Developing a Machine Learning Framework to Determine the Spread of COVID-19

Akash Gupta*, Amir Gharehgozli
David Nazarian College of Business and Economics, California State University, Northridge, CA
akash.gupta@csun.edu, amir.gharehgozli@csun.edu.

Abstract

Coronavirus disease of 2019 (COVID-19) has become pandemic in the matter of a few months, since the outbreak in December 2019 in Wuhan, China. We study the impact of weather factors including temperature and pollution on the spread of COVID-19. We also include social and demographic variables such as per capita Gross Domestic Product (GDP) and population density. Adapting the theory from the field of epidemiology, we develop a framework to build analytical models to predict the spread of COVID-19. In the proposed framework, we employ machine learning methods including linear regression, linear kernel support vector machine (SVM), radial kernel SVM, polynomial kernel SVM, and decision tree. Given the non-linear nature of the problem, the radial kernel SVM performs the best and explains 95% more variation than the existing methods. In align with the literature, our study indicates the population density is the critical factor to determine the spread. The univariate analysis shows that a higher temperature, air pollution, and population density can increase the spread. On the other hand, a higher per capita GDP can decrease the spread.
KEYWORDS: COVID-19; disease spread; social and demographic factors; machine learning; epidemiology; predictive modeling

1 Introduction

Coronavirus disease of 2019 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which belongs to the family of Coronaviruses [1]. It was first diagnosed, as a atypical case of pneumonia, in Wuhan, China in December 2019 and since then it has spread to all around the world. The intensity of the spread of this virus is higher than the previously known viruses such as of influenza, SARS (severe acute respiratory syndrome), MERS (Middle East respiratory syndrome) and Ebola [2]. Considering the highly contagious nature of the virus, it is critical to understand the factors that could potentially affect the spread of COVID-19.

In the United States (US), the number of confirmed COVID-19 cases and deaths have suddenly peaked and taken over the rest of the world. According to the Center for Disease Control and Prevention (CDC), from January 14, 2020 to April 16, 2020, the number of confirmed COVID-19 cases in the US has increased 3 to a whopping 685,686 cases [3]. The heatmap in Figure 1 shows

*Corresponding author: 18111 Nordhoff Street, Northridge, CA 91330-8378, the USA. Tel: +1 (818) 677-2413, Email: akash.gupta@csun.edu.

the distribution of confirmed COVID-19 cases in the US as of April 16, 2020. The majority of the investigated cases is due to close contact with an infected individual. In this study, we focus building predictive model using the COVID-19 case incidence data from the US.

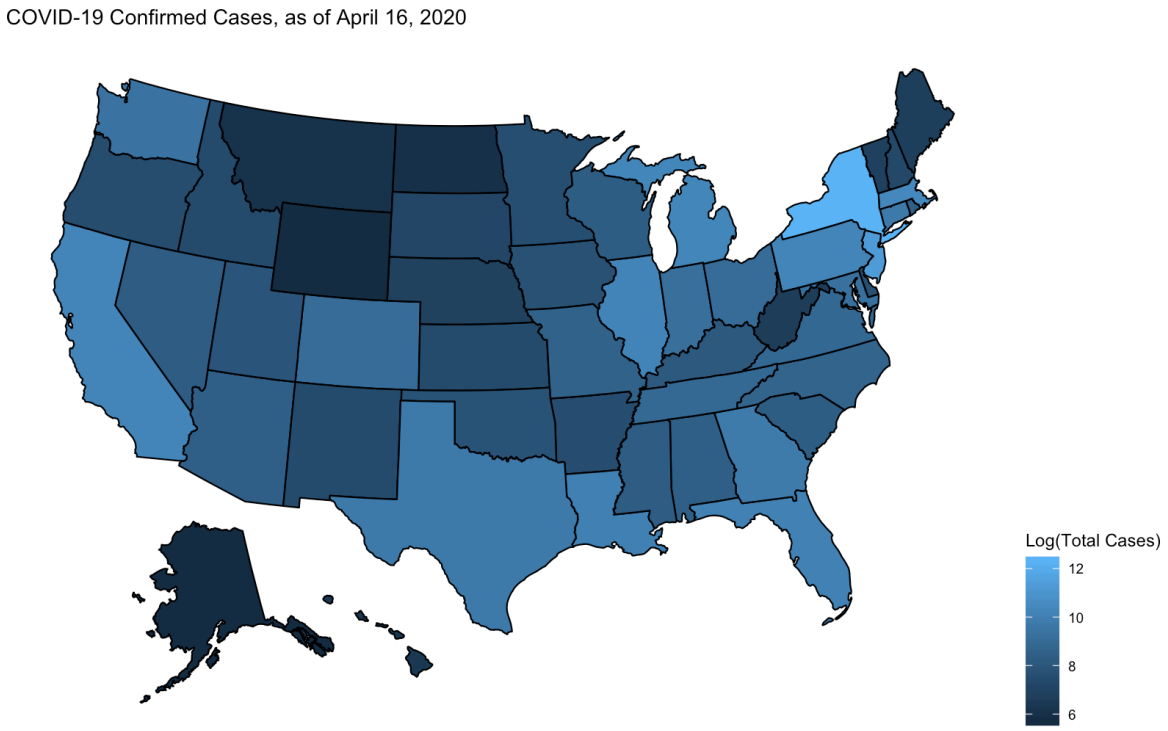

Figure 1: Distribution of confirmed COVID-19 cases in the United States, as of April 16, 2020

We adapt the SIR (i.e., susceptible (S), infected (I), or recovered (R)) compartmental approach from the field of epidemiology to determine the spread of COVID-19. The underline epidemiological phenomenon is depicted in Figure 2. In the beginning, we have the whole population that is susceptible to infection. A fraction of susceptible population gets infected due to the spread of the virus. The infected population receives treatment to recover. Depending on the severity of the illness, some of the infected people may recover and some may succumb to death. We apply the SIR model to approximate the incidence curve [4] and estimate the average number of infectee from a single infector, widely known as a *reproduction number* R_0. In other words, the reproduction number is an indicator of the spread of COVID-19.

Following the estimation of R_0, we develop analytical models to study the impact of several factors including weather, population density, per capita Gross Domestic Product (GDP) on the spread of COVID-19 in the US. These factors have proven to be critical in pandemics (see the

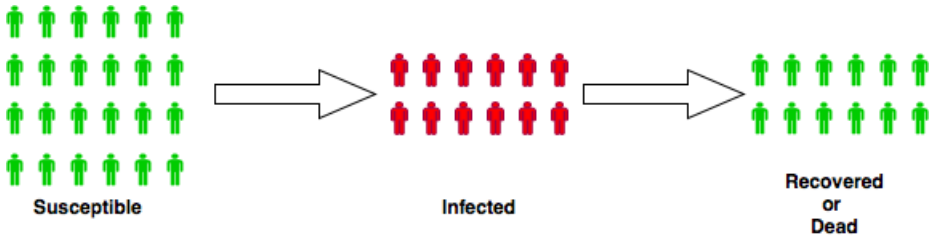

Figure 2: Epidemiological phenomenon

literature review section). The estimated R_0 is considered as a response variable. We used supervised machine learning methods such as linear regression, Support Vector Machine (SVM), and decision tree to develop analytical models. We collected data for our predictors from multiple sources such as CDC, World Health Organization (WHO), the Iowa Environmental Mesonet, AirNow, the US Census Bureau, and the Bureau of Economic Analysis. We have performed comprehensive numerical experiments to determine the best approach to predict the spread of virus. The contributions of the paper are as follows:

1. estimate the spread of COVID-19 across the states in the US.
2. develop a framework to derive a predictive model for the spread of COVID-19.
3. determine the important factors affecting the spread.
4. compare the performance of different supervised machine learning methods.

The paper is structured as follows. In Section 2 we briefly review the literature. In Section 3, we describe the analytical model. Section 4 is dedicated to discuss our findings and insights. Finally, Section 5 concludes the paper and deliberates directions for future research.

2 Literature review

Among disasters, pandemics are a national and international public health concern. In outbreaks of epidemics, the lives of people are at stake [5]. Furthermore, they have sever negative consequences on the (world) economy by disrupting and introducing risks, vulnerabilities, and uncertainties to today's supply chains spreading all over the globe [6]. As a matter of fact, naturally occurring disease outbreaks fall under the umbrella of public health security, as an essential part of national security, can be achieved by effectively preventing, detecting, and responding to events that affect

public health [7]. Therefore, there is an increasing body of research research on modeling the spread [8].

[9] categorizes the research efforts as follows : (1) modeling the spread and path of a pandemic [10; 11; 12], (2) studying the impact of (non-)pharmaceutical interventions [13; 14; 15; 16], (3) response and preparedness planning [17; 18; 19]. The latter two categories can benefit from the findings of the models developed in the first category to understand the spread of a disease geographically and over time. For example, [20] integrate a spread pattern model with a facility location and resource allocation network model to study food distribution planning during an influenza pandemic.

Due to the scope of this paper, which is on the rate of spread, we focus on the first category in the rest of the literature review section. COVID-19 is often transmitted from human to human by viral particle transmission from an infected individual to a susceptible individual by various mechanisms, often through sneezing or coughing, similar to influenza [21; 22]. In general, four common methods can be used to model the spread [20; 23]: (1) differential equations [24; 25], (2) (agent-based) simulation [26; 27; 28], (3) random graphs [29], and (4) difference equations [29; 30]. The most common approach is to use differential equations and model the spread following a SIR compartmental approach where each person in the population is susceptible (S), infected (I), recovered (R), or dead [31]. These models are mainly based on the basic reproductive number R_0 which is defined as the average number of people to whom an infected person spreads in a fully susceptible population [32]. If $R_0 > 1$, then the disease can spread [33]. Furthermore, $R(t)$ is the average number of secondary infected individuals on generation t. The models which incorporate R_0 generally use the following approaches: express R_0 in terms of parameters that describe the virulence and morbidity [34], fit the parameter to data [35], or use endemic equilibrium data to derive the parameter [36]. One of the main limitations of these approaches is that they are negligent to the dynamic epidemic situation, which can vary significantly among populations [37].

Last but not least, there is an extensive literature on risk factors of outbreaks [8; 38; 39; 40]. Relevant to our study, weather is recognized among the factors impacting the spread [41; 42]. High concentration of air pollutants such as PM10 and O3 [43; 44], and low temperature can increase the spread and mortality [45; 46]. In contrast, some emphasize that although a higher temperature blocks aerosol transmission but it does not decrease contact transmission [47; 48]. Furthermore, [49]

maintain that social and demographic factors which include both individual characteristics (such as age, gender residence, and the level of general interaction with other members of the population), population-wide characteristics (such as population size, population density, age distribution, and distribution of household size) can impact the spread. They also assert that behavioral factors including daily commutes and attendance at schools, workplaces, and hospitals need to be as well considered among the predictors of the spread. [50] combine demographic predictors (e.g. age, gender) and surrounding environment conditions (e.g. location, weather, nearby terrain features) to predict a person's risk of being infected by a disease.

3 Analytical method

In this section, we describe the framework used to develop the models (see Figure 3). First, in Section 3.1, for each state of the US, we estimate reproduction number R_0 from the daily COVID-19 incidence data collected from John Hopkins University web portal [51]. Second, we extracted information about the predictors from multiple sources explained in Section 3.2. Finally, in Section 3.3, considering R_0 as a response, we develop predictive models and then compare the performance.

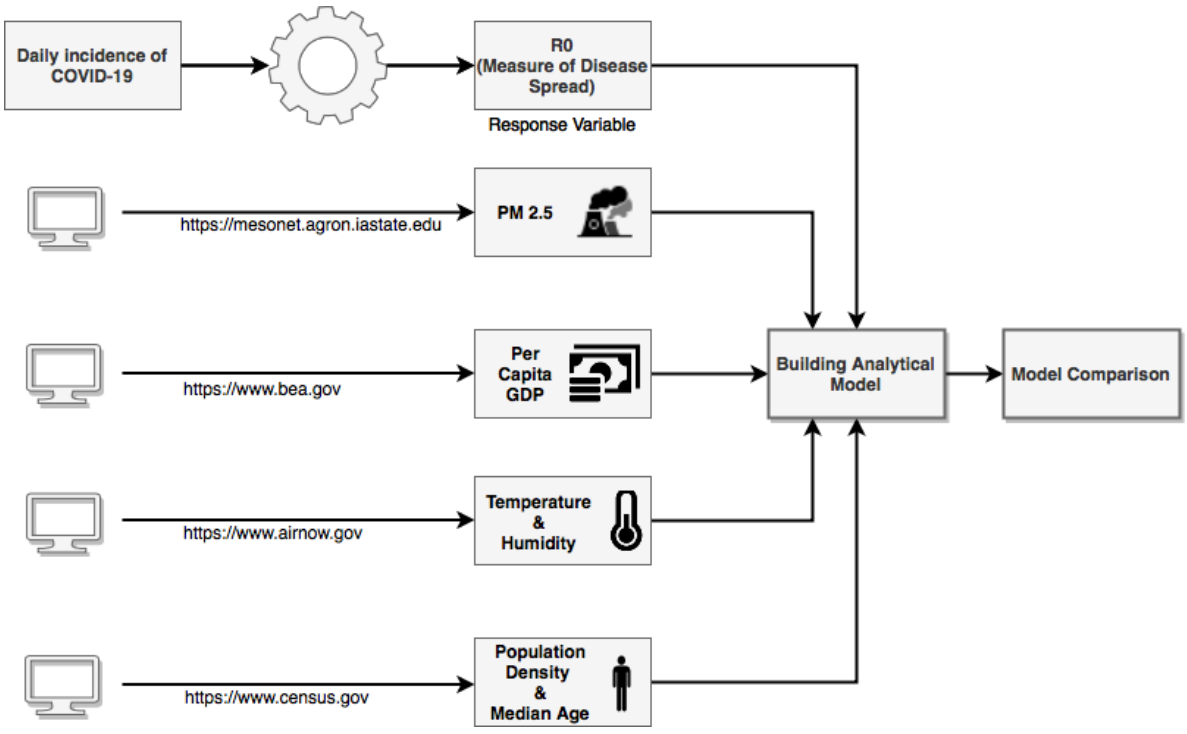

Figure 3: Framework to build predictive model

3.1 Estimating Reproduction Number R_0

The epidemiological phenomenon for each state in the US can be graphically illustrated using Figure 4. The three nodes represent three phases of infection progression in the SIR compartmental approach: susceptible, infected and recovered/died. The arcs indicate the rate of change from one phase to another. The notations are explained in Table 1.

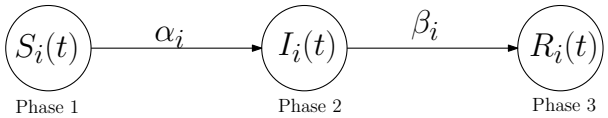

Figure 4: Graphical representation of epidemiological model

Table 1: Terminology

Notation	Explanation
i	index denoting states in the US
$S_i(t)$	susceptible population size at time t for State i
$I_i(t)$	infected population size at time t for State i
$R_i(t)$	recovered or dead population size at time t for State i
α_i	transmission rate from susceptible to infected phase
β_i	recovery rate from infected to recovery phase
$\mathbf{x}_i$	vector representing input variables for State i
$R_{0,i}$	reproduction number at State i
$\mathbf{R_0}$	vector representing all reproduction numbers $\{R_{0,Alabama}, R_{0,Alaska}, \ldots\}$

The rate of change of the susceptible population for State i is given by Equation 1. The size of the susceptible population decreases as more people get infected. $S_i(t)I_i(t)$ represents all the possible interactions between the susceptible and infected populations. Multiplying the product by the transmission rate (α_i) results in a population that moves from susceptible to infected.

At Node 1, representing Phase 1 of the disease (susceptible population):

$$\begin{aligned} S_i(t+\delta t) =& S_i(t) - \alpha S_i(t) I_i(t) \delta t \\ \frac{dS_i}{dt} =& -\alpha S_i(t) I_i(t) \end{aligned} \tag{1}$$

The rate of change of the infected population for State i is given by Equation 2. The change in the infected population is reflected by the newly infected population subtracted by the population that gets recovered or died. β_i denotes the recovery rate for State i. The percentage of recovered

people is relatively higher than the percentage of people who died, hence we consider β_i as the recovery rate.

At Node 2, representing Phase 2 of the disease (infected population):

$$\begin{aligned} I_i(t+\delta t) =& I_i(t) + \alpha_i S_i(t) I_i(t) \delta t - \beta I(t) \delta t \\ \frac{dI_i}{dt} =& \alpha_i S_i(t) I_i(t) - \beta_i I_i(t) \end{aligned} \tag{2}$$

The rate of change of the recovery population is given by Equation 3. The change in the recovery population is governed by transition from the infected phase to the recovery phase.

At Node 3, representing Phase 3 of the disease (recovered/died population):

$$\begin{aligned} R_i(t+\delta t) =& R_i(t) + \beta_i I_i(t) \delta t \\ \frac{dR_i}{dt} =& \beta_i I_i(t) \end{aligned} \tag{3}$$

We collected the COVID-19 incidence data for each state in 24-hour intervals. At t=0, $I_i(0) = I_i^0$ is the initial infected population for State i; $S_i(0) = S_i^0 - I_i^0$, where S_i^0 is the population of the State i. The state population data is collected from the US Census Bureau (https://www.census.gov/).

In Equation 2, term $\frac{\alpha_i S_i(t)}{\beta_i}$ is known as a *reproduction number* for State i ($R_{0,i}$). The reproduction number can be estimated at different time instances during the outbreak. The reproduction number computed using the data from the beginning of the outbreak is specifically known as the initial reproduction number, which reflects the severity of the spread of the infection. An effective estimate for the reproduction number (reflection of disease spread) R_0 can be derived by minimizing the effect of intervention. To accomplish this, we use the incident data of early days (till March 24, 2020) that had a minimal effect of enforcement of social distancing guidelines recommended by the CDC on March 14, 2020. If R_0 is greater than one, the infection spreads. On the contrary, if R_0 is less than one, the infection phases out.

To estimate the initial reproduction number, we employed three methods: Exponential Growth [52], Maximum Likelihood Estimation (ML) [53] and Time Dependent Reproduction Numbers (TD) [54]. These methods require prior information of the distribution of *serial interval*. The serial

interval is the time difference between the symptoms onset of the primary case (infector) and the secondary case (infectee). From [55], we adapted Weibull distribution for serial intervals with mean and standard deviation of 7.4 and 5.2, respectively.

3.2 Extracting Predictors

To develop the models, we extracted the information about following variables for each state in the US: median age, PM2.5, population density, per capita GDP, average temperature and average humidity. The weather data (temperature and humidity) was collected from AirNow (https://airnow.gov/). We used the state capital weather station to extract the temperature and humidity time series data. The mean of all the data points were used to reflect the average temperature of the state.

We included a pollution variable in the model. Specifically, we included PM2.5, fine inhalable particles, with diameters that are generally 2.5 micrometers and smaller. The data for this variable is collected from the Iowa Environmental Mesonet, which collects environmental data from cooperating members with observing networks (https://mesonet.agron.iastate.edu/).

We also incorporated population density and per capital GDP for developing the analytical models. The data for the per capita GDP is collected from the Bureau of Economic Analysis (https://www.bea.gov/). We believe that population density could be an important factor because of the human to human transmission of the disease. The statistical summary of all six variables is shown in Table 2.

Table 2: Statistical summary of variables

Variable	Mean	Standard deviation	Minimum	Maximum
Median Age	38.32	2.38	31	45
PM2.5	5.88	1.48	2	8
Population Density	154.90	184.83	1	867
Per Capita GDP	47634.00	8871.10	31504	67278
Temperature (°F)	47.24	10.84	31	74
Humidity	34.85	11.67	16	64

3.3 Predictive models

In this section, we briefly describe the machine learning techniques applied in this paper including linear regression, SVM and decision tree.

3.3.1 Linear regression

Linear regression is used to obtain a linear relationship between a dependent variable (spread of COVID-19 reflected by R_0) and independent variables (median age, PM2.5, population density, per capita GDP, temperature and humidity) [56]. The linear regression computes weights corresponding to each variable to determine the variable significance. The greater the absolute value of weight, the greater the significance of the variable.

The multiple regression model shown in Equation 4 describes how the dependent variable R_0 is related to the independent variables.

$$\begin{aligned} R_0 =& w_0 + w_1 \times \text{median age} + w_2 \times \text{PM2.5} + w_3 \times \text{population density} + w_4 \times \text{per capita GDP} + \\ & w_5 \times \text{temperature} + w_6 \times \text{humidity} + \epsilon \end{aligned} \tag{4}$$

Here, $w_0, w_1, w_2, \ldots, w_6$ are the slope coefficients (weights) and ϵ is an error term, which is a normally distributed random variable with a mean of zero and a constant variance across all observations. The slope coefficient w_j represents the change in the mean value of the dependent variable R_0 that corresponds to a one-unit increase in the independent variable, holding the values of all other independent variables in the model constant.

3.3.2 Support vector machine

SVM, initially proposed by [57], is a supervised machine learning techniques that can model non-linear relationship between the dependent and independent variables. The idea behind SVM is to develop a hyperplane to partition the data into classes. The SVM can be applied for different types of dependent variables. When the dependent variable is categorical the SVM is used for classification, whereas when the dependent variable is continuous (i.e., R_0 in our problem), the SVM is used for regression.

The objective function for our problem can be explained using Equation 5. In contrast to linear objective function where we minimize squared error, the objective function in SVM minimizes the L-2 norm of the coefficient vector ($||\mathbf{w}||^2$). The error term ($R_{0,i} - \mathbf{w}\mathbf{x_i}$) is managed in a constraint. $\mathbf{x_i}$ is a vector of the input variables listed in Table 2. ϵ is an acceptable threshold of error. ξ_i is a slack variable which accounts for an error that is larger than the margin ϵ. C is the inverse of the

regularization parameter.

$$\min \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}|\xi_i| \tag{5}$$

$$\text{subject to } |R_{0,i} - \mathbf{w}^T\mathbf{x_i}| \leq \epsilon + |\xi_i| \tag{6}$$

The model presented in Equation 5 and 6 can be solved after transforming it into its dual [58]. The solution involves using a kernel function, which is used for mapping the non-linear separable data into a higher dimensional space where we can find a hyperplane that can separate the samples. In our study, we employed three different kernels: linear, radial and polynomial. The results corresponding to each kernel is explained in Section 4.

3.3.3 Decision tree

Decision tree is one of the most popular data modeling techniques used for both regression and classification [59; 60; 61]. In the decision tree, the data is recursively partitioned into two subgroups by a variable. The selection of splitting variables is performed based on the variable that reduces the standard deviation most. The Standard Deviation Reduction (SDR) is defined as the difference in standard deviation before and after the split.

$$\text{SDR}_i = S_i(R_0) - S_i(R_0, X) \tag{7}$$

where $S_i(R_0)$ is the standard deviation of R_0 (response variable) at node i before the split, $S_i(R_0, X)$ is the standard deviation at node i after splitting the node by Variable X.

$$S_i(R_0, X) = \sum_{c \in X} p(c)S(c) \tag{8}$$

where $p(c)$ is the probability of class c and $S(c)$ is the standard deviation of R_0 within class c.

The decision tree algorithm can be summarized as:

Algorithm 1 DT algorithm

Require: Data of the Dependent and independent variables;
Ensure: A decision tree;
1: **procedure** DT
2: Start with the node including all the data points;
3: Select variable that provides the highest SDR;
4: Split data using the selected variable;
5: Repeat Step 2 and 3 until SDR is less than the pre-defined threshold;
6: **end procedure**

4 Results and discussions

In this section we present our results and discuss our findings. First, we estimate R_0 in different states of the US. Then we use different machine learning techniques to predict the spread. furthermore, we compare these models and evaluate their accuracy.

As discussed in Subsection 3.1, we use three methods including time-dependant, exponential growth, and maximum likelihood to estimate the reproductive number, R_0. This is an essential step, since all the other steps of the analyses depend on R_0. Figure 5 shows the estimated reproduction number, along with 95% confidence interval, for one of the states of the US, $R_{0,California}$. We observe that the estimates of $R_{0,California}$ from all three approaches are close to each other.

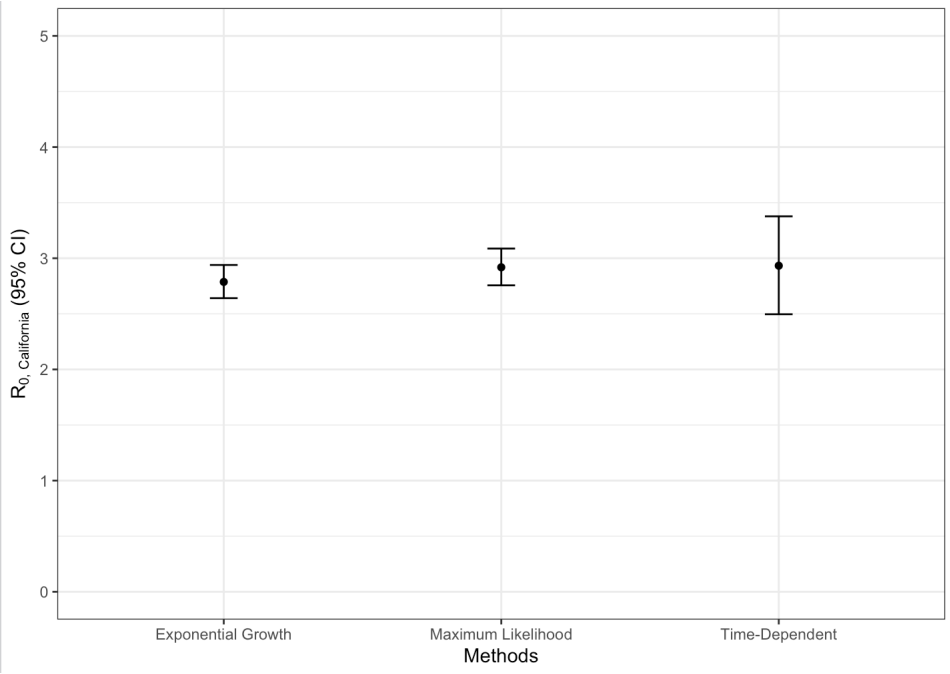

Figure 5: Reproduction number for the state of California using three different approaches

Furthermore, Figure 6 shows the estimation of epidemic curve for the State of California using the three approaches. The red, green and blue lines indicate the approximation of COVID-19 incidence using the exponential growth, maximum-likelihood and time-dependent methods, respectively. Similarly, we estimate the reproduction number for each state and present a statistical summary of R_0 in Table 3.

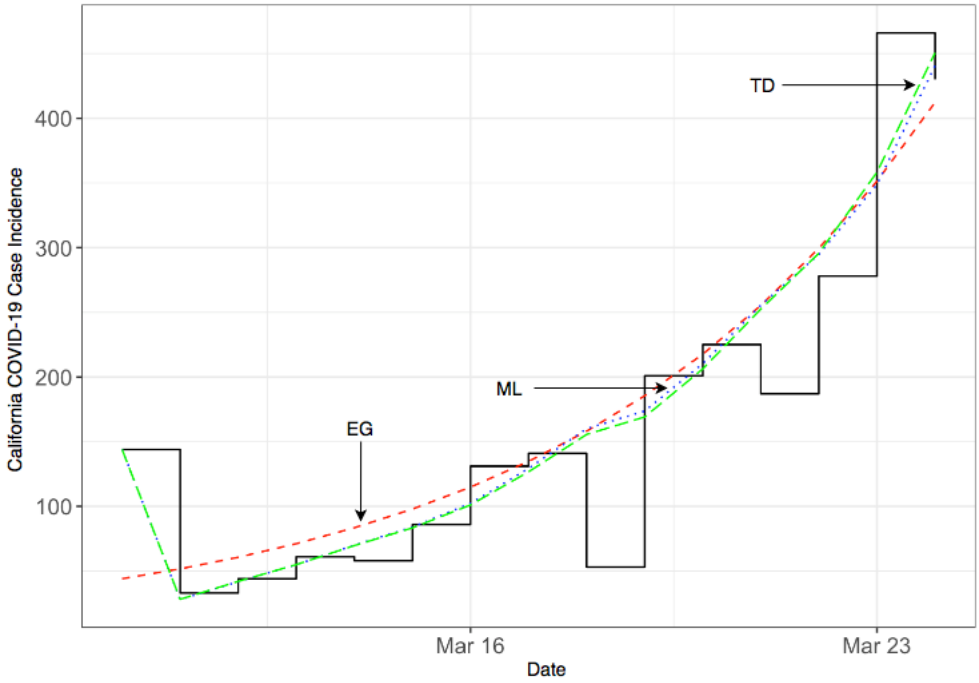

Figure 6: Approximation of the epidemic curve of California (EG: Exponential Growth, ML: Maximum Likelihood, TD: Time dependent)

Table 3: Statistical summary of R_0 estimated using three approaches (SD: Standard Deviation), Q1: First Quartile, Q3: Third Quartile

Methods	R_0						
	Minimum	Q1	Median	Mean	Q3	Maximum	SD
Time-dependent	1.30	2.93	3.69	3.81	4.71	6.97	1.16
Exponential growth	1.31	3.17	3.99	4.07	4.88	7.99	1.36
Maximum likelihood	1.89	3.24	3.87	3.90	4.50	6.63	0.98

Figure 7 shows the estimated R_0, using the time-dependent approach, on a heatamp across the US. The other methods result in similar heatmaps, so for the sake of brevity, we do not present them anymore. Michigan shows the highest value of R_0 (6.97), while South Dakota shows the minimum R_0 (1.30). The mean R_0 is 3.81. In other words, on average, each patient transmits the infection to

an additional 3.81 individuals. This is in line with the literature. For example, according to [62], the average R_0 for COVID-19 is 3.28. The study covers 12 studies from 1 January 2020 to 7 February 2020, which estimated the basic reproductive number for COVID-19 from China and overseas. The results show that the estimates range from 1.4 to 6.49, with a mean of 3.28, a median of 2.79 and interquartile range (IQR) of 1.16. It needs to be noted that R_0 can be significantly impacted by countermeasures. For example, on the Diamond Princess cruise ship, an initial estimated R_0 was 14.8 (approximately, 4 times higher than the average mentioned here). However, it reduced to 1.78 after implementing on-board isolation and quarantine measures [63].

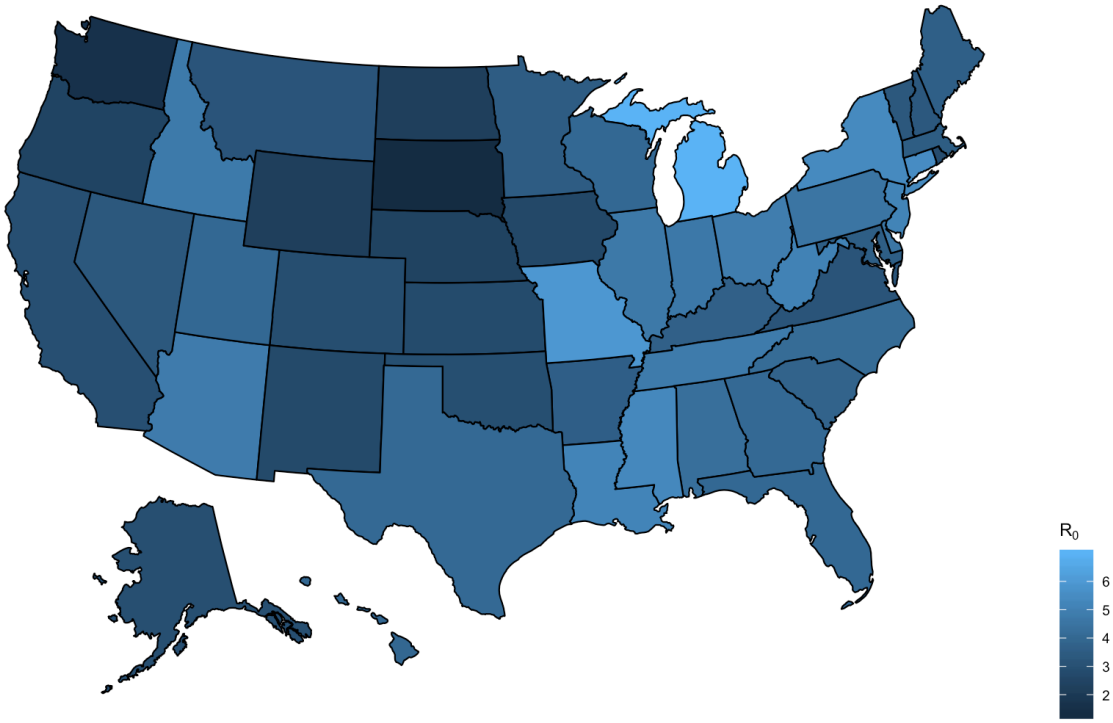

Figure 7: Heatmap illustrating estimated R_0 across states in the US

Table 4 shows modeling results. The performance of each model integrated with each R_0 estimation method is measured using three numeric measures including Root Mean Square Error (RMSE), R-Squared (coefficient of determination), and Mean Absolute Error (MAE). R-Squared measures the goodness of fit for regression models. The higher the R-Squared, the better the performance. The models are trained using 10-fold cross-validation to avoid overfitting. From the results presented in Table 4, we can draw following insights:

- **Insight 1:** The first insight is that all three approaches to approximate the epidemic curve

(time-dependent, exponential and maximum-likelihood) integrate well with radial kernel SVM (RMSE and MAE are the lowest and R-squared is the highest). (shown in **bold** fonts)

- **Insight 2:** The second insight is that the combination of time-dependent approach to estimate R_0 and radial kernel SVM modeling techniques perform the best. (shown in **red bold** font)

- **Insight 3:** Finally, the third insight is that the best modeling approach explains approximately 47.3% of the variation of R_0. our model outperforms the existing models in terms of explaining the variations in the data. R-squared of our model is 95% higher than the existing model proposed by [64].

Table 4: Modeling results (LK: Linear Kernel, RK: Radial Kernel and PK: Polynomial Kernel, RMSE: Root Mean Square Error, MAE: Mean Absolute Error)

R_0 Estimation method	Modeling technique	RMSE	R-Squared	MAE
Time-dependent	Linear Regression	1.009	0.405	0.857
	SVM (LK)	1.085	0.343	0.892
	SVM (RK)	**0.897**	**0.473**	**0.715**
	SVM (PK)	1.028	0.392	0.826
	Decision Tree	1.084	0.383	0.893
Exponential growth	Linear Regression	1.250	0.340	1.053
	SVM (LK)	1.446	0.273	1.229
	SVM (RK)	**1.063**	**0.465**	**0.859**
	SVM (PK)	1.286	0.335	1.052
	Decision Tree	1.406	0.237	1.156
Maximum Likelihood	Linear Regression	0.919	0.333	0.775
	SVM (LK)	1.050	0.260	0.885
	SVM (RK)	**0.795**	**0.436**	**0.642**
	SVM (PK)	0.900	0.273	0.747
	Decision Tree	1.012	0.252	0.841

As discussed above, the SVM with radial kernel indicated best performance. Hence, we used this model to understand the relative importance of predictors (Figure 8). Population density is the most important factor [11; 65]. Furthermore, air pollution and daily temperature follow population density. As discussed in the literature review section, this is in line with the previous studies on pandemics where these factors are introduced as predicting factors (see, for example, 41; 42; 43; 44; 45; 46). Furthermore, as suggested in the literature on pandemics [66; 67; 68], next to population density, we consider the power of another socio-economic factor, namely per capita

GDP, in predicting the spread. Our analysis shows that per capita GDP has a positive impact on the spread of virus.

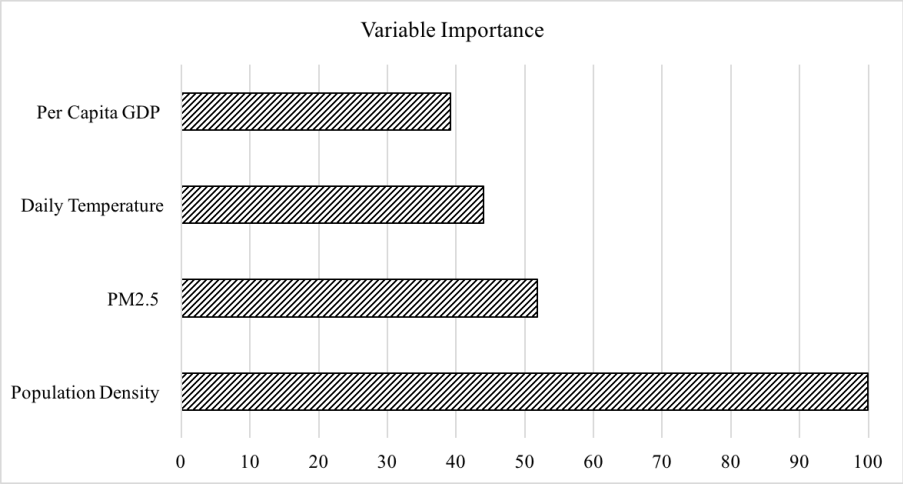

Figure 8: Variable importance (SVM with radial kernel)

In Figure 9, we investigate the linear relationship between the predictors considered in our paper and the spread of the disease measured as R_0. The dots, the blue line, and the shaded area show the data points the fitted line, and the 95% confidence interval, respectively. As expected, there is a positive linear relationship between population density and R_0. In other words, as population density increases, R_0 also increases. The same relationship can be seen between the air pollutant, PM2.5, and R_0 as well as temperature and R_0. In other words, a higher air pollution and temperature increase the spread. Some studies in the literature have shown that a higher temperature decreases the spread [45; 46], whereas in the others, the reverse has been shown [47; 48]. Our finding is in line with the latter. One explanation is that in the initial stages of the spread, temperature is not the only predicting factor. There are multiple factors that determine the spread. Obviously, in a state such as California, although the temperature is higher than the other states, due to its population, population density, population diversity, and more resources to detect the infected cases, the spread can be higher in the beginning. As the time passes, in the later stages of the disease, temperature can play a more important role. Last but not least, there is a negative linear relationship between per capita GDP and R_0. It is that when per capita GDP grows, the spread decreases.

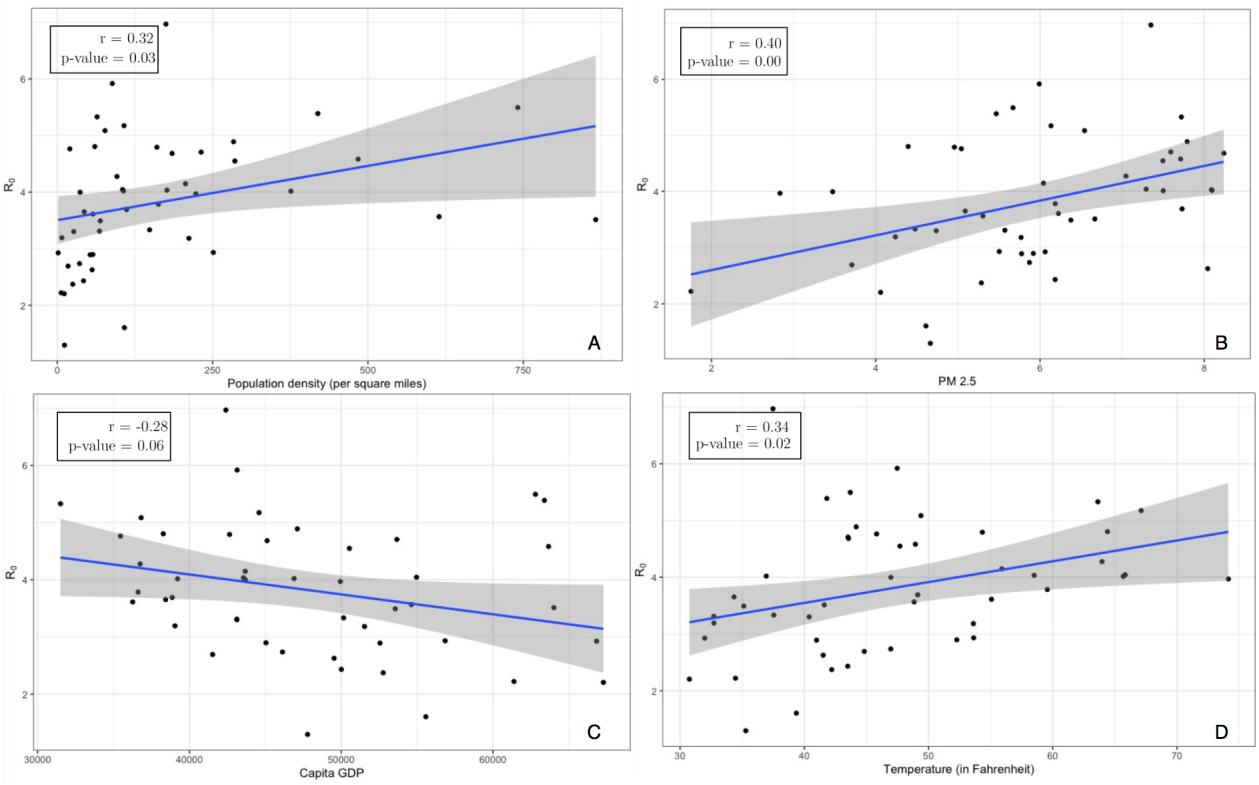

Figure 9: Linear trend between R_0 and A) population density B) PM2.5 C) per capita GDP and D) Temperature

5 Conclusions and future research

COVID-19 is one of the fastest spreading diseases in history. It has become pandemic with millions of confirmed cases and hundreds of thousands of deaths in about three months with a huge impact on public health, mortality and economy worldwide. In this paper, we adapted the theory from epidemiological literature and integrated it with machine learning approaches to predict the spread of COVID-19. Our model performs 95% better than the existing model in terms of explaining the variation in the reproduction number. The results show that a higher temperature, air pollution, and population density have a positive impact on the spread, whereas the per capita GDP has a negative effect. The proposed analytical method has the potential to be used as a tool in the future to determine the potential hot spots in the event of the second wave of virus.

This paper can be extended in several directions. First, one can consider data from several countries and compare the results or build a more accurate model. Furthermore, adding more variables to the models can add more depth to the analyses and can result in interesting findings.

References

[1] A. Gorbalenya, S. Baker, R. Baric, R. de Groot, C. Drosten, A. Gulyaeva, B. Haagmans, C. Lauber, A. Leontovich, B. Neuman, *et al.*, "The species severe acute respiratory syndrome-related coronavirus: classifying 2019-ncov and naming it sars-cov-2," *Nature Microbiology*, 2020.

[2] E. Callaway, D. Cyranoski, S. Mallapaty, E. Stoye, and J. Tollefson, "The coronavirus pandemic in five powerful charts.," *Nature*, 2020.

[3] Centers for Disease Control and Prevention (CDC), "Coronavirus disease 2019 (covid-19)." Retrieved March 30, 2020, from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html, 2020.

[4] F. Brauer, P. d. DRIESSCHE, and J. Wu, *Lecture notes in mathematical epidemiology*. Berlin, Germany: Springer, 2008.

[5] T. K. Dasaklis, C. P. Pappis, and N. P. Rachaniotis, "Epidemics control and logistics operations: A review," *International Journal of Production Economics*, vol. 139, no. 2, pp. 393 – 410, 2012.

[6] K. P. Scheibe and J. Blackhurst, "Supply chain disruption propagation: a systemic risk and normal accident theory perspective," *International Journal of Production Research*, vol. 56, no. 1-2, pp. 43–59, 2018.

[7] M. L. Brandeau, "Or forum—public health preparedness: Answering (largely unanswerable) questions with operations research—the 2016–2017 philip mccord morse lecture," *Operations Research*, vol. 67, no. 3, pp. 700–710, 2019.

[8] Z. Xu, W. Hu, G. Williams, A. C. Clements, H. Kan, and S. Tong, "Air pollution, temperature and pediatric influenza in brisbane, australia," *Environment International*, vol. 59, pp. 384 – 388, 2013.

[9] O. M. Araz, T. Lant, J. W. Fowler, and M. Jehn, "Simulation modeling for pandemic decision making: A case study with bi-criteria analysis on school closures," *Decision Support Systems*, vol. 55, no. 2, pp. 564 – 575, 2013. 1. Analytics and Modeling for Better HealthCare 2. Decision Making in Healthcare.

[10] G. Chowell, C. Ammon, N. Hengartner, and J. Hyman, "Transmission dynamics of the great influenza pandemic of 1918 in geneva, switzerland: assessing the effects of hypothetical interventions," *Journal of theoretical biology*, vol. 241, no. 2, pp. 193–204, 2006.

[11] N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke, "Strategies for mitigating an influenza pandemic," *Nature*, vol. 442, no. 7101, pp. 448–452, 2006.

[12] R. Grais, J. H. Ellis, A. Kress, and G. Glass, "Modeling the spread of annual influenza epidemics in the us: The potential role of air travel," *Health care management science*, vol. 7, no. 2, pp. 127–134, 2004.

[13] S. Enayati and O. Y. Özaltın, "Optimal influenza vaccine distribution with equity," *European Journal of Operational Research*, vol. 283, no. 2, pp. 714 – 725, 2020.

[14] L. E. Duijzer, W. [van Jaarsveld], and R. Dekker, "Literature review: The vaccine supply chain," *European Journal of Operational Research*, vol. 268, no. 1, pp. 174 – 192, 2018.

[15] L. E. Duijzer, W. [van Jaarsveld], and R. Dekker, "The benefits of combining early aspecific vaccination with later specific vaccination," *European Journal of Operational Research*, vol. 271, no. 2, pp. 606 – 619, 2018.

[16] N. M. Ferguson, D. A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke, "Strategies for containing an emerging influenza pandemic in southeast asia," *Nature*, vol. 437, no. 7056, pp. 209–214, 2005.

[17] M. Sharifyazdi, K. A. Navangul, A. Gharehgozli, and M. Jahre, "On- and offshore prepositioning and delivery mechanism for humanitarian relief operations," *International Journal of Production Research*, vol. 56, no. 18, pp. 6164–6182, 2018.

[18] H. Arora, T. Raghu, and A. Vinze, "Resource allocation for demand surge mitigation during disaster response," *Decision Support Systems*, vol. 50, no. 1, pp. 304 – 315, 2010.

[19] A. Janiak and M. Y. Kovalyov, "Scheduling in a contaminated area: a model and polynomial algorithms," *European Journal of Operational Research*, vol. 173, no. 1, pp. 125–132, 2006.

[20] A. Ekici, P. Keskinocak, and J. L. Swann, "Modeling influenza pandemic and planning food distribution," *Manufacturing & Service Operations Management*, vol. 16, no. 1, pp. 11–27, 2014.

[21] H. Mamani, S. E. Chick, and D. Simchi-Levi, "A game-theoretic model of international influenza vaccination coordination," *Management Science*, vol. 59, no. 7, pp. 1650–1670, 2013.

[22] D. Yamin and A. Gavious, "Incentives' effect in influenza vaccination policy," *Management Science*, vol. 59, no. 12, pp. 2667–2686, 2013.

[23] K. R. Nigmatulina and R. C. Larson, "Living with influenza: Impacts of government imposed and voluntarily selected interventions," *European Journal of Operational Research*, vol. 195, no. 2, pp. 613 – 627, 2009.

[24] I. C.-H. Fung, R. Antia, and A. Handel, "How to minimize the attack rate during multiple influenza

outbreaks in a heterogeneous population," *PloS one*, vol. 7, no. 6, 2012.

[25] E. Hansen and T. Day, "Optimal antiviral treatment strategies and the effects of resistance," *Proceedings of the Royal Society B: Biological Sciences*, vol. 278, no. 1708, pp. 1082–1089, 2011.

[26] N. B. Dimitrov, S. Goll, N. Hupert, B. Pourbohloul, and L. A. Meyers, "Optimizing tactics for use of the us antiviral strategic national stockpile for pandemic influenza," *PloS one*, vol. 6, no. 1, 2011.

[27] B. Y. Lee, S. T. Brown, G. W. Korch, P. C. Cooley, R. K. Zimmerman, W. D. Wheaton, S. M. Zimmer, J. J. Grefenstette, R. R. Bailey, T.-M. Assi, *et al.*, "A computer simulation of vaccine prioritization, allocation, and rationing during the 2009 h1n1 influenza pandemic," *Vaccine*, vol. 28, no. 31, pp. 4875–4879, 2010.

[28] J. T. Wu, S. Riley, C. Fraser, and G. M. Leung, "Reducing the impact of the next influenza pandemic using household-based public health interventions," *PLoS medicine*, vol. 3, no. 9, 2006.

[29] F. Carrat, C. Pelat, D. Levy-Bruhl, I. Bonmarin, and N. Lapidus, "Planning for the next influenza h1n1 season: a modelling study," *BMC infectious diseases*, vol. 10, no. 1, p. 301, 2010.

[30] R. C. Larson, "Simple models of influenza progression within a heterogeneous population," *Operations research*, vol. 55, no. 3, pp. 399–412, 2007.

[31] A. Teytelman and R. C. Larson, "Modeling influenza progression within a continuous-attribute heterogeneous population," *European Journal of Operational Research*, vol. 220, no. 1, pp. 238 – 250, 2012.

[32] O. Diekmann, J. A. P. Heesterbeek, and J. A. Metz, "On the definition and the computation of the basic reproduction ratio r 0 in models for infectious diseases in heterogeneous populations," *Journal of mathematical biology*, vol. 28, no. 4, pp. 365–382, 1990.

[33] H. W. Hethcote, "The mathematics of infectious diseases," *SIAM review*, vol. 42, no. 4, pp. 599–653, 2000.

[34] J. M. Hyman and J. Li, "An intuitive formulation for the reproductive number for the spread of diseases in heterogeneous populations," *Mathematical biosciences*, vol. 167, no. 1, pp. 65–86, 2000.

[35] N. G. Becker, *Analysis of infectious disease data*, vol. 33. CRC Press, 1989.

[36] K. Dietz and J. Heesterbeek, "Daniel bernoulli's epidemiological model revisited," *Mathematical biosciences*, vol. 180, no. 1-2, pp. 1–21, 2002.

[37] S. Eubank, H. Guclu, V. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, "Modelling disease outbreaks in realistic urban social networks," *Nature*, vol. 429, no. 6988, pp. 180–184, 2004.

[38] W. Hu, G. Williams, H. Phung, F. Birrell, S. Tong, K. Mengersen, X. Huang, and A. Clements, "Did socio-ecological factors drive the spatiotemporal patterns of pandemic influenza a (h1n1)?," *Environment International*, vol. 45, pp. 39 – 43, 2012.

[39] J. Peng, W. Kong, D. Guo, M. Liu, Y. Wang, H. Zhu, B. Pang, X. Miao, B. Yu, T. Luo, *et al.*, "The epidemiology and etiology of influenza-like illness in chinese children from 2008 to 2010," *Journal of medical virology*, vol. 84, no. 4, pp. 672–678, 2012.

[40] S. Thiberville, L. Ninove, V. V. Hai, E. Botelho-Nevers, C. Gazin, L. Thirion, N. Salez, X. de Lamballerie, R. Charrel, and P. Brouqui, "The viral etiology of an influenza-like illness during the 2009 pandemic," *Journal of medical virology*, vol. 84, no. 7, pp. 1071–1079, 2012.

[41] R. E. Davis, C. E. Rossier, and K. B. Enfield, "The impact of weather on influenza and pneumonia mortality in new york city, 1975–2002: a retrospective study," *PloS one*, vol. 7, no. 3, 2012.

[42] W. Yang and L. C. Marr, "Dynamics of airborne influenza a viruses indoors and dependence on humidity," *PloS one*, vol. 6, no. 6, 2011.

[43] I. Jaspers, J. M. Ciencewicki, W. Zhang, L. E. Brighton, J. L. Carson, M. A. Beck, and M. C. Madden, "Diesel exhaust enhances influenza virus infections in respiratory epithelial cells," *Toxicological Sciences*, vol. 85, no. 2, pp. 990–1002, 2005.

[44] M. J. Kesic, M. Meyer, R. Bauer, and I. Jaspers, "Exposure to ozone modulates human airway protease/antiprotease balance contributing to increased influenza a infection," *PloS one*, vol. 7, no. 4, 2012.

[45] A. C. Lowen, S. Mubareka, J. Steel, and P. Palese, "Influenza virus transmission is dependent on relative humidity and temperature," *PLoS Pathog*, vol. 3, no. 10, p. e151, 2007.

[46] J. Shaman and M. Kohn, "Absolute humidity modulates influenza survival, transmission, and seasonality," *Proceedings of the National Academy of Sciences*, vol. 106, no. 9, pp. 3243–3248, 2009.

[47] B. D. Elderd and J. R. Reilly, "Warmer temperatures increase disease transmission and outbreak intensity in a host–pathogen system," *Journal of Animal Ecology*, vol. 83, no. 4, pp. 838–849, 2014.

[48] A. C. Lowen, J. Steel, S. Mubareka, and P. Palese, "High temperature (30°c) blocks aerosol but not contact transmission of influenza virus," *Journal of Virology*, vol. 82, no. 11, pp. 5650–5652, 2008.

[49] D. M. Aleman, T. G. Wibisono, and B. Schwartz, "A nonhomogeneous agent-based simulation approach to modeling the spread of disease in a pandemic outbreak," *INFORMS Journal on Applied Analytics*,

vol. 41, no. 3, pp. 301–315, 2011.

[50] R. A. Vinarti and L. M. Hederman, "A personalized infectious disease risk prediction system," *Expert Systems with Applications*, vol. 131, pp. 266 – 274, 2019.

[51] J. H. CSSE, "https://github.com/cssegisanddata/covid-19," 2020.

[52] J. Wallinga and M. Lipsitch, "How generation intervals shape the relationship between growth rates and reproductive numbers," *Proceedings of the Royal Society B: Biological Sciences*, vol. 274, no. 1609, pp. 599–604, 2007.

[53] L. Forsberg White and M. Pagano, "A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic," *Statistics in medicine*, vol. 27, no. 16, pp. 2999–3016, 2008.

[54] J. Wallinga and P. Teunis, "Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures," *American Journal of epidemiology*, vol. 160, no. 6, pp. 509–516, 2004.

[55] J. Wang, K. Tang, K. Feng, and W. Lv, "High temperature and high humidity reduce the transmission of covid-19," *Available at SSRN 3551767*, 2020.

[56] J. D. Camm, J. J. Cochran, M. J. Fry, J. W. Ohlmann, and D. R. Anderson, *Essentials of Business Analytics (Book Only)*. Nelson Education, 2014.

[57] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[58] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[59] S. Lee, "Using data envelopment analysis and decision trees for efficiency analysis and recommendation of b2c controls," *Decision Support Systems*, vol. 49, no. 4, pp. 486 – 497, 2010.

[60] S. Piri, D. Delen, T. Liu, and H. M. Zolbanin, "A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a model ensemble," *Decision Support Systems*, vol. 101, pp. 12 – 27, 2017.

[61] A. Gupta, T. Liu, and S. Shepherd, "Clinical decision support system to assess the risk of sepsis using tree augmented bayesian networks and electronic medical record data," *Health informatics journal*, p. 1460458219852872, 2019.

[62] Y. Liu, A. A. Gayle, A. Wilder-Smith, and J. Rocklöv, "The reproductive number of COVID-19 is higher compared to SARS coronavirus," *Journal of Travel Medicine*, vol. 27, 02 2020.

[63] J. Rocklöv, H. Sjödin, and A. Wilder-Smith, "COVID-19 outbreak on the Diamond Princess cruise ship: estimating the epidemic potential and effectiveness of public health countermeasures," *Journal of Travel Medicine*, 2020.

[64] J. Wang, K. Tang, K. Feng, and W. Lv, "High temperature and high humidity reduce the transmission of covid-19," *Available at SSRN 3551767*, 2020.

[65] A. Holko, M. Mdrek, Z. Pastuszak, and K. Phusavat, "Epidemiological modeling with a population density map-based cellular automata simulation system," *Expert Systems with Applications*, vol. 48, pp. 1 – 8, 2016.

[66] P. Hosseini, S. H. Sokolow, K. J. Vandegrift, A. M. Kilpatrick, and P. Daszak, "Predictive power of air travel and socio-economic data for early pandemic spread," *PLoS One*, vol. 5, no. 9, 2010.

[67] N. Dattani and A. Jiang, "The diabetic pandemic: globalization, industrialization, and type 2 diabetes," *The Meducator*, vol. 1, no. 15, 2009.

[68] N. Rybnikova, A. Haim, and B. A. Portnov, "Does artificial light-at-night exposure contribute to the worldwide obesity pandemic?," *International Journal of Obesity*, vol. 40, no. 5, pp. 815–823, 2016.