Bulletin of Mathematical Biology (2022) 84:57
https://doi.org/10.1007/s11538-022-01012-8

Society for Mathematical Biology

ORIGINAL ARTICLE

A Hypothesis-Free Bridging of Disease Dynamics and Non-pharmaceutical Policies

Xiunan Wang[1,2] · Hao Wang[1] · Pouria Ramazi[3] · Kyeongah Nah[1,4] · Mark Lewis[1,5]

Received: 26 November 2021 / Accepted: 8 March 2022 / Published online: 8 April 2022

Abstract

Accurate prediction of the number of daily or weekly confirmed cases of COVID-19 is critical to the control of the pandemic. Existing mechanistic models nicely capture the disease dynamics. However, to forecast the future, they require the transmission rate to be known, limiting their prediction power. Typically, a hypothesis is made on the form of the transmission rate with respect to time. Yet the real form is too complex to be mechanistically modeled due to the unknown dynamics of many influential factors. We tackle this problem by using a hypothesis-free machine-learning algorithm to estimate the transmission rate from data on non-pharmaceutical policies, and in turn forecast the confirmed cases using a mechanistic disease model. More specifically, we build a hybrid model consisting of a mechanistic ordinary differential equation (ODE) model and a gradient boosting model (GBM). To calibrate the parameters, we develop an "inverse method" that obtains the transmission rate inversely from the other variables in the ODE model and then feed it into the GBM to connect with the policy data. The resulting model forecasted the number of daily confirmed cases up to 35 days in the future in the USA with an averaged mean absolute percentage error of 27%. It can identify the most informative predictive variables, which can be helpful in designing improved forecasters as well as informing policymakers.

✉ Hao Wang
hao8@ualberta.ca

1 Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada

2 Department of Mathematics, University of Tennessee at Chattanooga, Chattanooga, TN 37403, USA

3 Department of Mathematics and Statistics, Brock University, St. Catharines, ON L2S 3A1, Canada

4 National Institute for Mathematical Sciences, Daejeon 34047, Korea

5 Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada

Keywords Hypothesis-free · Non-pharmaceutical policies · COVID-19 · Inverse method · Machine Learning · Generalized boosting model

1 Introduction

The world has experienced a devastating pandemic of COVID-19, a novel coronavirus disease caused by SARS-CoV-2. As of November 16, 2021, the COVID-19 pandemic is still affecting 224 countries and territories, causing about 254,901,115 cases and 5,127,051 deaths worldwide (Worldometers 2021). The first case in the USA was reported on January 23, 2020 (Wikipedia 2021), and the first death in the USA was reported on February 29, 2020 (Worldometers 2021). The confirmed cases and deaths kept increasing in the USA in 2020, making it the epicenter. In the beginning several months of the pandemic, pharmaceutical interventions such as vaccination and drugs are not available, and containing the spread of SARS-CoV-2 largely depends on government policies including school closing, workplace closing, cancellation of public events, restrictions on gatherings, public transport closing, stay at home requirements, international travel controls, public information campaigns, testing, contact tracing, facial coverings, protection of elderly people, etc. (Ritchie et al. 2021). Most of these policies directly affect human mobility which further influence the transmission of the virus. Revealing the quantitative relationship between the transmission rate and policies and human mobilities is critical in forecasting the pandemic.

There has been an overwhelming number of research papers about the transmission dynamics of COVID-19 (e.g., Coletti et al. 2021; Mukandavire et al. 2020; Sun et al. 2020; Liu et al. 2020; IHME 2020; Serina et al. 2021; Calvetti et al. 2020; Ramazi et al. 2021a). Nonetheless, intuitive modeling and accurate forecasting of the spread of COVID-19 remain a challenge. On the one hand, the traditional epidemiological models are fully mechanistic and intuitive. They nicely capture the disease spread yet heavily rely on the transmission rate parameter which in turn depends on variables such as preventive policies and human mobility, whose relation to the disease dynamics is too complex to be accurately modelled mechanistically. Therefore, to forecast the future, the transmission rate is considered constant or piecewise linear, or some restrictive hypothesis is made about its future values. The mechanistic models are, thus, often not competent enough in prediction. On the other hand, the data-based machine learning models are powerful in prediction but typically non-intuitive, and perhaps less reliable, especially if trained with few data instances. We bridge the gap by developing a hybrid model combining a compartmental epidemiological model that captures the disease spread with a time-varying transmission rate and a machine-learning model that links the transmission rate to data on preventive policies whose future values are known a priori. The epidemiological model consists of an ordinary differential equations (ODE) and a machine-learning algorithm—a gradient boosting model (GBM). We use part of the available data to *train* the GBM, by first, developing an "inverse method" that estimates the values of the transmission rate from the other variables of the ODE, and next, fitting the estimated transmission rate values to the policy data using the GBM. The trained hybrid model can then be used to generate predictions of the number of daily confirmed cases by using the future values of the

preventive policies to estimate the transmission rate by the GBM and in turn, the daily cases using the ODE. To examine the role of human mobility on the disease spread, we run a separate series of simulations where in addition to the preventive policies, human mobility data are used to estimate the transmission rate. We apply the model to the case study of COVID-19 in the USA and then find those variables whose inclusion improved the model performance most.

The rest of the paper is organized as follows. Section 2 explains the data used in this study. In Sect. 3, we develop the compartmental epidemiological model for COVID-19 and introduce the inverse method to estimate the transmission rate. In Sect. 4, we formulate the generalized boosting model, show the training and testing results and make predictions of daily confirmed cases using the ordinary differential equation (ODE) model. We also explore the relative importance of each variable in training the model. In Sect. 5, we investigate the prediction performance when mobility and part of the policies are additionally included as the predictor variables. Section 6 provides a brief summary of the method and results as well as suggestions for future work.

2 Data Availability

The data used in this study include the total number of daily confirmed cases of COVID-19 in the USA and policy indices in each state collected from the official website of the flagship project *Our World in Data* of Global Change Data Lab Ritchie et al. (2021) (https://ourworldindata.org/coronavirus), the six categories of human mobility data in the USA from the official website of Google Team (2021) (https://www.google.com/covid19/mobility/), and deaths, recovered and active cases in the USA from the worldometer website Worldometers (2021) (https://www.worldometers.info/coronavirus/country/us/), on a daily basis from April 4, 2020, to December 19, 2020.

We obtain the time-series indices for school closing (denoted by C1), workplace closing (C2), cancel public events (C3), restrictions on gatherings (C4), close public transport (C5), stay at home requirements (C6), restrictions on internal movement (C7), international travel controls (C8), public information campaigns (H1), testing policies (H2), contact tracing (H3), facial coverings (H6), and protection of elderly people (H8) in the USA by taking an average of the corresponding policy indices over all the 50 US states as well as Washington D.C., weighted by their populations. Here the policies beginning with "C" represent containment policies, whereas those beginning with "H" represent health policies. The emergency investment in healthcare (H4) and investment in vaccines (H5) are not available. Since we focus on the pre-vaccination case in this paper, we do not take into account the vaccination delivery policy (H7) either. Human mobility data include changes of mobility trends (%) in retail and recreation (M1), grocery and pharmacy (M2), parks (M3), transit stations (M4), workplaces (M5), and residential (M6), compared to the baseline level (0). These policy and mobility data are shown in Fig. 1.

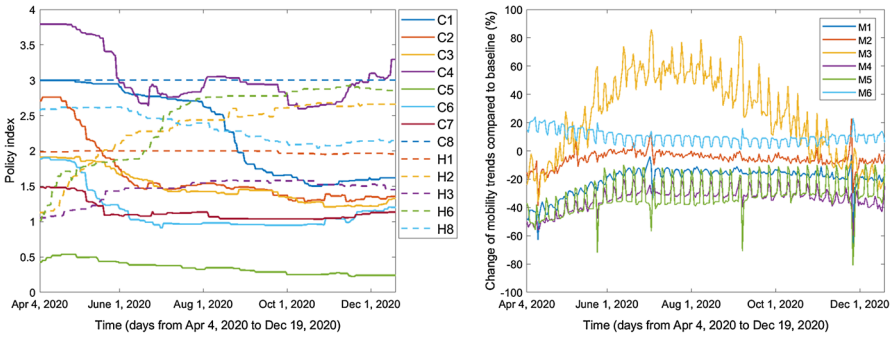


Fig. 1 Policy and mobility data in the USA from April 4, 2020, to December 19, 2020 (Color figure online)

3 Mechanistic Model and Inverse Method

We use a susceptible-exposed-infectious-recovered framework to model the transmission dynamics of COVID-19. The model divides the human population into five compartments: the susceptible (denoted by S), the exposed (E), the symptomatic infected (I), the asymptomatic infected (A), and the recovered individuals (R). Our SEIAR model is described by the following system of differential equations:

$$\begin{aligned}
\frac{\mathrm{d}S(t)}{\mathrm{d}t} &= -\frac{\beta(t)S(t)(I(t)+\theta_{\mathrm{E}}E(t)+\theta_{\mathrm{A}}A(t))}{N},\\
\frac{\mathrm{d}E(t)}{\mathrm{d}t} &= \frac{\beta(t)S(t)(I(t)+\theta_{\mathrm{E}}E(t)+\theta_{\mathrm{A}}A(t))}{N} - \delta E(t),\\
\frac{\mathrm{d}I(t)}{\mathrm{d}t} &= (1-p)\delta E(t) - (\mu(t)+r_{\mathrm{I}})I(t),\\
\frac{\mathrm{d}A(t)}{\mathrm{d}t} &= p\delta E(t) - r_{\mathrm{A}}A(t),\\
\frac{\mathrm{d}R(t)}{\mathrm{d}t} &= r_{\mathrm{I}}I(t) + r_{\mathrm{A}}A(t).
\end{aligned} \tag{1}$$

The susceptible individuals enter the incubation period if they are infected with SARS-CoV-2. The incubation period has an average duration of $1/\delta$ days. Upon the incubation period ends, the exposed individuals enter either the symptomatic infected compartment (I) or the asymptomatic infected compartment (A), depending on whether symptoms occur or not. We assume that a proportion p of all the infectives are asymptomatic and hence the symptomatic infections account for a proportion of $1-p$. The transmission rate is $\beta(t)$. As exposed individuals and asymptomatic infected individuals can also spread the virus at reduced probabilities compared to symptomatic infected individuals (Zhang et al. 2020), we assume that the relative transmissibility of exposed and asymptomatic infected individuals are θ_{E} and θ_{A}, respectively ($0 \le \theta_{\mathrm{E}} \le 1, 0 \le \theta_{\mathrm{A}} \le 1$). The disease induced death rate is $\mu(t)$. It takes an average of $1/r_{\mathrm{I}}$ days and $1/r_{\mathrm{A}}$ days for symptomatic and asymptomatic infected individuals to recover, respectively.

We obtain the values of the constant parameters from the literature. The total US population N is taken as 331,449,281 which is estimated on April 1, 2020, by US Census Bureau (Bureau 2021). The incubation period could vary greatly among patients. The current official estimated range for the incubation period is 2 to 14 days. However, more recent reports show that the incubation period can extend beyond 14 days (https://www.news-medical.net/news/20201025/COVID-19-incubation-period-potentially-much-longer-than-previously-thought.aspx). We take $\delta = 1/14$ per day. The time to recover from COVID-19 may vary from 1.5 to 30 days among different patients Kumar et al. (2021), depending on their infection severity, overall health and age. We assume the average recovery period for both symptomatic and asymptomatic infected individuals is 14 days, which leads to $r_{\mathrm{I}} = r_{\mathrm{A}} = 1/14$ per day. Asymptomatic infections contribute substantially to community transmission together with presymptomatic ones. Even if asymptomatic infections transmit poorly, presymptomatic and asymptomatic cases together comprise at least 50% of the force of infection (Subramanian et al. 2021). We set $p = 0.7$ to represent that approximately 70% of the infections are asymptomatic in our model. We estimate the relative transmissibilities of exposed and asymptomatic infected individuals as $\theta_{\mathrm{E}} = 0.1$ and $\theta_{\mathrm{A}} = 0.5$, respectively. The values and interpretations of all constant parameters are given in Table 1.

The time-varying death rate is estimated using the following formula where $\mu[k]$ represents the disease induced death rate of symptomatic infected individuals on day k:

$$\mu[k] = \frac{\#\text{new deaths on day } k}{\#\text{currently infected individuals on day } k}.$$

Motivated by Kong et al. (2015), Pollicott et al. (2012), we create an inverse method to estimate the time-varying transmission rate. The starting point is to derive the time series $E(t)$ by utilizing the notification data. The real incidence data will be between $\delta E(t)$ and $(1 - p)\delta E(t)$, but most asymptomatic individuals are not tested due to unawareness of their infections. Although some special individuals such as sports players or frontline health workers may be forced to be tested, this accounts for a

Table 1 Parameter interpretation and values

Parameter	Interpretation	Value
$\beta(t)$	Transmission rate	See Fig. 3
N	Total population of USA	331,449,281
θ_{E}	Relative transmissibility of exposed individuals	0.1
θ_{A}	Relative transmissibility of asymptomatic individuals	0.5
$1/\delta$	Incubation period	14 days
p	Proportion of asymptomatic infections	0.7
$\mu(t)$	Death rate of symptomatic infected individuals	See Fig. 2
r_{I}	Recovery rate of symptomatic infected individuals	1/14 day^{-1}
r_{A}	Recovery rate of asymptomatic infected individuals	1/14 day^{-1}

tiny portion of the total population. Some regions in China (e.g., Wuhan, Shenyang, Guangzhou) tested everyone once several new cases were reported locally. However, this never happened in the USA. Hence, we use the values of $(1-p)\delta E(t)$ as an approximation of the notification data.

We use $S[k]$, $E[k]$, $I[k]$, $A[k]$ and $R[k]$ to represent the values of variables in model (1) and $y[k]$ to be the notification data on the kth day of study. In addition, we use $D[k]$ to represent the cumulative death number on the kth day. Then, we have

$$E[k] = \frac{y[k]}{(1-p)\delta}, \quad k = 1, 2, 3, ..., K,$$

where K is the length of the vector of the notification data. We estimate the initial values $I[1]$, $R[1]$ and $D[1]$ from reporting data (Ritchie et al. 2021; Worldometers 2021): $I[1] = 21{,}637$, $R[1] = 14{,}813$, $D[1] = 10{,}595$. Moreover, we assume that $A[1] = 2I[1]$ considering that most infected people are asymptomatic (Subramanian et al. 2021). Then, $S[1] = N - E[1] - I[1] - A[1] - R[1] - D[1]$. It follows that

$$\begin{aligned}
I[k] &= I[k-1] + (1-p)\delta E[k-1] - (\mu[k-1] + r_{\mathrm{I}})I[k-1],\\
A[k] &= A[k-1] + p\delta E[k-1] - r_{\mathrm{A}}A[k-1],\\
R[k] &= R[k-1] + r_{\mathrm{I}}I[k-1] + r_{\mathrm{A}}A[k-1],\\
D[k] &= D[k-1] + \mu[k-1]I[k-1],\\
S[k] &= N - E[k] - I[k] - A[k] - R[k] - D[k],\\
\beta[k-1] &= -\frac{N(S[k] - S[k-1])}{(S[k-1](\theta_{\mathrm{E}}E[k-1] + \theta_{\mathrm{A}}A[k-1] + I[k-1]))},
\end{aligned}$$

for $k = 2, 3, ...K$. Approximately, we have $\beta[K] \approx \beta[K-1]$. The idea is that once we get the time series values of $E(t)$, we are able to obtain the time series values of $I(t)$, $A(t)$, $R(t)$, and hence, $S(t)$. Then, from the first equation of system (1), we can solve for $\beta(t)$. Note that the inverse method used in this study is different from that in Kong et al. (2015), Pollicott et al. (2012), although the essential idea is similar, that is, to solve for the transmission rate inversely.

4 Machine Learning and Prediction

Human mobility can affect the transmission rate, and policies from the government may affect human mobility. Therefore, the transmission rate can be indirectly affected by the policies. Indeed, some policies such as facial coverings may even directly affect the transmission rate. We use a GBM to estimate the transmission rate from the policy predictor variables: C1 $\sim$ C8, H1, H2, H3, H6, H8.

Having estimated the transmission rate in Sect. 3, we can fit $\log(\beta(t))$ with mobility and policy data using the GBM. We partition the data into a *training dataset*, used to calibrate the parameters, and a *testing dataset*, used to test the model performance in making predictions. The partitioning should be temporal: Since the model is supposed to make predictions in the future, it should be tested on a dataset that is "in the future"

compared to the dataset used for estimating the model parameters, where the values of the number of confirmed cases are unavailable. More specifically, the data instances from time 0 to T are used for training and from time $T + 1$ to $T + T'$, for some $T, T' > 0$, is used for testing the model. We may, otherwise, obtain misleadingly high-performance results (Ramazi et al. 2021a, c).

We fix the start date of the training at April 4, 2020, and let the training duration increase from 105 to 224 days by a step size of 7 days (see Table 2). The training dataset consists of the transmission rate on each day obtained by the inverse method as the response variable and all the 13 types of policy data C1 $\sim$ C8, H1, H2, H3, H6 and H8 on each day as predictor variables for the GBM. We fix the test duration at 35 days right after each training duration (see Table 2). The trained GBMs will predict the transmission rate based on the policy data provided during the test duration. The `gbm` package and the `predict` function in `R` are used.

Then we can plot the curve of $(1 - p)\delta E(t)$ of the SEIAR model (1) by using the time series of trained and tested daily transmission rates to compare with notification data of COVID-19 confirmed cases. To evaluate the fitting results, we use the mean absolute error (MAE) and the mean absolute percentage error (MAPE) to compute the differences between the transmission rates predicted by the GBM and those obtained by the inverse method as well as the differences between the predicted and actual numbers of daily COVID-19 confirmed cases. The formulas of MAE and MAPE are

Table 2 Training and testing durations

Train length (days)	Train duration	Test duration
105	Apr 4, 2020 to Jul 17, 2020	Jul 18, 2020 to Aug 21, 2020
112	Apr 4, 2020 to Jul 24, 2020	Jul 25, 2020 to Aug 28, 2020
119	Apr 4, 2020 to Jul 31, 2020	Aug 1, 2020 to Sept 4, 2020
126	Apr 4, 2020 to Aug 7, 2020	Aug 8, 2020 to Sept 11, 2020
133	Apr 4, 2020 to Aug 14, 2020	Aug 15, 2020 to Sept 18, 2020
140	Apr 4, 2020 to Aug 21, 2020	Aug 22, 2020 to Sept 25, 2020
147	Apr 4, 2020 to Aug 28, 2020	Aug 29, 2020 to Oct 2, 2020
154	Apr 4, 2020 to Sept 4, 2020	Sept 5, 2020 to Oct 9, 2020
161	Apr 4, 2020 to Sept 11, 2020	Sept 12, 2020 to Oct 16, 2020
168	Apr 4, 2020 to Sept 18, 2020	Sept 19, 2020 to Oct 23, 2020
175	Apr 4, 2020 to Sept 25, 2020	Sept 26, 2020 to Oct 30, 2020
182	Apr 4, 2020 to Oct 2, 2020	Oct 3, 2020 to Nov 6, 2020
189	Apr 4, 2020 to Oct 9, 2020	Oct 10, 2020 to Nov 13, 2020
196	Apr 4, 2020 to Oct 16, 2020	Oct 17, 2020 to Nov 20, 2020
203	Apr 4, 2020 to Oct 23, 2020	Oct 24, 2020 to Nov 27, 2020
210	Apr 4, 2020 to Oct 30, 2020	Oct 31, 2020 to Dec 4, 2020
217	Apr 4, 2020 to Nov 6, 2020	Nov 7, 2020 to Dec 11, 2020
224	Apr 4, 2020 to Nov 13, 2020	Nov 14, 2020 to Dec 18, 2020

given by

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|, \quad \text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - x_i}{x_i}\right|,$$

where x_i is the ith component of the vector of actual values, y_i is the ith component of the vector of prediction values, and n is the total number of data instances.

Gradient boosting trains models gradually, additively, and sequentially by minimizing the loss function via the number of trees. Alongside with the number of trees, the other parameters including the distribution of response variable, the stochastic gradient descent, the learning rate, the depth of interaction, and the minimum number of observations allowed in the trees' terminal nodes can all directly affect the performance of the model (Mayr et al. 2014; Zhang et al. 2019). We select the parameter values for GBM based on the averaged MAE and MAPE over the different training durations in Table 2. We apply the `summary` function with the default method of relative influence in R to investigate the variable importance in training the model.

After trying different combinations of the GBM parameters, we decide to employ 1000 trees with a Gaussian distribution of the response variable, 0.9 stochastic gradient descent, 0.01 learning rate, 30 depth of interaction and a minimum number of 10 observations allowed in the trees' terminal nodes, which results in a smaller averaged MAE and MAPE for predictions based on the various training durations in Table 2 (Figs. 2, 3).

The prediction performance of the GBM for the daily confirmed cases of COVID-19 is summarized in Tables 3 and 4. We can see that small MAE and MAPE are obtained when the GBM is trained for 126 days, 147 days, 154 days and 175 days. The corresponding training and testing (prediction) results of the transmission rate together with those of confirmed cases are presented in Fig. 4, supplementary Figs. 10, 12 and 14 (see "Appendix"). The trained transmission rates (i.e., the orange curves in the left panels of these figures) generally fit well with the ones obtained from the inverse method (i.e., the blue curves in the left panels of these figures). However, the

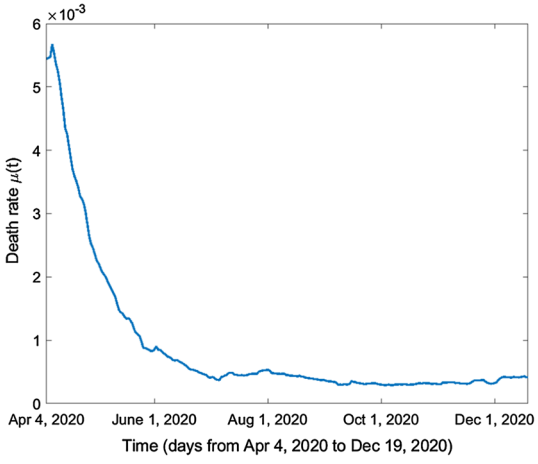


Fig. 2 Disease induced death rate from April 4, 2020, to December 19, 2020 (Color figure online)

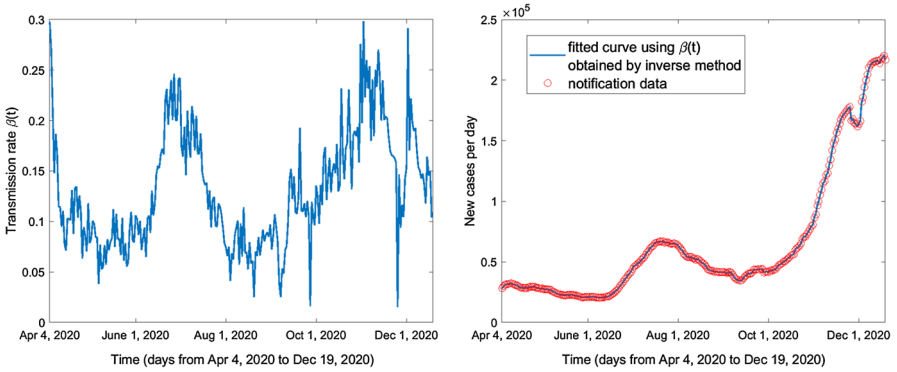


Fig. 3 Transmission rate obtained by the inverse method and the fitting with notification data from April 4, 2020, to December 19, 2020 (Color figure online)

Table 3 MAE and MAPE of predictions of notification data based on model (1) and the GBM in Sect. 4 corresponding to different training durations

Train length (days)	Train duration	MAE	MAPE (%)
105	Apr 4, 2020 to Jul 17, 2020	36616.64	70.07
112	Apr 4, 2020 to Jul 24, 2020	27423.54	58.03
119	Apr 4, 2020 to Jul 31, 2020	17786.35	39.88
126	Apr 4, 2020 to Aug 7, 2020	2059.83	4.92
133	Apr 4, 2020 to Aug 14, 2020	6063.13	15.61
140	Apr 4, 2020 to Aug 21, 2020	5714.62	13.85
147	Apr 4, 2020 to Aug 28, 2020	2451.90	6.47
154	Apr 4, 2020 to Sept 4, 2020	2540.38	6.45
161	Apr 4, 2020 to Sept 11, 2020	11753.02	26.39
168	Apr 4, 2020 to Sept 18, 2020	9150.12	17.76
175	Apr 4, 2020 to Sept 25, 2020	4268.39	8.53
182	Apr 4, 2020 to Oct 2, 2020	19195.41	24.87
189	Apr 4, 2020 to Oct 9, 2020	25871.46	25.30
196	Apr 4, 2020 to Oct 16, 2020	41907.22	33.51
203	Apr 4, 2020 to Oct 23, 2020	35601.46	23.23
210	Apr 4, 2020 to Oct 30, 2020	59973.39	37.70
217	Apr 4, 2020 to Nov 6, 2020	56162.49	30.63
224	Apr 4, 2020 to Nov 13, 2020	70696.84	36.14

tested transmission rate (i.e., the yellow curves in the left panels of these figures) do not fit well with the peaks or troughs of the blue curves of the transmission rate. In the right panels of Fig. 4 and supplementary Figs. 10, 12, 14, the orange curves (i.e., trained part) fit almost perfectly with the real notification data of confirmed cases. The yellow curve of prediction in the right panel of Fig. 4 also fits well with the blue circles of real data, with the MAPE equal to 4.92%. In the right panel of supplementary Figs. 10 and 12, the yellow prediction curve does not show a good fitting with the local

Table 4 Averaged MAE and MAPE for the prediction of daily confirmed cases by using model (1) and the GBM in Sect. 4

Data used in GBM	Averaged MAE	Averaged MAPE
Policy data C1 ∼ C8, H1, H2, H3, H6, H8	24179.79	26.63%

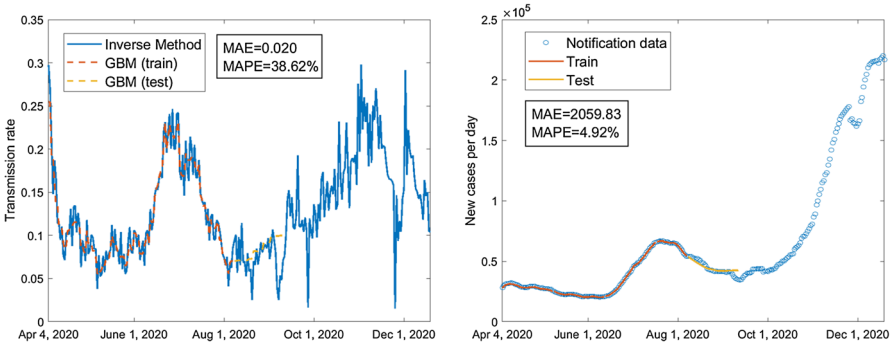


Fig. 4 Using policy data C1 ∼ C8, H1, H2, H3, H6 and H8, train 126 days from April 4, 2020, to August 7, 2020; test 35 days from August 8, 2020, to September 11, 2020 (Color figure online)

minimum point around September 11 although the MAPE is as small as 6.47% and 6.45%, respectively.

The relative influence of a variable in a single tree is the sum of the empirical improvement by splitting on the variable at those points. Friedman extended it to boosting models by averaging the relative influence of each variable across all the trees generated by the boosting algorithm (Friedman 2001). The relative influence of mobility and policy variables for the GBM based on the different training durations of 126 days, 147 days, 154 days and 175 days are shown in Table 5, supplementary Tables 9, 10 and 11, respectively. Among these policies, restrictions on gatherings always have the highest weight of relative influence which is as large as 42.46% when trained for 154 days. Other important predictors are testing policies, facial coverings, school closing, protection of elderly people and workplace closing. Public information campaigns and international travel controls are the least important policies with a weight of at most 0.43% for public information campaigns when trained for 147 days and zero influence from international travel controls (see Table 5, supplementary Tables 9, 10 and 11). As can be seen from Fig. 5 and supplementary Figs. 11, 13 and 15, the rankings of the relative influence of some policy variables have changed when trained for different lengths of days.

5 Machine Learning with Policy and Mobility Data

While fitting the transmission rate with policy data is helpful for prediction, it would be interesting to see how the transmission rate can be affected by mobility as well since human mobility is considered to have direct impact on the transmission rate. Among

Table 5 Relative influence of policy variables when trained for 126 days from April 4, 2020, to August 7, 2020

Variable	Relative influence (%)
Restrictions on gatherings	31.0489157
Testing policies	16.7287051
Protection of elderly people	15.2217062
School closing	7.9329945
Facial coverings	6.0243431
Workplace closing	5.0720110
Restrictions on internal movement	4.5464197
Close public transport	4.0595600
Cancel public events	3.4738417
Stay at home requirements	3.2881507
Contact tracing	2.4412499
Public information campaigns	0.1621026
International travel controls	0.0000000

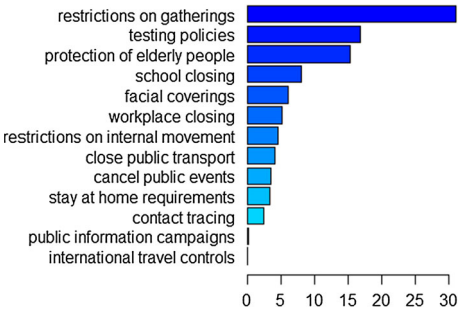


Fig. 5 Relative influence of policy variables when trained for 126 days from April 4, 2020, to August 7, 2020 (Color figure online)

all the policies that we have investigated in Sect. 4, testing policies (H2), contact tracing (H3) and facial coverings (H6) normally do not affect human mobility. Thus, it is reasonable to set H2, H3, H6 and mobility variables M1 $\sim$ M6 as the predictor variables and to keep all the mobility variables unchanged while changing some of these policies when we explore the effects of these three policies on the transmission rate. In this section, We use GBM to connect the transmission rate with mobility data in the presence or absence of policy data. We perform two GBMs with different predictor variables: one involves the mobility variables M1 $\sim$ M6 only; the other consists of both the mobility variables M1 $\sim$ M6 and the policy variables H2, H3, H6.

We use the same values of parameters as those in Sect. 4, train the two GBMs for different training durations increasing from 105 days to 224 days by 7 days, and test the models for 35 days following each training duration (see Table 2). The training dataset consists of the transmission rate on each day obtained by the inverse method as the response variable and all the six types of mobility data M1 $\sim$ M6 on each day as predictor variables for both GBMs. Additionally, the training dataset of the GBM involving both mobility and policy variables includes the

Table 6 Averaged MAE and MAPE for the prediction of daily confirmed cases by using model (1) + the GBM with mobility only and model (1) + the GBM with mobility and policy as predictors

Data used in GBM	Averaged MAE	Averaged MAPE
Mobility data	26188.58	36.22%
Mobility data + policy data H2, H3, H6	20408.11	25.67%

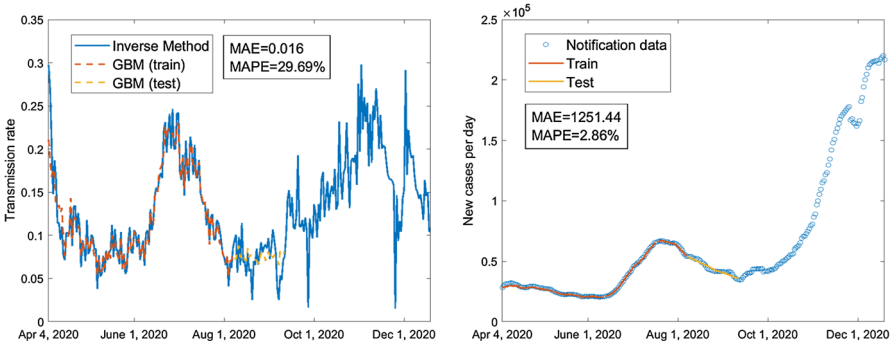


Fig. 6 Using mobility data M1 ∼ M6 and policy data H2, H3, H6, train 126 days from April 4, 2020, to August 7, 2020; test 35 days from August 8, 2020, to September 11, 2020 (Color figure online)

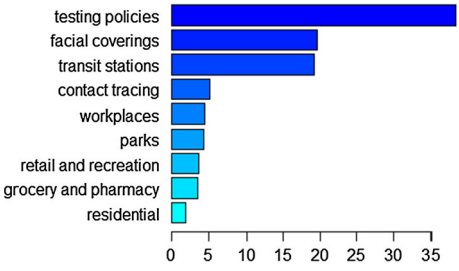


Fig. 7 Relative influence of mobility and H2, H3, H6 policy variables when trained for 126 days from April 4, 2020, to August 7, 2020 (Color figure online)

three types of policy data H2, H3 and H6 on each day as predictor variables as well. The trained GBMs will give a prediction for the transmission rate based on the mobility and/or policy data provided during the test duration. Then, we can plot the curve of $(1-p)\delta E(t)$ of the SEIAR model (1) by using the time series of trained and tested daily transmission rates to compare with notification data of confirmed cases.

The prediction performance for the confirmed cases of COVID-19 is summarized in Table 6. We can see that the averaged MAE and MAPE of the GBM with both mobility and policy predictors are smaller than those of the GBM with mobility predictors only, which indicates that involving policy data can produce better prediction results. In particular, very small MAEs and MAPEs are obtained for the prediction results of daily confirmed cases when the GBM involving both mobility and policy variables is trained for 126 days, 133 days, and 217 days. The corresponding training and testing (prediction) results of the transmission rate

Table 7 Relative influence of mobility and H2, H3, H6 policy variables when trained for 126 days from April 4, 2020, to August 7, 2020

Variable	Relative influence (%)
Testing policies	38.370821
Facial coverings	19.574882
Transit stations	19.215391
Contact tracing	5.093295
Workplaces	4.396082
parks	4.316401
Retail and recreation	3.614830
Grocery and pharmacy	3.556656
Residential	1.861642

together with the fitted curves of confirmed cases are presented in Fig. 6, supplementary Figs. 16 and 18. The trained transmission rates (i.e., the orange curves in the left panels of these figures) generally fit well with the ones obtained from the inverse method (i.e., the blue curves in the left panels of these figures). However, the tested transmission rate (i.e., the yellow curves in the left panels of these figures) do not fit well with the peaks or troughs of the blue curves of the transmission rate. In the right panels of Fig. 6 and supplementary Figs. 16, 18, the orange curves (i.e., trained part) fit almost perfectly with the real notification data of confirmed cases. The yellow curves of prediction in the right panels of Fig. 6 and supplementary Fig. 16 also fit quite well with the blue circles of real data, with the MAPE equal to 2.86% and 4.66%, respectively. In the right panel of supplementary Fig. 18, the yellow prediction curve does not show a good fitting with the local minimum point around November 30 although the MAPE is as small as 5.64%. This may be because it is near the Thanksgiving holiday during which people get together and may not have many testings as usual (Fig. 7). For the GBM which involves only mobility variables as predictors, smaller MAE and MAPE are obtained when the model is trained for 224 days as shown in Fig. 8. In this case, the predicted result is able to show the local minimum of daily confirmed cases around November 30 (see the yellow curve in the right panel of Fig. 8).

The relative influence of the variables for the GBM involving both mobility and policy based on the different training durations of 126 days, 133 days, and 217 days are shown in Table 7, supplementary Tables 12, 13, and Fig. 7, supplementary Figs. 17, 19, respectively. Among the three policies, the testing policy H2 always has the highest weight of relative influence which is as large as 38.37% when trained for 126 days. The second most important predictor is the facial covering policy H6 which weighs from about 18.36% to 21.50% corresponding to the above three training durations. The contact tracing policy H3 is the least important, with a weight ranging from about 5.09% to 13.64%. As can be seen from Fig. 7 and supplementary Figs. 17, 19, the rankings of the relative influence of the mobility and policy variables have changed when trained for different lengths of days. When the policy variables are dropped, the ranking of the relative influence of mobility variables in Fig. 9 is also different from those in Fig. 7, supplementary Figs. 17 and 19.

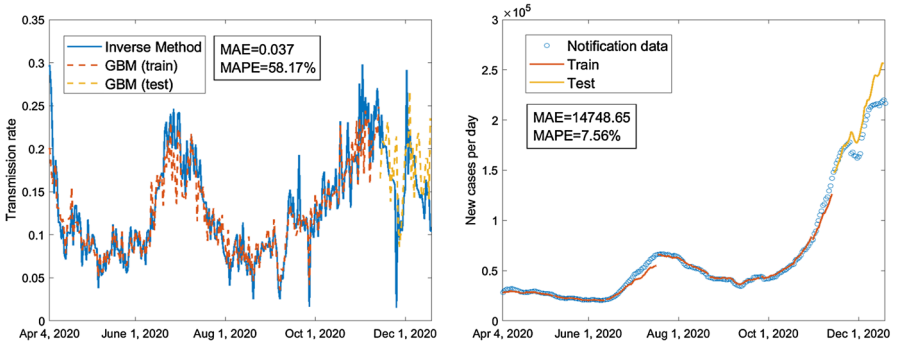


Fig. 8 Using mobility data M1 ∼ M6, train 224 days from April 4, 2020, to November 13, 2020; test 35 days from November 14, 2020, to December 18, 2020 (Color figure online)

Table 8 Relative influence of mobility variables when trained for 224 days from April 4, 2020, to November 13, 2020

Variable	Relative influence (%)
Parks	29.631102
Workplaces	16.835859
Grocery and pharmacy	16.086562
Transit stations	14.995899
Retail and recreation	12.799867
Residential	9.650711

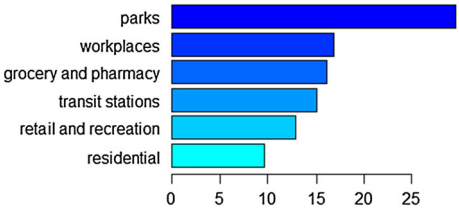


Fig. 9 Relative influence of mobility variables when trained for 224 days from April 4, 2020, to November 13, 2020 (Color figure online)

6 Discussion

We proposed a new framework for making predictions, that is, a hybrid model combining a mechanistic SEIAR model and gradient boosting models (GBM) with policy and mobility variables as predictors. We created an inverse method to estimate the time-varying transmission rate of COVID-19. This inverse method allows us to directly deal with time series data of daily confirmed cases without needing to get a smooth curve of the notification data at first or to substitute the integral form of any compartmental variables as the authors did in Kong et al. (2015), Pollicott et al. (2012), which greatly simplifies the process of deriving the transmission rate. Using the transmission rate obtained by the inverse method can give an almost perfect fit with the notification data, which obviously outcompetes the traditionally used method of least squares. The tree-based method used by GBM increases the accuracy of prediction by turning "weak learners" into "strong learners" in a gradual, additive and sequential way (Friedman

2001). Both MAE and MAPE are used for evaluating the prediction performance of the GBMs on the transmission rate as well as the fitting result of the number of confirmed cases by the SEIAR model. The selected GBM is capable of capturing the correlation between the transmission rate obtained from the inverse method and the policy as well as mobility variables so that accurate predictions of daily confirmed cases are made based on the SEIAR model and notification data. The bar plots of relative influence show that the most important policy is always restrictions on gatherings.

The method presented in this paper for connecting policy/mobility and transmission rate is data-driven and hypothesis-free. This is different from some other methods such as the least-squares method where one needs to make simplifying assumptions on the form of the transmission rate in the future, e.g., it is constant, piecewise constant, or a combination of sigmoid functions (López and Rodo 2021; Balcha 2020; Sahoo and Sapra 2020; Tátrai and Várallyay 2020; Ianni and Rossi 2020; Choi and Ki 2020; Pluchino et al. 2021; Zhou et al. 2020). The least-squares method makes future predictions based on either a pre-assumed form of the transmission rate function with respect to time or the "current" (i.e., using the transmission rate on the last day of the training set as the transmission rate on each day of the prediction period), whereas machine-learning models make predictions based on the "past" (i.e., the trained experience), and typically without making restrictively simplifying assumptions. In particular, our hybrid model that is based on only preventive policies and/or mobilities, is trained on "past" data to link the policies/mobilities to transmission rate, and uses "future" data on policies/mobilities to estimate the future values of the transmission rate. Given a set of "future" policy/mobility data for the 35-day test window, we can get corresponding predicted values of the transmission rate during that window. As such, the model can be used to compare the dynamics under different future NPIs or mobility trends. Non-pharmaceutical preventive policies are often a priori known and available for making predictions. In situations where the data are unavailable regularly or contains missing values, Bayesian networks can be used instead of GBM (Ramazi et al. 2021b).

Strikingly 90% of the world's data have been generated in the past several years; thus, machine learning has become more efficient in making predictions; however, mechanistic models can provide the causality missing from machine-learning approaches (Baker et al. 2018). Our hybrid model could provide more reliable predictions, especially when future policies have dramatic changes and enough amount of data are provided for training. Logically, our method has similar accuracy as machine learning approaches, but the disease spread compartment of our method includes a mechanistic model that captures established epidemiological causal relationships between the disease variables. In addition, there is no need to compare our method with the least-squares method because the inverse method has perfect data fitting for transmissibility without making any assumptions.

Since machine learning requires sufficient amount of data in order to obtain effective training, our hybrid model may not be competent in making predictions in the initial stage of an epidemic/pandemic caused by a novel pathogen. In addition, in our model we simply assume that human individuals in the USA are homogeneously mixed and obtain policy data by averaging the policy data over different states together with Washington D.C. weighted by their populations. Indeed, different states or regions usually have different epidemic progress and different preventive and control policies.

Even within a small region, different people may have different immunity abilities and hence different recovery or death rates, etc. To incorporate the role of heterogeneity in disease transmission, we can either apply our model and method to different smaller regions with region-specific parameters and then compare the prediction results or develop a patchy ODE model or PDE model with nonlocal dispersal. We can also divide the population into more compartments according to their ages, health states or activity levels such as in different exposed periods, hospitalized, quarantined, on travel, working in medical frontlines, etc., and assume parameter values to be group-specific accordingly.

Our method can be applied to the study of other infectious diseases or future newly emerged pandemics in early stages. It can identify the most influential variables in predicting the disease spread and predict disease dynamics under different policies, which may guide policy makers to design mitigation measures. Our next step is to apply the inverse method plus machine learning approach to make predictions on daily new cases for the post-vaccination period and uncover the role of vaccination policies in future pandemic waves (Table 8).

Acknowledgements This work was funded by Alberta Innovates and Pfizer via project number RES0052027. HW gratefully acknowledges support from Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2020-03911 and NSERC Accelerator Grant RGPAS-2020-00090. KN gratefully acknowledges support from National Institute for Mathematical Sciences (NIMS) grant funded by the Korean Government (NIMS-B22910000). MAL is a Canada Research Chair in Mathematical Biology and gratefully acknowledges support from NSERC Discovery Grant.

Appendix: Supplementary figures and tables

In this Appendix, we present supplementary figures and tables. The selected training and testing results about the transmission rates and the fittings with notification data of daily confirmed cases are displayed in Figs. 10, 12 and 14 for the model with policy as the only predictors, in Figs. 16 and 18 for the model with both policy and mobility as the predictors. After each of these figures, we present a table and a figure of the relative influence of the involved predictor variables in training the model. Tables 9, 10, 11 and Figs. 11, 13, 15 show the relative influence of the policy variables when the model is trained for 147 days, 154 days, 175 days, respectively. Tables 12, 13 and Figs. 17, 19 give the relative influence of the mobility and part of policy variables when the model is trained for 133 days and 217 days, respectively.

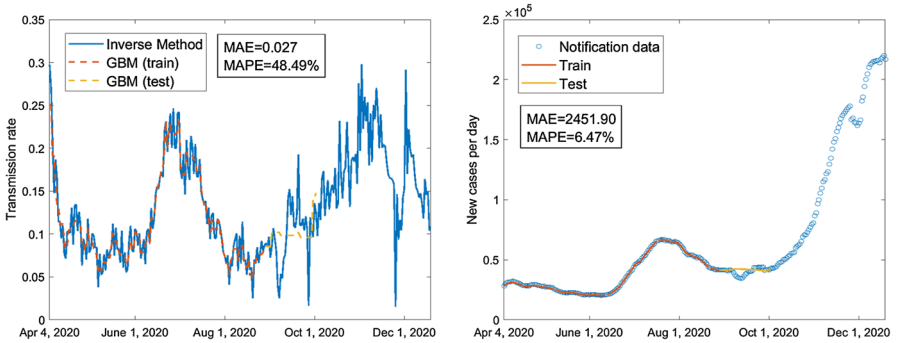
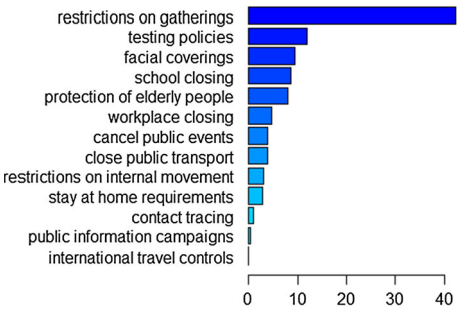


Fig. 10 Using policy data C1 ∼ C8, H1, H2, H3, H6 and H8, train 147 days from April 4, 2020, to August 28, 2020; test 35 days from August 29, 2020, to October 2, 2020 (Color figure online)

Table 9 Relative influence of policy variables when trained for 147 days from April 4, 2020, to August 28, 2020

Variable	Relative influence (%)
Restrictions on gatherings	42.3668084
Testing policies	11.9967018
Facial coverings	9.4461173
school closing	8.7000103
Protection of elderly people	7.8984697
workplace closing	4.6655947
Cancel public events	3.9032770
Close public transport	3.7699113
Restrictions on internal movement	3.1179248
Stay at home requirements	2.8130527
Contact tracing	0.8900337
Public information campaigns	0.4320982
International travel controls	0.0000000

Fig. 11 Relative influence of policy variables when trained for 147 days from April 4, 2020, to August 28, 2020 (Color figure online)



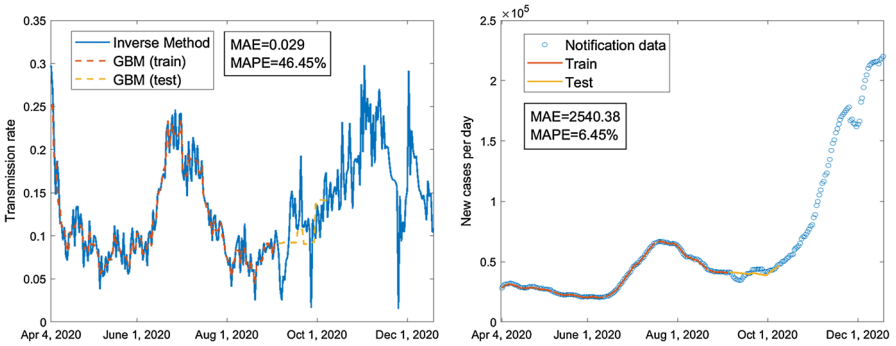


Fig. 12 Using policy data C1 ∼ C8, H1, H2, H3, H6 and H8, train 154 days from April 4, 2020, to September 4, 2020; test 35 days from September 5, 2020, to October 9, 2020 (Color figure online)

Table 10 Relative influence of policy variables when trained for 154 days from April 4, 2020, to September 4, 2020

Variable	Relative influence (%)
Restrictions on gatherings	42.4561382
Testing policies	12.8734097
School closing	8.5351249
Facial coverings	8.0608618
Protection of elderly people	6.6531206
Workplace closing	5.6067289
Cancel public events	4.3324652
Close public transport	3.4850619
Restrictions on internal movement	3.4346518
Stay at home requirements	3.1492062
Contact tracing	1.0979493
Public information campaigns	0.3152815
International travel controls	0.0000000

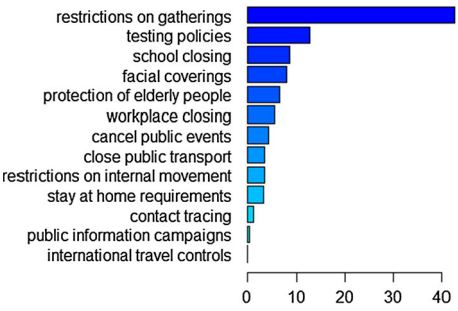


Fig. 13 Relative influence of policy variables when trained for 154 days from April 4, 2020, to September 4, 2020 (Color figure online)

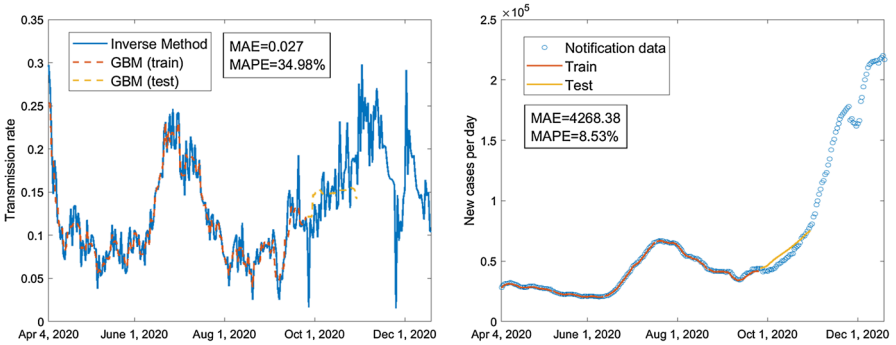


Fig. 14 Using policy data C1 ∼ C8, H1, H2, H3, H6 and H8, train 175 days from April 4, 2020, to September 25, 2020; test 35 days from September 26, 2020, to October 30, 2020 (Color figure online)

Table 11 Relative influence of policy variables when trained for 175 days from April 4, 2020, to September 25, 2020

Variable	Relative influence (%)
Restrictions on gatherings	38.0716839
Testing policies	12.3025438
School closing	8.4113758
Facial coverings	8.4021886
Workplace closing	7.9880485
Close public transport	6.1772695
Protection of elderly people	5.9374740
Cancel public events	4.4452982
Stay at home requirements	3.1632396
Restrictions on internal movement	2.9548620
Contact tracing	1.9726367
public information campaigns	0.1733794
International travel controls	0.0000000

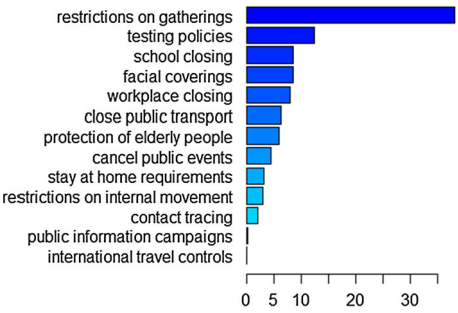


Fig. 15 Relative influence of policy variables when trained for 175 days from April 4, 2020, to September 25, 2020 (Color figure online)

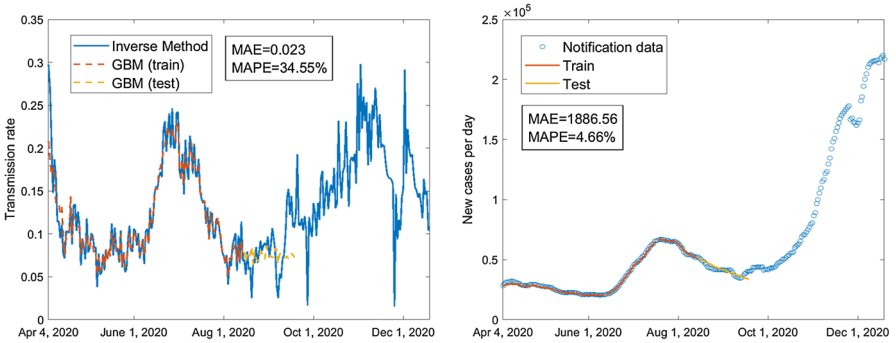
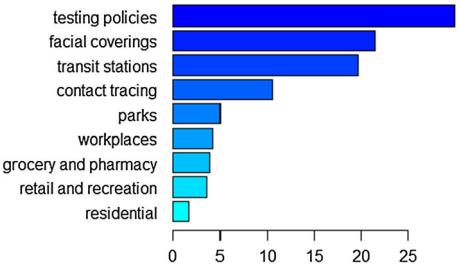


Fig. 16 Using mobility data and H2, H3, H6 policy data, train 133 days from April 4, 2020, to August 14, 2020; test 35 days from August 15, 2020, to September 18, 2020 (Color figure online)

Table 12 Relative influence of mobility and H2, H3, H6 policy variables when trained for 133 days from April 4, 2020, to August 14, 2020

Variable	Relative influence (%)
Testing policies	29.935771
Facial coverings	21.503403
Transit stations	19.706089
contact tracing	10.468970
Parks	4.930258
Workplaces	4.243045
Grocery and pharmacy	3.940721
Retail and recreation	3.599258
Residential	1.672485

Fig. 17 Relative influence of mobility and H2, H3, H6 policy variables when trained for 133 days from April 4, 2020, to August 14, 2020 (Color figure online)



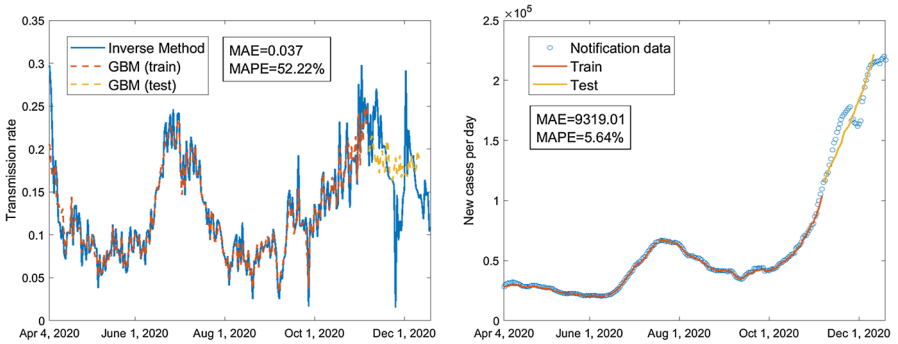
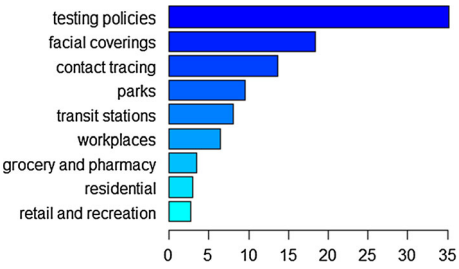


Fig. 18 Using mobility data and H2, H3, H6 policy data, train 217 days from April 4, 2020, to November 6, 2020; test 35 days from November 7, 2020, to December 11, 2020 (Color figure online)

Table 13 Relative influence of mobility and H2, H3, H6 policy variables when trained for 217 days from April 4, 2020, to November 6, 2020

Variable	Relative influence (%)
Testing policies	35.148781
Facial coverings	18.359430
Contact tracing	13.641922
Parks	9.487232
transit stations	7.922605
Workplaces	6.320635
Grocery and pharmacy	3.452438
Residential	3.007597
Retail and recreation	2.659359

Fig. 19 Relative influence of mobility and H2, H3, H6 policy variables when trained for 217 days from April 4, 2020, to November 6, 2020 (Color figure online)



References

Baker RE, Pena J-M, Jayamohan J, Jérusalem A (2018) Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biol Lett 14(5):20170660

Balcha AA et al (2020) Curve fitting and least square analysis to extrapolate for the case of COVID-19 status in Ethiopia. Adv Infect Dis 10(03):143

Bureau US Census (2021) Quick facts United States. https://www.census.gov/quickfacts/fact/table/US/PST045219. Last accessed 15 May 2021

Calvetti D, Hoover Alexander P, Rose J, Somersalo E (2020) Metapopulation network models for understanding, predicting, and managing the coronavirus disease COVID-19. Front Phys 8:261

Chang S, Pierson E, Koh PW, Gerardin J, Redbird B, Grusky D, Leskovec J (2021) Mobility network models of COVID-19 explain inequities and inform reopening. Nature 589(7840):82–87

Choi S, Ki M (2020) Estimating the reproductive number and the outbreak size of COVID-19 in Korea. Epidemiol Health 42:e2020011. https://doi.org/10.4178/epih.e2020011

Coletti P, Libin P, Petrof O, Willem L, Abrams S, Herzog SA, Faes C, Kuylen E, Wambua J, Beutels P et al (2021) A data-driven metapopulation model for the Belgian COVID-19 epidemic: assessing the impact of lockdown and exit strategies. BMC Infect Dis 21(1):1–12

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232

Google Team (2021) Google COVID-19 community mobility reports, 2021. https://www.google.com/covid19/mobility/. Last accessed 1 May 2021

Ianni A, Rossi N (2020) Describing the COVID-19 outbreak during the lockdown: fitting modified SIR models to data. Eur Phys J Plus 135(11):1–10

IHME COVID-19 Forecasting Team (2020) Modeling COVID-19 scenarios for the United States. Nat Med 27:94–105

Kong JD, Jin C, Wang H (2015) The inverse method for a childhood infectious disease model with its application to pre-vaccination and post-vaccination measles data. Bull Math Biol 77:2231–2263

Kumar N, Abdul Rahman AK, AlAli S, Otoom S, Atkin SL, AlQahtani M (2021) Time till viral clearance of severe acute respiratory syndrome coronavirus 2 is similar for asymptomatic and non-critically symptomatic individuals. Front Med 8

Liu Z, Magal P, Seydi O, Webb G (2020) A COVID-19 epidemic model with latency period. Infect Dis Model 5:323–337

López L, Rodo X (2021) A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: simulating control scenarios and multi-scale epidemics. Res Phys 21:103746

Mayr A, Binder H, Gefeller O, Schmid M (2014) The evolution of boosting algorithms: from machine learning to statistical modelling. Methods Inf Med 53(6):419–427

Mukandavire Z, Nyabadza F, Malunguza NJ, Cuadros DF, Shiri T, Musuka G (2020) Quantifying early COVID-19 outbreak transmission in South Africa and exploring vaccine efficacy scenarios. PLOS ONE 15(7):e0236003

Pluchino A, Biondo AE, Giuffrida N, Inturri G, Latora V, Le Moli R, Rapisarda A, Russo G, Zappala C (2021) A novel methodology for epidemic risk assessment of COVID-19 outbreak. Sci Rep 11(1):1–20

Pollicott M, Wang H, Weiss H (2012) Extracting the time-dependent transmission rate from infection data via solution of an inverse ODE problem. J Biol Dyn 6(2):509–523

Ramazi P, Haratian A, Meghdadi M, Oriyad AM, Lewis MA, Maleki Z, Vega R, Wang H, Wishart DS, Greiner R (2021) Accurate long-range forecasting of COVID-19 mortality in the USA. Sci Rep 11(1):1–11

Ramazi P, Kunegel-Lion M, Greiner R, Lewis MA (2021) Exploiting the full potential of Bayesian networks in predictive ecology. Methods Ecol Evol 12(1):135–149

Ramazi P, Kunegel-Lion M, Greiner R, Lewis MA (2021) Predicting insect outbreaks using machine learning: A mountain pine beetle case study. Ecol Evol 11(19):13014–13028

Ritchie H, Ortiz-Ospina E, Beltekian D, Mathieu E, Hasell J, Macdonald B, Giattino C, Appel C, Rodés-Guirao L, Roser M (2021) Coronavirus Pandemic (COVID-19). Our World in Data, 2020. https://ourworldindata.org/coronavirus. Last accessed 5 June 2021

Sahoo BK, Sapra BK (2020) A data driven epidemic model to analyse the lockdown effect and predict the course of COVID-19 progress in India. Chaos Solitons Fractals 139:110034

Subramanian R, He Q, Pascual M (2021) Quantifying asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity. PNAS 118(9):e2019716118

Sun J, Chen X, Zhang Z, Lai S, Zhao B, Liu H, Wang S, Huan W, Zhao R, Ng MTA et al (2020) Forecasting the long-term trend of COVID-19 epidemic using a dynamic model. Sci Rep 10(1):1–10

Tátrai D, Várallyay Z (2020) COVID-19 epidemic outcome predictions based on logistic fitting and estimation of its reliability. arXiv preprint arXiv:2003.14160

Wikipedia (2021) COVID-19 pandemic in North America, 2021. https://en.wikipedia.org/wiki/COVID-19_pandemic_in_North_America. Last accessed 1 May 2021

Worldometers.info. United States coronavirus cases, deaths, recovered, 2021. https://www.worldometers.info/coronavirus/country/us/. Last accessed 22 April 2021

Zhang C, Zhang Y, Shi X, Almpanidis G, Fan G, Shen X (2019) On incremental learning for gradient boosting decision trees. Neural Process Lett 50(1):957–987

Zhang K, Tong W, Wang X, Lau JY-N (2020) Estimated prevalence and viral transmissibility in subjects with asymptomatic SARS-CoV-2 infections in Wuhan, China. Precis Clin Med 3(4):301–305

Zhou W, Wang A, Xia F, Xiao Y, Tang S (2020) Effects of media reporting on mitigating spread of COVID-19 in the early phase of the outbreak. Math Biosci Eng 17(3):2693–2707