Science of the Total Environment 758 (2021) 144151

Contents lists available at ScienceDirect

Science of the Total Environment

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

Evaluating the impact of mobility on COVID-19 pandemic with machine learning hybrid predictions

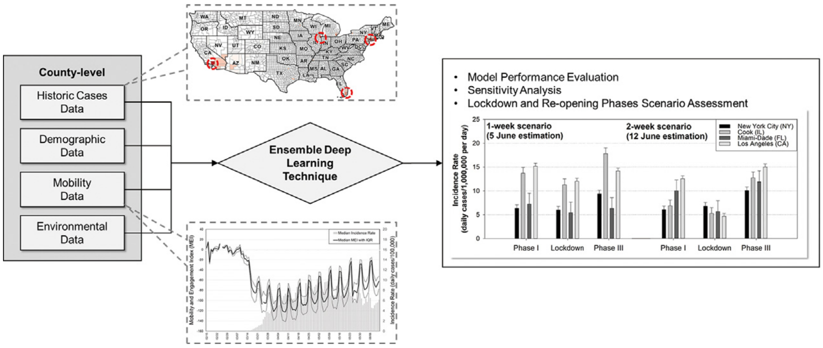
Cheng-Pin Kuo, Joshua S. Fu *

Department of Civil and Environmental Engineering, University of Tennessee Knoxville, Knoxville, TN, USA

HIGHLIGHTS

- County-level data were used to build up the COVID-19 prediction model by a machine learning hybrid framework.
- Weekly pattern of mobility and infections proved the incubation days is 4-5 days and high infections on the weekend.
- Compared with Phase I re-open, a 1-week and a 2-week lockdown can reduce 4-29% and 15-55% infections in the future week.

GRAPHICAL ABSTRACT


County-level
Historic Cases Data
Demographic Data
Mobility Data
Environmental Data
Ensemble Deep Learning Technique
Model Performance Evaluation
Sensitivity Analysis
Lockdown and Re-opening Phases Scenario Assessment
1-week scenario (5 June estimation)
2-week scenario (12 June estimation)
New York City (NY)
Cook (IL)
Miami-Dade (FL)
Los Angeles (CA)
Incidence Rate (daily cases/1,000,000 per day)
0
5
10
15
20
25
Phase I
Lockdown
Phase III

ARTICLE INFO

Article history:
Received 7 September 2020
Received in revised form 24 November 2020
Accepted 24 November 2020
Available online 28 November 2020

Editor: Jianmin Chen

Keywords:
Pandemic
County-level
Forecasting
Lockdown
Re-opening

ABSTRACT

COVID-19 pandemic had expanded to the US since early 2020 and has caused nationwide economic loss and public health crisis. Until now, although the US has the most confirmed cases in the world and are still experiencing an increasing pandemic, several states insisted to re-open business activities and colleges while announced strict control measures. To provide a quantitative reference for official strategies, predicting the near future trend based on finer spatial resolution data and presumed scenarios are urgently needed. In this study, the first attempted COVID-19 case predicting model based on county-level demographic, environmental, and mobility data was constructed with multiple machine learning techniques and a hybrid framework. Different scenarios were also applied to selected metropolitan counties including New York City, Cook County in Illinois, Los Angeles County in California, and Miami-Dade County in Florida to assess the impact from lockdown, Phase I, and Phase III re-opening.

Our results showed that, for selected counties, the mobility decreased substantially after the lockdown but kept increasing with an apparent weekly pattern, and the weekly pattern of mobility and infections implied high infections during the weekend. Meanwhile, our model was successfully built up, and the scenario assessment results indicated that, compared with Phase I re-opening, a 1-week and a 2-week lockdown could reduce 4%–29% and 15%–55% infections, respectively, in the future week, while 2-week Phase III re-opening could increase 16%–80% infections. We concluded that the mandatory orders in metropolitan counties such lockdown should last longer than one week, the effect could be observed. The impact of lockdown or re-opening was also county-dependent and varied with the local pandemic. In future works, we expect to involve a longer period of data, consider more county-dependent factors, and employ more sophisticated techniques to decrease the modeling uncertainty and apply it to counties nationally and other countries.

* Corresponding author at: Department of Civil and Environmental Engineering, the University of Tennessee, 851 Neyland Drive, Knoxville, TN 37996, USA.
E-mail address: jsfu@utk.edu (J.S. Fu).

https://doi.org/10.1016/j.scitotenv.2020.144151
0048-9697/© 2020 Elsevier B.V. All rights reserved.

1. Introduction

The coronavirus disease 2019 (COVID-19) epidemic, which originally started from Wuhan, China in December 2019, had expanded to the United States (US) in early 2020. Until May 31, a total of 1.8 million confirmed cases have been reported in the US where has the highest number of confirmed cases in the world. Although the US is experiencing an increasing number of pandemic cases now, several states have announced their re-opening schedule for businesses and universities while implemented strict control strategies at the same time. To control the pandemic, predicting the trend can provide a quantitative reference for planning the official strategies and better allocating the medical resources.

Among previous forecasting researches, several studies in China had employed simply observed data to predict the pandemic of main cities in China at the early stage of this worldwide pandemic (Ma et al., 2020; Xie and Zhu, 2020; Yang et al., 2020). To better understand the transmission characteristics of COVID-19 and predict the trend of pandemic, collecting more data and further investigation are still urgently needed. Technically, either using the statistical models or the epidemic models could predict the trend of the pandemic and its transmission scenario. One study used the statistical model and historical pandemic data to predict confirmed cases, deaths, and recoveries in the US and the study forecasted the impact in the US will double from 24 April to 7 July 2020 (Singh et al., 2020). Another study utilized an epidemic model to predict the national-wide trend of the pandemic in the US based on assumed parameters, and the proposed model successfully captured the trend in the future two weeks (Xu et al., 2020), but it was not for the state- or county-level transmission of the pandemic. Meanwhile, using assumed constant parameters such as a contact rate in an epidemic model or historical pandemic data in the statistical model could neglect the change of transmission parameters such as community mobility in reality, and the parameters could vary with time and states or counties. Moreover, since the polices varied with states, the predicting model with a finer spatial scale should be developed. For example, since 23 March 2020, several US state governments announced their statewide stay-at-home and lockdown order gradually, while states such as Arkansas, Iowa, Nebraska, North Dakota, Oklahoma, South Dakota, Utah, and Wyoming did not have a statewide mandatory order. Furthermore, each state also has a different re-opening schedule starting from late April to early June. Thus, the community mobility must vary within the states and date during the lockdown and after re-opening.

In this study, our objective is to develop a COVID-19 case predicting model based on county-level data with machine learning techniques, and the next-1-day (N1D), 4-day (N4D), and 7-day (N7D) averages of daily cases and cumulative cases would be used as a response respectively. A further sensitivity analysis for New York City and the other top 12 counties with the most confirmed cases in the US were conducted to analyze their near future trend within a week. Meanwhile, we also applied our model to assess the impact of Phase I re-opening, lockdown, and Phase III re-opening for the selected metropolitan counties during the pandemic.

2. Material and methods

2.1. Study settings

A total of 3219 counties in the US based on the definition of the US Census Bureau was considered. For each county, we considered input variables potentially related to the vulnerability of residents, the transmission of COVID-19, and exposure duration or frequency in community. First, for the vulnerability of residents in each county, county-dependent demographic characteristics such as population density, household income, labor force rate and unemployment rate which would affect the vulnerability of households were used. For example, those counties that have a lower population density usually have less developed health care systems, and lower-income or unemployed household would be more vulnerable to this crisis (Coibion et al., 2020; Mckibbin and Fernando, 2020). Second, for the transmission of COVID-19, environmental factors such as ambient temperature, precipitation, relative humidity, and wind speed will affect the life-time and transportation of coronavirus (Wang et al., 2020; Xie and Zhu, 2020). For each county, daily environmental factors were retrieved from gridMET estimations (Abatzoglou, 2013). Meanwhile, in terms of exposure duration or frequency in community, community mobility data which reflects individual exposure duration or population contact frequency with others was used. Moreover, community mobility could be lowered by mandatory orders such as city lockdown or stay-at-home order, thus the transmission of the disease could be reduced (Mahato et al., 2020). County-level mobility data from Feb 15 provided by Google was retrieved, the data contains the percent of mobility change from baseline for different places such as retail and recreation sites, grocery and pharmacy stores, parks, transit stations, workplaces, and residence according to the movement of Google users. Another aggregated mobility index, Mobility and Engagement Index (MEI), developed by the Federal Reserve Bank of Dallas was included in our model as well (Atkinson et al., 2020). Finally, we also considered historic COVID-19 cases as our input variables, because theoretically, the future trend of the pandemic will be affected by the number of confirmed cases (Yang et al., 2020). Detailed included variables and their sources, periods, and types were presented in Table 1.

In order to evaluate the impact of mobility change on the COVID-19 pandemic, for each county, data was only retrieved after state-wide lockdown announcement, and each county could have a different lockdown schedule. For example, the New York state announced state lockdown from March 23, while Georgia state announced from April 6. Daily data until May 31 were complied, and the input data set includes the COVID-19 daily confirmed cases and cumulative cases; county-dependent characteristics such as population density, labor force rate, unemployment rate, household median income, metro system or not (1 = metro, 0 = non metro), statewide stay-at-home order or not (0 = state without the statewide stay-at-home order, 1 = state with the statewide stay-at-home order); daily environmental variables such as maximum and minimum temperature, maximum and minimum relative humidity, precipitation, surface down-welling solar radiation, wind speed; daily community mobility data including retail and recreation percent change from the baseline which was from January 3 to February 22, grocery and pharmacy percent change from baseline, parks percent change from baseline, transit stations percent change from baseline, workplaces percent change from baseline, residential percent change from baseline, and MEI; time-series variables such as weekdays (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday), weekend or not (1 = weekend, 0 = weekday).

2.2. Data pretreatment

In the complied dataset, to add the link between present COVID-19 cases and historic trend of continuous variables, lag 1 day (the day before the present day) and moving averages of lag 1–7 day and lag 1–14 day were also used as input. Concerning the median incubation period for COVID-19 is 4–5 days from exposure to symptoms onset, only lag 1 and lag 1–7 day moving averages were used (CDC, 2020). In this study, we included a total of 52 input variables including a historic number of daily confirmed or cumulative cases (2 variables), county-dependent variables (6 variables), environmental variables (21 variables), community mobility variables (21 variables), and time-series variables (2 variables). The detailed of these input variables are illustrated in Table S1. Categorical variables would be included as dummy variables before modeling. If the response was the number of daily confirmed cases, a historic number of daily confirmed cases would be used, while the historic number of cumulative cases would be applied to

Table 1
Selected variables, data sources, periods and types used in this study.

Variable	Source	Data period	Data type
Cumulative cases	New York Times[a]	2020/01/01–2020/05/31	Continuous
Daily increased cases	New York Times[a]	2020/01/01–2020/05/31	Continuous
Population density	USDA ERA[b]	2019	Continuous
Labor force rate	USDA ERA[b]	2019	Continuous
Unemployment rate	USDA ERA[b]	2019	Continuous
Household median income	USDA ERA[b]	2019	Continuous
Metro system or not	USDA ERA[b]	2013	Categorical
Maximum and minimum temperature	gridMET[c]	2020/01/01–2020/05/31	Continuous
Maximum and minimum relative humidity	gridMET[c]	2020/01/01–2020/05/31	Continuous
Precipitation	gridMET[c]	2020/01/01–2020/05/31	Continuous
Surface downwelling solar radiation	gridMET[c]	2020/01/01–2020/05/31	Continuous
Wind speed	gridMET[c]	2020/01/01–2020/05/31	Continuous
Retail and recreation percent change from baseline	Google[d]	2020/02/15–2020/05/31	Continuous
Grocery and pharmacy percent change from baseline	Google[d]	2020/02/15–2020/05/31	Continuous
Parks percent change from baseline	Google[d]	2020/02/15–2020/05/31	Continuous
Transit stations percent change from baseline	Google[d]	2020/02/15–2020/05/31	Continuous
Workplaces percent change from baseline	Google[d]	2020/02/15–2020/05/31	Continuous
Residential percent change from baseline	Google[d]	2020/02/15–2020/05/31	Continuous
Mobility and Engagement Index (MEI)	Federal Reserve Bank of Dallas[e]	2020/01/03–2020/05/31	Continuous
Statewide stay-at-home order or not	New York Times[f]	2020/01/01–2020/05/31	Categorical
Weekday	–	2020/01/01–2020/05/31	Categorical
Weekend or not	–	2020/01/01–2020/05/31	Categorical

[a] New York Times, https://github.com/jeffcore/covid-19-usa-by-state/tree/d2fa4b2596889bac4687cdabb97a3967eb541392.
[b] United States Department of Agriculture, Economic Research Service, https://www.ers.usda.gov/data-products/county-level-data-sets/.
[c] gridMET (Abatzoglou, 2013), https://github.com/jbayham/gridMETr.
[d] Google, COVID-19 Community Mobility, https://www.google.com/covid19/mobility/.
[e] Federal Reserve Bank of Dallas, https://www.dallasfed.org/research/mei.
[f] New York Times, https://www.nytimes.com/interactive/2020/us/states-reopen-map-coronavirus.html.

predict the future cumulative cases. To select variables for modeling, each machine learning technique would have discarding or down-weighing procedure to select variables during modeling, and the final selected variables for modeling depends on the algorithm and model settings of each technique. The detailed selection or weighing criteria of each technique is detailed in the next section. Meanwhile, for predicting near-future trend, the N1D, N4D, and N7D averages of daily and cumulative cases would be used as a response respectively.

To assure input data quality, because not all counties have complete mobility data, only counties having all variables and > 80% valid data during the study period would be used, and missing values were replaced by the median of that variable. To focus on counties with this serious pandemic, only counties with cumulative cases more than 1000 on May 31 were included. Finally, a total of 172 counties was selected and used as modeling input in this study.

2.3. Applied machine learning and hybrid techniques

The workflow to apply machine learning techniques and generate hybrid estimations is illustrated in Fig. 1. We employed 8 types of basic machine learning techniques as basic learners, including (1) elastic net (EN) model, (2) principal components regression (PCR) model, (3) partial least squares regression (PLSR) model, (4) k-nearest neighbors regression (KNN) model, (5) regression tree (RT) model, (6) random forest (RF) model, (7) gradient boosted tree models (GBM), and (8) 2-layer artificial neural network (ANN) model to predict the

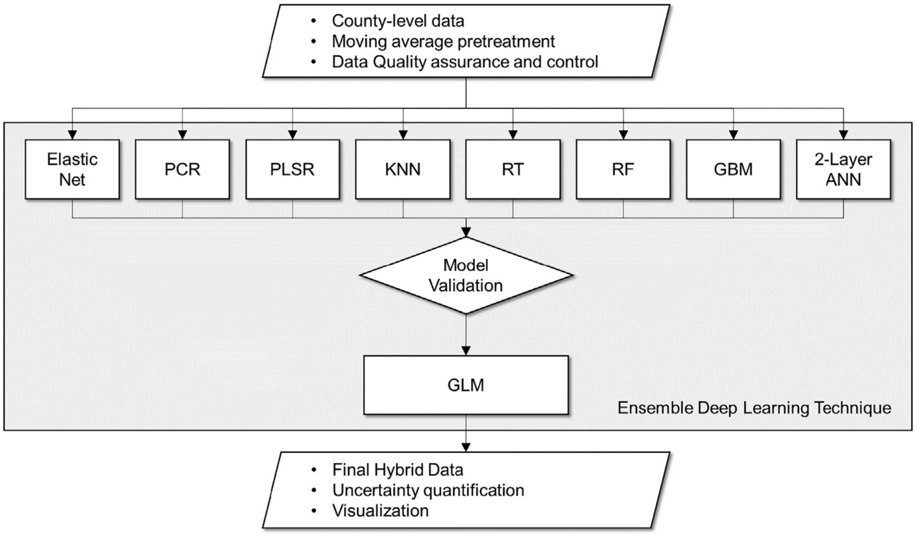

County-level data
Moving average pretreatment
Data Quality assurance and control
Elastic Net
PCR
PLSR
KNN
RT
RF
GBM
2-Layer ANN
Model Validation
GLM
Ensemble Deep Learning Technique
Final Hybrid Data
Uncertainty quantification
Visualization

Fig. 1. Predicting work flow based on ensemble deep learning technique.

near-future trend. Each learner has a different algorithm to build up and train the model, and all input variables were scaled to a mean of 0 and a standard deviation of 1 before modeling. Detailed model settings for each basic learner is presented in Table S2, and each technique has one or several parameters to tune the modeling performance and thus obtain the final model with the best performance.

Among 8 basic learners, the EN model takes advantage of the ridge and lasso regression model, which can retain correlated but important variables by assigning equal weights and also discard unimportant variables. Different mixing parameter (0–0.9) and penalty parameter (10^{-4}-10^{0}) was tested to obtain the best-fitting model with the lowest mean square error (MSE). Both PCR and PLSR are dimension reduction regression models. While PCR extracts potential components by maximizing the variance of each component without regard to the response, PLSR uses the response as a variable to extract the components one by one regarding the residuals. The number of components from 1 to 12 was tested to obtain the least-bias model. KNN model is a non-parametric method and predicts the response by averaging the values of *k* nearest neighbors. The number of *k* ranging from 5 to 10 was tested to obtain the best-tuned KNN model. RT model can divide the variable spaces into several non-overlapping spaces and estimate the prediction with the greatest reduction in errors for each space. We performed top-down recursive binary splitting and tuned the complexity parameter (10^{-5}–10^{-1}). RF model fits a set of decision trees and uses averages from decision trees that are trained on a randomly selected subsample of the training data by the bagging approach. To make the individual tree as independent as possible, randomly choosing a subset of variables and testing the number of subset variables ranging from 1 to 5 was adopted, and the number of tree was set as 500. GBM is also an ensemble of decision trees but uses a boosting approach to fit the model. To be more specific, GBM builds up consecutive trees where solve the net error of prior trees. We tested the number of trees from 0 to 2500 with a learning rate of 0.01 or 0.001 and minimum number of training set samples in a node of 5, and different number of splits (maximum tree depth) including 1,3, and 5 were tested as well. ANN is based on a collection of parallel and interconnected neurons, and the training process uses synaptic weights to store the acquired information in each hidden layer. We used 2-hidden-layer ANN to extract the potential relationship between selected variables and response. Because the performance of ANN is sensitive to its number of nodes in each layer, before ANN modeling, we conducted a sensitivity analysis to test the robustness of the model, and the procedure and results are presented in Appendix A and Table S3. The selected settings with least estimation errors for the numbers of nodes of 2 layers were 29 and 15, respectively.

The accuracy, generalizability, and uncertainty of the trained model was evaluated before its application. To assure modeling accuracy, for each learner, the 10-fold cross-validation was conducted to quantify the uncertainty of modeling performance. Furthermore, the data was split to training data and testing data. The training data was used for constructing models, and testing data was used for validation of the predicted results from the trained model. In this study, 60% of data was used for training data and 40% of data was used for test data. Meanwhile, we also used out-of-samples to validate the model generalizability by applying June 1 data of the selected counties which was not used during training the model.

For each learner, the basic model for daily confirmed cases (D_i) and cumulative cases (C_i) is shown in the following equations.

$$D_{t,i} = \mathrm{f}(V_1, V_2, \cdots, V_n)_j + \delta_{ij} = \widehat{D}_{t,ij} + \delta_{ij}$$

$$C_{t,i} = \mathrm{f}(V_1, V_2, \cdots, V_n)_j + \delta_{ij} = \widehat{C}_{t,ij} + \delta_{ij}$$

$D_{t,\ i}$ or $C_{t,\ i}$ is the future *t*-periodical average (N1D, N4D, or N7D) of cumulative cases or daily cases number at day *i*, $V_1, V_2, \cdots, V_n$ is selected variables for prediction, $\widehat{D}_{t,ij}$ or $\widehat{C}_{t,ij}$ is the *t*-periodical estimation predicted by the machine learning technique *j* at day *i*, and δ_{ij} is the remained error for the machine learning technique *j* at day *i*. Then modeled results with correlation >0.7 would be used as input variables in a further general linear model (GLM) with linear link function and intercept of 0. The GLM was defined as:

$$D_i = \beta_1 \widehat{D}_{i1} + \cdots + \beta_8 \widehat{D}_{i8} + \delta_i = \widehat{D}_i + \delta_i$$

$$C_i = \beta_1 \widehat{C}_{i1} + \cdots + \beta_8 \widehat{C}_{i8} + \delta_i = \widehat{C}_i + \delta_i$$

where D_i and C_i are the final hybrid predictions for daily cases and cumulative cases, respectively, for each county, the input variables $\widehat{D}_{i1}, \cdots, \widehat{D}_{i8}$ and $\widehat{C}_{i1}, \cdots, \widehat{C}_{i8}$ are predictions from previously mentioned 8 basic learners, and $\beta_1 \cdots \beta_8$ are the weighting factors of estimation for 8 basic learners, and δ_i is the remaining error. The modeling performance would be assessed by R-square, root mean square error (RMSE) and mean absolute error (MAE), and R^2, RMSE, and MAE were calculated by:

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \widehat{y}_i)^2}{n}}$$

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y}_i|$$

where y_i is daily confirmed cases or cumulative cases at day *i*, $\widehat{y}_i$ is predicted cases at day *i*, and $\overline{y}_i$ is averaged number of cases among all included counties at day *i*. SAS statistical software (SAS 9.4; SAS Institute Inc., Cary, NC, USA) was used to perform data pretreatment and post-treatment, and R software (version 3.6.2) was used to train the models and yield the final predictions.

3. Results and discussions

3.1. Mobility changes after lockdown

The temporal variation of MEI and daily averaged incidence rate (daily cases per 100,000) for 172 selected counties from February 15 to June 12 are presented in Fig. 2. Among these selected counties, most of them are classified as Urban Influence Codes level-1 (>1 million residents) ($n = 113$) and level-2 (<1 million residents) metropolitan counties ($n = 59$) according to the definition of the US Economic Research Service (Ghelfi and Parker, 2004). Since March 23, states such as New York or California with a serious pandemic announced statewide lockdown order, followed by other states. In late April, states with less impact like Texas started to announce Phase I re-opening and states such as New Jersey and Pennsylvania with more confirmed cases re-opened in early June.

In selected metropolitan counties, MEI dramatically decreased in mid-March due to mandatory stay-at-home order and remained an increasing the trend with an apparent weekly pattern, and the incidence rate has substantially increased from late March until late April and early May. After early May, the increase of mobility but decreasing pandemic in these metropolitan counties indicated that the public has increased awareness and efficiently took the personal protection measures such as social distancing or wearing a mask to decrease the transmission of COVID-19 while maintained or even increased their time spent on outdoor activities.

Meanwhile, after lockdown, both MEI and incidence rate had apparent weekly variation pattern, which MEI had peaks on the weekend and the incidence rate had peaks on Thursday or Friday. The lagged peak of the incidence rate also proved that the median of the incubation period for COVID-19 is 4–5 days from exposure on the weekend to symptoms onset on Thursday or Friday. Also, the weekly pattern of MEI and

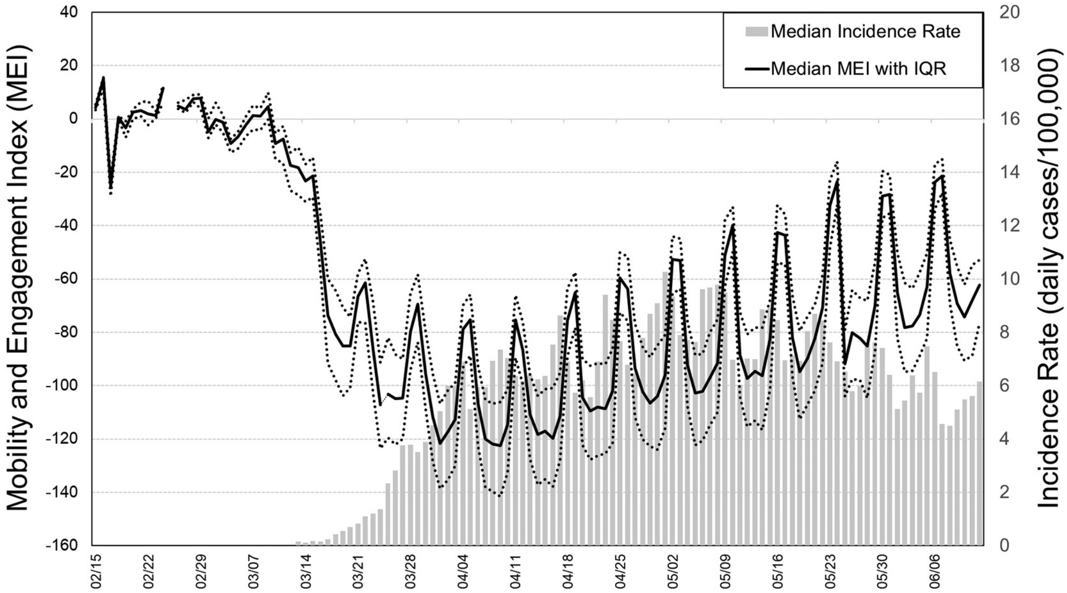


Fig. 2. Temporal variation of averaged daily incidence rate (per 100,000) and Mobility and Engagement Index (MEI) from February 15 to June 12 for selected metropolitan counties (n = 172).

incidence rate suggested that high opportunity for getting infected during the weekend.

Detailed community mobility changes for different places during lockdown are shown in Fig. 3. During the lockdown, the mobility to all public places except for parks decreased for most selected counties, and the time spent at home increased by an average of 18.8%. For those level-1 metropolitan counties such as New York City (including Bronx, Queens, Kings, New York (Manhattan), and Richmond (Staten Island) County) in New York, Cook County in Illinois, Miami-Dade County in Florida, and Los Angeles County in California, the time spent at public places or transit stations decreased more and up to 64.6%. Their time spent at home also increased and was even higher than the median of selected counties. This different mobility between selected counties was not only attributed to lockdown order, residents in these level-1 metropolitan counties could also have higher awareness to prevent transmission due to the denser population.

3.2. Performance of basic learners

The modeling performance of 8 basic learners is shown in Table 2, and the results of cross-validation are presented in Table S4. All base learners have predictions with high correlation with the reported number of daily cases ($R^2 > 0.81$) and cumulative cases ($R^2 > 0.92$). High R^2 for all learners not only resulted from their real correlation, the large sample size could also increase correlation and R^2, so judging with RMSE and MAE is a more appropriate and realistic way to assess the modeling performance.

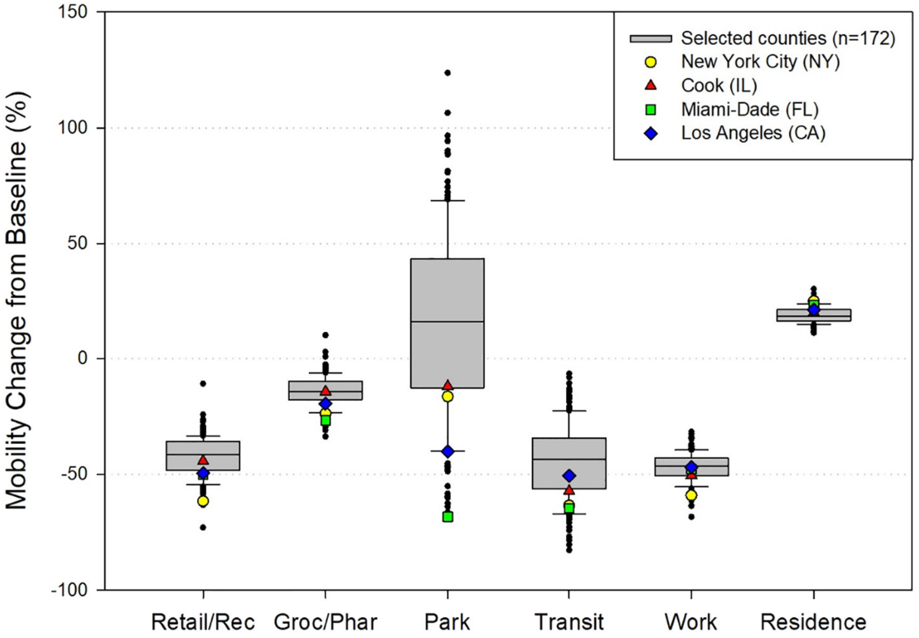


Fig. 3. Averages of community mobility change for selected counties (n = 172) and level-1 metropolitan counties during lockdown.

Table 2
Modeling performance of basic learners and GLM-hybrid results.

Index[b]		Model[a]																
		Training samples								Test samples								GLM hybrid
		EN	PCR	PLSR	KNN	RT	RF	GBM	ANN	EN	PCR	PLSR	KNN	RT	RF	GBM	ANN	
Daily cases																		
R^2	N1D	0.89	0.81	0.89	0.81	0.86	0.89	0.87	0.89	0.89	0.85	0.89	0.86	0.85	0.88	0.89	0.89	0.91
	N4D	0.94	0.84	0.93	0.87	0.91	0.94	0.96	0.93	0.92	0.87	0.92	0.90	0.89	0.94	0.95	0.94	0.96
	N7D	0.93	0.84	0.93	0.87	0.94	0.95	0.97	0.96	0.91	0.86	0.91	0.89	0.89	0.93	0.95	0.94	0.97
RMSE	N1D	97	125	98	122	116	107	101	96	110	128	111	122	129	122	108	110	105
	N4D	80	109	79	99	90	83	60	70	88	113	90	101	108	90	72	78	66
	N7D	74	106	73	95	71	76	52	59	95	118	96	105	106	93	70	79	62
MAE	N1D	38	60	39	56	42	40	37	38	39	59	39	56	41	40	38	41	39
	N4D	32	55	33	48	29	28	25	29	33	55	34	49	31	28	26	30	28
	N7D	32	54	32	46	26	24	22	27	34	56	34	48	28	25	23	29	27
Cumulative cases																		
R^2	N1D	1.00	0.93	1.00	0.94	0.99	0.99	1.00	1.00	1.00	0.95	1.00	0.95	0.99	0.99	1.00	1.00	1.00
	N4D	1.00	0.94	1.00	0.94	0.98	0.98	1.00	1.00	1.00	0.95	1.00	0.95	0.99	0.99	1.00	1.00	1.00
	N7D	1.00	0.93	0.99	0.92	0.99	0.99	1.00	1.00	1.00	0.95	0.99	0.95	0.99	0.99	1.00	1.00	1.00
RMSE	N1D	407	3160	534	3107	1153	2240	300	280	447	3091	541	3112	1196	2672	276	261	189
	N4D	538	3194	834	3172	1481	2204	339	514	556	3091	860	3203	1302	2591	263	499	187
	N7D	544	3251	1140	3309	1497	2013	403	768	461	3126	1183	3297	1742	2611	332	606	258
MAE	N1D	155	1894	222	1699	260	605	77	140	158	1858	219	1686	268	617	75	133	103
	N4D	227	1933	350	1763	323	601	110	238	226	1876	346	1746	310	634	106	236	100
	N7D	221	1956	477	1838	358	600	149	341	181	1900	474	1806	383	654	145	265	118

[a] **EN**: elastic net model; **PCR**: principal components regression model; **PLSR**: partial least squares regression model; **KNN**: k-nearest neighbors regression model; **RT**: regression tree model; **RF**: random forests model; **GBM**: gradient boosted tree models; **ANN**: 2-layer artificial neural network model; **GLM**: general linear model.

[b] **RMSE**: root mean square error; **MAE**: mean absolute error; **N1D**: future-1-day average; **N4D**: future-4-day average; **N7D**: future-7-day average.

For daily cases, the results showed that N7D has lower RMSE (52–106 cases) and MAE (22–54 cases) in training samples and better performance compared with N1D and N4D. The results of cross-validation also showed the similar tendency that N7D had lower uncertainty. On the other hand, for cumulative cases, N1D has lower RMSE (300–3160 cases) and MAE (77–1894 cases) compared with N4D and N7D. Similarly, the N1D had lower uncertainty based on its cross-validation results. Overall, the basic learners could predict better for N7D daily cases and N1D cumulative cases with the least average estimation error. The reason for this tendency of modeling performance is that the models perform well on more stable data and is easily biased with the peak value. For example, when the number of daily cases increases dramatically, the peak value during the next 1 to 7 days could be smoothed and averaged to N7D daily cases. On the contrary, the trend of cumulative cases is comparably stable and less affected by the sudden increase of cases, thus the N1D cumulative case is comparably easily predicted compared with the N7D cumulative cases.

To discuss about the strengths of include basic learners, among 8 basic learners, RF, GBM, and ANN had better performance for predicting daily cases, and EN and GBM predicted well for cumulative cases due to their lower RMSE and MAE. One of the strengths of these models is no assumption of a linear relationship between predicting variables and response, so the linear and non-linear relationship can be built through the models (Schonlau and Zou, 2020; Somers and Casal, 2009). Also, these models have the advantage to select important variables and discard the noise variables during the modeling (Zou and Hastie, 2005), thus the useful information can be filtered to obtain their potential relationships with the response without interference from noise information. Meanwhile, models such as RF, GBM, and ANN have no assumption of a normal distribution for predicting variables, response, and residuals (Schonlau and Zou, 2020; Somers and Casal, 2009). So even though the included variables are not normally distributed or need transformation, these models could still build up a stable model with acceptable accuracy. In addition, all of the basic learners have no significant overfitting when using test samples for validation.

3.3. Discussion of GLM hybrid technique performance

The results of 8 basic learners were further combined by GLM, and the GLM hybrid results are presented in Table 2. The GLM hybrid results enhanced the accuracy of prediction and lowered the estimated uncertainty compared with most basic learners. The strength of using GLM to hybrid the results from basic learners is that its retaining more accurate predictions by using higher weighting factors while down-weighing the poor-correlated predictions by lowering their weighting factors. Because each basic learner has their own strengths and weaknesses, and their modeling performance must be different when applying to different datasets. Using GLM hybrid technique could combine all modeling results, assemble their results by regulating their weights, and enhance the performance of hybrid results.

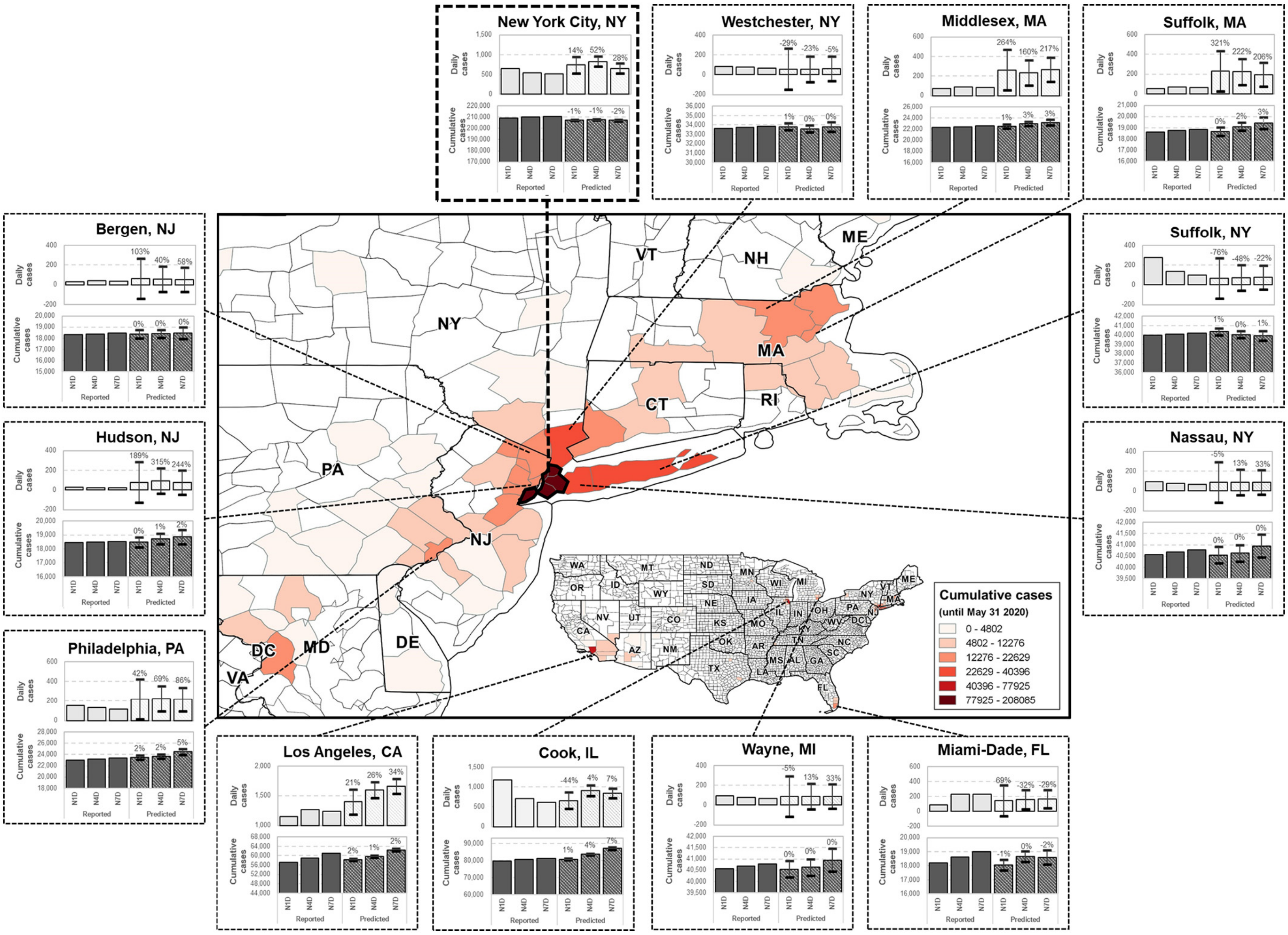


Fig. 4. Reported cases and predicted cases from June 2 with 95% confidence intervals based on June 1 and historic data for New York City and other top 12 counties with the most cumulative cases until May 31.

3.4. Generalizability of our model

Except for training and validating the model, to illustrate of generalizability of our proposed model, we selected New York City and the other top 12 counties with the most cumulative cases and used their June 1 data which was not included in training and testing data to predict the near-future trend of daily cases and cumulative cases from June 2. This validation can be viewed as out-of-sample validation, because the data we used to predicted was not included during training the model. The predicted results of these counties are shown in Fig. 4, and the bias of mean predicted values (%) compared with reported data was also illustrated on the predicted values in Fig. 4. For daily cases, the reported numbers located in the range of the predicted numbers with 95% confidence intervals (CIs) for most metropolitan counties such as Miami-Dade County in Florida. Concerning the prediction bias of N7D averages, the bias ranges from −59% to 244%, which represents the model still has limited capability to predict the near-future trend for some counties like Hudson in New Jersey, Middlesex and Suffolk in Massachusetts which had daily cases less than 100, and these counties also had large uncertainty. To discuss the reason for these biases, it could be due to the model cannot capture the trend without specific county-dependent characteristics that are related to the transmission of the pandemic, and other variables should be included for better performance. The estimation error in selected counties was also averaged out by the other counties with have fewer cases and lower estimation error, so the overall estimation error still remained low. For cumulative cases, the model has good performance in most of the selected counties, and the N1D bias ranges from −1% to 5%. It is worth noting that, for counties which had large population such as New York City, the bias could be alleviated by its large cumulative cases, thus the bias is much lower compared with predicted daily cases.

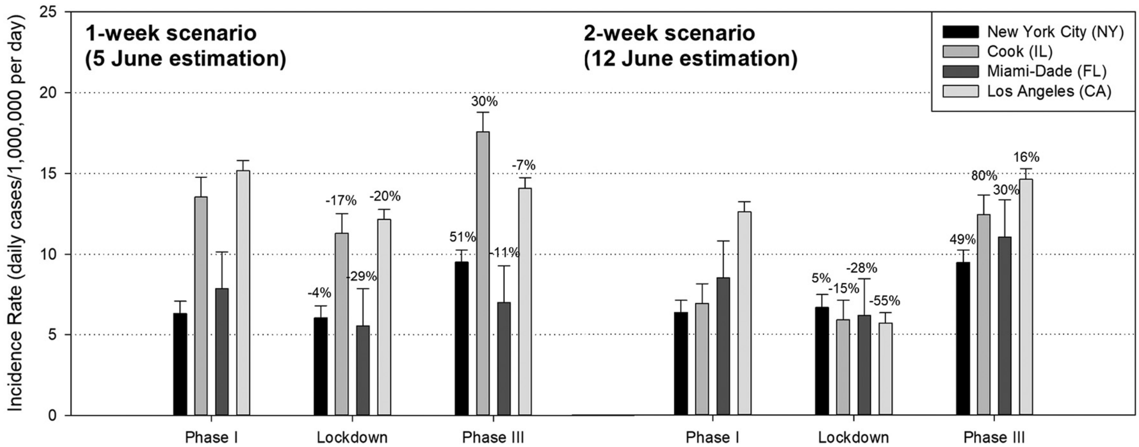
3.5. Discussion of the impact of lockdown, Phase I, and Phase III re-opening

To assess the impact of lockdown, Phase I and III re-opening, we applied different assumed mobility to our model for level-1 metropolitan counties including New York City in New York, Cook County in Illinois, Miami-Dade County in Florida, and Los Angeles County in California. We assessed how will lockdown and different phases lasting 1 week or 2 weeks affect the trend of N7D cases. The modeled results under different scenarios for 1 week or 2 weeks are shown in Fig. 5, and the differences of lockdown and Phase III re-opening compared with Phase I re-opening (%) are also shown on the predicted incidence rate in Fig. 5. For the Phase I re-opening scenario, real mobility, and MEI data were directly applied because during early June, all selected counties have announced Phase I re-opening. For the lockdown scenario, compared with baseline, we applied −50% to retail and recreation sites, transit stations, and workplaces; −25% to grocery and pharmacy stores; +25% for going to parks and staying at home; median of MEI during the lockdown. For the Phase III re-opening scenario, we assumed the mobility to retail and recreation sites, grocery and pharmacy stores, and transit stations was the same compared with the baseline; applied −20% to workplaces; +50% to parks; +20% for staying at home; real MEI during early June.

For a 1-week scenario since May 30, the modeled N7D results on June 5 showed that the 1-week lockdown has reduced the incidence rate in the following week for all selected counties, and the reduction rate ranges from 4% to 29%. For the 2-week scenario, the modeled N7D results on June 12 showed that the incidence rate for Cook County in Illinois, Miami-Dade County in Florida, and Los Angeles County in California have reduced 15%–55%, which is much higher than a 1-week lockdown. But for New York City, the incidence rate even bounces back and increases after a 2-week lockdown, and it could be due that residents in New York City had already efficiently took self-protection measures such as wearing masks or social distancing at the early stage of lockdown, so the change of mobility could only have limited impact on the incidence rate.

Meanwhile, if Phase III re-opening last 1 week, the incidence rate for New York City and Cook County in Illinois increase by 51% and 30% respectively, but for Miami-Dade County in Florida and Los Angeles County in California, there is no significant increase. For a 2-week Phase III re-opening scenario, all selected counties have significant increase of the incidence rate ranging from 16% to 80%.

The significant difference between lockdown, Phase I and III re-opening in 1-week and 2-week scenario also suggested that the mandatory orders should last longer than 1 week, the effectiveness of orders could be observed due to the change of community mobility, and the effectiveness of orders could be highly county-dependent and varied with the trend of the local pandemic. For example, even after 2-week lockdown, the reduced incidence rate in Cook County in Illinois (15%) is lower than Miami-Dade County in Florida (28%) and Los Angeles County in California (55%).



- **Phase I**: apply reported mobility values from Google and MEI, because all 4 areas underwent Phase I re-opening during this period.
- **Lockdown**: comparing wit baseline, apply -50% to retail and recreation sites, transit stations, and workplaces; -25% to grocery and pharmacy stores; +25% for going to parks and staying at home; median of MEI during lockdown.
- **Phase III**: comparing with baseline, apply no change to retail and recreation sites, grocery and pharmacy stores, and transit stations; -20% to workplaces; +50% to parks; +20% for staying at home; real MEI reported by Google.

Fig. 5. Modeled next-7-day (N7D) results under Phase I re-opening, lockdown, and Phase III re-opening lasting 1 week and 2 weeks for selected level-1 metropolitan counties.

3.6. Discussion of the strengths of the proposed method

Our method has several strengths. First, our method can deal with data with a non-linear relationship and non-normal distribution, thus the treatment of normal distribution transformation is not required for input data. With less assumption of a dataset and discarding noise information, the model still can extract useful information and build up their potential relationship between included variables and response. Second, the GLM hybrid technique can utilize the results of basic learners, and further lower the uncertainty and improve the modeling performance. Third, our model can be applied with different assumed scenarios to assess the effectiveness of the lockdown and different re-opening phases. For example, our assessment pointed out that if the lockdown order lasts 2 weeks, the incidence rate in the future week will decrease by 15%–55%. On the other hand, if the government announced Phase III re-opening, the infections will increase by 16%–80% after 2 weeks.

3.7. Limitation, uncertainty and future works

This study still has some limitations. First, in generalizability test, although our model has good performance on training data and testing data, the modeling still remained improvement to predict the near-future daily cases with high uncertainty for counties such as Hudson in New Jersey, Middlesex and Suffolk in Massachusetts. The reason for this inferiority could be due to lacking specific county-dependent characteristics that are related to the transmission of the pandemic, thus more county-dependent variables should be included in the future works. Second, only about 2-month data for each county was included during modeling. Indeed, longer-term data could increase the accuracy and robustness of the model while reduce the uncertainty. The impact of seasonality on COVID-19 pandemic in U.S. should be further investigated in the future works. Third, we only performed basic machine learning techniques to predict near-future trend for urgently needed suggestions to the officials and the public. and more time-series models such as Long Short-Term Memory (LSTM) model (V. Kumar et al., 2020), Autoregressive Integrated Moving Average (ARIMA) model (Benvenuto et al., 2020; A. Kumar et al., 2020), sophisticated models such as Convolutional Neural Network (CNN) (Zhao et al., 2017), or other techniques combining the different machine learning methods still remain great potential to develop.

4. Conclusion

In this study, the COVID-19 cases predicting model based on county-level demographic characteristics, environmental data, and mobility data with least-error predictions of the N7D daily case and N1D cumulative cases was successfully built up.

First, we pointed out that the community mobility in the metropolitan counties substantially dropped off since mid-March after the lockdown announcement and remained an increasing trend with an apparent weekly pattern. The lagged peak with weekly patterns for mobility and daily cases also implied high infections during the weekend. Second, our assessment for New York City, Cook County in Illinois, Miami-Dade County in Florida, and Los Angeles County in California showed that the mandatory orders should last longer than 1 week, the effectiveness of orders could be significant due to community mobility change. Compared with Phase I re-opening, a 1-week lockdown could reduce 4–29% infections in the future week. If the lockdown last 2 weeks, the infections could be reduced more by 15%–55% in the future week. On the contrary, the infections would increase by 16%–80% if Phase III re-opening last 2 weeks according to our modeled results. The impact of the lockdown or Phase III re-opening was also highly county-dependent and varied with the trend of the local pandemic.

In future work, we look forward to involving a longer period of data, considering more county-dependent factors, and employing more sophisticated models to decrease the modeling uncertainty and apply the model to counties nationally or other countries as well.

CRediT authorship contribution statement

Cheng-Pin Kuo: Writing – original draft, Writing – review & editing. **Joshua S. Fu:** Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2020.144151.

References

Abatzoglou, J.T., 2013. Development of gridded surface meteorological data for ecological applications and modelling. Int. J. Climatol. 33, 121–131. https://doi.org/10.1002/joc.3413.

Atkinson, T., Dolmas, J., Koch, C., Koenig, E., Mertens, K., Murphy, A., Yi, K.-M., 2020. Dallas fed mobility and engagement index gives insight into COVID-19's economic impact [WWW document]. URL. https://www.dallasfed.org/research/economics/2020/0521.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., Ciccozzi, M., 2020. Application of the ARIMA model on the COVID- 2019 epidemic dataset. Data Br. 29, 105340. https://doi.org/10.1016/j.dib.2020.105340.

CDC, 2020. Management of patients with confirmed 2019-nCoV | CDC. Coronavirus Dis. 2019.

Coibion, O., Gorodnichenko, Y., Weber, M., 2020. The cost of the COVID-19 crisis: lockdowns, macroeconomic expectations, and consumer spending. SSRN Electron. J. https://doi.org/10.2139/ssrn.3593848.

Ghelfi, L., Parker, T., 2004. Developing a county-level measure of urban influence. Amber Waves 12, 32–41.

Kumar, V., Chimmula, R., Zhang, L., 2020a. Time series forecasting of COVID-19 transmission in Canada using LSTM networks R. Chaos, Solitons Fractals 135. https://doi.org/10.1016/j.chaos.2020.109864.

Kumar, A., Rath, N., Sood, V., Pratap, M., 2020b. ARIMA modelling & forecasting of COVID-19 in top five affected countries. Diabetes Metab. Syndr. Clin. Res. Rev. 14, 1419–1427. https://doi.org/10.1016/j.dsx.2020.07.042.

Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., Yan, J., Niu, J., Zhou, J., Luo, B., 2020. Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. Sci. Total Environ. 724, 138226. https://doi.org/10.1016/j.scitotenv.2020.138226.

Mahato, S., Pal, S., Ghosh, K.G., 2020. Effect of lockdown amid COVID-19 pandemic on air quality of the megacity Delhi, India. Sci. Total Environ. 730, 139086. https://doi.org/10.1016/j.scitotenv.2020.139086.

Mckibbin, W., Fernando, R., 2020. Crawford School of Public Policy CAMA Centre for Applied Macroeconomic Analysis the Brookings Institution Centre of Excellence in Population Ageing Research the Global Macroeconomic Impacts of COVID-19: Seven Scenarios * 2, 12–22.

Schonlau, M., Zou, R.Y., 2020. The Random Forest Algorithm for Statistical Learning 3–29. https://doi.org/10.1177/1536867X20909688.

Singh, R.K., Rani, M., Bhagavathula, A.S., Sah, R., Rodriguez-Morales, A.J., Kalita, H., Nanda, C., Sharma, S., Sharma, Y.D., Rabaan, A.A., Rahmani, J., Kumar, P., 2020. Prediction of the COVID-19 pandemic for the top 15 affected countries: advanced autoregressive integrated moving average (ARIMA) model. JMIR Public Heal. Surveill. 6, e19115. https://doi.org/10.2196/19115.

Somers, M.J., Casal, J.C., 2009. Using Artificial Neural Networks to Model Nonlinearity 403–417.

Wang, J., Tang, K., Feng, K., Lv, W., 2020. High temperature and high humidity reduce the transmission of COVID-19. SSRN Electron. J. https://doi.org/10.2139/ssrn.3551767.

Xie, J., Zhu, Y., 2020. Association between ambient temperature and COVID-19 infection in 122 cities from China. Sci. Total Environ. 724, 138201. https://doi.org/10.1016/j.scitotenv.2020.138201.

Xu, C., Yu, Y., Yang, Q., Lu, Z., 2020. Forecast Analysis of the Epidemics Trend of COVID-19 in the United States by a Generalized Fractional-order SEIR Model 1–10.

Yang, Z., Zeng, Z., Wang, K., Wong, S.S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z., Liang, J., Liu, X., Li, S., Li, Y., Ye, F., Guan, W., Yang, Y., Li, F., Luo, S., Xie, Y., Liu, B., Wang, Z., Zhang, S., Wang, Y., Zhong, N., He, J., 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J. Thorac. Dis. 12, 165–174. doi:10.21037/jtd.2020.02.64.

Zhao, B., Lu, H., Chen, S., Liu, J., Wu, D., 2017. Convolutional Neural Networks for Time Series Classification 28, 162–169. doi:10.21629/JSEE.2017.01.18.

Zou, H., Hastie, T., 2005. Regularization and Variable Selection Via the Elastic Net 301–320.