Tackling COVID-19 in Peru with Econometrics, Statistics and Machine Learning

Renzo Guzmán* Miguel Paredes†

April 19, 2020

Abstract

In this paper, we conduct mathematical and statistical analyses for COVID-19. To predict the trend of COVID-19, we propose three methodological approaches: (i) an epidemiological approach that tracks the transmission rate and the recovering rate using SIR and SIER models, (ii) an econometric approach using ARIMA models and synthetic controls and (iii) a machine learning approach using SVM, Random Forest Regression, RNN and LSTM. We then discuss the assumptions and limitations of each methods and provide policy recommendations to allow for a better and more accurate measurement and understanding of the disease.

*Research Assistant and Teaching Assistant. Universidad del Pacífico.

†Vice President of Artificial Intelligence, Data Analytics and Behavioral Science. Rimac Seguros y Reaseguros.

1 Introduction

The 2019–20 Coronavirus Pandemic was reported to have spread to Peru on March 6th when a 25-year-old man who had traveled to Spain, France and the Czech Republic tested positive. On March 15th, President Martín Vizcarra announced a country-wide lockdown, closing borders, restricting domestic travel, and forbidding non-essential business operations, excluding health facilities, food vendors, pharmacies, and financial institutions.

A number of the statistical, dynamic and mathematical models of the COVID-19 outbreak including the SEIR model have been developed to analyze its transmission dynamics in various countries (Li et al., 2020; Wu et al., 2020; Zhao et al., 2020) as well as Peru (Bayes, Sal y Rosas and Valdivieso, 2020). Although epidemiological models are useful for estimating the dynamics of transmission, targeting resources and evaluating the impact of intervention strategies, the models require parameters and depend on many assumptions. Unlike system identification in engineering where the parameters in the models are estimated using real data, at the outbreak, estimated parameters using real time data are not readily available. Most analyses used hypothesized parameters and hence do not fit the data very well. The accuracy of forecasting the future cases of COVID-19 using these models may not be very high. Timely interventions are needed to control the serious impacts of COVID-19 on health.

Due to the recent development of the epidemic, we are interested in addressing the following important questions for COVID-19:

- Is it possible to contain COVID-19? Are the commonly used measures, such as city-wide lockdown, traffic halt and propaganda of health education knowledge, effective in containing COVID-19?
- If COVID-19 can be contained, when will be the peak of the epidemic, and when will it end?

2 Methods

Epidemiological models

The **SIR model** is a simple epidemiology compartmental model proposed by Kermack and McKendrick (1927), which assumes a fixed population with only three compartments or states:

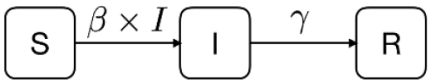
- S_t: number of susceptible, i.e. the number of individuals susceptible to the disease not yet infected at time t;
- I_t: number of infected, i.e. the number of individuals who have been infected at time t with the disease and are capable of spreading the disease to those in the susceptible category;
- R_t: number of recovered, i.e. those individuals who have been infected and then removed from the disease, either due to immunization or due to death. Members of this compartment are not able to be infected again or to transmit the infection to others.

Using a fixed population, i.e. with constant size $N = S_t + I_t + R_t$, Kermack and McKendrick (1927) derived the following system of quadratic ODEs:

$$\begin{aligned} \frac{\partial S_t}{\partial t} &= -\beta SI \\ \frac{\partial I_t}{\partial t} &= \beta SI - \gamma I \\ \frac{\partial R_t}{\partial t} &= \gamma I \end{aligned} \tag{1}$$

where $\beta > 0$ is the rate (constant for all individuals) at which an infected person infects a susceptible person, and $\gamma > 0$ is the rate at which infected people recover from the disease.

The flow of the SIR model can be represented in the following scheme:



where boxes represent the compartments and arrows indicate flows between compartments. Note that $\frac{\partial S_t}{\partial t} + \frac{\partial I_t}{\partial t} + \frac{\partial R_t}{\partial t} = 0$ then $S_t + I_t + R_t = N$, and the initial condition $S_0 > 0$, $I_0 > 0$ and $R_0 = 0$. Thus, the system can be reduced to a system of two ODEs.

We aim at estimating the values of β and γ based on the observed data by minimising the following loss function:

$$RSS(\beta, \gamma) = \sum_{t=1}^{T} \varepsilon_t = \sum_{t=1}^{T} \left(I_t - \hat{I}_t \right) \tag{2}$$

where I_t is the number of infected observed at time t, and $\hat{I}_t$ is the corresponding number of infected predicted by the model, which depends on the unknown parameters β and γ. Nonlinear least squares can be used to fit this model to data, but it strongly depends on the initial values as shown below. A more robust approach can be pursued by using Genetic Algorithms.

First of all, we define a function which computes the values of the derivatives in the ODE system at time t. This function is then used, together with the initial values of the system and the time sequence. The function RSS computes the predicted number of infected $\hat{I}_t$ from the solution of ODE system for the input parameters values, and returns the objective function in (2) to be minimized. Then, a GA function call can be used with local search to find the optimal values of parameters (β, γ) in SIR model. Finally, we use the following features for this optimizacion:

- Method: Limited-memory Broyden-Fletcher-Goldfarb-Shanno-Bird algorithm (L-BFGS-B) (Byrd et al., 1995; Zhu, et al., 1997)
- Initial values: $(0.5, 0.5)$
- Bounds on the variables: $(0, 1)$

- Population (N_t): 31,237,385

Furthermore, we employed an infectious disease dynamics model, **SEIR model**, for the purpose of modeling and predicting the number of COVID-19 cases in Peru. We assumed no new transmissions from animals, no differences in individual immunity, the time-scale of the epidemic is much faster than characteristic times for demographic processes (natural birth and death), and no differences in natural births and deaths. In this model, individuals are classified into four types: susceptible (S_t; at risk of contracting the disease), exposed (E; infected but not yet infectious), infectious (I_t; capable of transmitting the disease), and removed (R_t; those who recover or die from the disease). The total population size (N) is given by $N_t = S_t + E_t + I_t + R_t$. It is assumed that susceptible individuals who have been infected first enter a latent (exposed) stage, during which they may have a low level of infectivity. The differential equations of the SEIR model are given as:

$$\begin{aligned} \frac{\partial S_t}{\partial t} &= -\beta SI \\ \frac{\partial E_t}{\partial t} &= \beta SI - \sigma E \\ \frac{\partial I_t}{\partial t} &= \sigma E - \gamma I \\ \frac{\partial R_t}{\partial t} &= \gamma I \end{aligned} \tag{3}$$

where β is the transmission rate, σ is the infection rate calculated by the inverse of the mean latent period, and γ is the recovery rate calculated by the inverse of infectious period. For the calibration of the SEIR model we will use the following parameters:

- Initial number of cases caused by zoonotic exposure (I_0): 5 (Imai et al., 2020)
- Number potentially exposed by each initial case (E_0): 20 (Read et al., 2020)
- Infection rate (σ): 0.25 (Li et al., 2020)
- Recovery rate (γ): 0.3871 (Yang et al., 2020)

- Population (N_t): 31,237,385

Finally, we will perform a simulation under the assumption that the basic number of reproduction R_0 is variable in time. The variability of this parameter can be explained by the measures that the Peruvian government has taken in order to mitigate the spread of the disease. Analyzing the measures taken, five stages are identified which are described in Table 1. For each state, R_0 was computed using the SIR model.

Table 1. Stages identified to mitigate the expansion of the coronavirus in Peru

Stage	Government measures	R_0
Stage 1: March 6 - March 15	Initial stage where the government detected the first cases of imported COVID-19. The start of classes in schools was suspended and the population was recommended to stay at home, but it was not a mandatory measure.	2.83
Stage 2: March 16 - April 1	Government declared a 15-day quarantine effective from March 16th. What is more, two days later the government tightened the measures of quarantine, implementing a curfew from 8 PM-5 AM where citizens are not allowed to leave their homes.	1.77
Stage 3: April 2 - April 9	Government declared. that mobilization outside of the house will be limited by days. Only men will be able to leave the house to buy groceries, medicines, or go to the bank on Monday, Wednesday, and Friday. Only women are allowed outside on Tuesday, Thursday, and Saturday. No one is allowed on Sunday.	1.68
Stage 4: April 10 - April 26	Government once again extended the quarantine by 2 weeks, until the April 26th. Furthermore, government renounced the previously proposed gender rotation, and reinstated that only one member of a household can leave the home per week, from Monday through Saturday.	1.61
Stage 5: April 27 - June 13	In this last stage the population internalizes the imposed restrictions of social distancing.	0.90

Source: Information releases from Ministry of Health

Econometrics methods

Firstly, we use a **Synthetic Controls** approach. According to Abadie and Gardeazabal (2003) and Abadie et al. (2010), the synthetic control unit is created out of a "donor pool" of J control units. The comparability to the treated unit is determined by a set of predictors from T pre-intervention periods: M linear combinations of Y and r (other) covariates with explanatory power for Y. All k predictors (with $k = r + M$) are combined in a $(k \times 1)$ vector X_1 for the treated unit and in a $(k \times J)$ matrix X_0 for all control units. In one part of the optimization process (the inner optimization), one tries to find a linear combination of the columns of X_0 that represents X_1 best. The distance metric used to measure this difference is: $\|X1 - X_0W\|_v = \sqrt{(X1 - X_0W)'V\,(X1 - X_0W)}$, where the weights used to construct the synthetic control unit are denoted by the vector W, and the weights of the predictors are given by the non-negative diagonal matrix V. The inner optimization is then, for given predictor weights V, the task of finding non-negative control unit weights W, summing up to unity, such that:

$$\sqrt{(X1 - X_0W)'V\,(X1 - X_0W)} \xrightarrow{W} \min \tag{4}$$

The solution to this problem is denoted by $W^*(V)$, which typically contains many vanishing components as these cannot become negative. The second part of the optimization (the outer optimization) deals with finding optimal predictor weights. It usually follows a data-driven approach, where V is chosen among all positive definite and diagonal matrices such that the mean squared prediction error (MSPE) of the outcome variable Y is minimized over the pre-intervention periods. To this end, $Y_t^{(j)}$ denotes the value of Y for unit j at time $t = 1, ..., T$, where $j = 1$ denotes the treated unit, and $j = 2, ..., J + 1$ denote the control units. Using a discount factor $\beta \leq 1$ to put more weight on more recent observations, the outer optimization problem looks as follows:

$$\sum_{t=1}^{T} \beta^{T-t} \left(Y_t^{(1)} - \sum_{j=2}^{J+1} W^*(V)_j Y_t^{(j)} \right)^2 \xrightarrow{V} \min \tag{5}$$

As explained above, in traditional applications of SCM, the treated unit is synthesized by units from a so-called donor pool. For instance, when considering cases of COVID-19, the variable of interest is 'Number of cases of COVID-19' (Y_t), and for synthesizing Peru, one would look for COVID-19 data from countries where COVID-19 cases were presented before Peru (March 6). With regard to predictors, we consider linear combinations of lagged values of Y_t, as well as the number of recovered and deaths for COVID-19 and the number of molecular tests

On the other hand, we use an **Autoregressive Integrated Moving Average (ARIMA)** as proposed by Box and Jenkins (1976). In practice, most time series do not meet the requirements of stationarity. Some nonstationary time series have a particular trend; therefore, a differencing operator might transform them into a stationary series. For example, first-order differencing can transform a time series with a constant slope into a new series with constant mean. The ARIMA model expresses a time series using appropriate differencing and an ARMA model, which includes moving an average (MA) process and autoregressive (AR) model. This model is denoted by ARIMA (p, d, q) and is expressed as:

$$\Phi(L) \bigtriangledown^d y_t = \Theta(L)\epsilon_t \tag{6}$$

where L is the backward shift operator, $\bigtriangledown^d = (1-L)^d$, $\Theta(L) = 1 - \theta_1 L - ... - \theta_q L^q$ y $\Phi(L) = 1 - \phi_1 L - ... - \phi_p L^p$.

Based on the ARIMA model, **ARIMAX** model can take the impact of covariates into account by adding the covariate to the right hand of the ARIMA model equation. The equation of ARIMAX model is presented as follows:

$$\Phi(L) \bigtriangledown^d y_t = \mu + \Theta(L)x_t + \Theta(L)\epsilon_t \tag{7}$$

We use as covariates the number of recoveries and deaths from COVID-19 in Peru as well as the number of infected, recoveries and deaths from other countries.

Machine Learning methods

First of all, we use a **Support Vector Machine (SVM)** which is a supervised machine learning algorithm which can be used for both classification or regression challenges and gives high classification accuracy in many applications. An SVM is based on two ideas. The first idea is to map feature vectors to a high dimensional space with a nonlinear method and to use linear classifiers in this new space. The second idea is to separate the data with a high margin hyperplane. This plane is the best plane, which can separate the data as well as possible (Kulkarni and Harman, 2011). We use hyperparameter tuning using grid search, in which all possible combinations of given discrete parameter spaces are evaluated and the following features for this method:

- Train sample size: 90% of total sample
- Test sample size: 10% of total sample
- Cost parameter (C): (0.01, 0.1, 1, 10)
- Kernel coefficient (γ): (0.01, 0.1, 1)
- Distance epsilon (ϵ): (0.01, 0.1, 1)
- Number of repetitions of cross-validation: 5
- Number of iterations: 1000
- Number of jobs to run in parallel: Using all processors
- Scoring strategy: Negative mean squared error

Then, we apply a **Random Forest Regression model (RF)**. This is a kind of ensemble learning method, where training samples for building each decision tree are randomly selected (bootstrap samples) and the final result is obtained by majority voting or averaging the prediction results from multiple decision trees (Breiman, 2001; Podgorelee et al., 2002; Lindner et al,, 2013). Since RF is robust to noise and relatively insensitive to small amount

of training samples it has been widely used for classification and regression tasks in medical field. Similary to SVM, we use hyperparameter tuning an the following features for RF:

- Number of trees in the foreset: (10, 100, 10)
- Maximum number of levels in each decision tree: (3, 7, 10)
- Maximum number of features considered for splitting a node: (2, 20, 2)
- Minimum number of data points placed in a node before the node is split: (1, 20, 2)
- Minimum number of data points allowed in a leaf node: (0.5, 1)

Furthermore, we use a **Recurrent Neural Networks (RNN)** approach. RNN are the most commonly used Neural Networks architecture for sequence prediction problems. They have particularly gained popularity in the domain of natural language processing. Similar to Artificial Neural Networks (ANN), RNN are universal approximators as well. However, unlike ANN, the feedback loops of the recurrent cells inherently address the temporal order as well as the temporal dependencies of the sequences (Schafer and Zimmermann, 2006).

$$\begin{aligned} h_t &= \sigma(W_i \cdot h_{t-1} + V_i \cdot x_t + b_i) \\ \tilde{C}_t &= \tanh(W_o \cdot h_t + b_o) \end{aligned} \tag{8}$$

Our work implements a number of RNN architectures along with different RNN units. Firstly, we use an Elman Recurrent Unit (Elman, 1990). The structure of the basic Elman RNN cell is as shown in Figure 1. In equations (8), $h_t \in \mathbb{R}^d$ denotes the hidden state of the RNN cell (d being the cell dimension). This is the only form of memory in the Elman RNN cell. $x_t \in \mathbb{R}^m$ (m being the size of the input) and $z_t \in \mathbb{R}^d$ denote the input and output of the cell at time step t. $W_i \in \mathbb{R}^{d \times d}$ and $V_i \in \mathbb{R}^{d \times d}$ denote the weight matrices whereas $b_i \in \mathbb{R}^d$ denotes the bias vector for the hidden state. Likewise, $W_o \in \mathbb{R}^{d \times d}$ and $b_o \in \mathbb{R}^d$ signify the weight matrix and the bias vector of the cell output. The current hidden state depends on the hidden state of the previous time step as well as the current input. This is

supported with the feedback loops in the RNN cell connecting its current state to the next state. These connections are of extreme importance to consider past information in updating the current cell state. In the experiments, we use the sigmoid function (indicated by) as the activation of the hidden state and the hyperbolic tangent function (indicated by tanh) as the activation of the output.

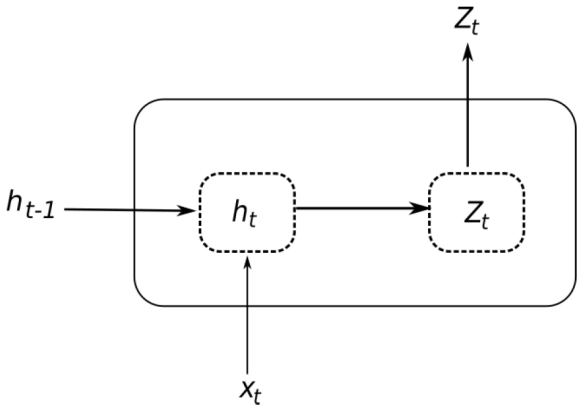


Figure 1: Elman Recurrent Unit

On the other hand, the **Long Short-Term Memory model (LSTM)** is an improvement introduced to RNN (Hochreiter and Schmidhuber, 1997). The difference between the RNN and LSTM is that for LSTM, a cell state is added to store long-term states. In the neural unit model structure in Figure 1, the internal structure of LSTM can be divided into the input gate, forget gate, and output gate. The principle of the LSTM input gate is expressed in the following formula:

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\ C_t &= f_t C_{t-1} + i_t \tilde{C}_t \end{aligned} \tag{9}$$

The first equation is used to decide which piece of information is to be added by passing h_{t-1} and x_t through the sigmoid layer. Subsequently, the second equation is used to pass h_{t-1} and x_t through the tanh layer to obtain new information $\tilde{C}_t$. The last equation is used to

combine the information of the current moment $\tilde{C}_t$ and long-term memory C_{t-1} into a new memory state C_t.

The forget gate of the LSTM uses a sigmoid layer and a dot product to allow information to pass through selectively. Equation (10) allows the LSTM to decide whether to forget the related information of the previous cell, at a certain probability, in which W_f is the weight matrix, and b_f is the offset term.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{10}$$

The output gate of LSTM decides which states are required to be maintained by the input h_{t-1} and x_t according to Equations (11). The final output results are obtained by passing the new information C_t through the tanh layer to multiply with state judgement vectors.

$$\begin{aligned} O_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= O_t \tanh(C_t) \end{aligned} \tag{11}$$

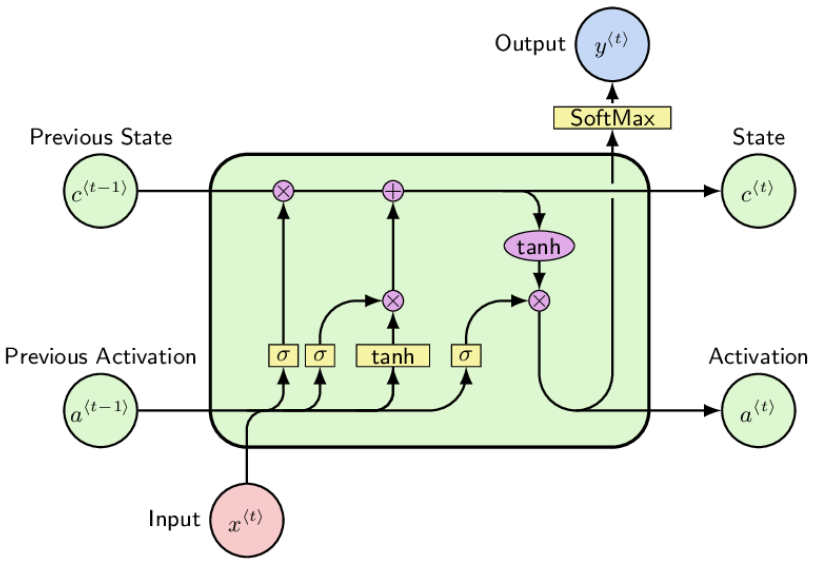


Figure 2: Structure of the deep neural network LSTM

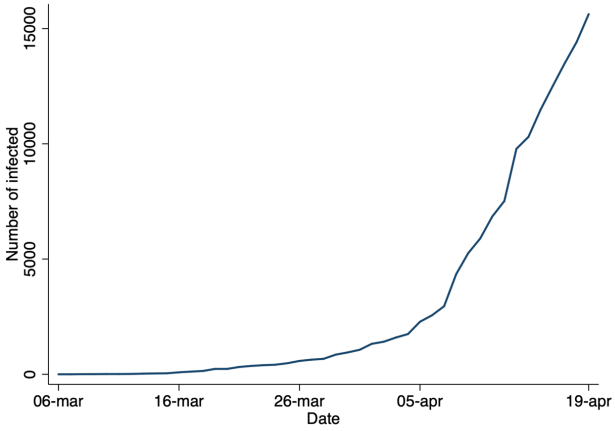
3 Data

Data is extracted from verified sources such as John Hopkins University, WHO and Ministry of Health of Peru. The sites reported confirmed COVID-19 cases, as well as recovered and deaths for affected countries and regions.

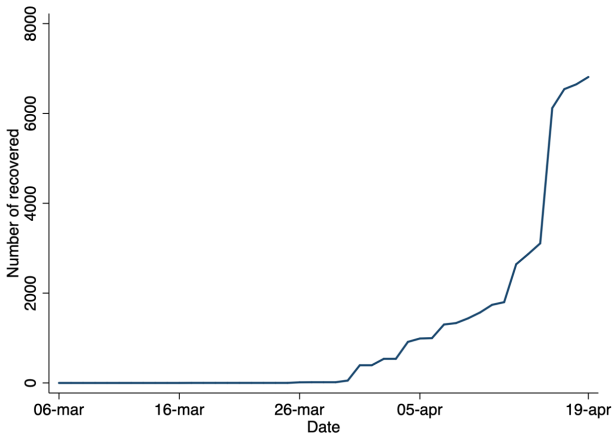
Ministry of Health reported to April 18th that samples for 135,895 people have been processed by COVID-19 (38,462 with molecular tests and 97,433 with serological tests), obtaining 15,628 positive and 121,475 negative results. Additionally, there are 1,268 patients hospitalized with COVID-19, of which 117 are in the ICU with mechanical ventilation. On the other hand, of the total of positive cases that completed their period of home isolation, 6,811 are already discharged. There have been 400 deaths nationwide from COVID-19 so this disease has a death rate equal to 2.55%.

At the regional level, Lima is the region with the highest number of infected by COVID-19 to date with 10,234. The following regions also present patients with COVID-19: Callao (1180), Lambayeque (642), Loreto (485), Piura (344), Ancash (217), La Libertad (211), Arequipa (146), Junín (130), Ica (124), Cusco (123), Tumbes (86), San Martín (81), Ucayali (73), Huánuco (64), Cajamarca (52), Amazonas (51), Apurímac (35), Ayacucho (29), Tacna (28), Moquegua (28), Madre de Dios (21), Huancavelica (18), Pasco (13) and Puno (05).

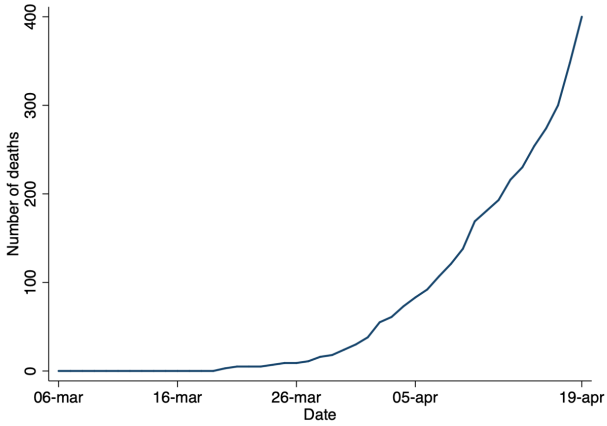
(i) Infected



(ii) Recovered



(iii) Deaths



(iv) Death rate

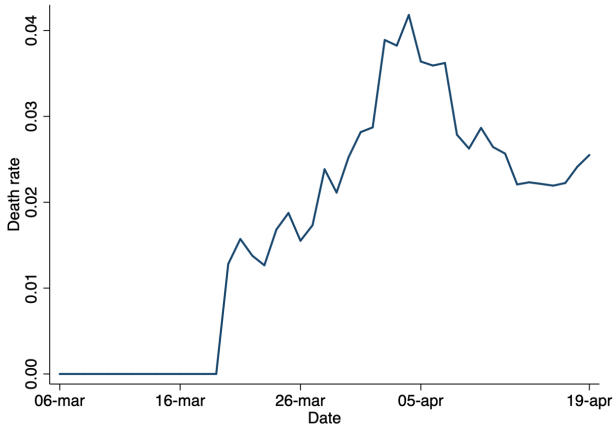


Figure 3: Infected, recovered, deaths and death rate due to COVID-19 Peru

4 Results

The projection was made for a 100-day horizon, counting from the first day an infected person presented (March 6th). Firstly, the SIR model suggest that the maximum number of infected people will be equal to 129,198 and will be reached on May 18 (Figure 4).

Based on the estimated parameters, average number of contacts per person per time ($\beta = 0.6128$) and rate of recovery ($\gamma = 0.3871$), other quantities of interest can be computed. For instance, $1/\gamma = 1/0.3871 \approx 2.58$ is the average recovery time which expresses the duration of infection (in days).

On the other hand, the basic reproductive rate (R_0) (i.e. the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection) is $R_0 = \beta/\gamma = 0.6128/0.3871 = 1.58$ is the infection's contact rate. Figure 5 shows the evolution of parameters R_0, β and γ, showing that the parameter stabilizes to the extent that there is a greater number of observations of infected.

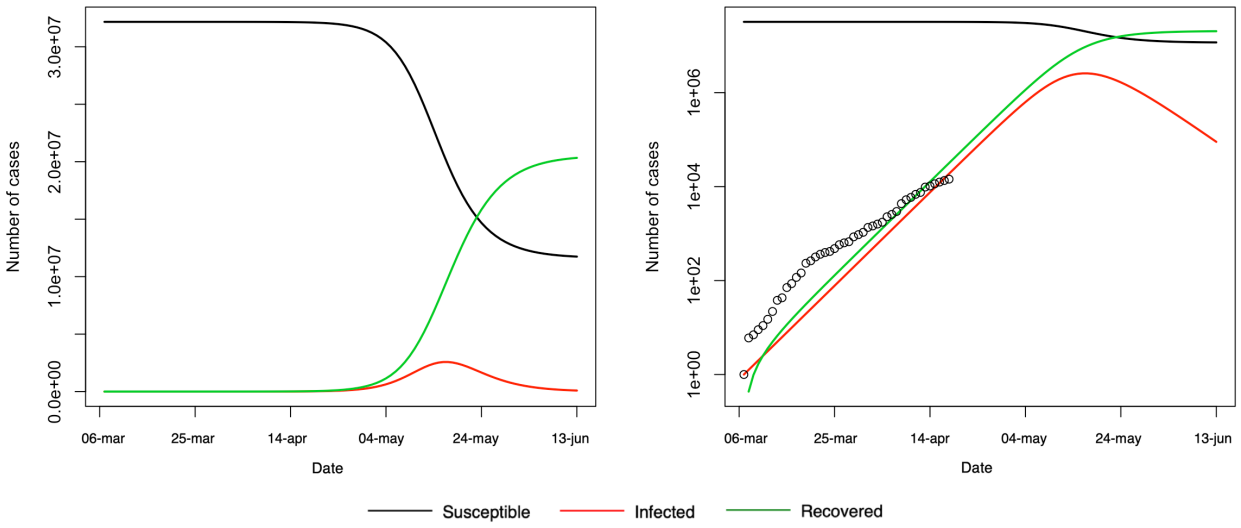
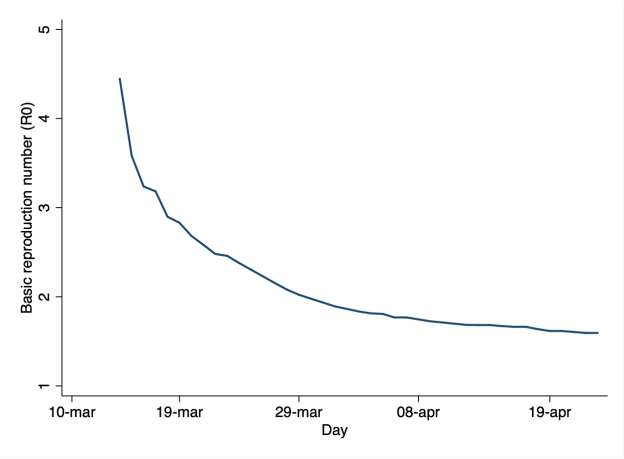
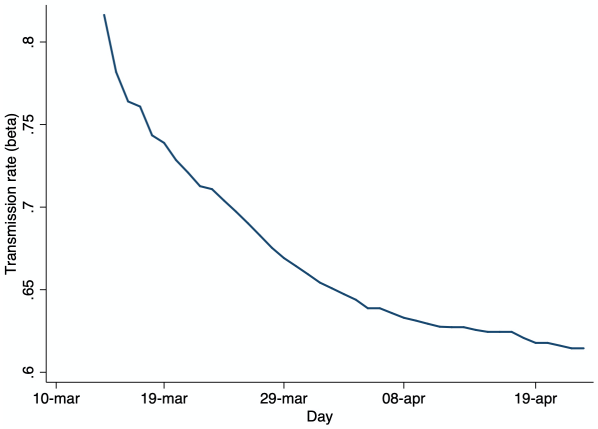


Figure 4: SIR model for COVID-19 cases in Peru (linear and logarithm scale)

(i) Basic reprodution number (R_0)



(ii) Transmission rate (β)



(iii) Recovery rate (γ)

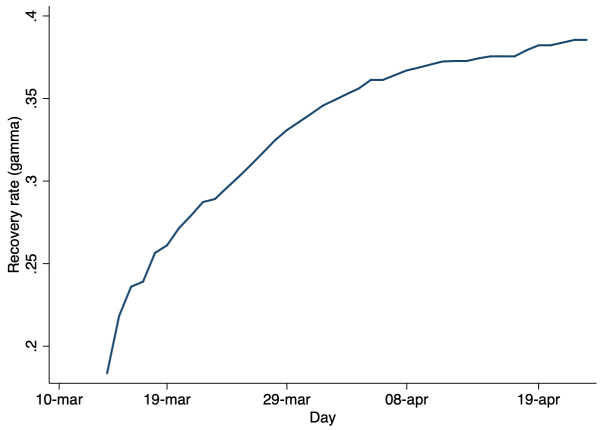


Figure 5: Evolution of parameters of the SIR model

Figure 6 shows the results of the SEIR model. It is important to mention that the parameters estimated in the SIR model and the available information on COVID-19 in other countries were used for the calibration of this model. The results show that a model with $R_0 = 1.6$ fits better with the series of infected cases in Peru. Probably the difference between the values predicted by the model and the real ones is probably due to a potential under-registration of cases in the first weeks of the pandemic, since the government did not have enough molecular tests to mislead of suspected cases.

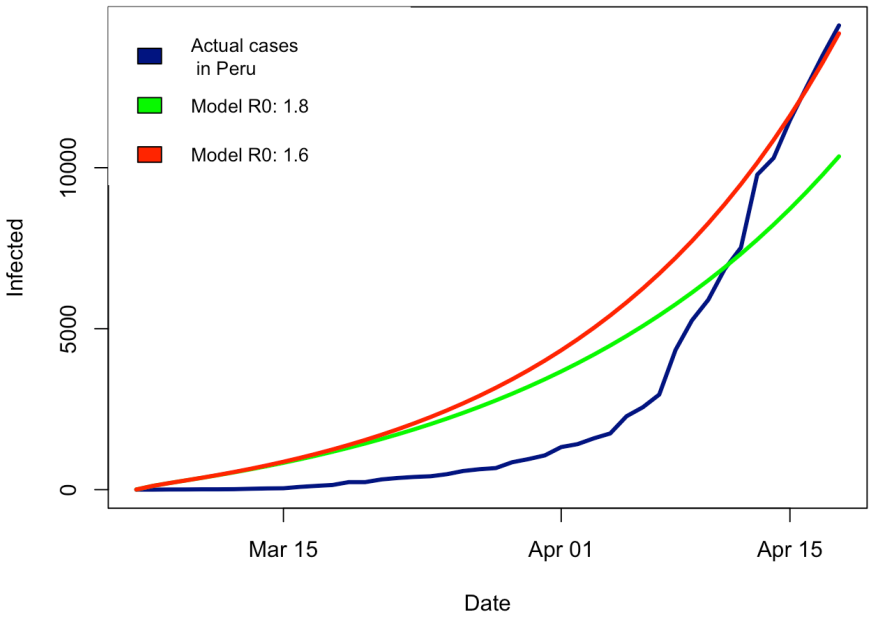


Figure 6: SEIR model for COVID-19 cases in Peru

Subsequently, we consider in the SIER model the interventions being carried out by the Peruvian government to reduce the propagation speed of COVID-19. For this, we identified 5 stages and in each of them the R_0 calculated with the SIR model was used (Figure 7). The results show that, considering the measures adopted by the government, during the 100-day horizon the number of infected with COVID-19 in Peru will be 35,261 and the highest point of detected cases will occur on April 30 (Figure 8).

It is important to mention that the scenario where the contagion dynamics would have occurred without any government intervention was also simulated, the results show that in that scenario the number of infected would have been 418,155 and the maximum point of contagion would have occurred on May 25.

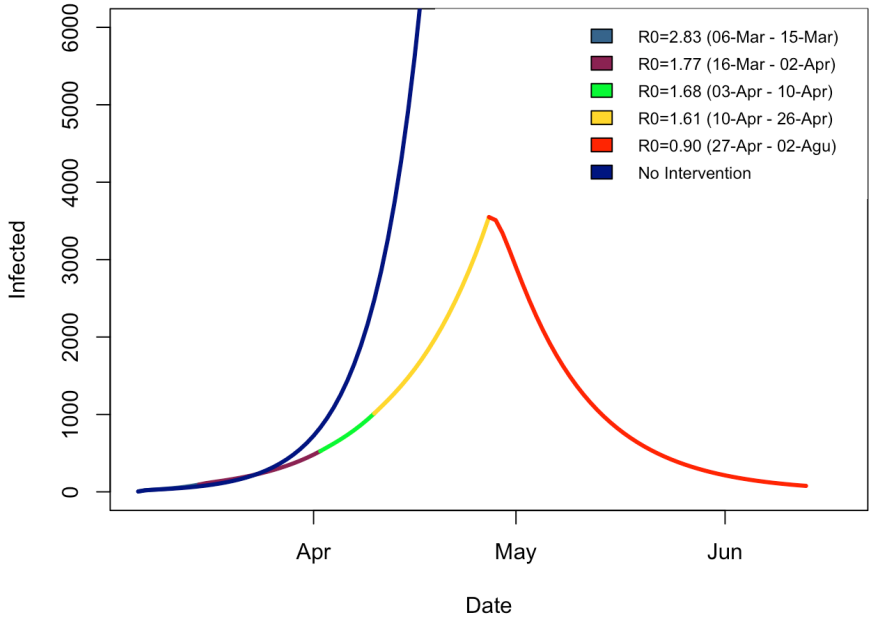


Figure 7: Number of people infected with interventions

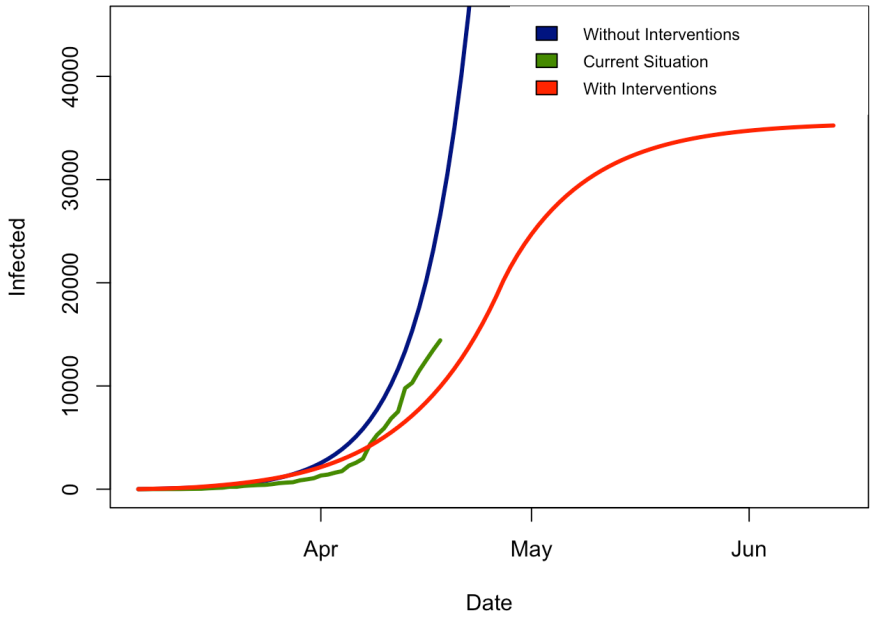


Figure 8: SEIR model with time-variant parameters for COVID-19 cases in Peru

Regarding the econometrics approach, Figure 9 shows the adjustment of the ARIMA models to the time series of detected cases of COVID-19 in Peru (Table 1 of Appendix shows the estimated parameters and their individual statistical significance). In general, the models fit well with the reported cases of COVID-19 to date.

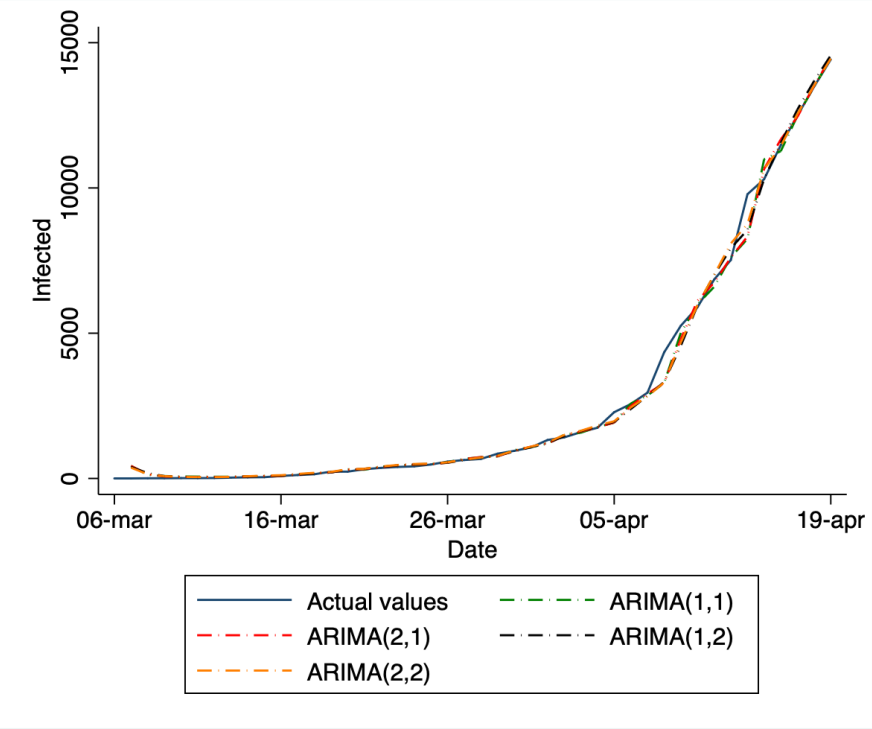


Figure 9: ARIMA models for COVID-19 cases in Peru

The results of synthetic control approach show that the maximum number of infected people during the analyzed period is equal to 31,906. Likewise, the curvature of the series constructed for synthetic Peru allows us to infer that the growth rate of the number of cases detected will begin to decrease as of May 5. The main limitation of this method is that it requires that all reference units for the construction of the synthetic control have the same number of observations in the projection horizon. Therefore, to make a projection on the 85-day horizon, the data from the following countries was used: United States, Spain, Italy, Germany, France, China, United Kingdom, Turkey, Belgium, Canada, Korea South, Russia, Sweden, Australia, India, Japan, Malaysia, Philippines, Finland, Thailand, Singapore, Vietnam and Cambodia (see Table 2 of Appendix).

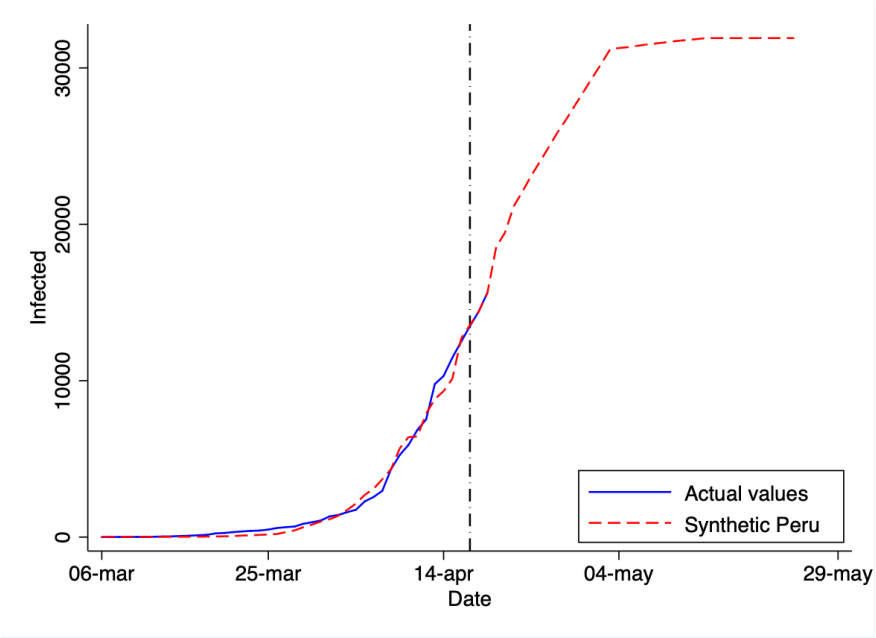


Figure 10: Synthetic controls approach for forecasting COVID-19 cases in Peru

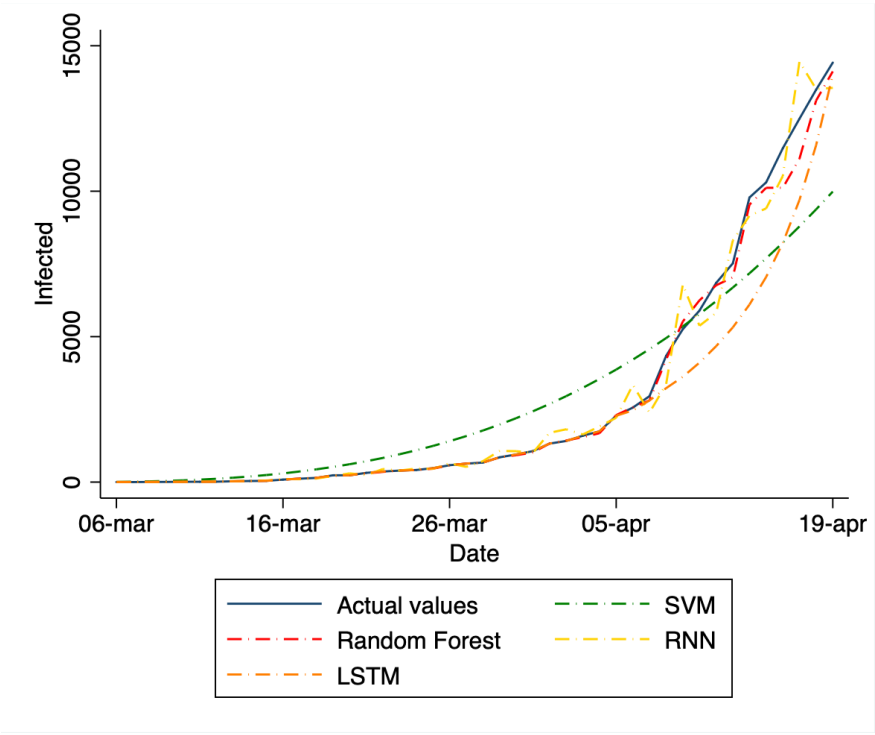


Figure 11: Machine Learning models for COVID-19 cases in Peru

On the other hand, Figure 11 presents the results of the Machine Learning models. In general, it can be seen that although all models accurately predict the trend of the contagion curve, the results of the Random Forest regression have a better fit with the current values, while the SVM estimates tend to underestimate the number of cases. detected from COVID-19 in Peru.

To evaluate the performance of the in-sample projection of the models used, Table 2 shows MAE, MSE and RMSE for each of the models. Overall, the Random Forest regression performs best for predicting COVID-19 cases, followed by the RNN, ARIMA, and SEIR with time variant parameters.

Table 2. Comparision between models performance in-sample

Model	MAE	MSE	RMSE
SIR	1,047	2,599,115	1,612
SIER with no intervention	967	2,407,962	1,552
SIER with time variant parameters	303	308,241	555
ARIMA(1,1)	150	105,856	325
ARIMA(2,1)	143	94,675	308
ARIMA(1,2)	148	86,191	294
ARIMA(2,2)	145	76,174	276
Synthetic controls	254	148,289	385
SVM	1,043	2,357,142	1,535
Random Forest regression	128	103,176	321
RNN	309	303,001	550
LSTM	556	1,432,817	1,197

Finally, Table 3 presents the results of the predictions in the number of detected cases of COVID-19 in Peru for a horizon of 100 days (until June 13). Considering the models that best fit the data available to date, we can affirm that the number of infections will range from 30,000 to 50,000 and the peak contagion day will occur between April 30 and May 10.

Table 3. Forecast of number of cases and peak contagion day of COVID-19 in Peru

Model	Infected	Peak contagion day
SIR	129,198	May 18
SIER with no intervention	418,155	May 25
SIER with time variant parameters	35,261	April 30
ARIMA(1,1)	80,392	May 19
ARIMA(2,1)	82,192	May 20
ARIMA(1,2)	79,975	May 15
ARIMA(2,2)	81,674	May 17
Synthetic controls	31,906	May 5
SVM	20,785	May 2
Random Forest regression	35,813	May 4
RNN	45,204	May 10
LSTM	42,816	May 5

5 Conclusions

The methods addressed in this paper show precision in predicting the evolution of the behavior of detected cases of COVID-19 in Peru. Although there is still limited information on this pandemic, so as more information becomes available, both nationally and internationally, more accurate predictions can be obtained on the extent of the pandemic.

The evidence suggests that the measures adopted by the government have had the expected impact to contain the flow of contagion in the population. However, predictions suggest that the number of people infected with COVID-19 in Peru in the first 100 days will be between 30,000 to 50,000 and the maximum point of infection will occur between April 30 and May 10 . Therefore, the next two weeks are crucial to control the spread of the pandemic and take targeted measures so that economic activities are gradually resumed.

References

[1] Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. American Economic Review, 93(1), pp. 113-132.

[2] Abadie, A.; Diamond, A.; Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association, 105(490), pp. 493-505.

[3] Bayes, C.; Sal y Rosas, V.; Valdivieso, L. (2020). Modelling death rates due to COVID-19: A Bayesian approach

[4] Box, G. and Jenkins, M. (1976). Time Series Analysis: Forecasting and Control.

[5] Breiman, L. (2001). Random forests. Machine learning, 45, pp 5-32.

[6] Byrd, R. H.; Lu, P.; Nocedal, J.; Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. Journal of Scientific Computing, 16 (5), pp. 1190-1208.

[7] Elman, J. (1990). Finding Structure in Time. CognitiveScience, 14(1), pp. 179-211.

[8] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), pp. 1735-1780.

[9] Imai, N.; Cori, A.; Dorigatti, C.; Baguelin, M.; Donnelly, C. (2020). Report 3: Transmissibility of 2019-nCoV.

[10] Kermack, W. and McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. Proceedings of the royal society of London. Series A, Containing papers of a mathematical and physical character, 115(772), pp. 700-721.

[11] Kulkarni, S. and Harman, G. (2011). Statistical learning theory: a tutorial. Wiley Interdisciplinary Reviews: Computational Statistics, 3(6), pp. 543-56.

[12] Li, Q.; Guan, X.; Wu, P.; Wang, M.; ... ; Xiang, P. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. The New England Journal of Medicine, 382 (1), pp. 1199-1207.

[13] Podgorelec, V.; Kokol, P.; Stiglic, B.; Rozman, I. (2002). Decision trees: an overview and their use in medicine. Journal of Medical Systems, 26 (1), pp. 445-463.

[14] Read, J.; Bridgen, J.; Cummings, D.; Ho, A.; Jewell, C. (2020). Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. medrxiv.

[15] Sammut, C. and Webb, G. (2017). Encyclopedia of Machine Learning and Data Mining: Springer US.

[16] Yang, Y.; Lu, Q.; Liu, M.; Wang, Y.; ... ; Fang, A. (2020). Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China.

[17] Zhu, C.; Byrd, R.; Lu, P.; Nocedal, J. (1997). L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. ACM Transactions on Mathematical Software, 23 (4), pp. 550-560.

Appendix

Table 1. ARIMA models

	(1)	(2)	(3)	(4)
	ARIMA(1,1)	ARIMA(2,1)	ARIMA(1,2)	ARIMA(2,2)
Infected_t				
Constant	430.9	428.5	396.2	368.9
	(0.70)	(0.70)	(0.63)	(0.67)
ARMA				
Infected_{t-1}	0.969***	0.562	0.961***	1.398***
	(10.54)	(1.67)	(10.43)	(4.43)
Infected_{t-2}		0.394		-0.445
		(1.42)		(-1.01)
ϵ_{t-1}	-0.652***	-0.472	-1.164***	-1.498***
	(-4.12)	(-1.42)	(-16.97)	(-5.18)
ϵ_{t-2}			0.593***	0.864**
			(3.43)	(2.67)
σ_t				
Constant	321.1***	302.8***	288.6***	271.4***
	(18.95)	(15.73)	(12.87)	(12.88)
N	44	44	44	44

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2. List of countries used for the synthetic control

Country	Unit weight (w_i)
United States	0.034
Spain	0.052
Italy	0.033
Germany	0.037
France	0.057
China	0.046
United Kingdom	0.031
Turkey	0.032
Belgium	0.044
Canada	0.039
Korea South	0.043
Russia	0.031
Sweden	0.049
Australia	0.05
India	0.059
Japan	0.051
Malaysia	0.049
Philippines	0.036
Finland	0.051
Thailand	0.04
Singapore	0.048
Vietnam	0.054
Cambodia	0.034