Machine learning-based modelling and forecasting of covid-19 under the temporally varying public intervention in the Chilean context

Yiye Jiang, Gaston Vergara-Hermosilla

MACHINE LEARNING-BASED MODELLING AND FORECASTING OF COVID-19 UNDER THE TEMPORALLY VARYING PUBLIC INTERVENTION IN THE CHILEAN CONTEXT

Yiye Jiang[1*] and Gaston Vergara-Hermosilla[2]

*Corresponding author
[1]Institut de Mathématiques de Bordeaux, Talence, France
[2]Department of Electronic Engineering, National University of Ireland Maynooth, Ireland
E-mails: yiyejiang93@gmail.com / gaston.v-h@outlook.com

Abstract

Objectives: In this paper, we aim to propose the efficient and interpretable models to study and predict the evolution of COVID-19 pandemic, which is in particular not limited to the setting the epidemic ends with one wave, and the public intervention is varying. In this setting, we firstly propose a novel method to infer the transmission rate on the top of the recently introduced SIRU model. Next, we establish the link between the transmission rate and the public intervention. Lastly, we propose a forecasting model for the cumulative daily reported cases, $CR(t)$.

Methods: Firstly, we incorporate the nonparametric estimation into SIRU system to obtain a precise reconstruction of the transmission rate dynamic. We then rely on the logistic regression to set up a prediction model for this dynamic given the variable public intervention. Lastly, we propose a regularized the polynomial approximation which considers the SIRU structure and the future public intervention to predict the cumulative daily reported cases.

Results: We demonstrated the proposed methods with the Chilean data. The inferred transmission rate exhibits the consistent thus interpretable dynamic with respect to the variable public intervention. To evaluate the performance of the $CR(t)$ predictor, we considered two time points after 9 months since the first breakout of the epidemic, and obtain the one-month prediction starting from these time points. The proposed predictor is able to give the accurate prediction, sometimes even with imperfect regularization hyperparameter.

Conclusions: This work provides a methodology to combine the machine learning methods with the compartmental models, to enhance the later with data. Firstly, the novel inference method for the transmission rate has made the SIRU model applicable to the setting with multiple waves and the varying public intervention. Moreover, the proposed technique is transferable to other compartmental models. Secondly, the proposed predictor of $CR(t)$ can consider the varying public intervention, and thus is able to provide valid prediction for one month long, which is a great advantage compared with its competitor in literature.

Keywords: SIRU model, transmission rate, cumulative daily reported cases, nonparametric estimation

Introduction

The mathematical systems modelling epidemiological phenomena have played a protagonist role in making decisions and controlling the current coronavirus epidemic around the world. The mathematical perspective can help to understand the underlying pattern of epidemic dynamic as well as the potential roles of government measures in the disease propagation. In this spirit, a timely epidemic model which can predict the development of the epidemic is favored by the public health authorities. Many approaches have been proposed to predict the COVID-19 epidemic evolution, see the overview paper [17]. In addition to the compartmental models, some works have considered the machine learning-based methods to leverage the information in the data, see [3, 9, 7]. In this paper, we combine the two points of views. We rely on the methods from the machine learning domain to exploit efficiently the information in the data set. Then we transfer this data knowledge to a compartmental model, so as to benefit the expert knowledge. We would like especially to consider the existence of unreported cases, which is a significant issue in the epidemiological analysis due to the low testing capacity of a country and the asymptomatic patients. Among several related models in literature, the recently introduced SIRU model [4] has been successfully used to describe the evolution of the epidemic during the first wave in various countries, such as China, South Korea, Italy and France. Therefore, in this paper, we rely on the SIRU model to explain the epidemic structure of COVID-19. On the other hand, we aim to analyse the dynamic of epidemic with respect to the varying public intervention. Nevertheless, the reference papers [4, 5, 16] on the SIRU model propose to add an additional equation on the dynamic of the transmission rate, with respect to the public intervention to consider its effects. This presumed parametric model is a strong hypothesis for real data of the various dynamics, which considers only one wave and the fixed public intervention intensity. Additionally the increased number of hyper-parameters in the system makes the follow-up estimation complicated and costly. Thus, our primary objective is to propose an initiative inference method which incorporates the nonparametric estimation in machine learning into the SIRU model, which can make the typical compartmental model fully benefit the data to therefore give a precise reconstruction of transmission rate dynamic. We then aim to link transmission rate dynamic to the public intervention changes to study the effects, which serve for the ultimate goal of this work, that is to propose a forecasting approach for $CR(t)$ taking into account the intervention plans for the future. Such forecasting models are of great interest for the decision makers.

In the rest of this section, we recall the readers the principles of SIRU model in the following. In Section , we propose a nonparametric method to estimate the transmission rate $\tau(t)$ as well as $I(t), R(t), S(t)$ and the unreported daily case $U(t)$. In Section , we rely on the logistic regression to predict the moment to appear an extreme value of $\tau(t)$ from the temporal variable public intervention $Q(t)$. In Section , we consider the prediction of the cumulative daily new cases. we propose the regularized polynomial approximation as the predictor for cumulative daily reported cases $CR(t)$. It is defined as a minimizer of an optimization problem which considers the historical data of $CR(t)$, meanwhile the predicted $\tau(t)$ dynamic which comes from the future information of $Q(t)$. Finally, in Section , we present our numeric results.

The SIRU model describes the dynamic of a pandemic situation by considering a system of ordinary differential equations (ODEs) involving four different states, which are denoted by S, I, R and U, and represent the susceptible individuals, infected individuals who do not yet have symptoms, reported infected individuals, and unreported infected individuals, respectively. This model can be presented by considering the following diagram flux [16, 4]:

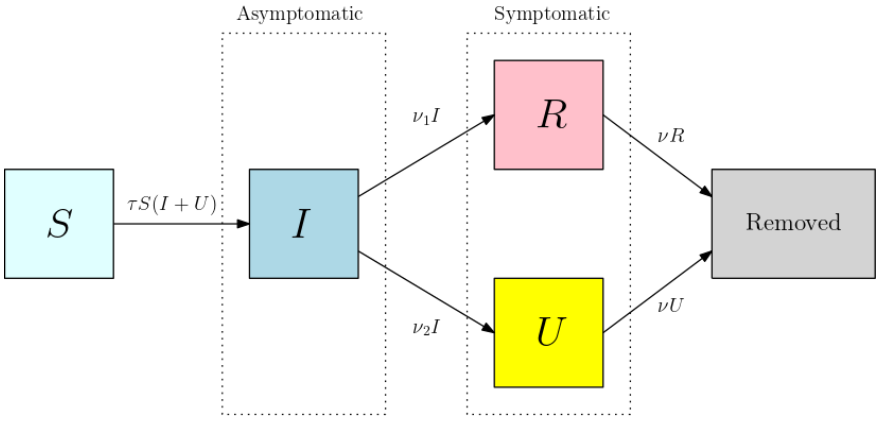

Figure 1: Diagram flux associated with the SIRU model.

The set of ODEs associated with the diagram flux read as:

$$\begin{cases} S'(t) = -\tau(t)S(t)(I(t) + U(t)) \\ I'(t) = \tau(t)S(t)(I(t) + U(t)) - \nu I(t) \\ R'(t) = \nu_1 I(t) - \eta R(t) \\ U'(t) = \nu_2 I(t) - \eta U(t) \end{cases} \tag{0.1}$$

SIRU model (0.1) is supplemented by initial data

$$S(t_0) = S_0 > 0, I(t_0) = I_0 > 0, R(t_0) = 0 \text{ and } U(t_0) = U_0 \geq 0. \tag{0.2}$$

The parameters attached with the model are listed in the table 1. In particular, note that $\nu = \nu_1 + \nu_2$. In addition, all the parameters that are considered $\tau, \nu, \nu_1, \nu_2, \eta$ are positive.

Methodology

Nonparametric estimation of the transmission rate

To help to predict the epidemic evolution, we would like to have a better understanding of the dynamic of transmission rate $\tau(t)$. More specifically, we wish to infer the values of transmission rate given the $CR(t)$ records. We require the inferred values to be reliable, in the sense that they are able to recover the historical $CR(t)$ records with good accuracy, when plugging them into the SIRU model. Specific to the Chile case, we

Symbol	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.
$\tau(t)$	Transmission rate of the disease.
$1/\nu$	Average time during which the infectious asymptomatic individuals remain in 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 become unreported infected individuals.
$1/\eta$	Average time during which an infected individual presents symptoms.

Table 1: Parameters and initial conditions of the model.

are additionally interested in inferring the dynamic of transmission rate during a long period, where there appears several epidemic waves and possibly changing public intervention policies. To cope with the gradually implemented public intervention case [6] adopts the piece-wise parametric functional design of $\tau(t)$ with multiple periods. Even though the large number of parameters are able to capture the complicated dynamic of transmission rate, they still pose the difficulty in estimation. Including [4, 5, 16], all of these cited work tune the parameter values by grid search. When it comes to longer period, the piece-wise parametric model of $\tau(t)$ will result in tremendous work of parameter tuning, when the high accuracy of reconstruction of $CR(t)$ is required. Thus for long term modelling of the transmission rate, we propose to employ non-parametric estimation.

Given the historical data of $CR(t)$ over the period $t_1, ..., t_N$, we propose to make $\tau(t)$ an unknown function, in return, to turn $I(t)$ to a known in SIRU (0.1). 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 $CR(t)$:

$$CR(t) = \nu_1 \int_{t_0}^{t} I(s)ds, \tag{0.3}$$

implies the estimators for $I(t)$ and $I'(t)$ write as

$$\widehat{I}(t) = \widehat{CR}'(t)/\nu_1,\ \widehat{I'}(t) = (\widehat{I})'(t). \tag{0.4}$$

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

$$\begin{cases} S'(t) = -\tau(t)S(t)(\widehat{I} + U(t)), \\ R'(t) = \nu_1\widehat{I} - \eta R(t), \\ U'(t) = \nu_2\widehat{I} - \eta U(t), \\ \widehat{I'}(t) = \tau(t)(\widehat{I} + U(t))S(t) - \nu\widehat{I}. \end{cases} \tag{0.5}$$

The above system is equivalent to:

$$\begin{cases} S'(t) = -\widehat{I'} - \nu\widehat{I}, \\ R'(t) = \nu_1\widehat{I} - \eta R(t), \\ U'(t) = \nu_2\widehat{I} - \eta U(t), \\ \tau(t) = \frac{\widehat{I'}+\nu\widehat{I}}{(\widehat{I}+U(t))S(t)}. \end{cases} \tag{0.6}$$

System (0.6) 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 [4]. It is worth to mention that, to be consistent with the initial values, before applying the nonlinear approximation, we fill the data points of $CR(t)$ generated by its exponential estimation used in the calculation of initial values, for interval $t_0, ..., 0$. Notably, the proposed inference method is valid for any length of period N.

The key point in the above estimation is to choose an admissible nonlinear method. Common nonlinear methods to reconstruct a data curve by a function are polynomial approximation, spline, and kernel smoother, see for example [1]. For Model (0.6), we propose to use the kernel smoother. On one hand, the polynomial approximation will usually introduce oscillation, which will furthermore be amplified after taking derivative, thus the final estimated $\tau(t)$ will exhibit multiple local extreme points which will misleading the interpretation of true dynamic contained in raw data. On the other hand, the $I'(t)$ expression given by the SIRU model (0.1) implies that $I(t)$ is likely to be a C^∞ function. Thus compared to spline function which is piece-wise polynomials of low order, the kernel smoother with Guassian kernel is preferable.

Logistic regression with the public intervention policies

In this section, we consider the case of the 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 measure 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 furthermore expected to help the prediction of future $CR(t)$.

In Chile, a significant varying public measure is the percentage of national population in quarantine. Such data is used in the work of [6] 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 web page [13] . We especially smooth the data points of quarantine percentage 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 the inferred transmission rate $\tau(t)$ obtained from the preceding section. We can observe that, from the aspect of the curve

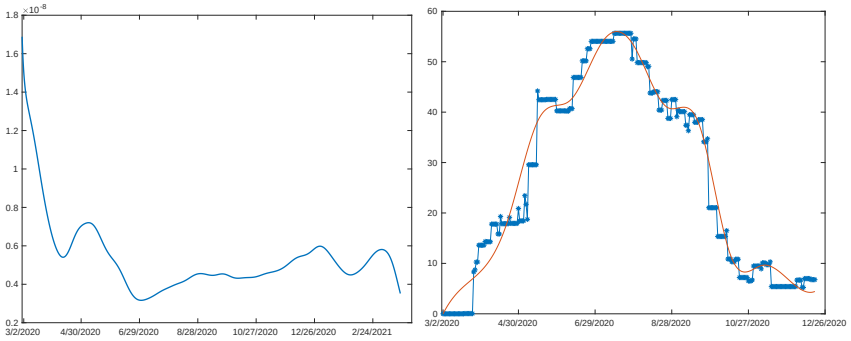

Figure 2: Inferred transmission rate (left) from the Chilean COVID data, Percentages of the Chilean population in quarantine (right). The red curve in the right subfigure shows the smoothing curve of the discrete data points.

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 introduce function $Q(t)$ to represent the quarantine percentage at time points 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 for the one of $\dot{\tau}$ to become small.

Recall that we aim to construct a model in terms of τ and Q, so that the fitted model can be used to predict the future behaviors of τ given the public intervention plans. More importantly, the predicted dynamic of τ can then help the forecasting of $CR(t)$, through the SIRU model. Thus, we propose to adopt the logistic regression [1], to predict the probability of the occurrence of event $\dot{\tau}(t) = 0$ for every moment. The

proposed model is given in Model (0.7).

$$P(\dot{\tau}(t) = 0|\ddot{Q}(t), \dot{Q}(t)) = Sig(\beta^\top[\mathbf{1}, \ddot{Q}(t), (\ddot{Q}(t))^2, \dot{Q}(t), (\dot{Q}(t))^2]), \tag{0.7}$$

where $Sig(\cdot)$ 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 smoothing and the model training is given in Appendix . Thus, the time instants $\hat{t}_E$, with high predicted probabilities $P(\dot{\tau}(\hat{t}_E) = 0|\ddot{Q}(\hat{t}_E), \dot{Q}(\hat{t}_E))$ (for example higher than 0.9) can be considered as the predicted moments for the transmission rate to reach local extreme values.

Model (0.7) 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 equation of $\tau(t)$, or regression model. Indeed, we have tested these ways of modelling [1]. However, the testing results imply that the quantitative dependency of τ 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 τ through the SIRU model.

Prediction of cumulative daily new cases

Recently, many works have considered the forecasting of cumulative new cases, for example [3][11][18][10]. However, most of them only study the phase before the appearance of second wave. By contrast, in this paper, we would like to propose the prediction method, which is suitable for any epidemic period. We also notice, many works only justify their methods in terms of fitting performance on the training data. Thus, in additional to the goodness of fitting, we will especially focus on reporting the prediction performance of the proposed method on the test set.

Instead of proposing specific quantitative models as in for example [11], [15][8] adopt the exponential smoothing models with errors and trends [2, Chapter 7] to extrapolate the $CR(t)$ trend to obtain the prediction. Exponential smoothing methods can be performed indifferently for the forecasting starting from any time instant. Nevertheless, since their predicted curves, namely the extrapolated trends, have the simple forms with little parametrization, exponential in [15] for EST(M,M,N) model, linear in [8] for EST(A,A,N) model, the performant prediction interval is very limited. Starting

[1] We fit the models with 80% of the historical data, and evaluate the prediction performance with the rest 20%.

from this point, we propose to use nonlinear function with adequate number of parameters to first fit the trend, and then extrapolate it with additional control. We consider polynomials, because its analytic facility enables us to relate the predicted behavior of τ to the predicted $CR(t)$ through the SIRU model. To avoid the Runge's phenomenon related to the polynomial approximation, especially the oscillation at the end of fitting interval. We sample the Chebyshev nodes in practice to fit the polynomial, which 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 for $\tau(t)$. 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 . We formalize the proposed method in the following optimization problem.

$$\begin{aligned} \widehat{CR} = \underset{P(\cdot;\Theta)\in \mathrm{P}_m}{\arg\min} \frac{1}{N}\sum_{i=1}^{N}(CR(t_i) - P(t_i;\Theta))^2, \\ \text{subject to: } (\tau_p'(\hat{t}_E;\Theta))^2 \leq \lambda_0. \end{aligned} \tag{0.8}$$

In the above problem, P_m is the family of polynomials of order m, $CR(t_i),\ i = 1, ..., N$ are the data points of fitting interval, λ_0 is a pre-given positive hyperparameter, $\tau_p(t;\Theta)$ is the deduced transmission rate defined by the SIRU model (0.6) where $\widehat{CR}(t)$ is given as $P(t;\Theta)$, $\hat{t}_E$ is the predicted moment after t_N for the extreme value of transmission rate. 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.

It is equivalent to Problem (0.9), when λ_0 in Problem (0.8) is tunable.

$$\widehat{CR} = \underset{P(\cdot;\Theta)\in \mathrm{P}_m}{\arg\min} \frac{1}{N}\sum_{i=1}^{N}(CR(t_i) - P(t_i;\Theta))^2 + \lambda(\tau_p'(\hat{t}_E;\Theta))^2, \tag{0.9}$$

where $\lambda > 0$ is the hyperparameter. Problem (0.8) 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. λ controls the influence of regularization term. The greater λ is, $\tau_p'(\hat{t}_E;\Theta^*)$ will be smaller.

Recall the comments at the end of Section , compared to forecasting $\tau(t)$ and use its deduced $CR(t)$ values as future prediction, forecasting $CR(t)$ directly as in Problem (0.8) 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 (0.8). When $CR(t)$ is given as the polynomial $P(t;\Theta) \in P_m$, solving directly the SIRU model gives the correspond-

ing transmission rate:

$$\tau_p' = \frac{\ddot{I} + v\dot{I}}{(I+U)S} - \frac{(\dot{I} + vI)\dot{S}}{(I+U)S^2} - \frac{(\dot{I} + vI)(\dot{I} + \dot{U})}{(I+U)^2 S}, \tag{0.10}$$

where

$$\begin{aligned} I &= \frac{1}{v_1}\dot{P},\ \dot{I} = \frac{1}{v_1}\ddot{P}, \\ S &= -I - \frac{v}{v_1}P + c_s, \dot{S} = -\dot{I} - vI, \\ U &= \frac{v_2}{\eta}\sum_{k=0}^{m-1}\frac{(-1)^k}{\eta^k}I^{(k)} + c_u\exp(-\eta t), \\ \dot{U} &= -v_2\sum_{k=1}^{m-1}\frac{(-1)^k}{\eta^k}I^{(k)} - \eta c_u\exp(-\eta t); \end{aligned} \tag{0.11}$$

Thus, τ_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 $S(t_i),\ i = 1, ..., N$, are the solutions of System (0.6) evaluating at the time t_i. Thus

$$\hat{c}_s = \frac{1}{N}\sum_{i=1}^{N}\left(S(t_i) + I(t_i;\Theta) + \frac{v}{v_1}P(t_i;\Theta)\right) \tag{0.12}$$

is also a function of Θ. Similarly, the least square estimation of c_u is

$$\hat{c}_u = \frac{1}{N}\sum_{i=1}^{N}\left(U(t_i)\exp(\eta t_i) - \frac{v_2}{\eta}\sum_{k=0}^{m-1}\frac{(-1)^k}{\eta^k}I^{(k)}(t_i;\Theta)\exp(\eta t_i)\right). \tag{0.13}$$

We use Equations (0.12) and (0.13) in Formula (0.11). In Section , we illustrate the performance of proposed predictor $\widehat{CR}$ using the data of Chile.

Numeric results

In this numeric study, we study the epidemic evolution in Chile for the period from March 2020 to December 2020, due to the availability of quarantine percentage information. We obtain the *CR* data from the daily reported new cases as its cumulative sum, and the quarantine percentage from [12, 13, 14] . We fix $f = 0.3$, $v = 1/7$, $n = 1/7$, and $S_0 = 19458310$ (population of Chile) throughout the experiments. We use the first 20 *CR* observations to fit the exponential growing and calculate the initial data, which are $t_0 = -0.6951$, $I_0 = 7.1934$, and $U_0 = 1.5945$. We first show the estimation results

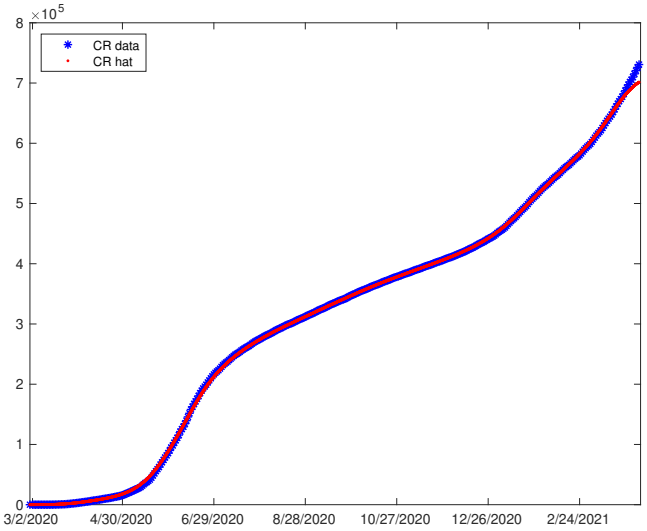

Figure 3: Nonlinear approximation of $CR(t)$ data. The kernel smoother used here writes as: $\widehat{CR}(t) = \frac{\sum_i z_i(t) CR(t_i)}{\sum_i z_i(t)}$, where $z_i(t) = \exp(-\frac{(t-t_i)^2}{128})$.

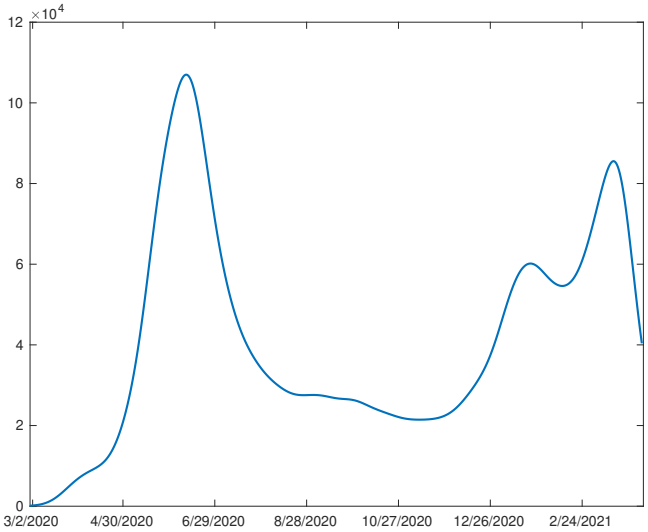

Figure 4: Estimation of $I(t)$ data based Equation (0.4.) from the estimation of $CR(t)$ in Figure 3.

from the methods proposed in Section . The CR data and its approximation by kernel smoother $\widehat{CR}(t)$ is given in Figure 3, with the corresponding $\hat{I}(t)$ is given in Figure 4.

The estimation of $R(t), U(t)$ and $S(t)$ as the solution of System (0.6) 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.

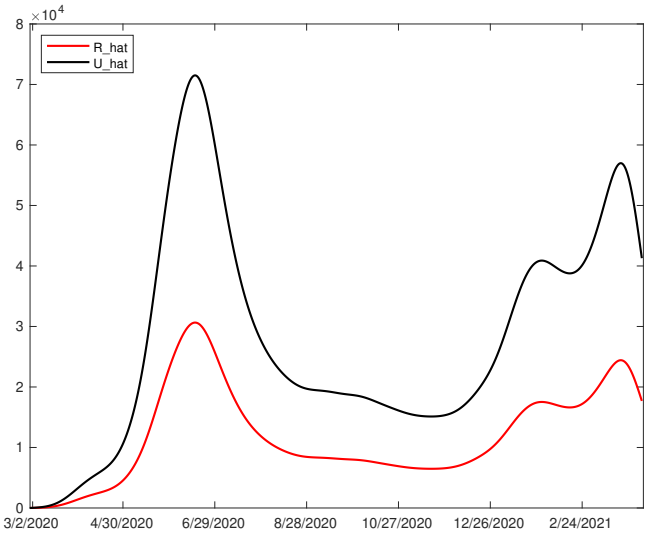

Figure 5: Estimation of $S(t)$ data based System (0.6).

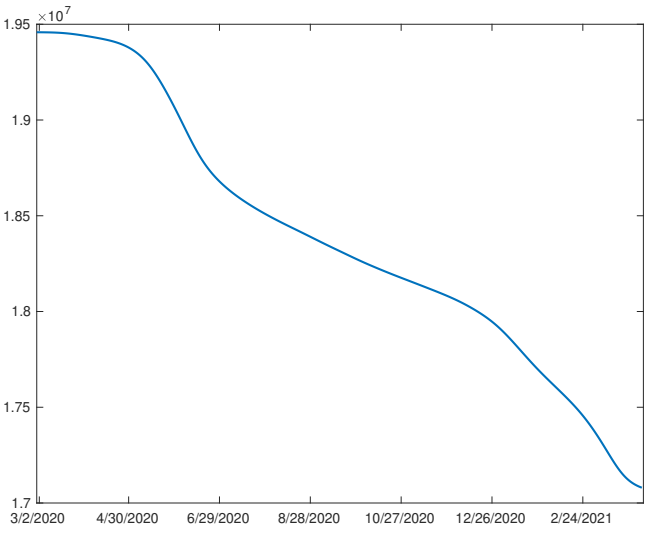

Figure 6: Estimation of $S(t)$ data based System (0.6).

Next, we show the result of Logistic regression. We use the inferred τ data and the quarantine percentage data until $9/02/2020$ to train the logistic model (0.7). Then the fitted model is used to forecast the probability of occurrence of $\dot{\tau} = 0$ during September

to December with the corresponding quarantine percentage data. We compare the prediction result with the true τ data in Figure 7, where the blue curve is the predicted probability. We can see that for the training data, the model successfully predicts 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 τ, while it predicts in October, it will be almost impossible to appear extreme points.

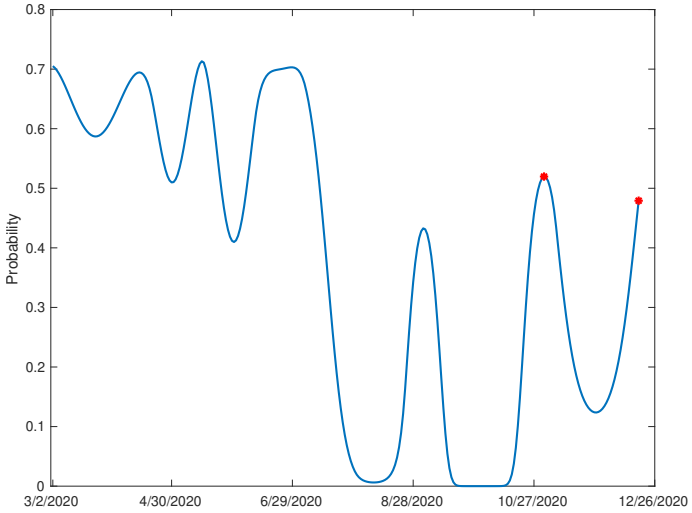

Figure 7: Prediction of the probability of $\dot{\tau}(t) = 0$ based on $Q(t)$ and $\dot{Q}(t)$. The model is fitted using the data until 9/02/2020. The two red points are the predicted future time instants $\hat{t}_E$, with the high predicted probabilities $P(\dot{\tau}(\hat{t}_E) = 0|\ddot{Q}(\hat{t}_E),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

We now use these predicted moments $\hat{t}_E$ to derive the predictor of CR as proposed in Equation (0.9), 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 τ which is well predicted from quarantine percentage, in Equation (0.9), we set the fitting interval $t_1, ..., t_N$ at least half a month further than the predicted extreme point of τ, and fit it with a polynomial without shape control as well as a polynomial with the proposed shape control given by the $\hat{t}_E$ prediction in order to compare the prediction improvement. 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 (0.9). While for $\hat{t}_E = 12/08/2020$, we test 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 λ values for each fitting intervals: $10^{36}, 50^{36}$, and 10^{37}. The numeric results are given in Figures 8 - 16. The blue curve is the true CR values, the red line is the predictions from the proposed predictor, where its thin part corresponds to the fitting interval, and the thick part is the forecasting of a month. The green line is the forecasting from the polynomial without shape control, which is fitted on the same

interval.

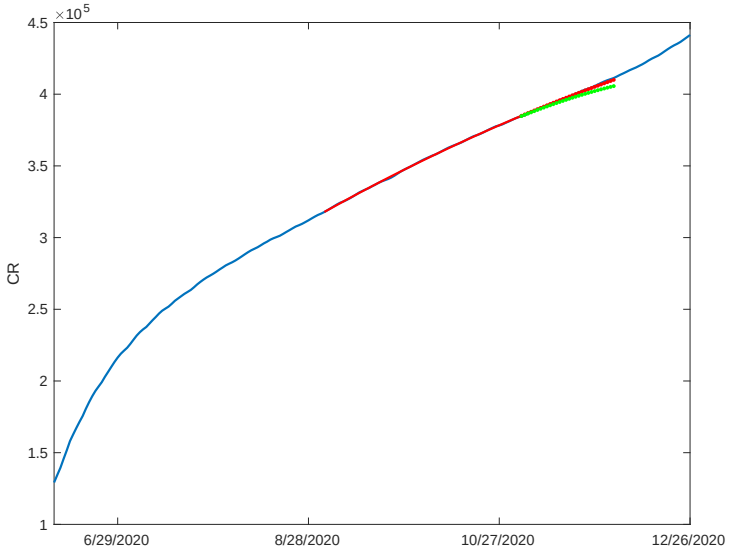

Figure 8: $\widehat{CR}$ predictor with fitting interval $t_1 = 9/02/2020$ to $t_N = 11/01/2020$, $m = 4$, $\lambda = 10^{36}$, and $\hat{t}_E = 12/08/2020$.

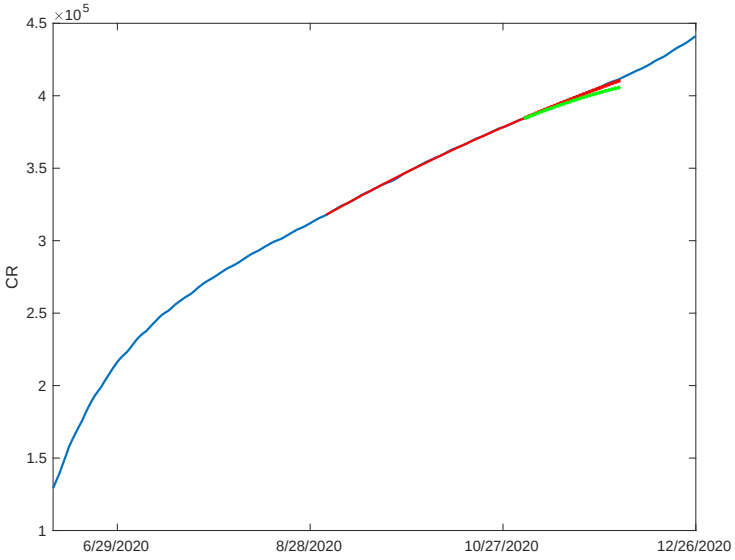

Figure 9: $\widehat{CR}$ predictor with fitting interval $t_1 = 9/02/2020$ to $t_N = 11/01/2020$, $m = 4$, $\lambda = 50^{36}$, and $\hat{t}_E = 12/08/2020$.

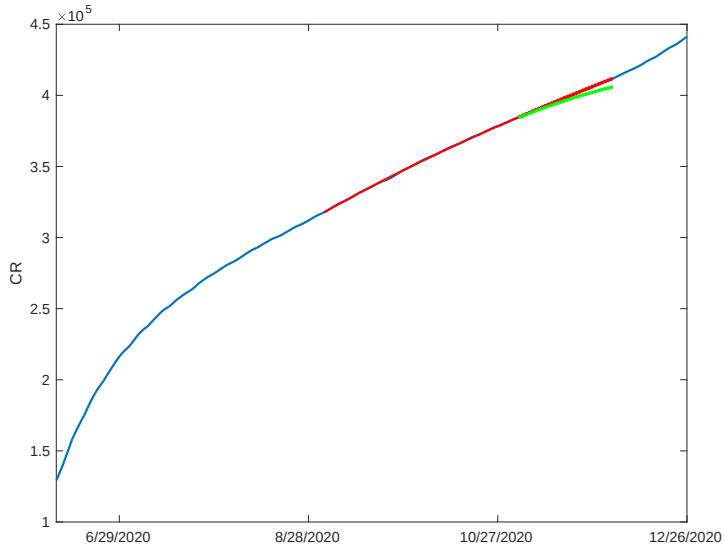

Figure 10: $\widehat{CR}$ predictor with fitting interval $t_1 = 9/02/2020$ to $t_N = 11/01/2020$, $m = 4$, $\lambda = 10^{37}$, and $\hat{t}_E = 12/08/2020$.

Figures 8 to 10 show 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 λ. We can see that in general, by incorporating the τ information, the forecasted values have been improved in all three λ cases, especially when $\lambda = 10^{37}$, the proposed method gives the perfect forecasting as shown in Figure 10.

Figures 11 to 13 show 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 Figure 13, where the λ value is set too large for this fitting interval. In this case, since the weight of regularization loss in the total loss is greater, the optimization problem (0.9) needs to search the polynomials of smaller regularization loss $\tau'_p(\hat{t}_E)^2$ which may however have relatively larger data term loss, to decrease the total loss. Thus, the optimal polynomial has a worse fitting performance.

Lastly, Figures 14 to 16 show 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 predictors 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 λ values are relatively low for this fitting interval.

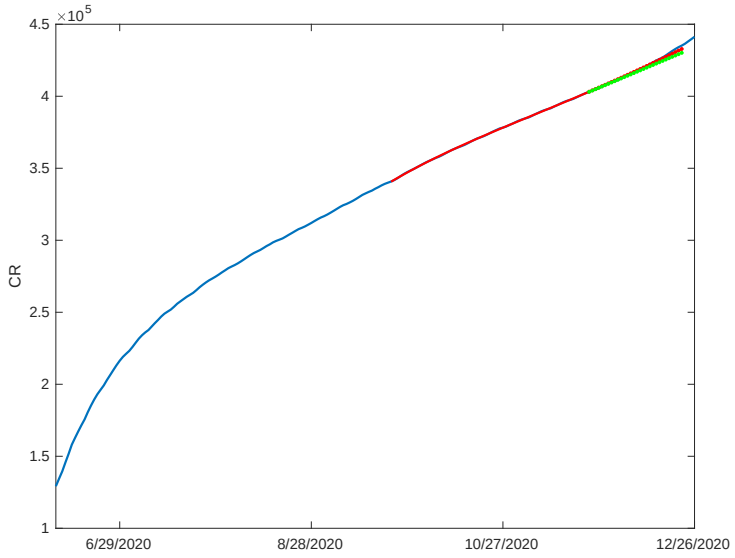

Figure 11: $\widehat{CR}$ predictor with fitting interval $t_1 = 9/22/2020$ to $t_N = 11/21/2020$, $m = 4$, $\lambda = 10^{36}$, and $\hat{t}_E = 12/08/2020$.

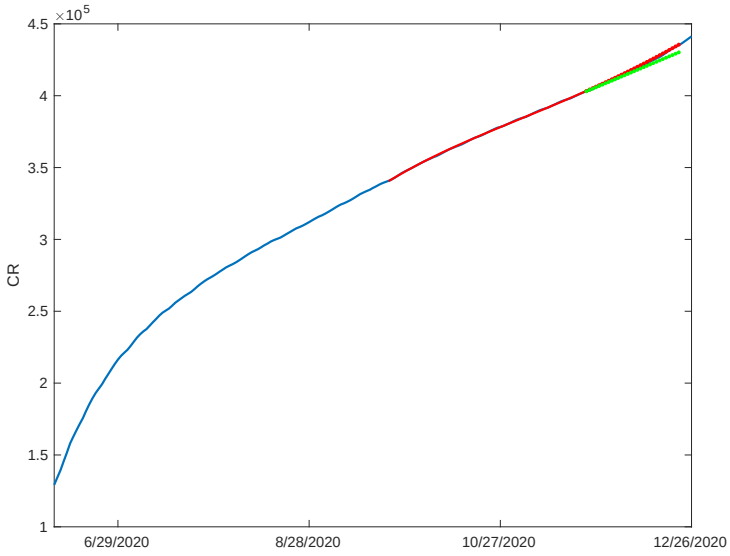

Figure 12: $\widehat{CR}$ predictor with fitting interval $t_1 = 9/22/2020$ to $t_N = 11/21/2020$, $m = 4$, $\lambda = 50^{36}$, and $\hat{t}_E = 12/08/2020$.

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, in

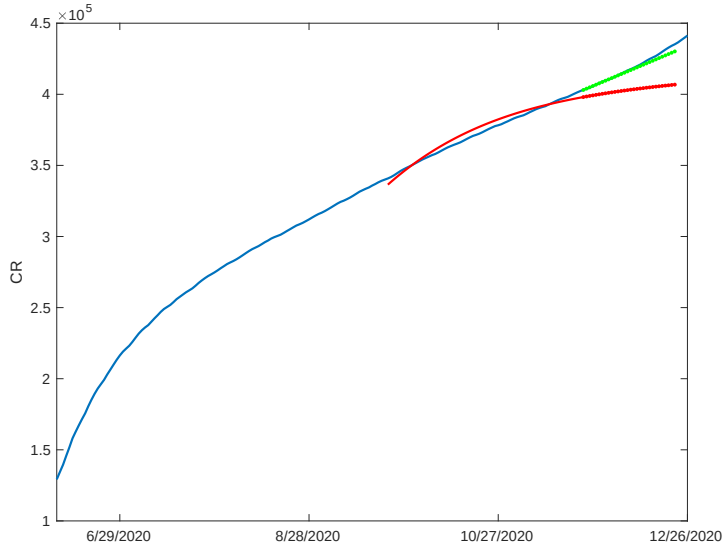

Figure 13: $\widehat{CR}$ predictor with fitting interval $t_1 = 9/22/2020$ to $t_N = 11/21/2020$, $m = 4$, $\lambda = 10^{37}$, and $\hat{t}_E = 12/08/2020$.

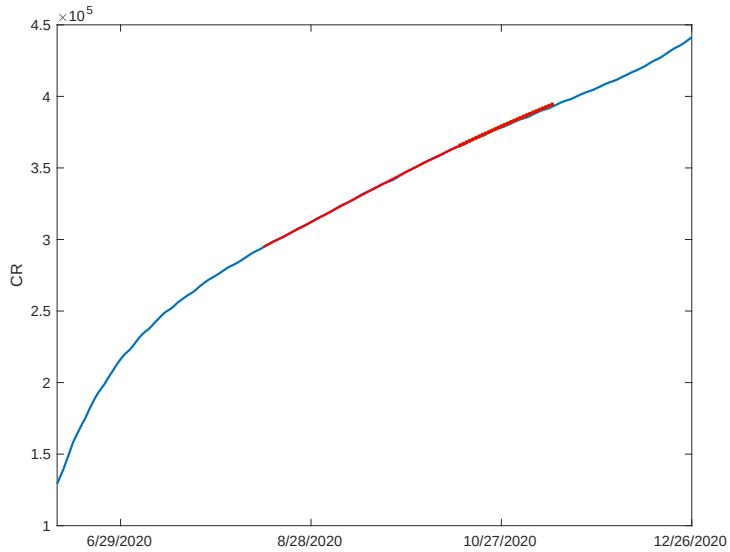

Figure 14: $\widehat{CR}$ predictor with fitting interval $t_1 = 8/13/2020$ to $t_N = 10/12/2020$, $m = 4$, $\lambda = 10^{36}$, and $\hat{t}_E = 11/01/2020$.

long term with the multiple epidemic waves and the changing public intervention, it is very difficult to find a good parametric functional design for transmission rate that can

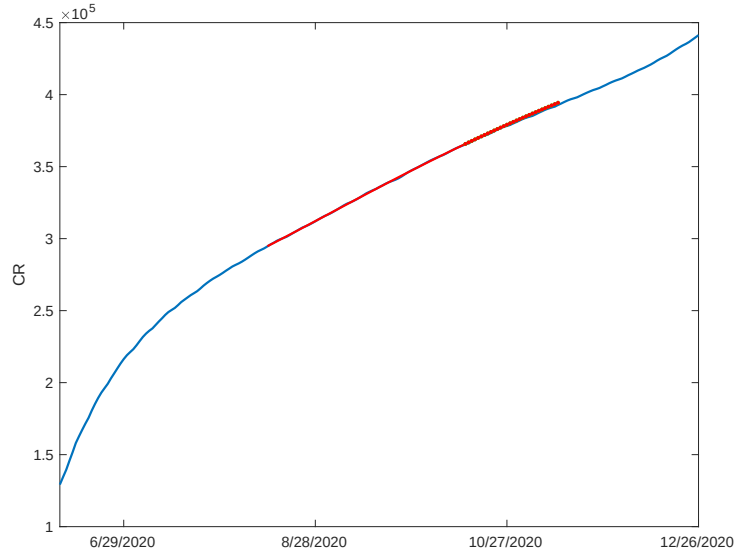

Figure 15: $\widehat{CR}$ predictor with fitting interval $t_1 = 8/13/2020$ to $t_N = 10/12/2020$, $m = 4$, $\lambda = 50^{36}$, and $\hat{t}_E = 11/01/2020$.

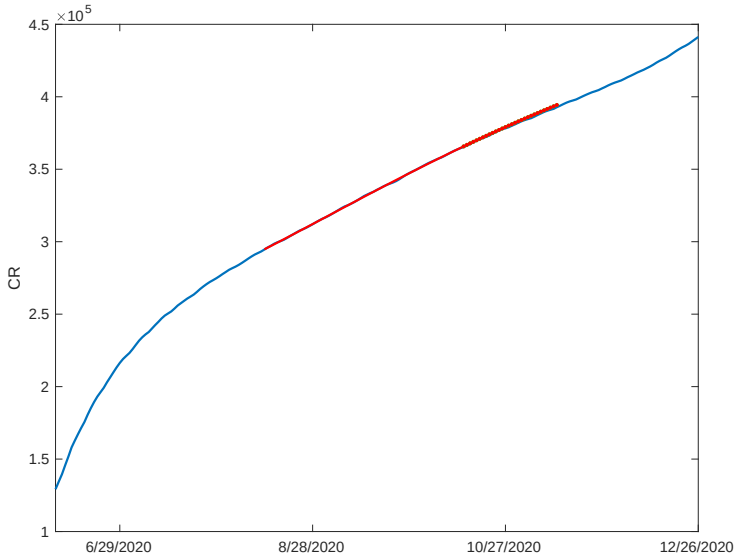

Figure 16: $\widehat{CR}$ predictor with fitting interval $t_1 = 8/13/2020$ to $t_N = 10/12/2020$, $m = 4$, $\lambda = 10^{37}$, and $\hat{t}_E = 11/01/2020$.

recover the true CR data. It has also considerably increased the use efficiency of the available data, instead of only using it in the hyperparameter tuning. The inferred trans-

mission rate function enables us to furthermore establish 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*. The numeric results have shown that the proposed *CR* predictor has a promising performance in terms of both accuracy and the efficient prediction interval, which can reach one month in our experiments.

Conflict of interest

The authors declare there is no conflict of interest.

Acknowledgment

The second author is supported by Science Foundation Ireland under Grant No. 12/RC-PhD/3486 for MaREI, the SFI research centre for energy, climate and marine research and innovation.

References

[1] Hastie, T., Tibshirani, R. and Friedman, J. 2009. *The elements of statistical learning: data mining, inference, and prediction*. 2^{nd} edition. Springer Series in Statistics, New York, USA.

[2] Hyndman, R. J. and Athanasopoulos, G. 2018. *Forecasting: principles and practice*. 2^{nd} edition. OTexts: Melbourne, Australia.

[3] Hu, Z., Ge, Q., Li, S., Jin, L. and Xiong, M. 2020. *Artificial intelligence forecasting of COVID-19 in China*. Preprint, ArXiV:2002.07112.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Official data about COVID-19 from the Chilean government, *(in Spanish)*, Online; , https://www.gob.cl/coronavirus/cifrasoficiales/ (accessed January 2, 2022).

[13] 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).

[14] 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).

[15] Petropoulos, F. and Makridakis, S. 2020. “Forecasting the novel coronavirus COVID-19". *PloS one* 15: e0231236.

[16] Webb, G., Magal, P. and Seydi, O. 2020. “Unreported cases for age dependent COVID-19 outbreak in Japan". *Biology* 9 (6): 132.

[17] 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

[18] 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 A

To approximate the data points of quarantine percentage with a function $Q \in C^2(\mathbb{R})$. 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 (0.7), 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.