SPECIAL ISSUE
Interdisciplinary Studies on Health

Forecasting COVID-19 Cases in the Philippines Using Various Mathematical Models

Edd Francis O. Felix[1,2] • Monica C. Torres[1,2]✉ • Christian Alvin H. Buhat[3] • Ben Paul B. Dela Cruz[1,2] • Eleanor B. Gemida[1,2] • Jonathan B. Mamplata[1,2]

[1] Institute of Mathematical Sciences and Physics, University of the Philippines Los Baños, PHILIPPINES
[2] University of the Philippines Resilience Institute, University of the Philippines System, PHILIPPINES
[3] Department of Mathematics, University of Houston, Texas, UNITED STATES OF AMERICA

Abstract

Due to the rapid increase of COVID-19 infection cases in many countries such as the Philippines, efforts in forecasting daily infections have been made to better manage the pandemic and respond effectively. In this study, we considered the cumulative COVID-19 infection cases in the Philippines from 6 March 2020 to 31 July 2020, and forecasted the cases from 1–15 August 2020 using various mathematical models—weighted moving average, exponential smoothing, Susceptible-Exposed-Infected-Recovered (SEIR) model, Ornstein-Uhlenbeck process, Autoregressive Integrated Moving Average (ARIMA) model, and random forest. We compared the results to the actual data using traditional error metrics. Our results showed that the ARIMA (1,2,1) model had the closest forecast values to the actual data. Policymakers can use this result in determining which forecast method to use for their community to have data-based information for the preparation of their personnel and facilities.

Keywords: forecasting • epidemics • moving average • exponential smoothing • ARIMA • Ornstein-Uhlenbeck • SEIR • random forest

Correspondence: MC Torres. Institute of Mathematical Sciences and Physics, University of the Philippines Los Baños, Philippines 4031. Email: mctorres4@up.edu.ph.

Author Contribution: BBDC: acquisition of data; EOF, MCT, CHB, BBDC, EBG, JBM: analysis and/or interpretation of data, preparation of draft manuscript and revision, and approval and review of final revision.

Editor: May Anne E. Mata, PhD, University of the Philippines Mindanao, PHILIPPINES

Received: 27 September 2021 **Accepted:** 31 March 2023 **Published:** 28 April 2023

Funding Source: Personally-funded research

Competing Interest: The authors have declared no competing interest.

Citation: Felix EFO, Torres MC, Buhat CAH, Dela Cruz BPB, Gemida EB, Mamplata JB. Forecasting COVID-19 cases in the Philippines using various mathematical models. Banwa B 18: art069.

Forecasting COVID-19 Cases in the Philippines Using Various Mathematical Models

Edd Francis O. Felix[1,2] • Monica C. Torres[1,2] • Christian Alvin H. Buhat[3] • Ben Paul B. Dela Cruz[1,2] • Eleanor B. Gemida[1,2] • Jonathan B. Mamplata[1,2]

[1] Institute of Mathematical Sciences and Physics, University of the Philippines Los Baños, PHILIPPINES
[2] University of the Philippines Resilience Institute, University of the Philippines System, PHILIPPINES
[3] Department of Mathematics, University of Houston, Texas, UNITED STATES OF AMERICA

Introduction

The coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is a kind of viral pneumonia (Li et al. 2020). Since the pandemic started, there has been a rapid increase in daily infection rates and death toll. Prevention and control are vital to reduce the burden of our healthcare system and prevent further decline of the economy, especially with the entry of new variants of COVID-19. Effective modeling and forecasting are needed for data-based decisions and policy making.

Since the COVID-19 outbreak, there have been many publications using various mathematical and machine learning (ML) models that forecast the spread and the epidemic peak globally (Dansana et al. 2020) for specific countries such as China, Taiwan, South Korea, Japan, Italy (Dansana et al. 2021), Brazil (Pereira et al. 2020), Russia, India, and Bangladesh (Nabi 2020), and the United States of America (Bertozzi et al. 2020).

The dynamics of the COVID-19 infection vary per country and depend on many factors such as the healthcare system's capacity, testing and contact tracing efforts, quarantine and lockdown impositions, and the public's reaction to the pandemic. In the Philippines, the first case of COVID-19 was reported on 30 January 2020, while the first local transmission was confirmed on 7 March 2020 (WHO 2020a). On 17 March 2020, the Philippine government declared an enhanced community quarantine (ECQ) in the entire island of Luzon and other parts of the country (Republic of the Philippines 2020). Months since the lockdown, the economy gradually reopened with less stringent quarantine regulations. Although the increase in the number of cases has slowed down, the threat of another surge is still present. Therefore, continuous effort to model and forecast the spread of the disease in the country is helpful in the fight against COVID-19.

Multiple mathematical methods such as Susceptible Exposed Infectious Recovered (SEIR) and computational models have been used to model and describe the COVID-19 dynamics in the country (Buhat et al. 2021a; Buhat et al. 2021b). The heterogeneous characteristics of different age populations were incorporated in studying the effects of the ECQ in reducing the exponential growth of the disease as well as the forecasts of the transmission rate (Dizon 2020). A SEIR model considering the symptomatic and asymptomatic populations was developed by Arcede et al. (2020) to describe the dynamics of the disease. The group used the data on the confirmed cases and death from several countries including France, the Philippines, Italy, Spain, the United Kingdom, China, and the USA to calibrate the model.

This study forecasts the cumulative daily cases of COVID-19 in the Philippines using various mathematical models. We examined six models and determined which best suits the considered Philippine data. The models under comparison were weighted moving average, exponential smoothing, SEIR, Ornstein-Uhlenbeck process, Autoregressive Integrated Moving Average (ARIMA), and random forest. The data set was fitted using each of the models and subsequently, obtained forecasts for 1–15 August 2020. These forecasts were then compared to the actual values using various error metrics.

Data Framework

Since 12 April 2020, COVID-19 data in the country have been made available through the COVID-19 Tracker (DOH 2020) to promote

transparency and accountability of data in the country. The tracker provides a daily COVID-19 data drop, which contains multiple COVID-19 infection-related data based on multiple cases such as specimen collection, the release of the result, and date of reporting/confirmation of positive COVID-19 result. We only consider the infections based on the date of reporting/confirmation of positive COVID-19 results, as this is the commonly reported data to the public. We noted that there is no adjustment made to the data reflecting the delays in daily reports. Figure 1 illustrates the actual March 6 to July 31 data.

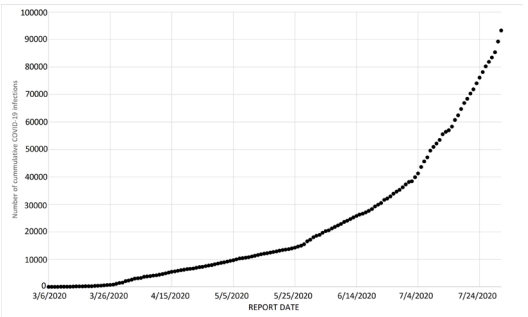

FIGURE 1 Actual cumulative COVID-19 cases of infection in the Philippines from 6 March 2020– 31 July 2020

Results and Discussion

Weighted Moving Average

Preliminaries

The simple moving average (SMA) is a very basic forecasting technique. The standard formula to get the forecast for the n^{th} day for an m-day SMA is given by

$$SMA_n = \frac{1}{m}\sum_{i=1}^{m} C_{n-i} \ ,$$

where C_i is the number of total confirmed cases at time i. It gives equal weights for all the data in a specific interval. On the other hand, the weighted moving average (WMA) gives custom weights for these data.

One of the most common weight functions used in WMA for the i^{th} day in the m-day interval is

$$w(i) = \frac{2i}{m(m+1)} \ .$$

The standard formula to get the n^{th} day forecast for an m-day WMA is given by

$$WMA_n = \sum_{i=1}^{m} w(m+1-i)\, C_{n-i} \ ,$$

where C_i is the number of total confirmed cases at time i. Note that the sum of all the weights in any WMA should be equal to 1.

There are different intervals used in the literature regarding COVID-19 such as 3-day, 5-day, 7-day, 10-day, and 14-day (Elmousalami and Hassanien 2020; He et al. 2020). In most cases, the 7-day interval is considered to cover both the incubation period and the time it takes from the first symptoms to occur to diagnosis (He et al. 2020).

We considered both SMA and WMA and notice that the forecasts for 1–15 August 2020 using WMA, with increasing weights, were closer to the actual data compared to SMA.

Numerical Implementation

We considered 3-day, 4-day, 7-day, and 10-day WMA to have a comparison among different intervals. We applied WMA to the cumulative cases, but the result was undesirable. For example, the actual number of cumulative cases on July 31 is higher than the forecast total number of COVID-19 cases on August 1 using 4-day WMA. This is inconsistent since the total cumulative cases must be increasing. Instead of applying WMA directly to the cumulative cases, we applied it to the daily cases. To get the forecast for the cumulative cases for August 1, we added the forecast for the daily cases on August 1 to the actual total cases as of July 31.

We considered several weight functions

$$w_r(i) = \frac{2(ir-r+1)}{m(mr-r+2)}$$

for an m-day WMA, where $r \in \{1, 2, 4, 10\}$. Note that when $r = 1$, our custom weight function is equal to the common weight function, i.e., $w_r(i) = w(i)$. Table 1 shows the different weights for the 4-day WMA.

After comparing the results for the different values of r, we noticed that the higher the value

of *r*, the closer the forecasts to the actual values. Below are the forecasts for the different time intervals using $r = 10$ (Table 2 and Figure 2).

Based on the results, the forecast closest to the actual data among the types of moving average is the 3-day WMA with weight function

$$w_{10}(i) = \frac{10i - 9}{33} \quad .$$

TABLE 1 Weights used for the 4-day weighted moving average (WMA)

4-day WMA	r = 1	r = 2	r = 4	r = 10
w_r (1)	10%	6.25%	3.57%	1.56%
w_r (2)	20%	18.75%	17.86%	17.19%
w_r (3)	30%	31.25%	32.14%	32.81%
w_r (4)	40%	43.75%	46.43%	48.44%

TABLE 2 15-day ahead forecasts on the cumulative COVID-19 cases in the Philippines using weighted moving average (WMA)

Date	Weighted moving average			
	3-day	4-day	7-day	10-day
08/01/2020	97297	96949	96272	96000
08/02/2020	101281	100727	99305	98704
08/03/2020	105255	104497	102441	101461
08/04/2020	109231	108244	105648	104269
08/05/2020	113207	112002	108878	107122
08/06/2020	117183	115758	112072	110013
08/07/2020	121159	119514	115243	112922
08/08/2020	125135	123270	118424	115828
08/09/2020	129111	127026	121611	118705
08/10/2020	133087	130782	124800	121567
08/11/2020	137063	134539	127987	124434
08/12/2020	141038	138295	131173	127307
08/13/2020	145014	142051	134359	130183
08/14/2020	148990	145807	137545	133061
08/15/2020	152966	149563	140732	135938

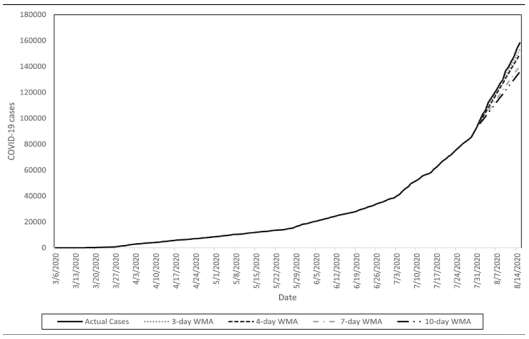

FIGURE 2 WMA with r = 10 forecast versus the actual data

Exponential Smoothing

Preliminaries

Exponential smoothing models are commonly used in time series forecasting. These models produce reliably accurate forecasts since they can also capture the trend, seasonality, or a combination of both. The simple exponential smoothing (SES) is the simplest smoothing model commonly used for data with no observable trend or seasonality. This method follows the following forecasting formula:

Forecast equation: $\hat{y}_{t+h|t} = l_t$

with $l_t = \alpha y_t + (1-\alpha) l_{t-1}$

where at time t, y_t, and $\hat{y}_{t+h}$ represent the actual and forecast values, respectively, and l_t represents the estimated level of the series with α as the smoothing factor.

The SES produces h-step ahead forecasts that are of the same value and are equal to the last level value. Holt's linear trend method (HLTM) extends the SES method by using two smoothing equations, both of which are dependent on the estimated level and estimated trend of the series at a particular time. The h-step ahead forecasting formula is given by the following:

Forecast equation: $\hat{y}_{t+h|t} = l_t + hb_t$

with $l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$

and $b_t = \beta(l_t - l_{t-1}) + (1-\beta) b_{t-1}$

where at time t, y_t and $\hat{y}_{t+h}$ represent the actual and the forecast values, respectively, while l_t and b_t represent the estimated level and estimated trend of the data, respectively. Furthermore, α and β are the smoothing factors of the level and the trend, respectively, with $0 \leq \alpha, \beta \leq 1$.

Other epidemic models include SI (Susceptible-Infected), SIS (Susceptible-Infected-Susceptible), that is suitable for data that exhibit trends. Its forecasts are also expected to exhibit the same. For longer forecast time horizons though, HLTM tends to over-forecast. To prevent this from happening, a parameter can be introduced to "dampen" the forecasts. The h-step ahead forecasting formula becomes

$$\text{Forecast equation: } \hat{y}_{t+h|t} = l_t + (\phi + \phi^2 + \ldots \phi^n) b_t$$
$$\text{with } l_t = \alpha y_t + (1-\alpha)(l_{t-1} + \phi b_{t-1})$$
$$\text{and } b_t = \beta(l_t - l_{t-1}) + (1-\beta)\phi b_{t-1}$$

Observe that when the damping parameter $\phi = 1$, the formula is the same as that of the HLTM. When $0 < \phi < 1$, the short-term forecasts are trended, and the long-term forecasts become constant.

We can observe from Figure 1 that the data has an increasing trend but has no seasonality. Furthermore, we are only interested in the short-term forecasts of 15 days ahead. Hence, we chose HLTM.

The confidence interval of the forecasts is given by

$$[\hat{y}_{t+h|t} - 1.96\sigma_h, \hat{y}_{t+h|t} + 1.96\sigma_h]$$

where

$$\overset{h}{\sigma^2} = \sigma^2 [1 + (\mathrm{h}-1)\{\alpha^2 + \alpha\beta h + \beta^2 h(2h-1)\}]$$

with σ as the variance of the forecast errors. Further discussions on these methods can be found in Hyndman and Athanasopoulos (2019).

Numerical Implementation

We set the initial values of the parameters to be $\alpha = 0.5$ and $\beta = 0.5$ with $l_1 = y_1$ and $b_1 = y_2 - y_1$. We then calculate the 1-day forecasts and the sum of squared errors (SSE). With this, we obtained the values of α and β that minimize the SSE subject to the restrictions of α and β. The values of the parameters that minimize the SSE are $\alpha = 1$ and $\beta = 0.5742$. These parameters are used to calculate the 15-day ahead forecasts as well as the forecast confidence interval of the cumulative daily COVID-19 cases in the Philippines. The summary is shown in Table 3.

As expected from the HLTM, we can see in Figure 3 that the 15-day ahead forecasts exhibit an increasing trend. Moreover, the majority of the actual values lie within the prediction interval, suggesting that the method produced reliable forecast values.

TABLE 3 15-day ahead forecasts on the cumulative COVID-19 cases in the Philippines using Holt's linear trend method

Date	Forecast	95% lower cl	95% upper cl
08/01/2020	96966.78	96296.45	97637.11
08/02/2020	100582.56	99332.42	101832.7
08/03/2020	104198.33	102291.31	106105.4
08/04/2020	107814.11	105174.54	110453.7
08/05/2020	111429.89	107987.37	114872.4
08/06/2020	115045.67	110734.78	119356.6
08/07/2020	118661.45	113421.01	123901.9
08/08/2020	122277.22	116049.67	128504.8
08/09/2020	125893.00	118623.84	133162.2
08/10/2020	129508.78	121146.17	137871.4
08/11/2020	133124.56	123618.95	142630.2
08/12/2020	136740.33	126044.22	147436.4
08/13/2020	140356.11	128423.79	152288.4
08/14/2020	143971.89	130759.28	157184.5
08/15/2020	147587.67	133052.13	162123.2

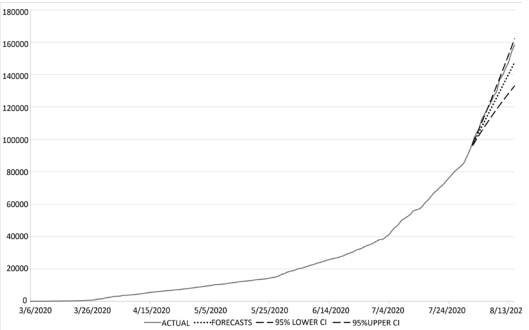

FIGURE 3 Holt's linear trend method forecasts versus the actual data

Susceptible Exposed Infectious Recovered (SEIR)

Preliminaries

Differential equation-based models like the susceptible, exposed, infectious, recovered (SEIR) model give information that is useful in controlling different infectious diseases. These models describe disease dynamics at a macroscopic level (Özmen et al. 2016). Other epidemic models include Susceptible-Infected (SI), Susceptible-Infected-Susceptible (SIS), Susceptible-Infected-Recovered (SIR), and Susceptible-Infected-Recovered-Susceptible (SIRS) models. A discrete Susceptible-Infected-Recovered-Dead (SIRD) model was used in the paper of Corcino et al. (2021). In this paper, the COVID-19 disease spread in Central Visayas in the Philippines was studied, and controls such as social distancing and enhanced community quarantine were included in the model.

The incubation period for COVID-19, which is the time between the transmission of infection and symptom onset, is on average 5–6 days and can be up to 14 days (WHO 2020b). To consider this period, we used the SEIR model to describe the dynamics of COVID-19 in the Philippines, considering the mortality due to infection. In this model, S, E, I, and R are the number of susceptible, exposed, infected, and recovered individuals, respectively. Susceptible individuals are those who can contract the disease and exposed individuals are those infected but are not yet infectious. Meanwhile, infected individuals are those who can transmit the disease, and recovered individuals are those who have recovered and are no longer infectious. Susceptible individuals may be in contact with infected individuals and are transferred to the exposed population at a rate β. Here, the parameter $\beta > 0$ is the transmission coefficient. A portion σ of the exposed individuals will become infected. Infected individuals die at rate d or recover from the disease at rate γ. The parameters $d > 0$ and $\gamma > 0$ are the death and recovery rates, respectively.

The COVID-19 transmission dynamics are governed by the following set of equations:

$$\frac{dS}{dt} = -\beta \frac{SI}{N}$$

$$\frac{dE}{dt} = \beta \frac{SI}{N} - \sigma E$$

$$\frac{dI}{dt} = \sigma E - \gamma I - dI$$

$$\frac{dR}{dt} = \gamma I$$

where

$$\beta = \frac{R_0}{\tau}\left(\frac{S(0)+E(0)+I(0)}{S(0)}\right)$$

is the transmission coefficient and N = S + E + I + R represents the total population size (Buhat et al. 2021a). Here, R_0 is the basic reproduction number which is the expected number of secondary infections produced by an infected individual during their entire infectious period (Diekmann et al. 1990). Table 4 shows the parameters used in the model and their values obtained from previous studies on COVID-19 while Figure 4 describes the inflows and outflows of individuals in each compartment.

TABLE 4 Parameters of the COVID-19 transmission model

Parameter	Description	Value	Reference
R_0	Reproduction number	1.15	Data fitted
σ	Progression rate from exposure to infection	0.2	Data fitted
γ	Recovery rate	0.0686	Data fitted
τ	Infectious period	14	Buhat et al. (2021a)
d	Death rate due to COVID-19	0.03/14	Chen (2020)

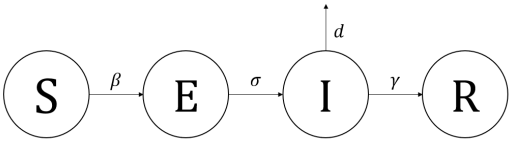

FIGURE 4 Flow diagram of the COVID-19 transmission model

Sensitivity analysis is used to identify the effect of each parameter in the model output I. In this study, a global sensitivity analysis technique called partial rank correlation coefficient (PRCC) analysis is used (Marino et al. 2008). In Figure 5, each bar corresponds to a PRCC value at an instance, specifically in days $t = 67 + 10k$, $k = 0, 1, 2, \ldots, 13$. A large absolute PRCC value implies a large correlation of the parameter with the output. The parameters with high PRCC values (>0.5 or <−0.5) are R_0, σ, γ, and τ. Moreover, R_0 and σ have positive PRCC values implying that an increase in the values of these parameters will increase I. In contrast, γ and τ have negative PRCC values which means that a positive change in these parameters will decrease *I*.

The output using the SEIR model is made close to the gathered data by estimating some of the parameters for which the model output is sensitive. Reported COVID-19 cases and recoveries in the Philippines are available on the COVID-19 tracker. We used these data sets to estimate R_0, σ, and γ. The model output was found to have high sensitivity values on these parameters, which indicate the parameters' influential effect on the outcome. We estimated these parameters by minimizing the error between the gathered data and the model output. The estimates are $R_0 = 1.5811$, $\sigma = 0.013$, and $\gamma = 0.0069$. Figure 6 shows the reported data and the corresponding best fit.

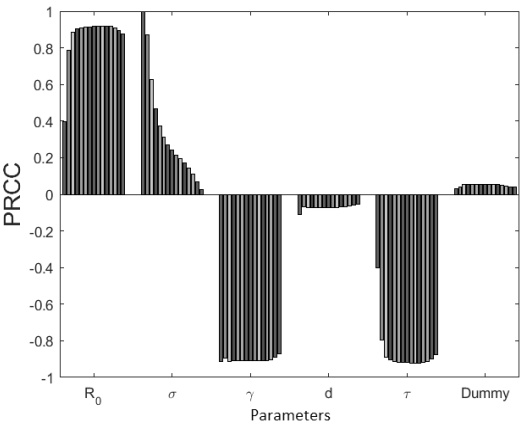

FIGURE 5 Partial rank correlation coefficient (PRCC) values showing the sensitivities of the model output I with respect to the parameters

Numerical Implementation

Using the values obtained in the parameter estimation and the values gathered from other studies on the disease, we forecasted the cumulative cases from August 1–15. Table 5 shows the 15-day forecast using the SEIR model and Figure 7 displays the forecast vs actual data.

Although a similar trend can be perceived in the estimates and the actual data, observe that the SEIR model gave low estimates, and the error increases as time passes. This may be attributed to the sudden rise of cases in the month of August as COVID-19 testing in the country improved.

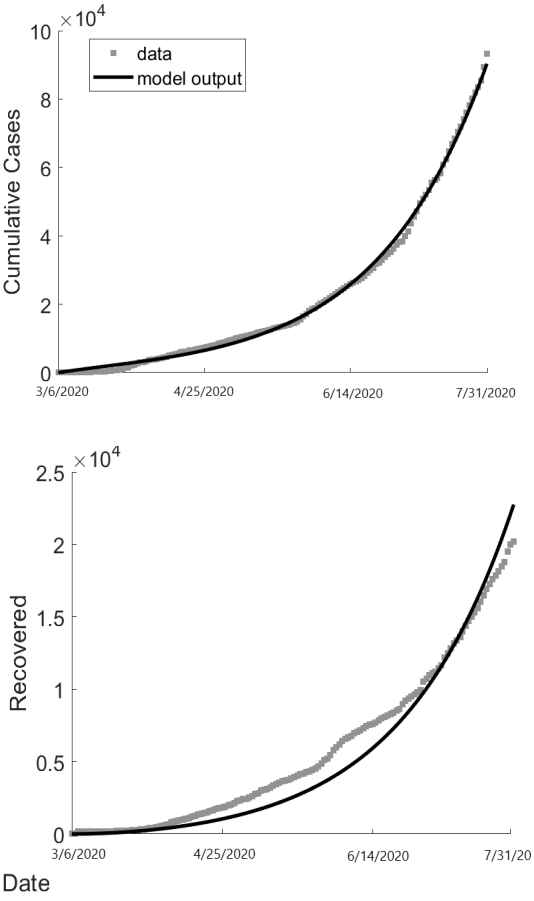

FIGURE 6 Model fitted curve vs actual cumulative cases and recoveries

TABLE 5 15-day ahead forecasts on the cumulative COVID-19 cases in the Philippines using the Susceptible-Exposed-Infected-Recovered (SEIR) model

Date	Forecast
08/01/2020	94319.40
08/02/2020	96819.00
08/03/2020	99381.97
08/04/2020	102009.77
08/05/2020	104703.86
08/06/2020	107465.76
08/07/2020	110296.98
08/08/2020	113199.09
08/09/2020	116173.66
08/10/2020	119222.28
08/11/2020	122346.60
08/12/2020	125548.25
08/13/2020	128828.91
08/14/2020	132190.28
08/15/2020	135634.08

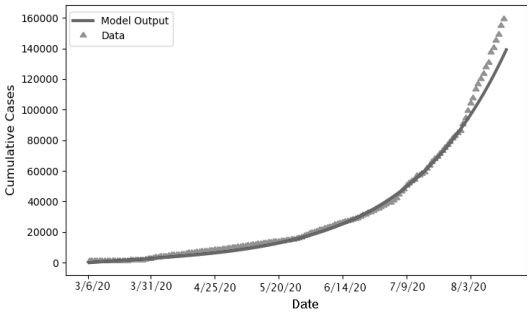

FIGURE 7 SEIR model forecasts versus the actual data

Ornstein-Uhlenbeck Process

Preliminaries

In this section, we assumed that the daily cases denoted by X_t, in the discrete sense, is governed by Ornstein-Uhlenbeck (OU) process by the stochastic differential equation

$$dX_t = \alpha\,(\beta - X_t)dt + \sigma dW_t \qquad (1)$$

where α is the mean reversion, β is the drift of the process, σ is the volatility, and W_t is a standard Weiner process. The parameters of X_t are determined using the maximum likelihood estimation. The explicit solution of (1) is given by

$$X_t = X_s e^{-\alpha(t-s)} + \beta(1 - e^{-\alpha(t-s)}) + \sigma e^{-\alpha t}\int_s^t e^{\alpha u}\, dW_u \quad (2)$$

Using the Euler-Maruyama approximation, the discrete version of (2) is given by

$$X_{k+1} = X_k e^{-\alpha\Delta k} + \beta(1 - e^{-\alpha\Delta k}) + \sigma\sqrt{\frac{1 - e^{-2\alpha\Delta k}}{2\alpha}}\,\omega_{k+1} \quad (3)$$

where ω_{k+1} is a sequence of IID standard normal random variables.

Let $\Theta = (\alpha, \beta, \sigma)$ be the set of parameters needed, and using the maximum likelihood estimation, it can be derived that the best estimate for β and σ

$$\hat{\beta} := f(\alpha) = \frac{\sum_{k=1}^{n} \frac{X_k - X_{k-1}e^{-\hat{\alpha}\Delta k}}{1 + e^{-\hat{\alpha}\Delta k}}}{\sum_{k=1}^{n} \frac{1 - e^{-\hat{\alpha}\Delta k}}{1 + e^{-\hat{\alpha}\Delta k}}}$$

and

$$\hat{\sigma} := g(\hat{\beta}, \hat{\alpha}) = \sqrt{\frac{2\alpha}{n}\sum_{k=1}^{n} \frac{(X_k - \beta - (X_k - 1 - \beta)e^{-\alpha\Delta k})^2}{1 - e^{-2\alpha\Delta k}}}$$

where α can be derived from the optimization problem given by

$$\min_{\alpha}\left\{-\frac{n}{2}\log\left[\frac{g(f(\alpha),\,\alpha)}{2\alpha}\right] - \frac{1}{2}\sum_{k=1}^{n}\log(1 - e^{-2\alpha\Delta k}) - \frac{\alpha}{g(f(\alpha),\,\alpha)^2}\sum_{k=1}^{n}\frac{(X_k - f(\alpha) - (X_{k-1} - f(\alpha))e^{-\alpha\Delta k})^2}{1 - e^{-2\alpha\Delta k}}\right\}$$

Numerical Implementation

From the data set, we determine the parameters of the OU process using the previous subsection's methodology. The estimated parameters of the model in terms of daily cases are as follows: α = 0.05948791, β = 1085.746, and σ = 372.6143. Using these parameters, we determined a 15-day forecast depicted in Table 6.

Figure 8 shows the plot of the forecasted values generated by the OU process against the actual recorded values on August 1–15. If X_k is the forecasted value at time k,

$$X_k = E[X_k \mid X_0] = X_0 e^{-\alpha k} + \beta(1 - e^{-\alpha k})$$

TABLE 6 15-day ahead forecasts on the cumulative COVID-19 cases in the Philippines following Ornstein-Uhlenbeck process

Date	Forecast
08/01/2021	97242.05
08/02/2021	100971.09
08/03/2021	104547.47
08/04/2021	107980.01
08/05/2021	111277.02
08/06/2021	114446.32
08/07/2021	117495.28
08/08/2021	120430.87
08/09/2021	123259.62
08/10/2021	125987.70
08/11/2021	128620.94
08/12/2021	131164.80
08/13/2021	133624.45
08/14/2021	136004.76
08/15/2021	138310.30

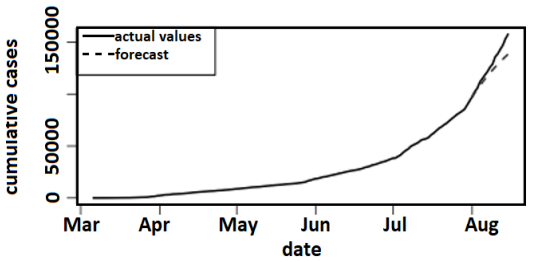

FIGURE 8 Ornstein-Uhlenbeck process forecast versus the actual data

where X_0 is the actual value on July 31. Notice from the plot that the forecast values are slightly bent compared to the actual values, and this is because as k increases, the white noise generated by the actual value from the forecast relatively increases. The noise attributed is

$$\sigma e^{-\alpha k}\int_s^t e^{\alpha u}dW_u$$

and becomes zero in getting the forecasts.

Autoregressive Integrated Moving Average (ARIMA)

Some of the widely used time series models are the autoregressive (AR) and moving average (MA) models. The AR model is in the same form as the multiple linear regression with the past values serving as the explanatory variables. The MA model tells us that the observation at a time t is a weighted average of past shocks. The autoregressive moving average (ARMA) model combines the AR and MA models into a compact form. The general form of an ARMA (p,q) model, p and q refer to the number of AR and MA parameters, respectively, is given by

$$x_t = \emptyset_0 + \sum_{i=1}^{p} \emptyset_i x_{t-i} + a_t - \sum_{i=1}^{q} \theta_i a_{t-i} \ ,$$

where $\{a_i\}$ is a white noise series, $\emptyset_i$, $i = 1, 2, \ldots, p$ are the AR parameters, $\{\theta_i\}$ are the MA parameters, and $\emptyset_0$ is a constant. In time series modeling, differencing is one method to transform a nonstationary time series into a stationary time series. A time series follows autoregressive integrated moving average (ARIMA) with parameters p, d, and q if the differenced data set of order d, that is, $y_t = (1-B)^d x_t$, follows an ARMA(p,q) model. Here, we use the Box-Jenkins backshift operator $B^m x_t = x_{t-m}$. Hence, the general form of an ARIMA (p,d,q) model is

$$\left(1 - \sum_{i=1}^{p} \emptyset_i B^i\right) y_t = \emptyset_0 + \left(1 - \sum_{i=1}^{q} \Theta_i B^i\right) a_t \ .$$

Please refer to Box et al. (1994) and Tsay (2010) for more discussion of these models.

Preliminaries

The increasing trend in the data set as observed in Figure 1 means that the data set is nonstationary. To remove the trend, we performed successive first-order differencing to achieve a stationary data set. Using the augmented Dickey-Fuller (ADF) test for stationarity, it was found that the second-order differenced data set is stationary at $\alpha = 0.05$ as illustrated in Figure 9.

Numerical Implementation

We executed a correlogram analysis on the second-order differenced data set and the autocorrelation (ACF) and partial autocorrelation

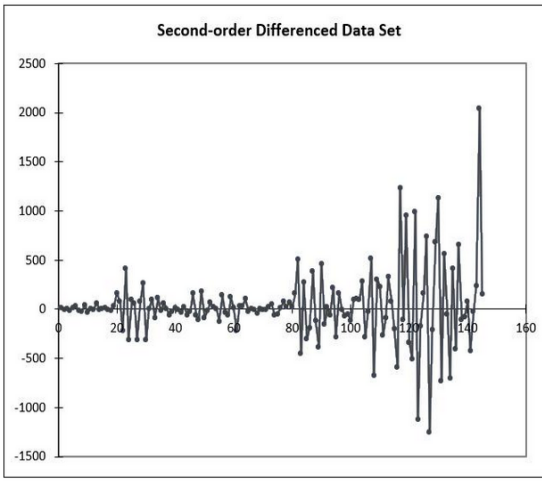

FIGURE 9 Second-order differenced data on the cummulative cases of infection in the Philippines

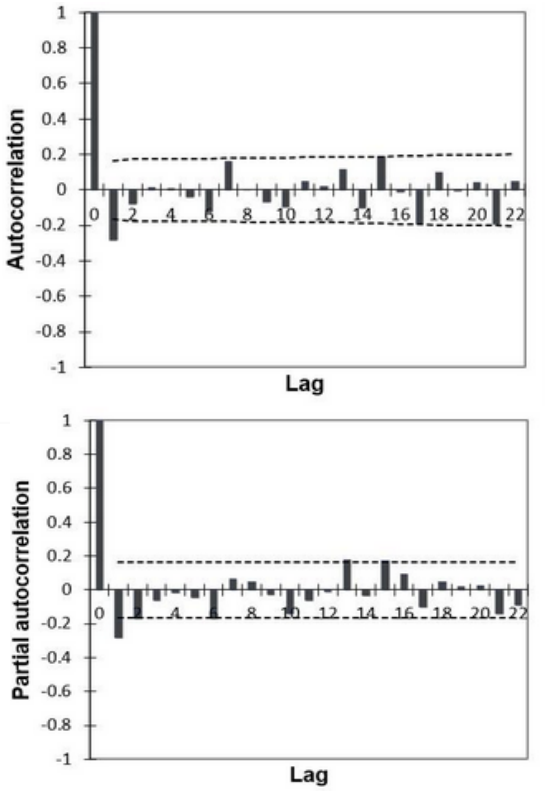

FIGURE 10 Autocorrelation (ACF) and partial autocorrelation (PACF) plots of the second-order differenced data

(PACF) plots are presented in Figure 10. Significant spikes at lag one in the ACF and PACF suggest that AR and MA processes at that lag may help explain the data. We examined several models for different values of p and q, capped at 5. We compared these models based on the Akaike information criterion (corrected for small sample sizes, AICc) and Schwarz's criterion (SBC). Based on the AICc, the best fit model is ARIMA(5,2,4), while on SBC is ARIMA(1,2,1). The comparison of the models based on AICc and SBC is shown in Table 7.

Using the goodness-of-fit statistics for the two models presented in Table 8, we perform the likelihood ratio test. At $\alpha = 0.05$, the ARIMA(5,2,4) model does not fit the data significantly better than ARIMA(1,2,1). Therefore, the final ARIMA model is ARIMA(1,2,1) and its parameters are presented in Table 9. In Figure 11, ARIMA(1,2,1) is plotted against the original time series.

Based on the model, the 15-day forecasts of the daily cumulative cases are presented in Table 10.

Random Forest

Preliminaries

Random forest (RF) is an ML method based on the classification and regression tree and bagging (Dudek 2015; Mei et al. 2014). A combination of decision trees is generated and is bootstrapped from the learning sample. These samples are obtained from a subset of random features from each chosen node. Every decision tree generated will be built on a subset of learning points and features that are considered from each chosen node to split on. Each tree will be grown to the largest extent possible, and there will be no pruning. After the trees are fitted, a new forecast is generated by averaging the forecasts of the trees (Dudek 2015) as shown in Figure 12.

Aside from model fitting, RF has been recently used to analyze, predict, and evaluate COVID-19 in India and COVID-19 patient health (Iwendi et al. 2020; Kolla 2020). Kolla's study (2020) showed COVID-19 predictions in India and results from RF outperformed other ML methods that were considered in their

TABLE 7 Comparison of the two ARIMA models based on AICc and SBC

p	q	AICc	SBC
1	1	1969.901	1978.526
1	2	1971.900	1983.340
1	3	1974.060	1988.281
1	4	1975.548	1992.519
1	5	1976.678	1996.364
2	1	1972.674	1984.114
2	2	1971.253	1985.474
2	3	1976.241	1993.211
2	4	1977.894	1997.581
2	5	1979.117	2002.718
3	1	1972.936	1987.157
3	2	1975.193	1992.164
3	3	1977.936	1997.623
3	4	1978.520	2000.888
3	5	1967.8840	2003.468
4	1	1975.091	1992.062
4	2	1977.234	1996.920
4	3	1973.700	1996.068
4	4	1974.510	1999.525
4	5	1979.347	2006.973
5	1	1980.538	1995.572
5	2	1979.330	2001.698
5	3	1981.525	2006.540
5	4	1967.334	1994.960
5	5	1978.153	2008.353

TABLE 8 Goodness-of-fit statistics of the two ARIMA models

Statistics	ARIMA(1,2,1)	ARIMA(5,2,4)
SSE	11073309.7	9359918.43
MSE	79664.0989	67337.5426
RMSE	282.248293	259.494783
WN Variance	79664.0989	67337.5426
MAPE(Diff)	174.524932	182.852761
MAPE	2.68510318	2.58111358
-2Log(Like.)	1963.72275	1945.61506
FPE	80818.651	72362.7324
AICc	1969.90053	1967.33381
SBC	1978.52617	1994.9598

TABLE 9 Parameters of ARIMA(1,2,1)

	AR(1)	MA(1)
Value	0.125	−0.722
Hessian standard error (HSE)	0.131	0.094
Lower bound for 95% confidence level (HSE)	−0.132	−0.905
Upper bound for 95% confidence level (HSE)	0.381	−0.538
Asymptotic standard error (ASE)	0.128	0.089
Lower bound for 95% confidence level (ASE)	−0.127	−0.897
Upper bound for 95% confidence level (ASE)	0.376	−0.547

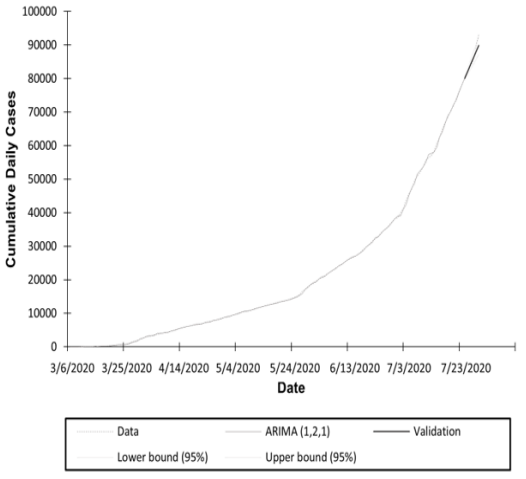

FIGURE 11 ARIMA(1,2,1) model versus the actual data

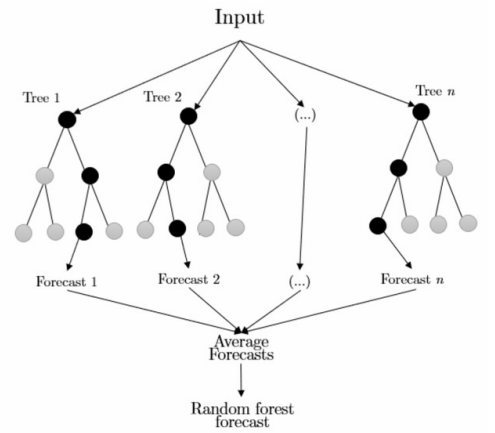

FIGURE 12. Forecasting process using random forest given n number of trees.

FIGURE 12 Forecasting process using random forest given n number of trees

TABLE 10 15-day ahead forecasts on the cumulative COVID-19 cases in the Philippines using ARIMA(1,2,1)

Date	Forecast
08/01/2020	97432.88
08/02/2020	101517.1
08/03/2020	105601.6
08/04/2020	109686.2
08/05/2020	113770.8
08/06/2020	117855.3
08/07/2020	121939.9
08/08/2020	126024.5
08/09/2020	130109.1
08/10/2020	134193.6
08/11/2020	138278.2
08/12/2020	142362.8
08/13/2020	146447.3
08/14/2020	150531.9
08/15/2020	154616.5

study. Thus, we implemented RF to forecast the COVID-19 cumulative cases of infection in the Philippines.

Using the RF package in R (Liaw and Wiener 2002), we first set the training data and testing data to 148 (number of days from March 6 to July 31) and 15 (number of days to be forecasted), respectively. After which, we fitted RF on the training data set with the following parameter inputs:

Training data: COVID-19 cumulative cases of infection in PH from March 6 to July 31, 2020
Type of RF: Regression
Number of trees: 1000
Number of candidate-split variables at each split: 1

We then set up a data frame to hold the RF model predictions since we are forecasting cases that have not been observed yet. We set up a 97.5% confidence interval for our predictions and compute the forecasts (point) with an upper and lower bound and its standard deviation.

Numerical Implementation

We simulate 15-day ahead forecasts for the RF method. Table 11 shows the results.

From Figure 13, notice that RF produced underestimates compared to the actual values for the August 1–15 COVID-19 infection cases. Some of the initial values were almost captured on the interval, but the further increase of actual cases was not captured. An abundance of data is necessary for a better forecast for an ML algorithm, which was not evident in this case. Increasing the number of trees did not also differ from the result generated with 1000 trees.

TABLE 11 15-day ahead forecasts on the cumulative COVID-19 cases in the Philippines usingrandom forest

Date	Forecast	97.5% lower cl	97.5% upper cl
08/01/2020	86572.08	72374.94	93351
08/02/2020	87229.36	72699.45	93351
08/03/2020	86722.70	72374.94	93351
08/04/2020	86623.09	73127.31	93351
08/05/2020	86660.43	72374.94	93351
08/06/2020	86623.09	73127.31	93351
08/07/2020	86778.93	73451.29	93351
08/08/2020	86778.93	73451.29	93351
08/09/2020	86711.77	72699.45	93351
08/10/2020	86778.93	73451.29	93351
08/11/2020	86615.32	72814.91	93351
08/12/2020	86615.32	72814.91	93351
08/13/2020	86595.98	72374.94	93351
08/14/2020	87161.19	74044.07	93351
08/15/2020	87161.19	74044.07	93351

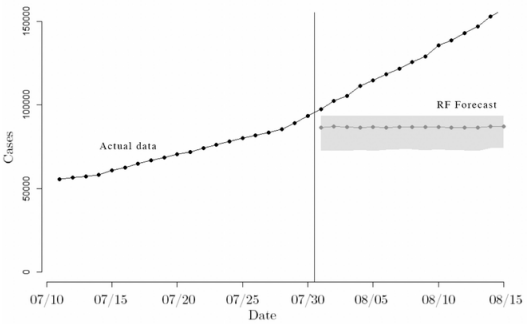

FIGURE 13. RF forecasts vs. the actual data

FIGURE 13 Random forest forecasts vs the actual data

Analysis of Forecast Errors

We plot the 15-day ahead forecasts for 1–15 August 2020 from the six models: 3-day WMA, HLTM, SEIR, OU process, ARIMA, and RF. We then compare them to the actual cumulative cases of infection data in the Philippines.

From Figure 14, notice that all models except RF were able to follow the trend of the actual data. The RF model failed to observe an increasing trend compared to the other models and was not able to forecast the data well. Of the models that continued to increase, the ARIMA model, 3-day WMA, and HLTM were able to maintain a difference of fewer than 10,000 cases from the actual data, given the considered forecast period. As the 15th-day forecast was approached, the ARIMA(1,2,1) model is closest to the actual values.

An error analysis of the deviations among the model forecasts is shown in Table 12. The error metrics used are root mean square error (RMSE) given by

$$\sqrt{\frac{1}{K}\sum_{k=1}^{K}(c_k-\hat{c}_k)^2} \quad ,$$

absolute mean error (MAE) computed as

$$\frac{1}{K}\sum_{k=1}^{K}|\,c_k-\hat{c}_k\,| \quad ,$$

and relative absolute error (RAE) obtained by

$$\frac{\sum_{k=1}^{K}|\,c_k-\hat{c}_k\,|}{\sum_{k=1}^{K}|\,\bar{c}-\hat{c}_k\,|} \quad .$$

From all the error metrics, the ARIMA(1,2,1) model forecasts have the least errors while the random forest forecasts have the highest.

TABLE 12 Error metrics for the forecasted values

Error metric	Root mean square error	Mean absolute error	Relative absolute error
3-day WMA	2150.65	1701.00	0.113736
HLTM	5265.00	4536.51	0.325803
SEIR Model	14036.70	12973.74	0.856841
OU	9357.66	7389.56	0.622168
ARIMA (1,2,1)	1330.69	1002.01	0.065484
RF	43934.17	40038.51	1.000000

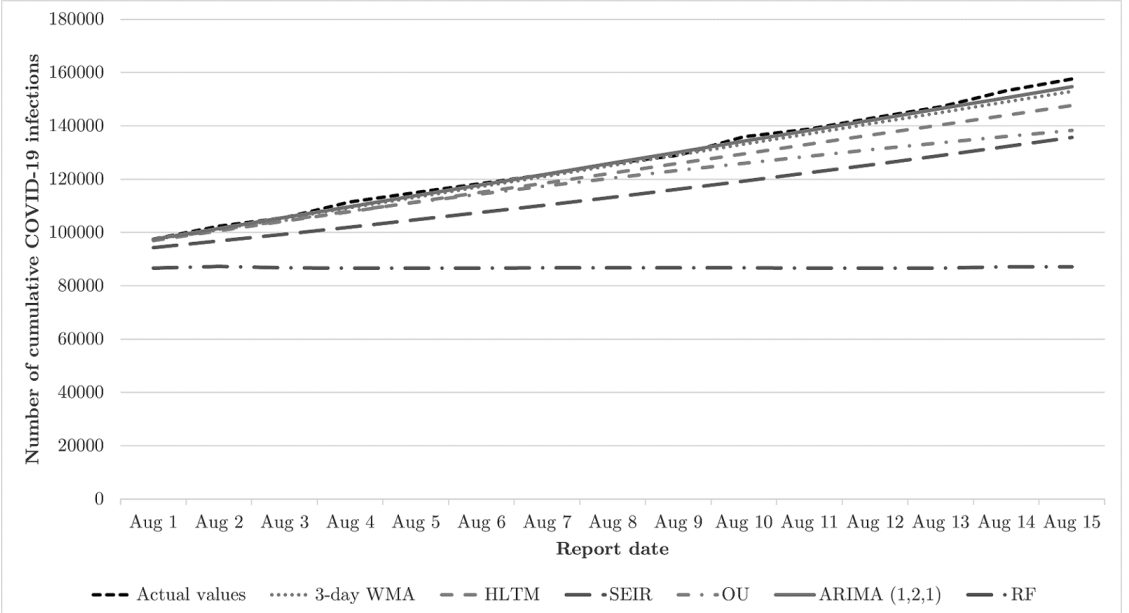

FIGURE 14 Plot of 15-day-ahead forecast points of the cumulative COVID-19 cases in the Philippines of various mathematical models versus the actual data

Therefore, among the six models used in this study, ARIMA(1,2,1) has the best forecasts of cumulative COVID-19 cases in the Philippines from 1–15 August 2020.

Conclusions and Recommendations

We obtained 15-day ahead forecasts for the daily cumulative COVID-19 cases in the Philippines using mathematical models namely 3-day weighted moving average, exponential smoothing using Holt's linear trend method, SEIR model, Ornstein-Uhlenbeck process, ARIMA(1,2,1) model, and random forest. We considered the 6 March to 31 July 2020 COVID-19 infection data retrieved from the DOH COVID-19 Data Drop. All models except the random forest produced forecasts that exhibit an increasing trend, with ARIMA(1,2,1) having the closest forecast values to the actual August 1–15 data based on the error analysis.

Results from the study can be used by policymakers in determining which forecast method to use/adapt in their community, especially for those with case behavior similar to the Philippines. These 15-day ahead forecasts provide data-based information for the preparation of the personnel and facilities, and delivery of an effective response. Others may also replicate the methods used and may have a different "best model" for their community. We do note that all models used in the study were based on the retrieved data drop only. Thus, factors that might affect the actual data such as delay or error in reporting were not considered and might affect the result of this study. For future studies, a paper that focuses on delays from symptom onset to public confirmation may be done. This can have a big impact on improving the forecast values. Furthermore, a larger amount of data with updates on parameters can improve the model and may produce better forecast results (Buhat 2021). We only considered cumulative cases based on the public confirmation date. Thus, further studies may consider other fields in the data drop, and may look into providing daily forecasts rather than cumulative. We also recommend doing pattern recognition to determine which models or ensemble of models can be more accurate, e.g., another ML method since the RF model did not do well on our data.

Notes

The codes/programs used in the study can be found here: https://github.com/alvinbuhat/forecasts

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgment

The authors would like to thank the University of the Philippines Los Baños's Office of the Vice Chancellor for Research and Extension for their assistance in our project.

References

Arcede J, Caga-Anan R, Mentuda C, Mammeri Y. 2020. Accounting for symptomatic and asymptomatic in a SEIR-type model of COVID-19. MMNP 15: 34. https://doi.org/10.1051/mmnp/2020021

Bertozzi A, Franco E, Mohler G, Short M, Sledge D. 2020. The challenges of modeling and forecasting the spread of COVID-19. Proceedings of the National Academy of Sciences of the United States of America 117(29): 16732-16738. https://doi.org/10.1073/pnas.2006520117

Box GEP, Jenkins GM, Reinsel GC. 1994. Time series analysis: Forecasting and control [3rd ed]. New Jersey: Prentice Hall, Englewood Cliff.

Buhat C. 2021. The impact of adherence to minimum health standards in the Philippines during the COVID-19 pandemic. Lancet Reg Health West Pac. 14: 100248. https://doi.org/10.1016/j.lanwpc.2021.100248

Buhat C, Torres M, Olave Y, Gavina M, Felix E, Gamilla G, Verano K, Babierra A, Rabajante J. 2021a. A mathematical model

of COVID-19 transmission between frontliners and the general public. Netw. Model. Anal. Health Inform. Bioinform. 10(1): 1–12. https://doi.org/10.1007/s13721-021-00295-6

Buhat C, Lutero D, Olave Y, Torres M, Rabajante J. 2021b. Community transmission of respiratory infectious diseases using agent-based and compartmental models. Mindanao J. Sci. Tech. 19: 164–183. https://www.researchgate.net/publication/357649066_Community_Transmission_of_Respiratory_Infectious_Diseases_using_Agent-based_and_Compartmental_Models

Chen J. 2020. Pathogenicity and transmissibility of 2019-NCov – A quick overview and comparison with other emerging viruses. Microbes Infect.22(2): 69-71. https://doi.org/10.1016/j.micinf.2020.01.004

Corcino R, Elnar A, Maglasanag G, Casas K. 2021. Estimation, control, and forecast of COVID-19 disease spread in Central Visayas, Philippines. Palawan Scientist: 114–131. https://www.researchgate.net/publication/360228345_Estimation_control_and_forecast_of_COVID-19_disease_spread_in_Central_Visayas_Philippines

Dansana D, Kumar R, Adhikari J, Mohapatra M, Sharma R,Priyadarshini I, Le D. 2020. Global forecasting confirmed and fatal cases of COVID-19 outbreak using autoregressive integrated moving average model. Front. Public Health 8. https://doi.org/10.3389/fpubh.2020.580327

Dansana D, Kumar R, Parida A, Sharma R, Adhikari J, Le H, Pham B, Singh K, Pradhan B. 2021. Using Susceptible-Exposed-Infectious-Recovered Model to forecast Coronavirus outbreak. Comput. Mater. Contin. 67(2): 1595–1612. https://doi.org/10.32604/cmc.2021.012646

Diekmann O, Heesterbeek J, Metz J. 1990. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. J. Math. Biol. 28(4): 365–382. https://doi.org/10.1007/BF00178324

Dizon R. 2020. The heterogeneous age-mixing model of estimating the COVID cases of different local government units in the National Capital Region, Philippines. CEGH 9: 12–16. https://doi.org/10.1016/j.cegh.2020.06.003

[DOH] Department Of Health. 2020. Covid-19 tracker; [accessed 2020 August]. https://www.doh.gov.ph/covid19tracker

Dudek G. 2015. Short-term load forecasting using random forests. Adv. Intell. Syst. 323. http://dx.doi.org/10.1007/978-3-319-11310-4_71

Elmousalami H, Hassanien A. 2020. Day level forecasting for coronavirus disease (COVID-19) spread: Analysis, modeling and recommendations. arXiv. https://arxiv.org/abs/2003.07778

Republic of the Philippines. 2020. Proclamation no. 929s. 2020. Official Gazette; [accessed October 3, 2020]. https://www.officialgazette.gov.ph/2020/03/16/proclamation-no-929-s-2020/

He Y, Wang X, He H, Zhai J, Wang B. 2020. Moving average based index for judging the peak of the COVID-19 epidemic. IJERPH 17(1). https://doi.org/10.3390/ijerph17155288

Hyndman R, Athanasopoulos G. 2019. Forecasting: Principles and practice (3rd ed). OTexts: Melbourne, Australia; . Kolassa-review.pdf (otexts.com); [accessed 2021 December].

Iwendi C, Bashir A, Pasupuleti N, Radha S, Chatterjee A, Peshkar A, Mishra R, Pillai S, Jo O. 2020. COVID-19 patient health prediction using boosted random forest algorithm. FFront. Public Health8. https://doi.org/10.3389/fpubh.2020.00357

Kolla B. 2020. Analysis, prediction and evaluation of COVID-19 datasets using machine learning algorithms. IJETER 8. https://dx.doi.org/10.30534/ijeter/2020/117852020

Li X, Geng M, Peng Y, Meng L, Lu S. 2020. Molecular immune pathogenesis and diagnosis

of COVID-19. J. Pharm. Anal. 10(2): 102–108. https://doi.org/10.1016/j.jpha.2020.03.001

Liaw A, Wiener M. 2002. Classification and regression by random forest. R News: 2(3), 18–22; [accessed 2021 December]. https://cran.r-project.org/doc/Rnews/

Marino S, Hogue I, Ray C, Kirschner D. 2008. A methodology for performing global uncertainty and sensitivity analysis in systems biology. J. Theor. Biol. 254(1): 178–196. https://doi.org/10.1016/j.jtbi.2008.04.011

Mei J, He D, Harley R, Habetler T, Qu G. 2014. A random forest method for real-time price forecasting in New York electricity market. National Harbor, MD, USA: IEEE PES General Meeting | Conference & Exposition 1-5; 2014. https://doi.org/10.1109/PESGM.2014.6939932

Nabi K. 2020. Forecasting COVID-19 pandemic: A data-driven analysis. Chaos Solit. Fractals. 139: 110046. https://doi.org/10.1016/j.chaos.2020.110046

Özmen Ö, Nutaro J, Pullum L, Ramanathan A. 2016. Analyzing the impact of modeling choices and assumptions in compartmental epidemiological models. Simulation 92. https://doi.org/10.1177/0037549716640877

Pereira I, Guerin J, Silva A, Distante C, Garcia G, Goncalves A. 2020. Forecasting COVID-19 dynamics in Brazil: A data driven approach. Int. J. Environ. Res. Public Health 17(14): 5115. https://doi.org/10.3390/ijerph17145115

Tsay RS. 2010. Analysis of Financial time series [3rd ed]. Hoboken, New Jersey: John Wiley & Sons, Inc.

[WHO] World Health Organization. 2020a. Coronavirus disease (COVID-19) in the Philippines; [accessed 2020 September].https://www.who.int/philippines/emergencies/covid-19-in-the-philippines

[WHO] World Health Organization. 2020b. Coronavirus disease 2019 (COVID-19) Situation Report – 73; [accessed 2-22 April]. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200402-sitrep-73-covid-19.pdf?sfvrsn=5ae25bc7_6