, $\eta_t = \sum_{i=1}^{t}(\lambda_i\tau)\, d_{i\to t}$, represents the average number of infectees or the mean number of observed cases $(\mathrm{H_t})$. Also, it is the η_t which includes the three components as reproduction number, underreporting parameter and delay parameter, which are to be estimated.

Symbolically, the effective mean due to underreporting is weighed by the accumulated cases in the time interval [*i*,*t*] being reported at time *t*. Physically, often the cases from temporal underreporting and from the delay in reporting are not distinguishable. Hence, to resolve such identifiability[10,11] usage of composite link function **L** is made to map transition from $\{\lambda_t\}$ to $\{\lambda_t\tau\}$. Next, we describe the different delay structures as follows.

(i) *One-day delay*: A fraction of the new cases on day t is reported on day $t+1$, denoted by $d_{t\to t+1}$

(ii) *Two-day delay*: Fraction of new cases on day t is respectively reported partially on days t +1 and t +2 respectively.

(iii) *Weekend Delay*: Due to weekly off on Saturday and Sunday at the reporting health centres, no reporting is recorded on these days. Wednesday, Thursday and Friday will have the same delay pattern as in case (*i*) above. Monday and Tuesday will have reporting of additional fraction which is carried over from the preceding Saturday and Sunday.

We use a penalized spline model to allow R_t to change over time t. Furthermore, it is well known that penalized splines can be embedded in the linear mixed model framework where the selection of the smoothing parameter is provided by estimating the random-effects variance component.[12,13] We model stochastic R_t in the epidemic renewal equation (5) by using a penalized spline as,

$$P(H_t / \eta_t) \sim \text{Poisson}(\eta_t)$$

$$\eta_t = \sum_{i=1}^{t} (\lambda_i \tau) d_{i\to t} = \tau \sum_{i=1}^{t} d_{i\to t} R_i \left(\sum_{s=1}^{k} g_s A_{i-s} \right) \tag{5}$$

$$log(R_t) = \alpha_0 + \alpha_1 t + \sum_{j=1}^{E} u_j \left|t - \varepsilon_j\right|^3 \tag{6}$$

where $\theta = (\alpha_0, \alpha_1, u_1, \ldots\ldots\ldots, u_E)^\intercal$ is the vector of regression coefficients and $\varepsilon_{j:E}$ is an ordered set of fixed knots such that $\varepsilon_1 < \varepsilon_2 < \ldots.. < \varepsilon_E$ and $j = 1,2,3 \ldots,E$. Spline component in (6) represents the non-parametric part, which is amenable to mathematical computations. The statistical description of the considered model and its estimation through computational software is well documented[12] in literature and application of the adapted model is being utilised well for H1N1 pandemic in USA.[10] To ensure the desired flexibility, the number of knots $\boldsymbol{E}$ should be sufficiently large enough. The choice for number of knots is fixed to 15 and is based on past implementations[10,12] of similar model. The vector of random coefficients $\boldsymbol{U} = \{u_1, u_2, \ldots\ldots, u_E\}$ is assumed to be independent and normally distributed with $\mathrm{E}(\boldsymbol{U}) = 0$ and $\mathrm{cov}(\boldsymbol{U}) = \sigma_u^2 \Lambda_E^{-1}$ where Λ_{E} is a matrix in which the $(\mathrm{m}, \mathrm{n})^{\mathrm{th}}$ entry is $|\varepsilon_m - \varepsilon_n|^3$. The class of splines used in the present research is well formulated and decoded[12] in past studies. We are modelling $log(R_t)$ instead of R_t because R_t cannot be negative and can take any value between 0 to infinity. The right hand side of equation (6) can be negative after estimation of model parameters. Hence, $log(R_t)$ instead of R_t.

We assign diffused informative priors as α_0 , $\alpha_1 \sim \mathrm{Normal}(0, 10^4)$; $\sigma_\mathrm{u}^2 \sim \mathrm{Gamma}(10^5, 10^5)$ and postulating flat prior Beata(1,1) for τ and $d_{i\to j}$ each. The choice of priors has been considered after understanding the application of the considered model under Bayesian setup in past studies[10,12]

3. Results

Considering, $E = 3$ with fixed generation distribution $g = \{0.2, 0.3, 0.5\}$ we estimate time varying R_t, with fixed τ_t for $t = 1,\ldots,30$ days. Smoothed estimate of R_t thus obtained are plotted in Fig. 1. We use R version 4.1.0 and OpenBUGS to execute the proposed theoretic formulations for estimation of $d_{i\to j}$ and R_t and used Deviance Information Criterion[14] (DIC) to assess model adequacy and complexity (Tables 1–3) under fixed $\tau = 0.1, 0.5, 0.9$. Posterior mean and posterior standard deviation are obtained based on Markov Chain Monte Carlo (MCMC) approximations. We run MCMC simulations for 10^5 iterations and attain convergence with respect to MCMC error (<0.05). Each simulation is repeated 1000 times to obtain the corresponding posterior estimate. First, we generate epidemic data on the basis of a "reference", which we will refer to as the true underlying delay pattern, under $\tau = 0.5$. Next, we estimate R_t ,$t = 1, 2, 3, \ldots\ldots, 30$, on the basis of four other delay patterns. Five combinations or scenarios of underreporting and delay structures are considered.

(i) Underreporting, one-day-delay ($\mathrm{D_{UO}}$).
(ii) Underreporting, two day-delay ($\mathrm{D_{UT}}$).
(iii) Underreporting weekend-delay ($\mathrm{D_{UW}}$).
(iv) Underreporting, no-delay ($\mathrm{D_{UN}}$).
(v) No-underreporting with no-delay ($\mathrm{D_{NN}}$).

We perform the next stage of simulation to assess the impact of assuming a wrong delay pattern on the estimation of the R_t ,$t = 1, 2, 3, \ldots\ldots, T$. We calculate the Mean Average Squared Error (MASE) with N = 1000 replications for T = 30 days to assess model efficacy as under,

$$MASE = \frac{1}{N} \sum_{n=1}^{N} \left[\frac{1}{T} \sum_{t=1}^{T} \left(\widehat{R}_t - R_t\right)^2 \right]$$

MASE and its bias-variance components for R_t for each true delay pattern vis-à-vis other considered delay-underreporting combinations (Table 4 and Figs. 2–6). We conduct sensitivity analysis to examine the impact of varying reporting probability $\tau = 0.15, 0.2, 0.4$ and 0.6. on the R_t estimates (Table 5). Estimates of τ are plotted under one-day, two-day and weekend delay misspecifications in Figs. 7–9.

Table 1 exhibited overall decrease in estimated one-day-delay probability for increasing reporting probability. DIC shows best, when reporting probability is half the actual occurrence (i.e., $\tau = 0.5$)

Table 2 displayed mixed trend of estimated delay in probabilities due to increase in reporting probability under two day set-up. Again, DIC is

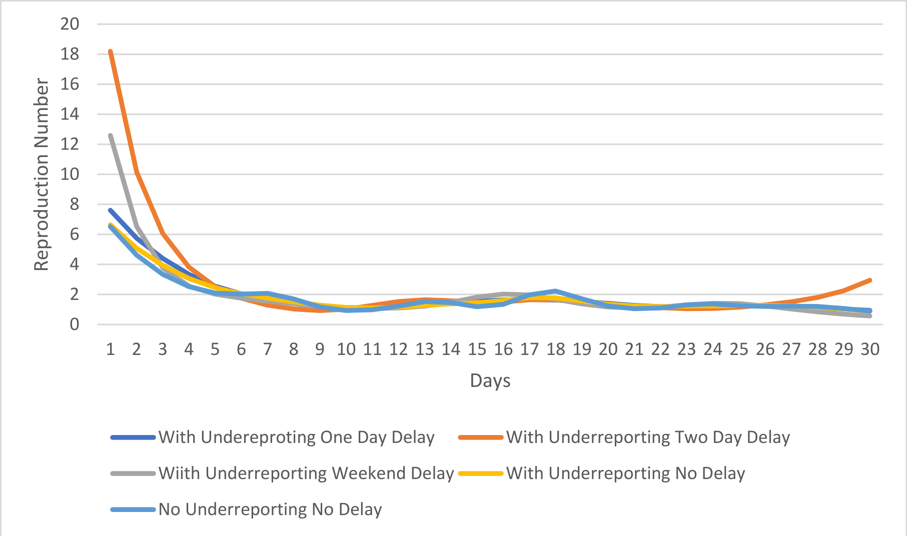


Fig. 1. Smoothed mean estimates of the reproduction numbers for the one-day, two-day and weekend delay patterns with reporting probability of $\tau = 0.5$.

Table 1
Posterior Summary for delay parameters under the case D_{UO}.

Delay Pattern	One-Day Delay					
Reporting Probability	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
Delay Parameters	Mean	SD	Mean	SD	Mean	SD
$d_{M\to Tu}$	0.219324	0.147898	0.166086	0.121979	0.151388	0.097722
$d_{Tu\to W}$	0.298863	0.145566	0.259786	0.122743	0.266019	0.106074
$d_{W\to Th}$	0.214809	0.14309	0.204369	0.118502	0.213984	0.103941
$d_{Th\to F}$	0.124743	0.139731	0.130401	0.115515	0.131481	0.098246
$d_{F\to Sa}$	0.161193	0.140694	0.164848	0.115484	0.15224	0.095382
$d_{Sa\to Su}$	0.268178	0.141742	0.257018	0.115537	0.23405	0.093469
$d_{Su\to M}$	0.159934	0.14653	0.122153	0.118508	0.096841	0.092972
τ	0.863145	0.120171	0.868446	0.109787	0.891494	0.088073
	DIC = 402.9		**DIC= 396.5**		**DIC=530.8**	

Table 2
Posterior Summary for the delay parameters under the case D_{UT}.

Delay Pattern	Two-Day Delay					
Reporting Probability	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
Parameters	Mean	SD	Mean	SD	Mean	SD
$d_{M\to Tu}$	0.822216	0.077909	0.851664	0.05909	0.364046	0.197813
$d_{M\to W}$	0.103838	0.078686	0.075073	0.059053	0.457852	0.190282
$d_{Tu\to W}$	0.850603	0.101463	0.846367	0.098535	0.369343	0.20879
$d_{Tu\to Th}$	0.622934	0.102148	0.564462	0.100854	0.272469	0.153726
$d_{W\to Th}$	0.261771	0.18307	0.31279	0.196005	0.435811	0.237962
$d_{W\to F}$	0.592006	0.177213	0.544739	0.179981	0.317714	0.219379
$d_{Th\to F}$	0.531526	0.188338	0.555002	0.19651	0.581797	0.210321
$d_{Th\to Sa}$	0.157624	0.125782	0.161249	0.123321	0.065932	0.062097
$d_{F\to Sa}$	0.867998	0.107157	0.857839	0.12485	0.109026	0.095322
$d_{F\to Su}$	0.142462	0.103317	0.156039	0.108198	0.799582	0.149343
$d_{Sa\to Su}$	0.402618	0.114068	0.354112	0.112702	0.109246	0.088297
$d_{Sa\to M}$	0.689788	0.051635	0.720571	0.05094	0.200929	0.112594
$d_{Su\to M}$	0.030094	0.028455	0.028453	0.026709	0.575438	0.128044
$d_{Su\to Tu}$	0.61971	0.062831	0.565232	0.059512	0.168654	0.107355
τ	0.857895	0.093755	0.884783	0.066689	0.87776	0.063583
	DIC = 388.6		**DIC = 359.0**		**DIC = 347.5**	

Table 3
Posterior Summary for the delay parameters under the case D_{UW}.

Delay Pattern	Week-Day Delay					
Reporting Probability	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
Parameters	Mean	SD	Mean	SD	Mean	SD
$d_{M\to Tu}$	0.812555	0.060303	0.793482	0.063261	0.825231	0.057408
$d_{Tu\to W}$	0.948164	0.039235	0.941562	0.040913	0.972514	0.023348
$d_{W\to Th}$	0.554671	0.048658	0.592335	0.046873	0.623621	0.037133
$d_{Th\to F}$	0.238958	0.042592	0.290453	0.046945	0.282972	0.035502
$d_{F\to Sa}$	0.039745	0.033162	0.058016	0.042654	0.024623	0.021552
$d_{Sa\to Su}$	0.070755	0.067146	0.091619	0.080396	0.06322	0.059453
$d_{Su\to M}$	0.646307	0.10453	0.637999	0.119541	0.602709	0.092993
τ	0.8435	0.093607	0.861299	0.073137	0.876365	0.070275
	DIC = 6.914e+13		**DIC = 6.914e+13**		**DIC=6.914e+13**	

observed to be lowest when number of reported cases are half of the actual cases.

Estimates of delay in probabilities are same for all the considered reporting probabilities. Overall, among all delay structures considered, the lowest DIC is seen for $\tau = 0.5$ under two-day delay setup for the considered data set.

Table 4 shows the corresponding MASE and its bias–variance decomposition for estimates of the reproduction numbers based on different delay and reporting patterns. Accounting for delay and underreporting is seen to increase the accuracy for estimates of R_t. Lowest MASE is obtained for the correct delay pattern, which validates the model. Largest MASEs are obtained for misspecification in D_{UT}. In general, Models with underreporting show less bias than those which incorrectly ignore underreporting. MASE's for the two considered scenarios of no-delay structures are closest to each other.

Table 5 yields the smaller MASE's for high reporting probabilities (0.4 and 0.6) for two-day and weekend delay patterns as compared to one-day delay patterns. Figs. 2–6 depict the estimated daily reproduction numbers are shown for different reference and fitted delay pattern combinations. Figs. 7–9 show the estimated reproduction numbers under different delay patterns for different values of reporting probability. MASE is seen to be highest for D_{UW} and lowest for D_{UO} for all the reporting probability situations. MASE's are seen to decrease with increase in reporting probability for D_{UT} and D_{UW}.

Table 4
MASE and its bias–variance decomposition for the estimates of the reproduction numbers based on the sensitivity analysis of different delay patterns.

True Delay Pattern	Fitted Pattern					
		D_{UO}	D_{UT}	D_{UW}	D_{UN}	D_{NN}
D_{UO}	**MASE**	3.96E-05	0.002195	0.000597348	9.81E-05	9.78E-05
	Var	1.388269	0.04243	4.084802	1.639246	2.187404
	(Bias)^2	7.08E-06	0.311071	0.03845118	0.001879091	0.00971007
D_{UT}	**MASE**	0.003539	6.27E-05	0.015843	0.00774084	0.008196599
	Var	3.209562	11.20448	0.159435	1.039951	1.325875
	(Bias)^2	0.34943	0.01076476	2.002158	0.6281018	0.5307091
D_{UW}	**MASE**	0.006787	2.31E-05	0.00808879	0.004091565	0.00454935
	Var	4.884714	0.044722	6.090793	2.966199	3.347738
	(Bias)^2	1.10E-01	0.431851	1.34E-01	0.001808002	0.006190591
D_{UN}	**MASE**	0.000771	0.001736	0.000747193	6.08E-05	8.00E-05
	Var	4.35536	0.037945	4.325529	2.21497	1.887889
	(Bias)^2	0.03646	0.249005	0.03246108	0.00595966	0.003908461
D_{NN}	**MASE**	0.000246	0.001586	0.003560694	5.88E-04	0.000183114
	Var	2.428834	0.030526	8.755043	3.494922	2.384256
	(Bias)^2	0.02197	0.230976	0.202911	0.02573087	0.00244879

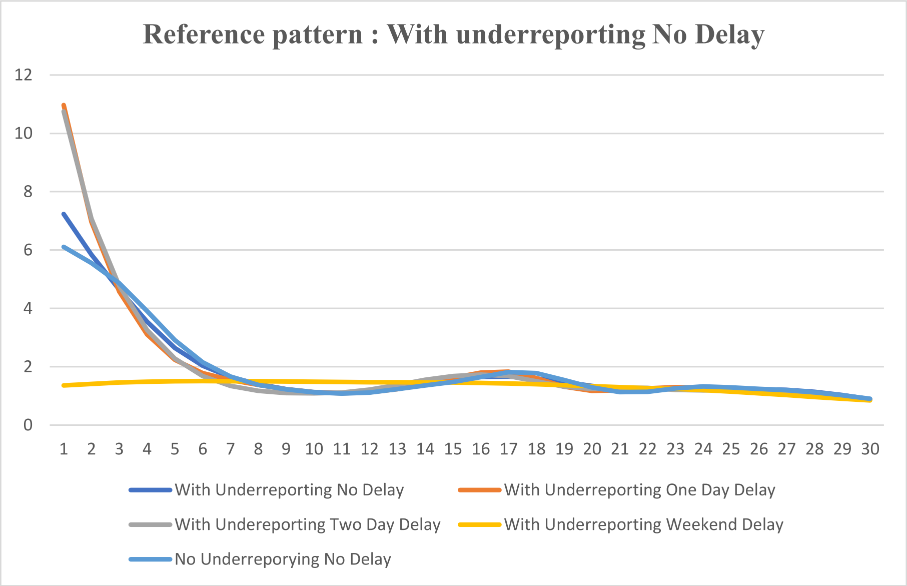


Fig. 2. Estimation of R_t taking D_{UN} as reference scenario.

4. Discussion

Estimation of reproduction number, usually referred to as "Ro" in most of the studies, comprises of sufficient number of classical statistical methods utilised in recent past in India. Exponential growth rate model[15] in India for Covid-19 has gained decent popularity in estimation reproduction number by using "Ro pacakage"[15,16] and "incidence package"[16] in R software. Estimates for reproduction number ranged from 1.6 to 2.7 in studies for Covid-19 in India and its states.[15–18] A series of studies[19] was carried out for COvid-19 in China for estimation of reproduction number using all possible statistical techniques based on maximum likelihood estimation, SEIRD, and MCMC. The estimated reproduction number for Covid-19 ranged from 2 to 7 in most of the studies.[19]

Renewal equations and their role in stochastic process[20] is well known in understanding the time-based random phenomenon observed in real life applications. In context of epidemic modelling, renewal theory based[21] applications are limited in number. Embedding renewal equation into spline regression for capturing the dynamics of pandemic caused by the novel corona virus under different scenarios is the soul of this research. More specifically, the present research was to estimate reproduction number incorporating the effect of underreporting and delay in reporting through innovative approach of penalized spline regression through Bayesian toolkit which isn't being explored yet in context of Covid-19 in India.

We note that there is a large variability in the estimation of the reproduction number for the first few observations. Often Information is scarce at the beginning of any epidemic break-out, hence R_t curve shows higher probability for the initial recorded cases. It can also imply that the initial infectees are responsible for larger number of secondary cases

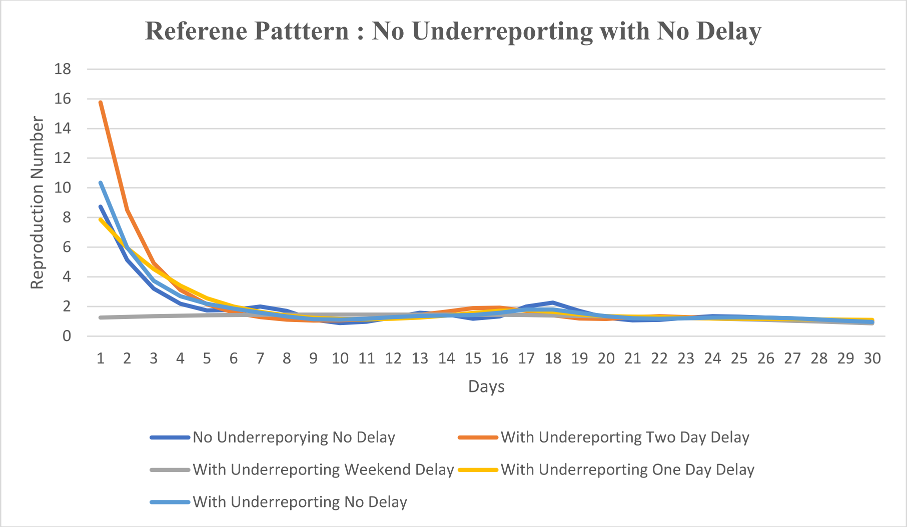


Fig. 3. Estimation of R_t taking $\mathrm{D_{NN}}$ as reference scenario.

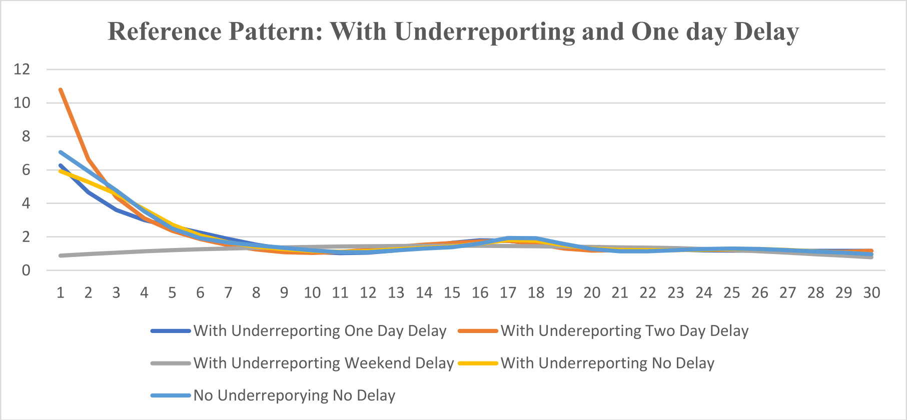


Fig. 4. Estimation of R_t taking $\mathrm{D_{UO}}$ as reference scenario.

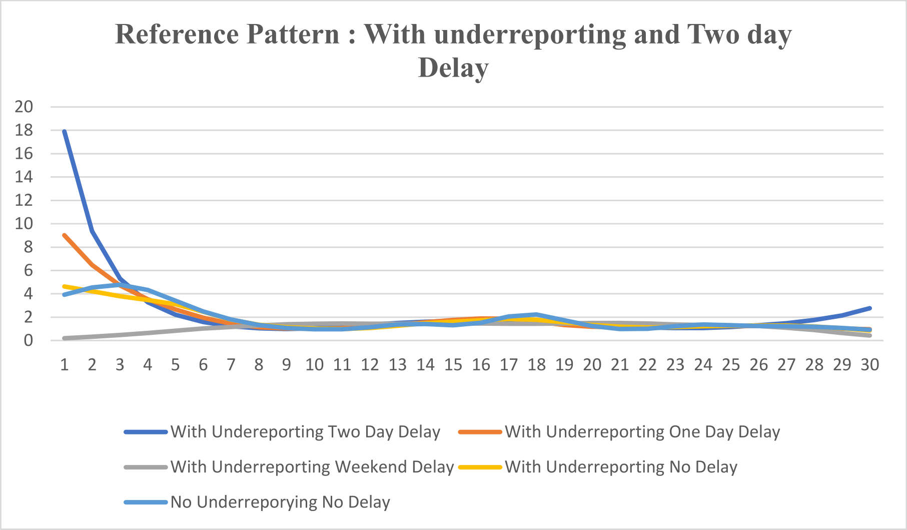


Fig. 5. Estimation of R_t taking $\mathrm{D_{UT}}$ as reference scenario.

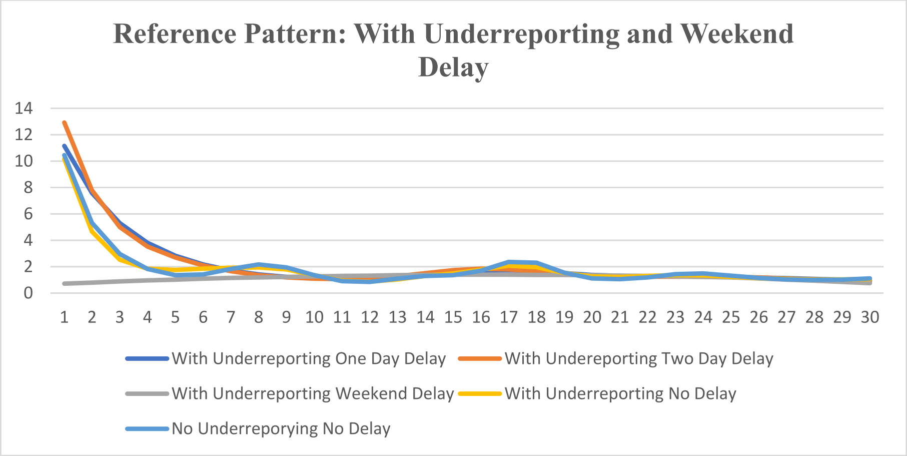


Fig. 6. Estimation of R_t taking $\mathrm{D_{UW}}$ as reference scenario.

Table 5
Sensitivity analysis for non-constant reporting probability (τ).

Delay Pattern	One-Day			
Reporting Probability	$\tau = 0.15$	$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.6$
MASE	1.38E-05	0.000979	8.57E-05	5.01E-05
Var	1.559231	1.996115	2.196816	1.856806
(Bias)^2	0.000746	0.064864	0.002298	0.003574
Delay Pattern	**Two-Day**			
Reporting Probability	$\boldsymbol{\tau = 0.15}$	$\boldsymbol{\tau = 0.2}$	$\boldsymbol{\tau = 0.4}$	$\boldsymbol{\tau = 0.6}$
MASE	0.000218	5.89E-05	3.54E-05	3.24E-05
Var	8.736962	9.81039	10.3802	12.301
(Bias)^2	0.015307	0.001446	7.41E-06	0.000477
Delay Pattern	**Weekend**			
Reporting Probability	$\boldsymbol{\tau = 0.15}$	$\boldsymbol{\tau = 0.2}$	$\boldsymbol{\tau = 0.4}$	$\boldsymbol{\tau = 0.6}$
MASE	0.009235	0.008048	0.007351	0.007217
Var	7.788936	6.551135	5.816634	5.847377
(Bias)^2	1.067153	1.094236	1.10153	0.971853

in absence of either awareness or control measures or both.

DIC being a measure of model adequacy and complexity gave us the answers for identifying the best case scenario for underreporting and delay parameters. Evidence from the analysis (Table 2) shows that among all the considered delay structures for different reporting probabilities two-day delay pattern with a reporting probability of 50% was the most suitable scenario which follows the considered dataset closely.

Fig. (2 - 6) show that mis-specifying the delay pattern for $\mathrm{D_{UO}}$, $\mathrm{D_{UT}}$, $\mathrm{D_{UW}}$, $\mathrm{D_{UN}}$ and $\mathrm{D_{NN}}$ has a moderate impact on the estimated trend for the reproduction number. Note that the structure of the $\mathrm{D_{UT}}$ pattern is an extension of the $\mathrm{D_{UO}}$ pattern. There is a substantial impact when mis-specifying the $\mathrm{D_{UW}}$ pattern. MASE with the help of simulation gave us the answers on how model estimates vary on mis-specifying a particular scenario and reflected the stability and suitability of the adopted modelling procedure. Simulation study (Tables 4 and 5) justified the modelling strategy adopted for the stated problem under different cases of misspecification of underreporting and delay structures.

5. Conclusion

Reporting of R_t instead of daily cases, alerts both the government and public to start and regulate preventive measures accordingly. In the present paper, explanation of the logarithmic transform of R_t through penalized splines under Bayesian setup have generated stronger evidence in favour of the proposed model as is evident from following discussion. The present paper follows Bayesian paradigm under diffused informative priors by treating the reporting fractions, reposting delays and their interactive influences distinctly. Time dependent R_t are seen to be influenced by reporting fractions and systematic delays in reporting. The present study conclude that observed data accompanied by 50% underreporting for a two day lag in recording the incidence of covid-19 cases are closest to the empirical data. A precise estimate of R_t, therefore, is crucial in formulation of corrective or preventive policies towards handling the epidemic and towards efficacy of intervention

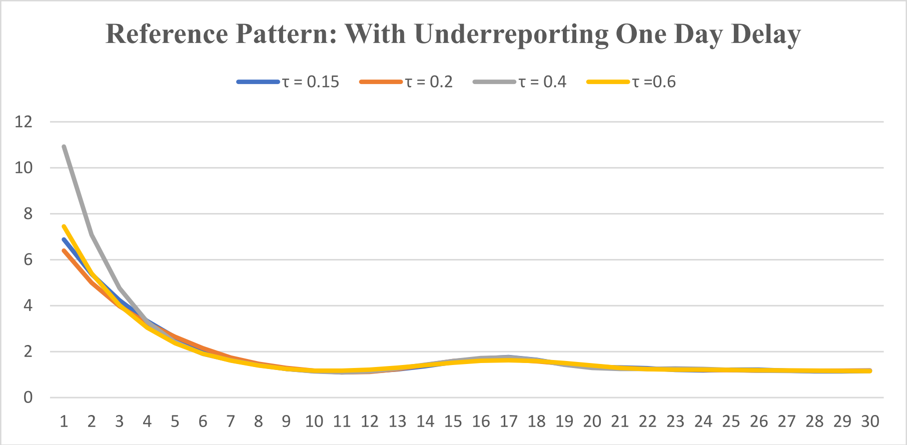


Fig. 7. Estimated R_t under different reporting probabilities.

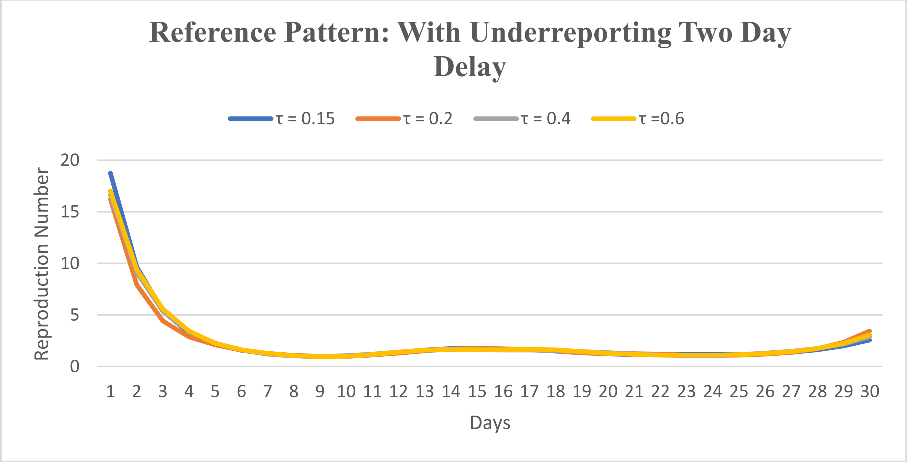


Fig. 8. Estimated R_t under different reporting probabilities.

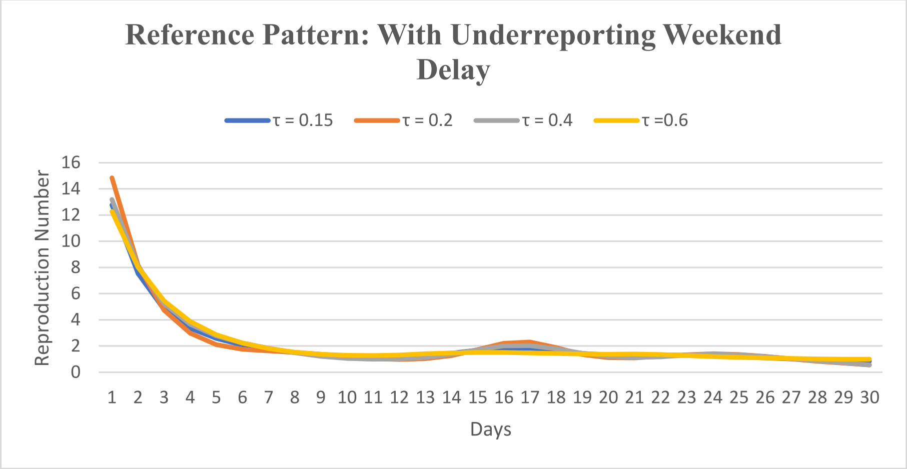


Fig. 9. Estimated R_t under different reporting probabilities.

indicators such as lockdown, quarantine and other related control measures. Future scope of study could include a stochastic serial distribution, correlated reproduction numbers and a different delay pattern than the weekly logistics studied in the present paper. Also, propagation of pandemic is considered for closed population in the present work, which could be deviated from, to understand the advantage and necessity of imposition of movement across spatial contours. Such a study could be useful in understanding cost-benefit bargains in economic activities. R_t can be computed geographically i.e., for each state, district or any other administrative boundary to understand the similarities and differences in dynamics of pandemic like Covid-19 for geospatial analysis and local level covariates responsible for pandemic situation.

Sources of funding

The first author gratefully acknowledges IoE grant from University of Delhi.

Declaration of competing interest

All the authors declared to have no conflict of interest.

Acknowledgements

Authors express their gratitude to all the anonymous reviwers and the editor for their constructive comments and suggestions which resulted in immense improvement in the original manuscript.

References

1 Agarwal GG. Splines in statistics. *Bull Allahabad Math Soc*. 1989;4:1–5.
2 Ganyani T, Faes C, Chowell G, Hens N. Assessing inference of the basic reproduction number in an SIR model incorporating a growth-scaling parameter. *Stat Med*. 2018: 1–17.
3 Korolev I. Identification and estimation of the SEIRD epidemic model for COVID-19. *J Econom*. 2021;220(1):63–85.
4 Dolton P. The statistical challenges of modelling COVID-19. *Natl Inst Econ Rev*. 2021; 257:46–82. https://doi.org/10.1017/nie.2021.22.
5 Mullah MAS, Yan P. A semi-parametric mixed model for short-term projection of daily COVID-19 incidence in Canada. *Epidemics*. 2022;38, 100537.
6 Cauchemez S, Boelle PY, Thomas G, Valleron AJ. Estimation in real time the efficacy of measures to control emerging communicable diseases. *Am J Epidemiol*. 2006;164: 591–597.
7 White LF, Wallinga J, Finelli L, et al. Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. *Influ Other Respir Virus*. 2009;3:267–276.
8 Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Royal Soc*. 2007;274:599–604.
9 Hens N, Van Ranst M, Aerts M, Robesyn E, Van Damme P, Beutels P. Estimating the effective reproduction number for pandemic influenza from notification data made publicly available in real time: a multi-country analysis for influenza A/H1N1v 2009. *Vaccine*. 2011;29(5):896–904.
10 Azmon A, Faes C, Hens N. On the estimation of the reproduction number based on misreported epidemic data. *Stat Med*. 2014;33(7):1176–1192.
11 Thompson R, Baker RJ. Composite link functions in generalized linear models. *Appl Stat*. 1981;30(2):125–131.

12 Crainiceanu CM, Ruppert D, Wand MP. Bayesian analysis for penalized spline regression using WinBUGS. *J Stat Software*. 2005;14:1–24.

13 Ruppert D, Wand MP, Carrol RJ. *Semiparametric Regression*. first ed. Cambridge: Cambridge University Press; 2003.

14 Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc B Stat Meth*. 2002;64(4):583–639.

15 Marimuthu S, Joy M, Malavika B, Nadaraj A, Asirvatham ES, Jeyaseelan L. Modelling of reproduction number for COVID-19 in India and high incidence states. *Clin Epiderm Global Health*. 2021;9:57–61.

16 Mitra A, Pakhare AP, Roy A, Joshi A. Impact of COVID-19 epidemic curtailment strategies in selected Indian states: an analysis by reproduction number and doubling time with incidence modelling. *PLoS One*. 2020;15(9), e0239026. https://doi.org/10.1371/journal.pone.0239026.

17 Patrikar SR, Kotwal A, Bhatti VK, et al. Incubation period and reproduction number for novel coronavirus (COVID-19) infections in India. *medRxiv*. 2020.

18 Rai B, Shukla A, Dwivedi LK. COVID-19 in India: predictions, reproduction number and public health preparedness. *medRxiv*. 2020.

19 Liu Ying, Gayle Albert A, Wilder-Smith Annelies, Rocklöv Joacim. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J Trav Med*. March 2020;27(Issue 2). https://doi.org/10.1093/jtm/taaa021. taaa021,.

20 Karlin S. *A First Course in Stochastic Processes*. Academic press; 2014.

21 Mishra S, Berah T, Mellan TA, et al. *On the Derivation of the Renewal Equation from an Age-dependent Branching Process: An Epidemic Modelling Perspective*. 2020. arXiv preprint arXiv:2006.16487.