



# Machine learning-based modelling and forecasting of COVID-19 under temporally varying public intervention

Yiye Jiang, Gaston Vergara-Hermosilla

## ► To cite this version:

Yiye Jiang, Gaston Vergara-Hermosilla. Machine learning-based modelling and forecasting of COVID-19 under temporally varying public intervention. 2023. hal-03680677v4

**HAL Id: hal-03680677**

**<https://hal.science/hal-03680677v4>**

Preprint submitted on 3 Oct 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# MACHINE LEARNING-BASED MODELLING AND FORECASTING OF COVID-19 UNDER TEMPORALLY VARYING PUBLIC INTERVENTION

Yiye Jiang<sup>1\*</sup> and Gaston Vergara-Hermosilla<sup>2</sup>

<sup>1</sup>Institut de Mathématiques de Bordeaux, UMR CNRS 5251, Université de Bordeaux, France.

<sup>2</sup>Laboratoire de Mathématiques et Modélisation d'Evry, UMR CNRS 8071, Université Paris-Saclay, France.

E-mails: gaston.v-h@outlook.com - yiyejiang93@gmail.com

## Abstract

By considering the recently introduced SIRU model, in this paper we study the dynamic of COVID-19 pandemic under the temporally varying public intervention in the Chilean context. More precisely, we propose a method to forecast cumulative daily reported cases  $CR(t)$ , and a systematic way to identify the unreported daily cases given  $CR(t)$  data. We firstly base on the recently introduced epidemic model SIRU (Susceptible, asymptomatic Infected, Reported infected, Unreported infected), and focus on the transmission rate parameter  $\tau$ . To understand the dynamic of the data, we extend the scalar  $\tau$  to an unknown function  $\tau(t)$  in the SIRU system, which is then inferred directly from the historical  $CR(t)$  data, based on nonparametric estimation. The estimation of  $\tau(t)$  leads to the estimation of other unobserved functions in the system, including the daily unreported cases. Furthermore, the estimation of  $\tau(t)$  allows us to build links between the pandemic evolution and the public intervention, which is modeled by logistic regression. We then employ polynomial approximation to construct a predicted curve which evolves with the latest trend of  $CR(t)$ . In addition, we regularize the evolution of the forecast in such a way that it corresponds to the future intervention plan based on the previously obtained link knowledge. We test the proposed predictor on different time windows. The promising results show the effectiveness of the proposed methods.

**Keywords:** SIRU model, Transmission rate, Cumulative daily reported cases, Nonparametric estimation.

## 1 Introduction

In the recent years, the modelling of epidemiological phenomena have played a protagonist role in taking decisions and controlling the COVID-19 pandemic around the world [7]. In particular, the mathematical approach has been an important input for understanding and predicting the underlying patterns of the epidemiological dynamics, as well as for understanding the potential roles of government measures in the disease propagation [11]. For instance, the mathematical COVID-19 models predicting key parameters involved in the temporal development of the pandemic, as the transmission rate or number of infected unreported in official statistics, has been a fundamental tool for public health authorities in order to plan interventions for controlling the propagation of the disease. In this spirit, the authors of [24] propose a review paper about recently introduced COVID-19 epidemic models dealing with these issues. In fact, one of the ways to treat the research challenges that have appeared in the context of the COVID-19 epidemic is by employing compartmental models, which are the classical systems used in the mathematical modelling of infectious diseases, for example, the authors of [3, 4] based their research on the SEIR (Susceptible, Exposed, Infectious, Removed) model, while the authors of [5, 6] rely on the SIRD (Susceptible, Infected, Recovered, Dead) model. On

---

\*Corresponding author.

the other hand, researchers of the machine learning community have also used their expertise in the modelling challenges of the COVID-19 pandemic, for instance, the authors in [2] applied tools from functional data analysis to model the trajectories of cumulative daily reported cases across countries, while the authors of [1] adopt the point process based approaches to model the infection/death cases which can be considered as events arriving at random times. Some other papers considering machine learning methods that we highlight are the following [10, 14, 16]. However, the methods and results of the two research approaches mentioned recently have seen proposed in a completely independent way of each other. In this paper, we consider machine learning techniques with a compartmental model in order to obtain efficient estimations and predictions of propagation of COVID-19 in a particular context that we proceed to explain. Considering that one of the main characteristics of the different strains of COVID-19 that have appeared is the heterogeneity of the symptoms of infected individuals, added to the low testing capacity in some realities around the world, motivates the study of epidemiological models involving compartments referring to the existence of unreported cases. Among the various models in the literature that consider this extra compartment, the recently introduced SIRU model [11] has been successfully implemented to describe the evolution of COVID-19 during the first and second pandemic waves in several countries, such as China, South Korea, Italy and France. In this work we rely on the SIRU model to understand the epidemiological dynamic of COVID-19. As this model does not propose a method to construct the appropriate transmission rate involved in the dynamic of the pandemic, which presents an evolution in time according to the different measures taken by the respective authorities in order to reduce the rise in the number of infections and deaths, we aim to propose a novel model setting which fill this gap. In the following we recall the principles of SIRU model.

In fact, this model describes the dynamic of a pandemic situation by considering a system of ordinary differential equations (ODEs) involving four different compartments: susceptible individuals, infected individuals who do not yet have symptoms, reported infected individuals, and unreported infected individuals, denoted by  $S$ ,  $I$ ,  $R$  and  $U$ , respectively. This model can be illustrated by the following diagram flux (see [11]):

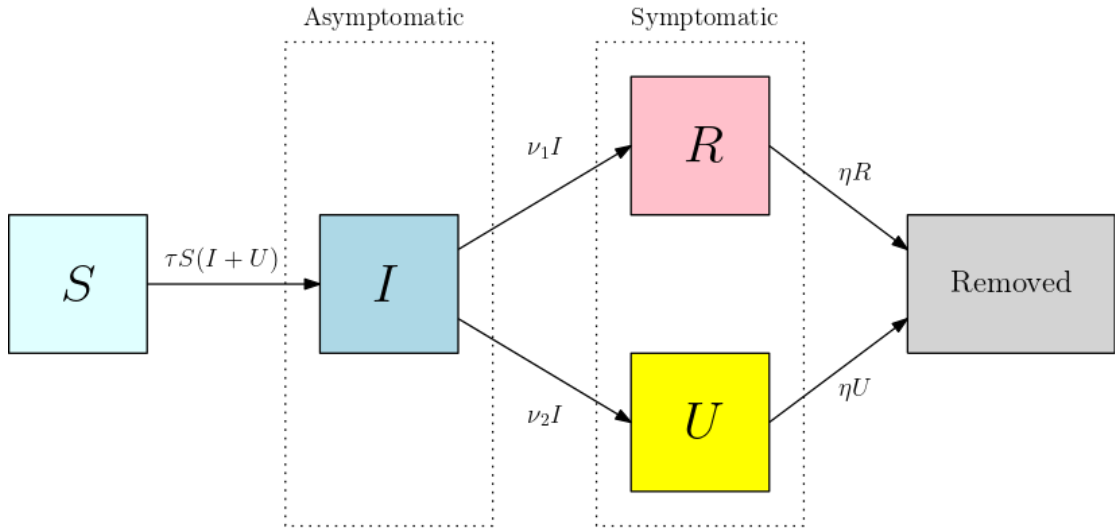


Figure 1: Diagram flux associated with the SIRU model.

The system of ODEs associated with the diagram flux reads as:

$$\begin{cases} S'(t) = -\tau S(t)(I(t) + U(t)) \\ I'(t) = \tau S(t)(I(t) + U(t)) - \nu I(t) \\ R'(t) = \nu_1 I(t) - \eta R(t) \\ U'(t) = \nu_2 I(t) - \eta U(t) \end{cases} \quad (1.1)$$

Parameters	Interpretation
$t_0$	Time at which the epidemic started.
$S_0$	Number of individuals susceptible to the disease at time $t_0$ .
$I_0$	Number of infected individuals without symptoms at time $t_0$ .
$R_0$	Number of reported infected individuals at time $t_0$ .
$U_0$	Number of unreported infected individuals at time $t_0$ .
$\tau$	Transmission rate of the disease.
$1/\nu$	Average time during which the infected asymptomatic individuals remain asymptomatic.
$f$	Fraction of asymptomatic infected individuals that become reported infected individuals.
$\nu_1 = f\nu$	Rate at which asymptomatic infected cases become reported symptomatic.
$\nu_2 = (1 - f)\nu$	Rate at which asymptomatic infected cases become unreported infected.
$1/\eta$	Average time during which an infected individual presents symptoms.

Table 1: Parameters of the SIRU model.

As is usual, the system of ODEs (1.1) is supplemented with initial data:

$$S(t_0) = S_0 > 0, I(t_0) = I_0 > 0, R(t_0) = 0 \text{ and } U(t_0) = U_0 \geq 0. \quad (1.2)$$

The parameters of the model are summarized in Table 1. In particular, we remark that  $\nu = \nu_1 + \nu_2$ , and that all the parameters that are considered  $\tau, \nu, \nu_1, \nu_2, \eta$  are positive. In the initial works on the SIRU models [23, 11], the initial data are derived by assuming the early stage of the system is exponential, and the parameters  $\nu, f, \eta$  are treated as hyperparameters, which means that they are pre-assigned but not learn from data. We use the same practice in our work. Thus, the transmission rate  $\tau$  will be the key parameter to control the epidemic propagation after the early stage. This implies a better  $\tau$  will make the SIRU system closer to real situation in the sense that the  $CR(t)$  recovered from the SIRU solution  $I(t)$  is closer to the real observation. In this case, the other SIRU solutions  $S(t), R(t), U(t)$  will be more reliable estimations of the unobserved data.

To adapt the SIRU model to the real COVID-19 data where the transmission rate is influenced by the public interventions, the authors of [12] and [23], propose to make the transmission rate  $\tau$  as a function of time, parametrized by the public intervention. More precisely, in [12] the authors propose the following structure of the transmission rate:

$$\tau(t) = \begin{cases} \tau_0, & \text{for } t \in [t_0, t_1], \\ 0, & \text{for } t \geq t_1, \end{cases} \quad (1.3)$$

while in [23] the authors consider:

$$\tau(t) = \begin{cases} \tau_0, & \text{for } t \in [t_0, t_1], \\ \tau_0 \exp(-\mu(t - t_1)), & \text{for } t \geq t_1, \end{cases} \quad (1.4)$$

where  $\tau_0$  is a scalar parameter which characterises the constant transmission rate during the time interval  $[t_0, t_1]$ , and  $\mu$  is a scalar parameter that characterises the constant intervention intensity. In both cases, even with these more detailed model design for  $\tau(t)$ , the numerical results obtained by considering these variants of the SIRU model still end with one pandemic wave, as the initially

proposed SIRU model [11]. However, the data obtained from the official statistics of all countries contain multiple waves. Thus, it is important to find a more powerful and flexible model setting for the transmission rate. This is one of the main aims of this work.

On the other hand, in [13] the authors proposed a more sophisticated form of the transmission rate which considers  $r$  intervention intensity values:

$$\tau(t) = \begin{cases} \tau_0, & \text{for } t \in [t_0, t_1], \\ \tau_1(t) = \tau_0 \exp(-\mu_1(t - t_1)), & \text{for } t \in [t_1, t_2], \\ \vdots & \\ \tau_r(t) = \tau_{r-1}(t) \exp(-\mu_{r-1}(t - t_{r-1})), & \text{for } t \in [t_{r-1}, t_r]. \end{cases} \quad (1.5)$$

Even though the large number of parameters is potentially able to capture the complicated dynamic of transmission rate, to find the ideal parameter values which can correctly reconstruct the real data became another problem. In fact, the values of  $t_i, i = 0, \dots, r$  need to be either manually tuned or extensively searched over grids, both by comparing with the data. Thus we were to search for a method which can bring high variability of  $\tau$  but also is easy to estimate. We proposed to make  $\tau$  a fully free function of time and to approximate this function using nonparametric estimation.

Considering the tuning difficulty is inevitable for any parametric forms imposed to  $\tau(t)$ , for data illustrating multiple waves, we propose to use the nonparametric estimation in the SIRU system to infer the shape directly from the data. This initiative makes the classical compartmental model fully benefit the data so as to give a precise reconstruction of transmission rate dynamic. This estimation of transmission rate is the primary result of this paper. We then link transmission rate dynamic to the public intervention changes to study its influence, which serve for the forecasting approach for the cumulative number of reported symptomatic infectious cases  $CR(t)$  which can take into account the future intervention plans. Such forecasting models are of great interest for the decision makers. In the following, we resume the main novelties of the proposed methods, and then we present the organization of this article.

**Contributions of the paper.** The main contributions of this article are twofold:

1. firstly, proposes a nonparametric method to estimate the transmission rate  $\tau(t)$  and the state related to the unreported daily cases  $u(t)$ ;
2. secondly, proposes a method to predict the cumulative number of reported symptomatic infectious cases  $CR(t)$  which take into account the varying public intervention and a long prediction period.

**Organization of the paper.** In Section 2.1, we propose the nonparametric method to estimate the transmission rate  $\tau(t)$  as well as  $I(t), R(t), S(t)$  and the unreported daily cases  $U(t)$ . In Section 2.2, we rely on the logistic regression to predict from the temporal variable public intervention  $Q(t)$  the dynamic of  $\tau(t)$ . In Section 2.3, we consider the prediction of  $CR(t)$ . we propose the regularized polynomial approximation as the predictor. It is defined as a minimizer of an optimization problem which considers simultaneously the historical data of  $CR(t)$ , and the predicted future  $\tau(t)$  dynamic. Finally, in Section 3, we present the numerical evidence of the proposed methods.

## 2 Methodology

### 2.1 Nonparametric estimation of the transmission rate

In this section, we propose a method to infer the curve of transmission rate  $\tau(t)$ . We require the inferred values to be reliable, in the sense that they are able to recover the historical of the cumulative

number of reported infectious cases  $CR(t)$  records with good accuracy, when plugged back into the SIRU model. Given the observations of  $CR(t)$  at time instants  $t_1, \dots, t_N$ , we propose to make  $\tau(t)$  an unknown function to be solved instead of providing an explicit form as in literature. In return, we make  $I(t)$  a known function in SIRU (1.1) system by estimating it outside the system. The resolution of this transformed SIRU model gives the inferred transmission rate, together with the precise reconstruction of  $S(t)$ ,  $R(t)$  and  $U(t)$ , in the sense they can precisely recover the  $CR(t)$  data.

To get a function, which is highly close to the “true”  $I(t)$  under the SIRU model assumption, we first apply an admissible nonlinear approximation on the  $CR(t)$  data to obtain the estimated curve  $\widehat{CR}(t)$ . Then the relationship between  $I(t)$  and  $CR(t)$ :

$$CR(t) = \nu_1 \int_{t_0}^t I(s) ds, \quad (2.1)$$

implies that the estimators for  $I(t)$  and  $I'(t)$  are defined as

$$\widehat{I}(t) = \widehat{CR}'(t)/\nu_1. \quad (2.2)$$

Therefore, we can plug the estimated functions  $\widehat{I}$ ,  $\widehat{I}'$  in the SIRU model, and consider the resulting ODE system as the system of  $S(t)$ ,  $R(t)$ ,  $U(t)$ ,  $\tau(t)$ , which reads as

$$\begin{cases} S'(t) = -\tau(t)S(t)(\widehat{I} + U(t)), \\ R'(t) = \nu_1 \widehat{I} - \eta R(t), \\ U'(t) = \nu_2 \widehat{I} - \eta U(t), \\ \widehat{I}'(t) = \tau(t)(\widehat{I} + U(t))S(t) - \nu_1 \widehat{I}. \end{cases} \quad (2.3)$$

The above system is equivalent to:

$$\begin{cases} S'(t) = -\widehat{I}' - \nu_1 \widehat{I}, \\ R'(t) = \nu_1 \widehat{I} - \eta R(t), \\ U'(t) = \nu_2 \widehat{I} - \eta U(t), \\ \tau(t) = \frac{\widehat{I}' + \nu_1 \widehat{I}}{(\widehat{I} + U(t))S(t)}. \end{cases} \quad (2.4)$$

System (2.4) is easy to solve given the initial values  $S_0, R_0, U_0$  and  $t_0$ . The initial data is obtained in the same manner as [11]. It is worth to mention that, to be consistent with the initial values, before applying the nonlinear approximation, we amend to the observations the data points  $CR(t_0), \dots, CR(0)$  that are generated by the exponential estimation of early stage of  $CR(t)$ . The exponential estimation is the one used in the calculation of initial values.

The key point in the above estimation is to choose an admissible nonlinear method. Common nonlinear methods that reconstruct a data curve by a function, are polynomial approximation, spline, and kernel smoother, see for example [8]. For model (2.4), we propose to use the kernel smoother. On one hand, the polynomial approximation usually introduces oscillations, which will furthermore be amplified after taking derivatives. Thus the final estimated  $\tau(t)$  will exhibit more local extreme points which will mislead the interpretation of true dynamic contained in raw data. On the other hand, the  $I'(t)$  expression given by the SIRU model (1.1) implies that  $I(t)$  is likely to be a  $C^\infty$  function. Thus compared to spline function which is piece-wise polynomials of low order, the kernel smoother with Gaussian kernel is preferable.

To close this subsection, we comment that the nonparametric estimation we proposed to  $\tau$  is transferable. That means, we can apply the same method over other hyperparameters as  $f, \nu, \eta$ . However, since only  $I(t)$  is turned known, in return, to main exactly 4 unknowns for 4 equations,

each time, we can only make one of  $\tau, f, \nu, \eta$  unknown (time-varying) with the rest presumed values which are fixed for all the time. At this point, we privilege a functional  $\tau$ . Because it is by nature more variable than other hyperparameters like average infectious time, since the latter are related to psychological facts. Thus it is less reasonable to let vary others but set the transmission rate fixed all the way. On the other hand, transmission rate is the key characteristic of the evolution of epidemic, thus it is important to investigate it deeply. In the next subsection, we will use the transmission rate function in the further studies on the impact of changing government measures.

## 2.2 Logistic regression with the public intervention policies

In this section, we consider the case of variable public intervention measure. We wish to study its impact on the evolution of epidemic and develop the analysis, given the transmission rate data, and additionally the historical intervention data. To this end, we first introduce a new temporal function which is able to represent the intervention measure. Then we propose a mathematical model which describes the relationship between the introduced measure function and the transmission rate. The resulting model is expected to furthermore help the prediction of unseen  $CR(t)$ .

In Chile, a significant varying public measure is the percentage of national population in quarantine. Such measurement is used in the work of [13] to motivate the design the epidemic model, and leads to a good fit of  $CR(t)$ . We therefore consider the same measurement as the representative of overall public intervention. In Figure 2, we show the evolution of national quarantine percentage. The data is obtained from official information about quarantines provided by the ministry of health of the Chilean government via the webpage [20]. We especially smooth the data points to facilitate the observation. We can see that generally, the dynamic of the measurement is complicated. There exists several accelerations and decelerations of the implementation of quarantine. On the left of Figure 2 is the inferred transmission rate  $\tau(t)$ <sup>1</sup> obtained from the preceding section. We can observe

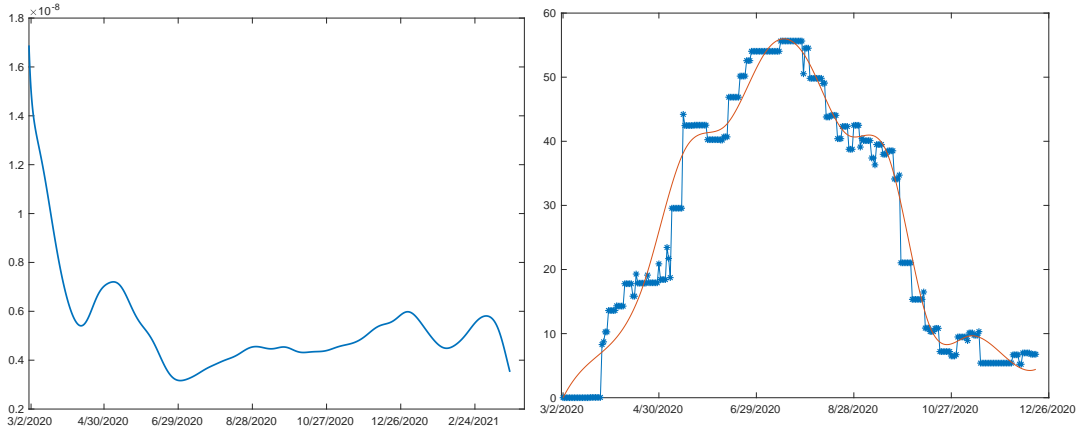


Figure 2: Inferred transmission rate (left) from the Chilean COVID-19 data, percentages of the Chilean population in quarantine (right). The red curve in the right plot shows the smoothing curve of the discrete data points.

<sup>1</sup>In the literature there is no mention nor discussion about the unit of transmission rate. Thus, here we give a reasonable interpretation, that is, transmission rate represents the probability of an individual from  $I$  (asymptomatic infectious) or  $U$  (unreported symptomatic infectious) infecting an individual from  $S$  (susceptible to be infected). The support of this interpretation is the equation of the SIRU model

$$S'(t) = -\tau S(t)(I(t) + U(t)),$$

which describes how much  $S$  individuals should become  $I$  individuals.

Thus given the population, the value of transmission needs to be very small. Furthermore, we made the transmission rate time-varying, thus we can furthermore interpret the transmission rate at time  $t$  as the probability at time  $t$  of an individual from  $I$  or  $U$  infecting an individual from  $S$ .



that, from the aspect of the two curve shapes, the extreme points of the transmission rate and the inflection points of the quarantine percentage coincide approximately in time, for example around 5/10/2020 and 6/29/2020. In order to furthermore study this potential link of dynamics, we denote the quarantine percentage at time instant  $t$  by  $Q(t)$ , whose values are located in  $[0, 100]$ . We require  $Q(t) \in C^2(\mathbb{R})$ . We also need to assume  $\tau(t) \in C^1(\mathbb{R})$ , notice that the inferred  $\tau(t)$  by the proposed method belongs to  $C^\infty(\mathbb{R})$ . Thus, the observation indicates that, when the absolute value of  $\ddot{Q}(t)$  is small, it is very likely that the absolute value of  $\dot{\tau}$  becomes small as well.

Recall that we aim to construct a model in terms of  $\tau(t)$  and  $Q(t)$ , so that the fitted model can be used to predict the future behaviors of  $\tau$  given the public intervention plans. Thus, we propose to adopt the logistic regression (see for example [8]), to predict the probability of the occurrence of event  $\dot{\tau}(t) = 0$  at every time instant  $t$ . The proposed model is:

$$P(\dot{\tau}(t) = 0 \mid \ddot{Q}(t), \dot{Q}(t)) = \text{Sig}\left(\beta^\top \left[1, \ddot{Q}(t), (\ddot{Q}(t))^2, \dot{Q}(t), (\dot{Q}(t))^2\right]\right), \quad (2.5)$$

where  $\text{Sig}(\cdot)$  is the Sigmoid function. In practice, we smooth the data points to obtain the approximating function of  $Q(t)$  (the red curve in Figure 2), so that we can calculate the derivatives. The details on this smoothing and the further training strategies<sup>2</sup> of Model 2.5 are given in the appendix. Therefore, the time instants  $\hat{t}_E$  with high predicted probabilities  $P(\dot{\tau}(\hat{t}_E) = 0 \mid \ddot{Q}(\hat{t}_E), \dot{Q}(\hat{t}_E))$  (for example bigger than 0.9) can be considered as the predicted moments for the transmission rate to reach local extreme values.

Model (2.5) assumes that the likelihood of  $\tau(t)$  to reach its local extreme values at time  $t$  depends on whether the government is changing the public intervention policies at that moment. To distinguish the impacts of changes between different public intervention policies, for example:

- from decelerating (accelerating) to accelerating (decelerating) the reinforcement of intervention,
- from accelerating (decelerating) to decelerating (accelerating) the relaxation of intervention,

we consider  $\ddot{Q}(t)$ ,  $\dot{Q}(t)$ , and  $(\dot{Q}(t))^2$  as dependent variables as well in the model. Note that we intentionally avoid quantitative models of  $\tau(t)$ , such as ordinary differential equation of  $\tau(t)$ , or regression model. Indeed, we have tested these ways of modelling<sup>3</sup>. However, the testing results imply that the quantitative dependency of  $\tau$  and  $Q$  can be very complicated. This brings to inevitable prediction errors. These errors will moreover be amplified in the retrieved  $CR(t)$ , when passing the predicted  $\tau$  through the SIRU model.

### 2.3 Prediction of cumulative daily cases

Recently, many works have considered the forecasting of cumulative daily cases, for example [10, 17, 18, 25]. However, some of them propose the methods only valid for the prediction over the first wave. At this point, we especially refer to the works in [15, 18, 22] adopt the exponential smoothing models with errors and trends [9, Chapter 7] to extrapolate the  $CR(t)$  trend to obtain the prediction. A big advantage of exponential smoothing methods is that they can be used indifferently for the forecasting of any time interval. Nevertheless, since their predicted curves, namely the extrapolated trends, have the simple forms with little parametrization: exponential in [22] for EST(M,M,N) model, linear in [15] for EST(A,A,N) model, the performant prediction interval is very limited. Starting from this point, we propose to use nonlinear function with adequate number of parameters to first fit the trend, and then extrapolate it with an additional control. We consider polynomials, because its analytic facility enables us to relate the predicted behavior of  $\tau$  to the prediction of  $CR(t)$  through SIRU model.

<sup>2</sup>The event of this logistic regression model is  $\dot{\tau} = 0$ , which is the turn points of disease transmission rate. However, there are only a few turn points, for example, approximately 6 – 7 turn points in Figure 2 (left panel). Thus, we propose the particular strategies to compose large enough and balanced training set.

<sup>3</sup>We fit the models with 80% of the historical data, and evaluate the prediction performance with the rest 20%.



To avoid Runge's phenomenon associated to the polynomial approximation, especially the oscillation at the end of fitting interval, we sample the Chebyshev nodes in practice to fit the polynomial. The use of Chebyshev nodes can reduce the oscillation. Moreover, we consider the shape control of the polynomial, especially in the trend extrapolation part. We propose to fit the polynomial under the constraint given by the predicted  $\tau(t)$  behavior. Namely, we require the optimal polynomial to have the consistent characteristics so that its deduced transmission rate reaches the extreme values around the previously predicted moments. Meanwhile we would like the optimal polynomial to be as similar as possible as  $CR(t)$  in the fitting interval. The performance of the resulting predictor polynomial has been significantly improved, where it recovers the  $CR(t)$  values precisely for an ongoing month, as shown in Section 3. We formalize the proposed method in the following optimization problem.

$$\begin{aligned} \widehat{CR} = \arg \min_{P(\cdot; \Theta) \in P_m} \frac{1}{N} \sum_{i=1}^N (CR(t_i) - P(t_i; \Theta))^2, \\ \text{subject to: } (\tau'_p(\hat{t}_E; \Theta))^2 \leq \lambda_0. \end{aligned} \quad (2.6)$$

In the above problem,  $P_m$  is the family of polynomials of order  $m$ ,  $CR(t_i)$ ,  $i = 1, \dots, N$  are the data points of fitting interval,  $\lambda_0$  is a pre-given positive hyperparameter,  $\tau_p(t; \Theta)$  is the deduced transmission rate defined by the SIRU model (2.4) where  $\widehat{CR}(t)$  is given as  $P(t; \Theta)$ ,  $\hat{t}_E$  is the predicted moment when the transmission rate reaches the first extreme value after time instant  $t_N$ . The constraint on the one hand addresses the oscillation problem of polynomial approximation, on the other hand, transfers the future information of transmission rate to the  $CR(t)$  predictor. Note that, we propose to control the magnitude of  $\tau_p(t; \Theta)$  at  $\hat{t}_E$  instead of the equality constraint  $\tau_p(\hat{t}_E; \Theta) = 0$ , so as to reduce the impact of prediction error in the preceding logistic regression.

When  $\lambda_0$  is tunable, Problem (2.6) is equivalent to the following formulation:

$$\widehat{CR} = \arg \min_{P(\cdot; \Theta) \in P_m} \frac{1}{N} \sum_{i=1}^N (CR(t_i) - P(t_i; \Theta))^2 + \lambda (\tau'_p(\hat{t}_E; \Theta))^2, \quad (2.7)$$

where  $\lambda > 0$  is the hyperparameter. Problem (2.6) is a classical composition of optimization problem for learning models, with a data term and a regularization term which aims to address the ill-posedness of the original problem or/and to endow the additional characteristics of the optimizer.  $\lambda$  controls the influence of regularization term. The greater  $\lambda$  is, the smaller  $\tau'_p(\hat{t}_E; \Theta^*)$  will be.

Recall the comments at the end of Section 2.2, compared to forecasting  $\tau(t)$  and use its deduced  $CR(t)$  values as future prediction, forecasting  $CR(t)$  directly as in Problem (2.6) with more accurate  $\tau(t)$  information will avoid the error accumulation in the SIRU model, hence lead to a more performant  $CR(t)$  forecasting.

We now provide the explicit formula of  $\tau'_p(t)$  in Problem (2.6). When  $CR(t)$  is given as the polynomial  $P(t; \Theta) \in P_m$ , solving directly the SIRU model gives the corresponding transmission rate:

$$\tau'_p = \frac{\ddot{I} + v\dot{I}}{(I + U)S} - \frac{(\dot{I} + vI)\dot{S}}{(I + U)S^2} - \frac{(\dot{I} + vI)(\dot{I} + \dot{U})}{(I + U)^2S}, \quad (2.8)$$

where

$$\begin{aligned} I &= \frac{1}{v_1} \dot{P}, \quad \dot{I} = \frac{1}{v_1} \ddot{P}, \\ S &= -I - \frac{v}{v_1} P + c_s, \quad \dot{S} = -\dot{I} - vI, \\ U &= \frac{v_2}{\eta} \sum_{k=0}^{m-1} \frac{(-1)^k}{\eta^k} I^{(k)} + c_u \exp(-\eta t), \\ \dot{U} &= -v_2 \sum_{k=1}^{m-1} \frac{(-1)^k}{\eta^k} I^{(k)} - \eta c_u \exp(-\eta t). \end{aligned} \quad (2.9)$$

Thus,  $\tau'_p$  can be essentially expressed in terms of  $P(t; \Theta)$ . Note that, to determine the constant  $c_s$ , usually we only need one function value  $S(t_s)$ . However, we would like to fully use the training data, and make the estimated model  $S$  as general as possible. Thus, we employ the least square estimation to evaluate constant  $c_s$  as:

$$\hat{c}_s := \arg \min_c \frac{1}{N} \sum_{i=1}^N \left( S(t_i) - \left( -I(t_i; \Theta) - \frac{v}{v_1} P(t_i; \Theta) + c \right) \right)^2,$$

where  $S(t_i)$ ,  $i = 1, \dots, N$ , are the solutions of System (2.4) evaluating at time  $t_i$ . Thus

$$\hat{c}_s = \frac{1}{N} \sum_{i=1}^N \left( S(t_i) + I(t_i; \Theta) + \frac{v}{v_1} P(t_i; \Theta) \right) \quad (2.10)$$

is also a function of  $\Theta$ . Similarly, the least square estimation of  $c_u$  is

$$\hat{c}_u = \frac{1}{N} \sum_{i=1}^N \left( U(t_i) \exp(\eta t_i) - \frac{v_2}{\eta} \sum_{k=0}^{m-1} \frac{(-1)^k}{\eta^k} I^{(k)}(t_i; \Theta) \exp(\eta t_i) \right). \quad (2.11)$$

Thus, we inject the terms in (2.10) and (2.11) into the Formula (2.9). In Section 3, we will illustrate the performance of the proposed predictor  $\widehat{CR}$  using the official COVID-19 data from the government of Chile.

### 3 Numerical evidence

In this numerical study, we consider the evolution of the COVID-19 pandemic in Chile for the period from March 2020 to December 2020, due to the availability of quarantine percentage information. We obtain the  $CR(t)$  data from the daily reported new cases as its cumulative sum, and the quarantine percentage from [19, 20, 21]. We fix  $f = 0.3$ ,  $v = 1/7$ ,  $n = 1/7$ , and  $S_0 = 19458310$  (total population of Chile) throughout the experiments. We use the first 20  $CR$  observations to fit the exponential growing and calculate the initial data. The initial data supplementing System (2.4) are:  $t_0 = -0.6951$ ,  $I_0 = 7.1934$ , and  $U_0 = 1.5945$ . We first show the estimation results from the methods proposed in Section 2.1. The  $CR(t)$  data and its approximation by kernel smoother  $\widehat{CR}(t)$  is given in Figure 3, with the corresponding  $\hat{I}(t)$  given in Figure 4.

The estimations of  $R(t)$ ,  $U(t)$  and  $S(t)$  as the solution of System (2.4) are given in Figures 5 and 6. Using these estimations, the inferred transmission rate has been shown in Figure 2, which has a consistent interpretation with respect to the quarantine percentage data.

Next, we show the result of Logistic regression. We use the inferred  $\tau(t)$  data and the quarantine percentage data until 9/02/2020 to train the logistic model (2.5). Then the fitted model is used to forecast the probability of occurrence of  $\tau = 0$  from September to December with the corresponding quarantine percentage data. We compare the prediction result in Figure 7 with the “true”  $\tau(t)$  data, where the blue curve is the predicted probability. We can see that for the training data, the model has successfully predicted its extreme points such as the one in the beginning of May 2020, and the one at the end of June 2020. For the test data, the model forecasts that around 11/01/2020 and 12/18/2020, there will likely appear extreme points of  $\tau(t)$ , while it predicts in October, it will be almost impossible to appear extreme points.

We now use these predicted moments  $\hat{t}_E$  to derive the predictor of  $CR(t)$  as proposed in Equation (2.7), and compare their values with the true data values. To evaluate the performance of predictor, especially to examine the improvement brought by the information of future  $\tau$  which is well predicted from quarantine percentage, we set the fitting interval  $t_1, \dots, t_N$  at least half a month earlier than the predicted extreme point of  $\tau(t)$ , and fit it with a polynomial without shape control as well as a polynomial with the proposed shape control by  $\hat{t}_E$ . For  $\hat{t}_E = 11/01/2020$ , we set the fitting interval as  $t_1 = 8/13/2020$  to  $t_N = 10/12/2020$  in Equation (2.7). While for  $\hat{t}_E = 12/08/2020$ , we test

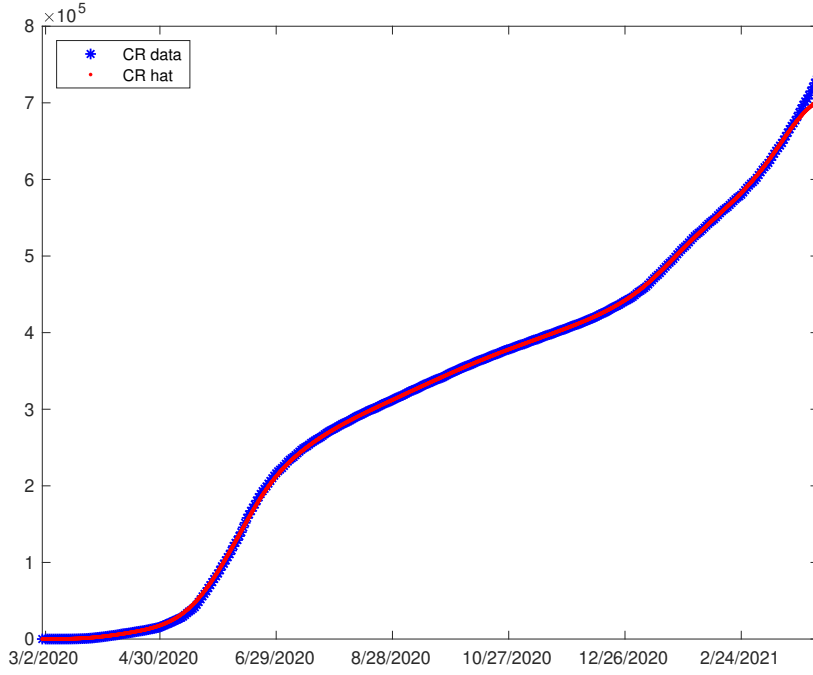


Figure 3: Nonlinear approximation of  $CR(t)$  data. The kernel smoother used here writes as:  $\widehat{CR}(t) = \frac{\sum_i z_i(t)CR(t_i)}{\sum_i z_i(t)}$ , where  $z_i(t) = \exp(-\frac{(t-t_i)^2}{128})$ .

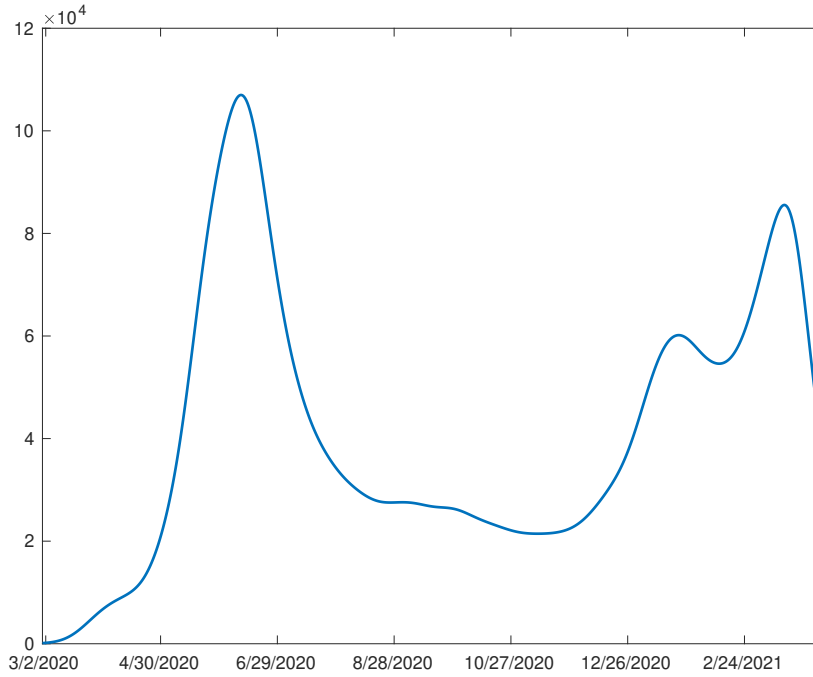


Figure 4: Estimation of  $I(t)$  based on Equation (2.2) from the estimation of  $CR(t)$  in Figure 3.

two fitting intervals, one spanning from  $t_1 = 9/02/2020$  to  $t_N = 11/01/2020$ , the other from  $t_1 = 9/22/2020$  to  $t_N = 11/21/2020$ . The order for all polynomials are fixed as  $m = 4$ . We try 3  $\lambda$  values for each fitting intervals:  $10^{36}$ ,  $50^{36}$ , and  $10^{37}$ . The numerical evidence of our methods are provided in Figure 8 and Figure 9. In both figures, the blue curves refer to the real  $CR$  data while the red curves refers to the predictions from the proposed predictor, where their thin part corresponds to the fitting interval, and their thick part refer to the forecasting of a month. The green curve is the forecasting from the polynomial without shape control, which is fitted on the same interval.

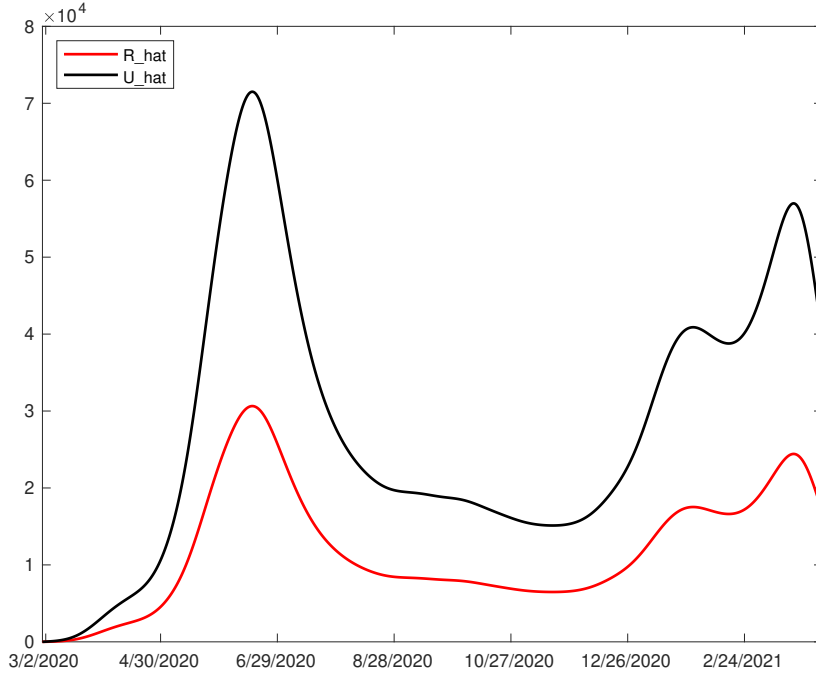


Figure 5: Estimation of  $R(t)$  and  $U(t)$  based on System (2.4).

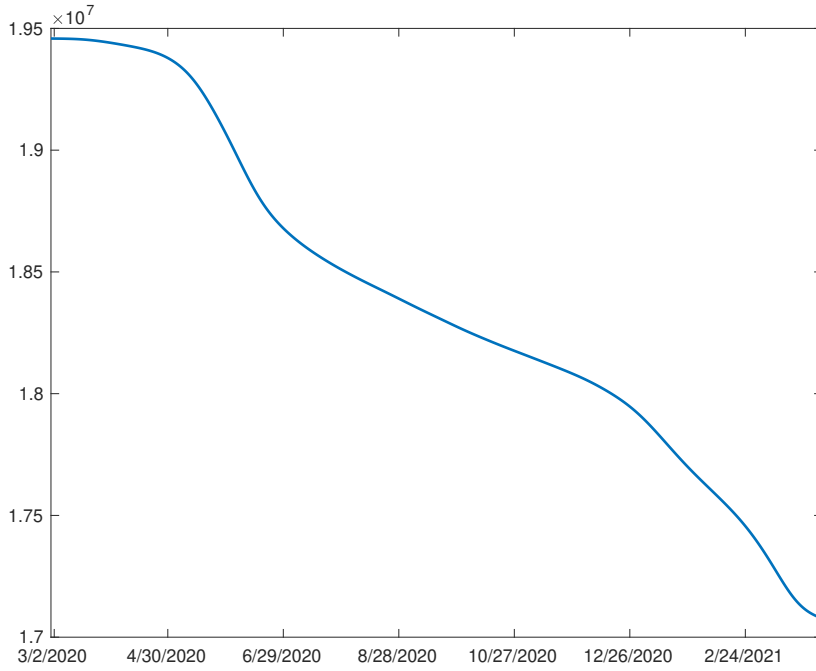


Figure 6: Estimation of  $S(t)$  based on System (2.4).

In Figure 8, the plots shows the one-month forecasting of  $CR$  from the proposed method incorporating the information of the forecasted extreme point  $\hat{t}_E = 12/08/2020$ , with different weights  $\lambda$ . We can see that in general, by incorporating the  $\tau$  information, the forecasted values have been improved in all three  $\lambda$  cases, especially when  $\lambda = 10^{37}$ , the proposed method gives the perfect forecasting.

In Figure 9, the plots shows the one-month forecasting of  $CR$  from  $\hat{t}_E = 12/08/2020$  with the other fitting and predicting intervals. We can see that in this case, the best forecasting result is given by  $\lambda = 50^{36}$ . We would also like to report the result in the bottom subfigure of Figure 9, where the  $\lambda$  value is set too large for this fitting interval. In this case, since the weight of regularization loss is greater, the optimization problem (2.7) needs to search the polynomials of smaller regularization

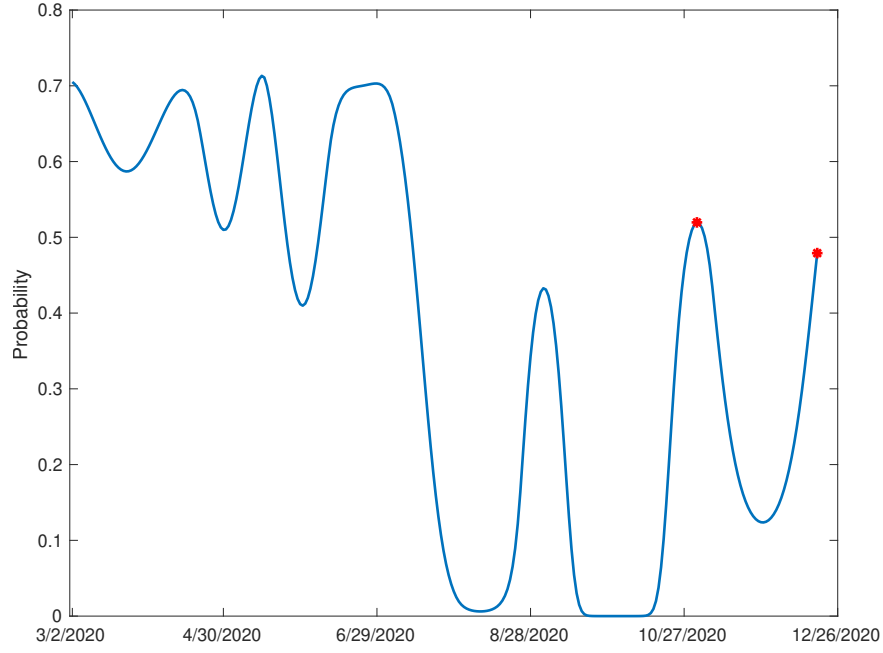


Figure 7: The blue curve is the prediction of the probability of  $\tau(t) = 0$  based on  $Q(t)$  and  $\dot{Q}(t)$ . The model is fitted using the data until 9/02/2020. The two red points denote the predicted future time instants  $\hat{t}_E$ , with the high predicted probabilities  $P(\tau(\hat{t}_E) = 0 \mid \ddot{Q}(\hat{t}_E), \dot{Q}(\hat{t}_E))$ . They are on the dates 11/01/2020 and 12/18/2020.

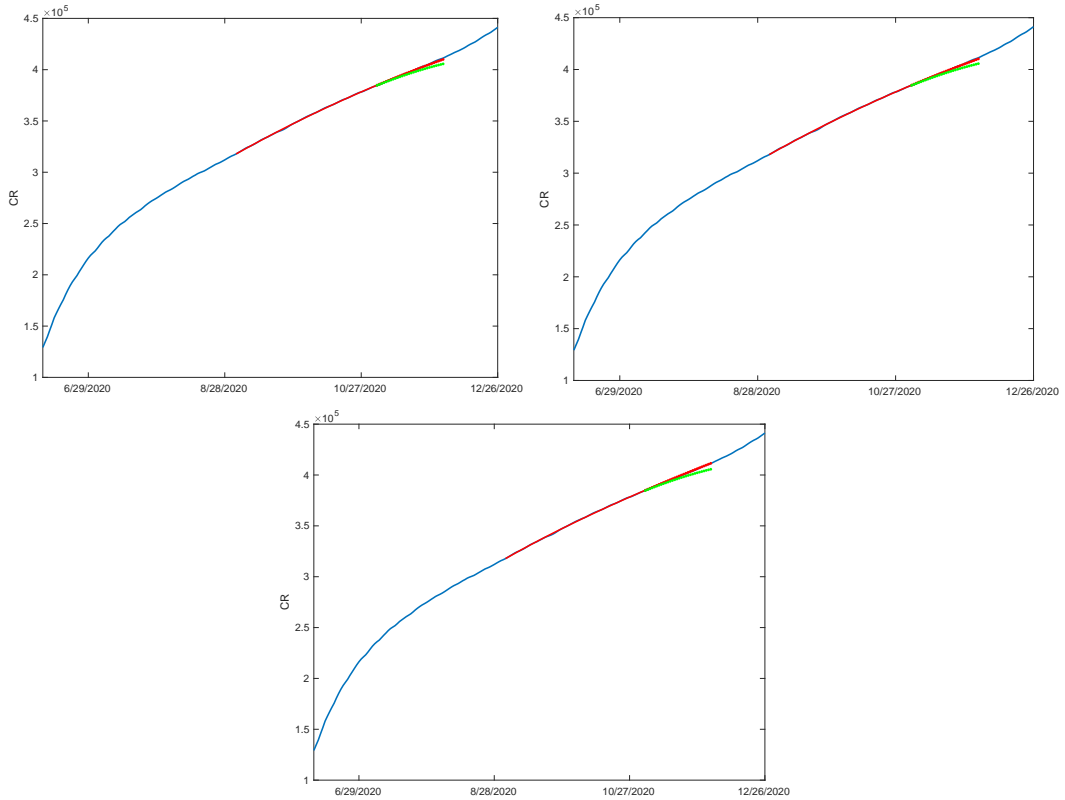


Figure 8:  $\widehat{CR}$  predictor with fitting interval  $t_1 = 9/02/2020$  to  $t_N = 11/01/2020$ ,  $m = 4$ ,  $\lambda = 10^{36}$ (left),  $50^{36}$ (right),  $10^{37}$ (bottom), and  $\hat{t}_E = 12/08/2020$ .

loss  $\tau'_p(\hat{t}_E)^2$ . Such polynomials may have in return relatively larger data term loss. Thus, the optimal

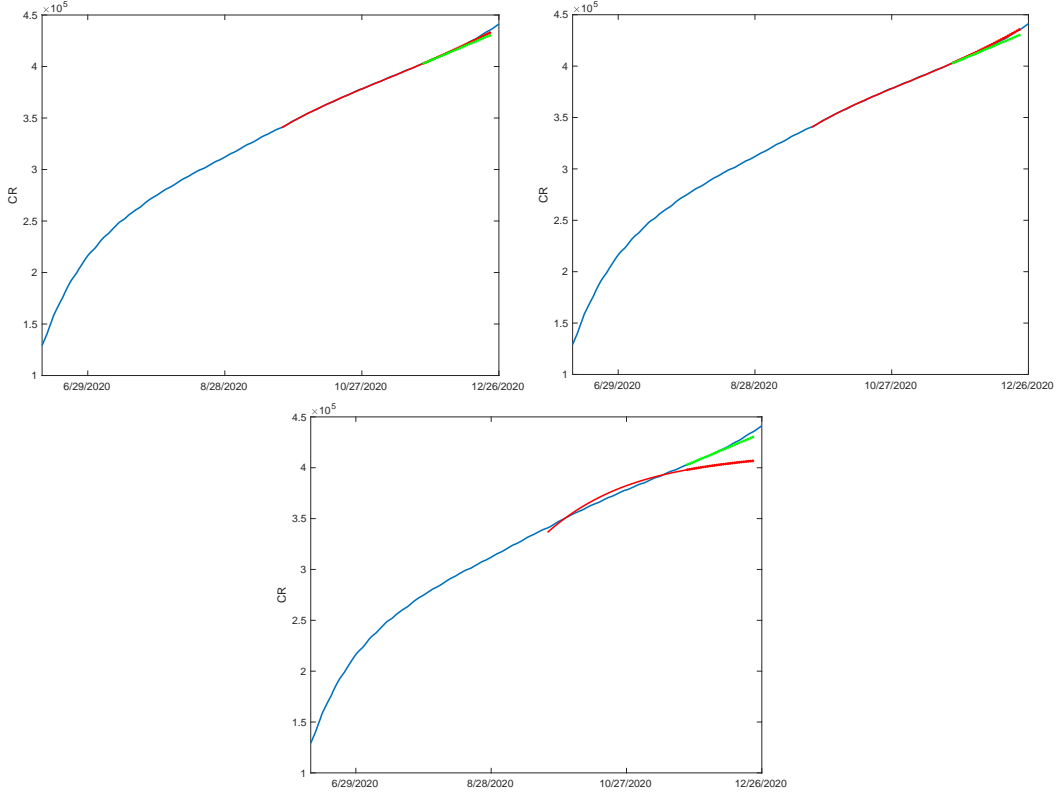


Figure 9:  $\widehat{CR}$  predictor with fitting interval  $t_1 = 9/22/2020$  to  $t_N = 11/21/2020$ ,  $m = 4$ ,  $\lambda = 10^{36}$ (left),  $50^{36}$ (right),  $10^{37}$ (bottom), and  $\hat{t}_E = 12/08/2020$ .

polynomial has a worse fitting performance.

Lastly, Figures 10 shows the forecasting performance of the  $CR$  predictor with  $\hat{t}_E = 11/01/2020$ . We can see that, in this case, the regularization terms have very little influence on the polynomial shapes, with all the forecasting from the proposed predictor overlapping the forecasting of the polynomial without shape control. The possible reason can be that, the polynomial without shape control has already a good prediction performance, namely, a low data term loss, meanwhile the  $\lambda$  values are relatively low for this fitting interval.

## 4 Conclusion

In this paper, we firstly propose a novel way to infer the transmission rate based on the nonparametric estimation. This proposed method has solved the problem that, with multiple epidemic waves, it is very difficult to find an accurate parametric form of transmission rate which can recover the true  $CR(t)$  data. It has also considerably increased the use efficiency of the available data, instead of only using it in the hyperparameter tuning. The inferred transmission rate function enables us to furthermore establish a more sophisticated model between the epidemic and the government control. Thus, the extra government control information, which is the quarantine percentage in our case, can be used to improve the prediction of  $CR(t)$ . The numeric results have shown that the proposed  $CR(t)$  predictor has a promising performance in terms of both accuracy and the efficient prediction interval, which can reach one month in our experiments.

## 5 Conflict of interest

The authors declare there is no conflict of interest.

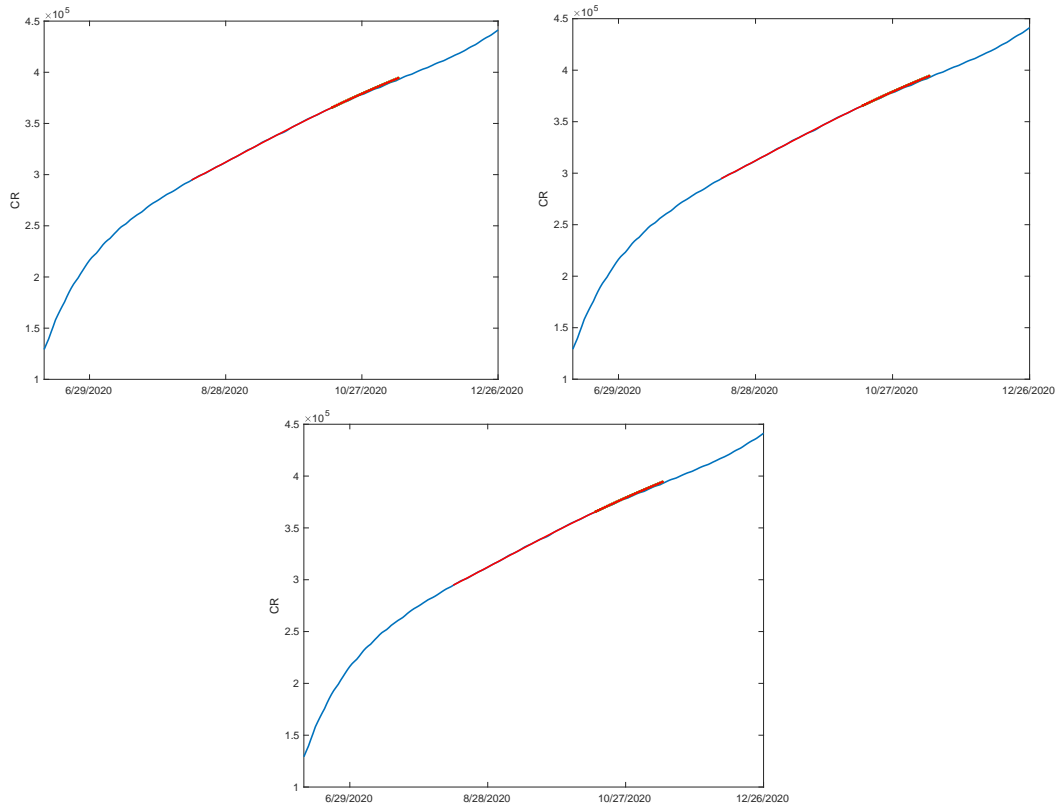


Figure 10:  $\widehat{CR}$  predictor with fitting interval  $t_1 = 8/13/2020$  to  $t_N = 10/12/2020$ ,  $m = 4$ ,  $\lambda = 10^{36}$ (left),  $50^{36}$ (right),  $10^{37}$ (bottom), and  $\hat{t}_E = 11/01/2020$ .

## References

- [1] Gajardo, Á., and Müller, H. G. 2021. "Point process models for COVID-19 cases and deaths". *Journal of applied statistics* 1-16.
- [2] Carroll, C., Bhattacharjee, S., Chen, Y., Dubey, P., Fan, J., Gajardo, A., Zhou, X., Müller, H.-G. and Wang, J. L. 2020. "Time dynamics of COVID-19". *Scientific reports* 10(1), 1-14.
- [3] Yang, Z., Zeng, Z., Wang, K., Wong, S., Liang, W., Zanin, M., Liu P., Cao, X., Gao, Z., Mai Z., Liang, J., Liu, Z., Li, S., Li, Y., Ye, F., Guan, W., Yang, Y., Li, F., Luo, S., Xie, Y., Liu, B., Wang Z., Zhang, S., Wang, Y., Zhong, N., and Jianxing He, 2020. "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions". *Journal of thoracic disease* 12(3), 165.
- [4] Hou C., Chen J., Zhou Y., Hua L., Yuan J., He S., Guo Y., Zhang S., Jia Q., Zhao C., Zhang J., Xu G., and Jia E. 2020. "The effectiveness of quarantine of Wuhan city against the Corona Virus Disease 2019 (COVID-19): A well-mixed SEIR model analysis". *Journal of medical virology* 92(7), 841-848.
- [5] Anastassopoulou, C., Russo, L., Tsakris, A., and Siettos, C. 2020. "Data-based analysis, modelling and forecasting of the COVID-19 outbreak". *PloS one* 15(3), e0230405.
- [6] Fanelli, D., and Piazza, F. 2020. "Analysis and forecast of COVID-19 spreading in China, Italy and France". *Chaos, Solitons & Fractals* 134, 109761.
- [7] Griette, Q., Demongeot, J., and Magal, P. 2021. "What can we learn from COVID-19 data by using epidemic models with unidentified infectious cases?". *Mathematical Biosciences and Engineering* 19(1), 537-594.



- [8] Hastie, T., Tibshirani, R. and Friedman, J. 2009. *The elements of statistical learning: data mining, inference, and prediction*. 2<sup>nd</sup> edition. Springer Series in Statistics, New York, USA.
- [9] Hyndman, R. J. and Athanasopoulos, G. 2018. *Forecasting: principles and practice*. 2<sup>nd</sup> edition. OTexts: Melbourne, Australia.
- [10] Hu, Z., Ge, Q., Li, S., Jin, L. and Xiong, M. 2020. *Artificial intelligence forecasting of COVID-19 in China*. Preprint, ArXiv:2002.07112.
- [11] Liu, Z., Magal, P., Seydi, O. and Webb, G., 2020. "Understanding unreported cases in the COVID-19 epidemic outbreak in Wuhan, China, and the importance of major public health interventions". *Biology* 9 (3): 50.
- [12] Liu, Z., Magal, P. and Webb, G. 2021. "Predicting the number of reported and unreported cases for the COVID-19 epidemics in China, South Korea, Italy, France, Germany and United Kingdom". *Journal of theoretical biology* 509: c110501.
- [13] Navas, A. and Vergara-Hermosilla, G. 2020. *On the dynamics of the Coronavirus epidemic and the unreported cases: the Chilean case*. Preprint, ArXiv:2006.02632.
- [14] Nikparvar, B., Rahman, M., Hatami, F. and Thill, J.C. 2021 "Spatio-temporal prediction of the COVID-19 pandemic in US counties: modeling with a deep LSTM neural network". *Scientific reports* 11: 1–12.
- [15] Rustam, F., Reshi, A., Mehmood, A., Ullah, S., On, B., Aslam, W. and Choi, G.S. 2020. "COVID-19 future forecasting using supervised machine learning models". *IEEE access* 8: 101489–101499.
- [16] Shawaqfah, M. and Almomani, F. 2021. "Forecast of the outbreak of COVID-19 using artificial neural network: Case study Qatar, Spain, and Italy" *Results in Physics* 27: 104484.
- [17] Suyel, N., Dhamodharavadhani, S. and Rathipriya, R. 2021. "Nonlinear neural network based forecasting model for predicting COVID-19 cases". *Neural Processing Letters* 5: 1–21.
- [18] Torrealba-Rodriguez, O., Conde-Gutiérrez, R. A. and Hernández-Javier, A. L. 2020. "Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models". *Chaos, Solitons & Fractals* 138: 109946.
- [19] Official data about COVID-19 from the Chilean government, (*in Spanish*), Online; , <https://www.gob.cl/coronavirus/cifrasoficiales/> (accessed January 2, 2022).
- [20] Official data about COVID-19 from the Ministry of Health, Chilean government, (*in Spanish*), <https://www.minsal.cl/nuevo-coronavirus-2019-ncov/> (accessed January 2, 2022).
- [21] Official data about COVID-19 from the Ministry of Science, Technology, Knowledge, and Innovation, Chilean government, (*in Spanish*), <https://github.com/MinCiencia/Datos-COVID19/> (accessed January 2, 2022).
- [22] Petropoulos, F. and Makridakis, S. 2020. "Forecasting the novel coronavirus COVID-19". *PloS one* 15: e0231236.
- [23] Webb, G., Magal, P. and Seydi, O. 2020. "Unreported cases for age dependent COVID-19 outbreak in Japan". *Biology* 9 (6): 132.
- [24] Xiang, Y., Jia, Y., Chen, L., Guo, L., Shu, B. and Long, E. 2021. "COVID-19 epidemic prediction and the impact of public health interventions: A review of COVID-19 epidemic models". *Infectious Disease Modelling* 6: 324–342

- [25] Yousaf, M., Zahir, S., Riaz, M., Hussain, S. and Shah, K. 2020. "Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan". *Chaos, Solitons & Fractals* 138: 109926

## Appendix

In the following we explain how to approximate a function  $Q \in C^2(\mathbb{R})$  from the data points of quarantine percentage (see the right plot in Figure 2). To primarily filter out the intense fluctuations, we first subsample the data points by a frequency of 25, then we use the kernel smoother with kernel  $\exp(-\frac{(t-t')^2}{18})$  on the subsampled points to generate the smooth approximation of data, denoted as  $\hat{Q}$ , which is the red curve in Figure 2. To furthermore obtain the functions of  $\dot{Q}$  and  $\ddot{Q}$ , we fit  $\hat{Q}$  with the cubic spline, and use the derivatives of the fitted spine as  $\dot{Q}$  and  $\ddot{Q}$ . Similarly, we fit the inferred transmission rate with cubic spline and use its derivative to obtain function  $\dot{\tau}$ .

To train Logistic model (2.5), we need to provide the balanced set which consists in the moments  $t_i^0$  whose  $\tau(t_i^0)$  are extremas as well as the  $t_j^1$  whose  $\tau(t_j^1)$  are not extremas. We also need the predictor quarantine percentage function values at these points, namely  $\dot{Q}(t_i^0)$ ,  $\ddot{Q}(t_i^0)$ ,  $\dot{Q}(t_j^1)$ , and  $\ddot{Q}(t_j^1)$ . Given  $\dot{\tau}$ , we use the bisection method to find its roots so as to determine the moments  $t_i^0$ . The root finding results show there are 4 extreme points before 9/02/2020, which are 4/11/2020, 5/07/2020, 6/30/2020, and 8/30/2020. We also consider their six nearest neighbouring dates as  $t_i^0$ , to increase the training samples also to compensate any errors during the calculation. For the moments  $t_i^1$ , we choose the dates 3/22/2020, 4/24/2020, 6/03/2020, 7/30/2020, and their six nearest neighbouring dates.

The code implementing the numeric studies in this work is available on: <https://github.com/yiyej>.